Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)
genehmigte Dissertation von
M.Sc. Aditya Mogadala

MULTI-VIEW REPRESENTATION LEARNING FOR UNIFYING LANGUAGES, KNOWLEDGE AND VISION

ADITYA MOGADALA

Tag der mündlichen Prüfung:	7. November, 2018
Referent:	PD Dr. Achim Rettinger
Korreferent:	Prof. Dr. Volker Tresp
Korreferent:	Prof. Dr. York Sure-Vetter

Karlsruhe, 2018

This thesis is decicated to my family and friends.

ABSTRACT

The growth of content on the web has raised various challenges, yet also provided numerous opportunities. Content exists in varied forms such as text appearing in different languages, entity-relationship graph represented as structured knowledge and as a visual embodiment like images/videos. They are often referred to as *modalities*. In many instances, the different amalgamation of modalities co-exists to complement each other or to provide consensus. Thus making the content either heterogeneous or homogeneous. Having an additional point of view for each instance in the content is beneficial for data-driven learning and intelligent content processing. However, despite having availability of such content. Most advancements made in data-driven learning (i.e., machine learning) is by solving tasks separately for the single modality. The similar endeavor was not shown for the challenges which required input either from all or subset of them.

In this dissertation, we develop models and techniques that can leverage multiple views of heterogeneous or homogeneous content and build a shared representation for aiding several applications which require a combination of modalities mentioned above. In particular, we aim to address applications such as content-based search, categorization, and generation by providing several novel contributions.

First, we develop models for heterogeneous content by jointly modeling diverse representations emerging from two views depicting text and image by learning their correlation. To be specific, modeling such correlation is helpful to retrieve cross-modal content. Second, we replace the heterogeneous content with homogeneous to learn a common space representation for content categorization across languages. Furthermore, we develop models that take input from both homogeneous and heterogeneous content to facilitate the construction of common space representation from more than two views. Specifically, representation is used to generate one view from another. Lastly, we describe a model that can handle missing views, and demonstrate that the model can generate missing views by utilizing external knowledge. We argue that techniques the models leverage internally provide many practical benefits and lot of immediate value applications.

From the modeling perspective, our contributed model design in this thesis can be summarized under the phrase Multi-view Representation Learning (MVRL). These models are variations and extensions of shallow statistical and deep neural networks approaches that can jointly optimize and exploit all views of the input content arising from different independent representations. We show that our models advance state of the art, but not limited to tasks such as cross-modal retrieval, cross-language text classification, image-caption generation in multiple languages and caption generation for images containing unseen visual object categories.

ACKNOWLEDGMENTS

"*Starting strong is good. Finishing strong is epic.*" – **Robin Sharma**.

The successful completion of this thesis would not have been possible without the support of several people.

First, I would like to express my gratitude to Prof. Dr. Rudi Studer for giving me an opportunity to join his research group at the Institute AIFB. His skill of asking challenging questions has further shaped the topics of my research. In addition, I am grateful to PD Dr. Achim Rettinger and Prof. York Sure-Vetter who gave unprecedented support in the research projects and immense freedom to explore topics of my interest. Without their guidance, it is hard to imagine the completion of this thesis today.

I am thankful for my co-referent, Prof. Dr. Volker Tresp who readily agreed to read my dissertation. Also, I am happy that Prof. Dr. Lexing Xie and her group hosted me at the Australian National University in Canberra. I was warmly welcomed there and was provided with all the resources to conduct my research.

My thanks to all colleagues and friends from the *Yordiverse* who has made the Ph.D. journey a truly remarkable experience. Also, I want to thank my family for their support during this time.

Last but not least, I am deeply indebted to my wife Manasa, who has given constant encouragement and was always there with me in bad and good moments to whom I dedicate this work.

PUBLICATIONS

The text as well as the pictures which are part of the thesis have been already published or under preparation. The thesis is written based on the following papers:

[1] Mogadala, Aditya and Rettinger, Achim. **Multimodal Correlated Centroid Space for Multilingual Cross-Modal Retrieval**. Proceedings of the European Conference on Information Retrieval (ECIR), 2015.

[2] Mogadala, Aditya and Rettinger, Achim. **Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-Language Text Classification**. Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2016.

[3] Mogadala, Aditya and Bista, Umanga and Xie, Lexing and Rettinger, Achim, **Knowledge Guided Attention and Inference for Describing Images Containing Unseen Objects**. Proceedings of the Extended Semantic Web Conference (ESWC), 2018.

Some of the ideas of this thesis have been presented in the doctoral consortium

[4] Mogadala, Aditya **Polylingual Multimodal Learning**. Proceedings of the Doctoral Consortium at The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2015.

Research contributions during the Ph.D. study that are not part of this thesis:

[5] Paul, Christian and Rettinger, Achim and Mogadala, Aditya and Knoblock, Craig A. and Szekely, Pedro. **Efficient graph-based document similarity**. Processing of the Extended Semantic Web Conference (ESWC), 2016. **Best Research Paper Nominee**.

[6] Zhang, Lei and Thalhammer, Andreas and Rettinger, Achim and Färber, Michael and Mogadala, Aditya and Denaux, Ronald. **The xLiMe system: Cross-lingual and cross-modal semantic annotation, search and recommendation over live-TV, news and social media streams**. Web Semantics: Science, Services and Agents on the World Wide Web , 2017.

[7] Mogadala, Aditya and Kanuparthi, Bhargav and Rettinger, Achim and Sure-Vetter, York. **Discovering Connotations as Labels for Weakly Supervised Image-Sentence Data**. The Web Conference (Cognitive Computing Track), 2018.

[8] Mogadala, Aditya and Jung, Dominik and Rettinger, Achim. **Linking Tweets with Monolingual and Cross-Lingual News using Transformed Word Embeddings**. International Journal of Computational Linguistics and Applications (IJCLA), To Appear.

CONTENTS

INTRODUCTION

1

INTRODUCTION

Context of this Thesis. In this thesis, we are interested in understanding heterogeneous and homogeneous content which has multiple views for any given instance. This kind of data can support several applications in diverse domains. We show that our research about building multi-view representation learning models are useful for tasks such as cross-modal retrieval, cross-language text classification, consistent multi-language image caption generation and generation of a caption for those images which contain unseen visual object categories.

1.1 MOTIVATION

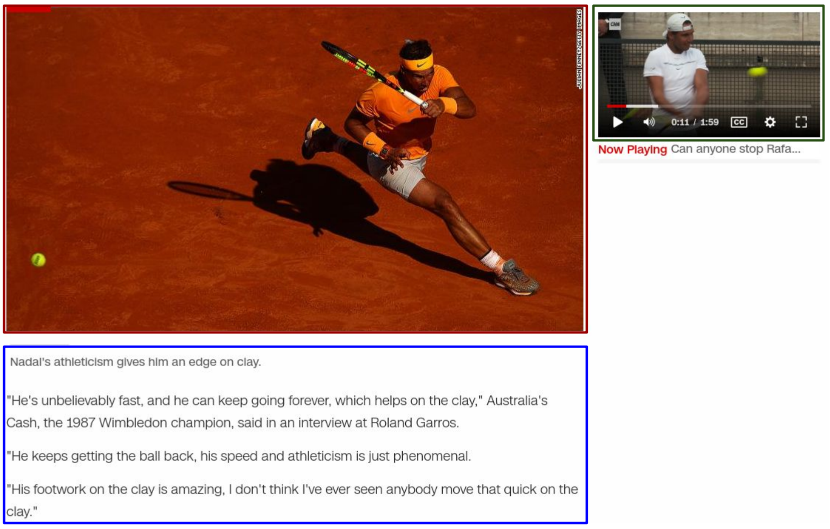
1.1.1 *Heterogeneous and Homogeneous Content*

There is tremendous growth in the usage of the Web over past decade. Almost 50% of the world population is already online[1] using different devices such as desktops and mobiles to disseminate and access content. Most of the online content created and obtained by individuals and enterprises belong to *varied forms* and serve different real-world objectives such as content search, content categorization and content generation. However, varied forms of the content are highly unstructured [74] and require machine comprehensible representation to facilitate such objectives.

> ✎ **Example 1**
>
> Consider an example page from the news website[a] in the Figure 1 which contain unstructured content in varied forms. A human reader understands that this page is about Tennis player *Rafael Nadal*[b] containing a textual description (Blue), image (Red Berry), and a video (Dark Green). However, for a machine, it is hard to comprehend such unstructured content in varied forms without sophisticated processing.
>
> [a] https://edition.cnn.com/
>
> [b] https://en.wikipedia.org/wiki/Rafael_Nadal

When we deep dive and infer about varied forms of the unstructured content observed in the page, we understand that they represent various *modalities*. To apprehend what are modalities, we present their clip art representation in the Figure 2. From the left, clip art in the first position represents a video modality and the second position denote audio. Similarly, modality in the third position

[1] https://thenextweb.com/insights/2017/01/24/digital-trends-2017-report-internet/



Nadal's athleticism gives him an edge on clay.

"He's unbelievably fast, and he can keep going forever, which helps on the clay," Australia's Cash, the 1987 Wimbledon champion, said in an interview at Roland Garros.

"He keeps getting the ball back, his speed and athleticism is just phenomenal.

"His footwork on the clay is amazing, I don't think I've ever seen anybody move that quick on the clay."

Figure 1: News article containing textual description (Blue), Image (Red Berry), and a Video (Dark Green).

represents an image, and the fourth position denotes text. The last position denotes linked data graph (e.g., DBpedia[2]) constituting entities and their relationships.

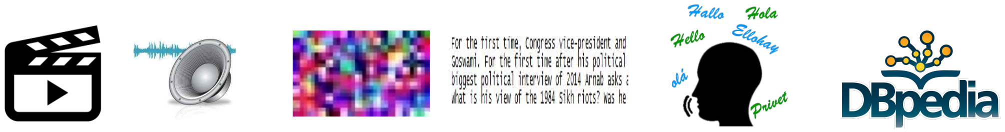

For the first time, Congress vice-president and Goswami. For the first time after his political biggest political interview of 2014 Arnab asks a what is his view of the 1984 Sikh riots? Was he
Hallo
Hola
Hello
Ellohay
Olá
Privet
DBpedia

Figure 2: Varied forms of the content denoted by different modalities (video, audio, image, text in different languages and entity-relationship graph).

As mentioned earlier, modalities are highly *unstructured* as they do not fit neatly into a relational database. Also, as pointed in many surveys[3] that the growth of unstructured content is more significant than the structured data, which is attributable to the ease of creation of unstructured content in contrast to the structured content.

✎ **Example 2**

In the Figure 3, we present an example comparing the structured vs. unstructured content. Most of the structured content represents discrete rows and columns with storage denoting relational databases (DB) and tables (e.g., spreadsheets). While, unstructured data usually have an unmanaged file structure having better semantics and require storage which is NoSQL[a] (e.g., MongoDB).

[a] https://en.wikipedia.org/wiki/NoSQL

Now to address the problem of representing each modality in a machine comprehensible manner. Automatic learning of representations from each modality

[2] http://wiki.dbpedia.org/
[3] http://blog.aylien.com/rapidminer-wisdom-recap-the-value-in-analyzing/

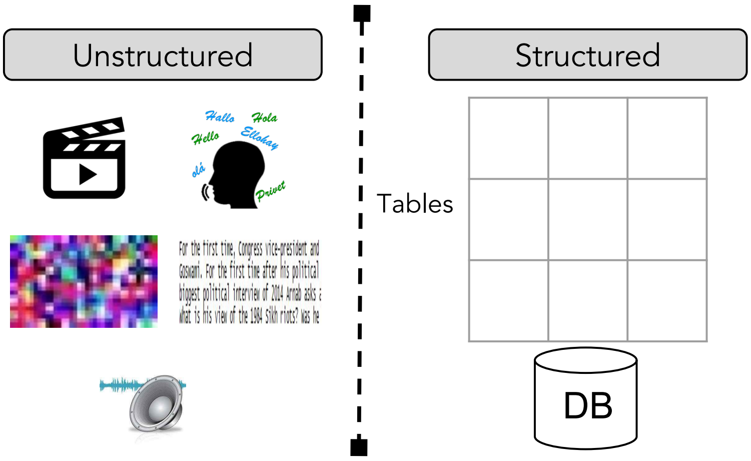


Figure 3: An example distinguishing structured against unstructured data.

is widely adopted with representation learning [25]. Howbeit, modalities do not occur in isolation, but they usually co-exist and blend between themselves to add multiple *views* in representing a topic or concept, thus making the content either *homogeneous* or *heterogeneous*. A homogeneous instance contains views from the same modality while heterogeneous instance contains views from different modalities.

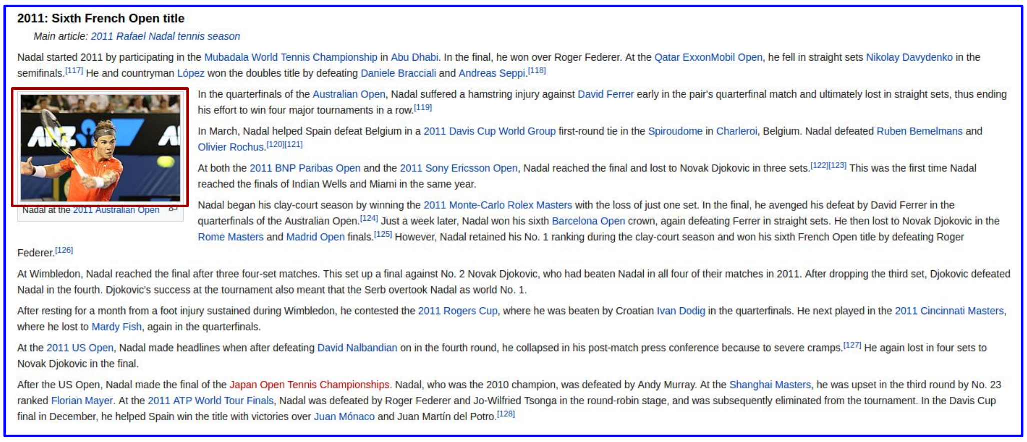
➳ Definition 1: View

A single view is a modality represented by either image, text, video or audio.

Research conducted earlier [156] have shown that leveraging *homogeneous* or *heterogeneous* content can exploit information that is more expressive than that of single view to learn representations having wide applicability. We find such content footprints on the Web at many places like news websites, e-commerce web pages, social media sites (SNS) (e.g., Facebook[4], Twitter[5], Pinterest[6]), Wikipedia[7], video streaming platforms (e.g., YouTube[8] and Vimeo[9]) and audio streaming platforms (e.g., SoundCloud[10], Spotify[11]). Example 3 and Example 4 shows sample of heterogeneous and homogeneous content respectively.

> **✎ Example 3**
>
> In the Figure 4, we present an example of *heterogeneous* content as observed in Wikipedia. It provides the picture of *Rafael Nadal* at French Open[a] and a write-up about his participation in it.
>
> [a] https://en.wikipedia.org/wiki/French_Open

[4] https://www.facebook.com/
[5] https://twitter.com/
[6] https://www.pinterest.com/
[7] https://www.wikipedia.org/
[8] https://www.youtube.com/
[9] https://vimeo.com/
[10] https://soundcloud.com/
[11] https://www.spotify.com



2011: Sixth French Open title

Main article: 2011 Rafael Nadal tennis season

Nadal started 2011 by participating in the Mubadala World Tennis Championship in Abu Dhabi. In the final, he won over Roger Federer. At the Qatar ExxonMobil Open, he fell in straight sets Nikolay Davydenko in the semifinals.[117] He and countryman López won the doubles title by defeating Daniele Bracciali and Andreas Seppi.[118]

Nadal at the 2011 Australian Open

In the quarterfinals of the Australian Open, Nadal suffered a hamstring injury against David Ferrer early in the pair's quarterfinal match and ultimately lost in straight sets, thus ending his effort to win four major tournaments in a row.[119]

In March, Nadal helped Spain defeat Belgium in a 2011 Davis Cup World Group first-round tie in the Spiroudome in Charleroi, Belgium. Nadal defeated Ruben Bemelmans and Olivier Rochus.[120][121]

At both the 2011 BNP Paribas Open and the 2011 Sony Ericsson Open, Nadal reached the final and lost to Novak Djokovic in three sets.[122][123] This was the first time Nadal reached the finals of Indian Wells and Miami in the same year.

Nadal began his clay-court season by winning the 2011 Monte-Carlo Rolex Masters with the loss of just one set. In the final, he avenged his defeat by David Ferrer in the quarterfinals of the Australian Open.[124] Just a week later, Nadal won his sixth Barcelona Open crown, again defeating Ferrer in straight sets. He then lost to Novak Djokovic in the Rome Masters and Madrid Open finals.[125] However, Nadal retained his No. 1 ranking during the clay-court season and won his sixth French Open title by defeating Roger Federer.[126]

At Wimbledon, Nadal reached the final after three four-set matches. This set up a final against No. 2 Novak Djokovic, who had beaten Nadal in all four of their matches in 2011. After dropping the third set, Djokovic defeated Nadal in the fourth. Djokovic's success at the tournament also meant that the Serb overtook Nadal as world No. 1.

After resting for a month from a foot injury sustained during Wimbledon, he contested the 2011 Rogers Cup, where he was beaten by Croatian Ivan Dodig in the quarterfinals. He next played in the 2011 Cincinnati Masters, where he lost to Mardy Fish, again in the quarterfinals.

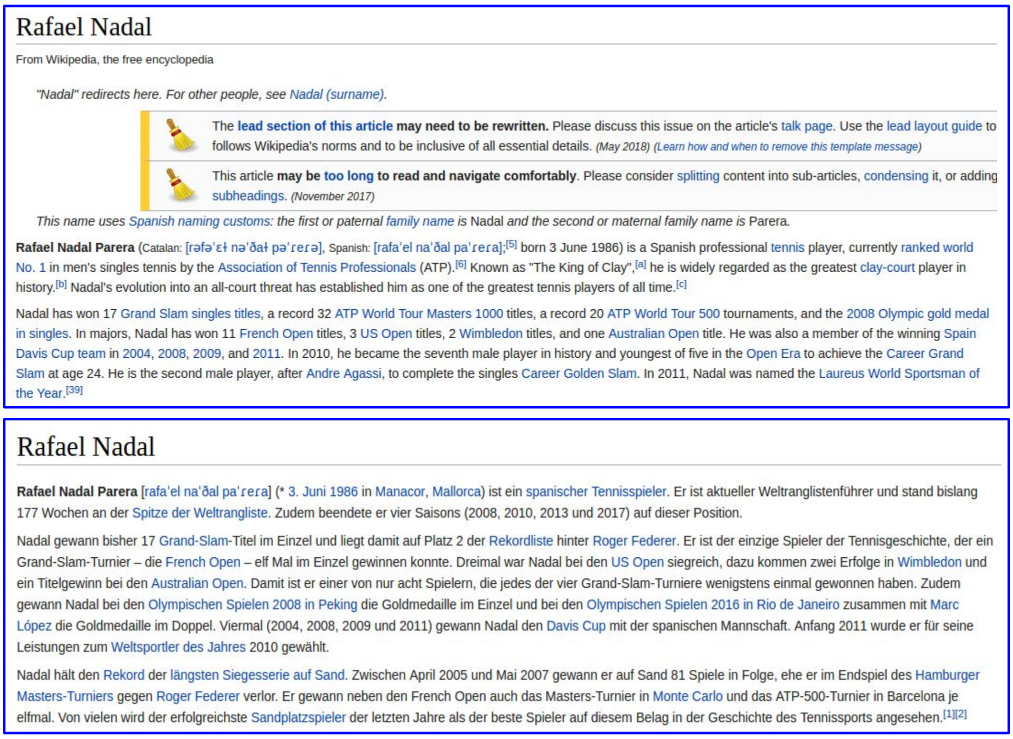
At the 2011 US Open, Nadal made headlines when after defeating David Nalbandian on in the fourth round, he collapsed in his post-match press conference because to severe cramps.[127] He again lost in four sets to Novak Djokovic in the final.

After the US Open, Nadal made the final of the Japan Open Tennis Championships. Nadal, who was the 2010 champion, was defeated by Andy Murray. At the Shanghai Masters, he was upset in the third round by No. 23 ranked Florian Mayer. At the 2011 ATP World Tour Finals, Nadal was defeated by Roger Federer and Jo-Wilfried Tsonga in the round-robin stage, and was subsequently eliminated from the tournament. In the Davis Cup final in December, he helped Spain win the title with victories over Juan Mónaco and Juan Martín del Potro.[128]

Figure 4: An example showing image and textual modalities about a common topic.

✎ **Example 4**

In the Figure 5, we present an example of *homogeneous* content as observed in Wikipedia. It provides a write-up in different languages (English and German) about *Rafael Nadal*.



Rafael Nadal

From Wikipedia, the free encyclopedia

"Nadal" redirects here. For other people, see Nadal (surname).

The **lead section of this article may need to be rewritten.** Please discuss this issue on the article's talk page. Use the lead layout guide to follows Wikipedia's norms and to be inclusive of all essential details. *(May 2018) (Learn how and when to remove this template message)*

This article **may be too long to read and navigate comfortably**. Please consider splitting content into sub-articles, condensing it, or adding subheadings. *(November 2017)*

This name uses Spanish naming customs: the first or paternal family name is Nadal *and the second or maternal family name is* Parera.

Rafael Nadal Parera (Catalan: [rəfəˈɛɫ nəˈðaɫ pəˈɾeɾə], Spanish: [rafaˈel naˈðal paˈɾeɾa];[5] born 3 June 1986) is a Spanish professional tennis player, currently ranked world No. 1 in men's singles tennis by the Association of Tennis Professionals (ATP).[6] Known as "The King of Clay",[a] he is widely regarded as the greatest clay-court player in history.[b] Nadal's evolution into an all-court threat has established him as one of the greatest tennis players of all time.[c]

Nadal has won 17 Grand Slam singles titles, a record 32 ATP World Tour Masters 1000 titles, a record 20 ATP World Tour 500 tournaments, and the 2008 Olympic gold medal in singles. In majors, Nadal has won 11 French Open titles, 3 US Open titles, 2 Wimbledon titles, and one Australian Open title. He was also a member of the winning Spain Davis Cup team in 2004, 2008, 2009, and 2011. In 2010, he became the seventh male player in history and youngest of five in the Open Era to achieve the Career Grand Slam at age 24. He is the second male player, after Andre Agassi, to complete the singles Career Golden Slam. In 2011, Nadal was named the Laureus World Sportsman of the Year.[39]

Rafael Nadal

Rafael Nadal Parera [rafaˈel naˈðal paˈɾeɾa] (* 3. Juni 1986 in Manacor, Mallorca) ist ein spanischer Tennisspieler. Er ist aktueller Weltranglistenführer und stand bislang 177 Wochen an der Spitze der Weltrangliste. Zudem beendete er vier Saisons (2008, 2010, 2013 und 2017) auf dieser Position.

Nadal gewann bisher 17 Grand-Slam-Titel im Einzel und liegt damit auf Platz 2 der Rekordliste hinter Roger Federer. Er ist der einzige Spieler der Tennisgeschichte, der ein Grand-Slam-Turnier – die French Open – elf Mal im Einzel gewinnen konnte. Dreimal war Nadal bei den US Open siegreich, dazu kommen zwei Erfolge in Wimbledon und ein Titelgewinn bei den Australian Open. Damit ist er einer von nur acht Spielern, die jedes der vier Grand-Slam-Turniere wenigstens einmal gewonnen haben. Zudem gewann Nadal bei den Olympischen Spielen 2008 in Peking die Goldmedaille im Einzel und bei den Olympischen Spielen 2016 in Rio de Janeiro zusammen mit Marc López die Goldmedaille im Doppel. Viermal (2004, 2008, 2009 und 2011) gewann Nadal den Davis Cup mit der spanischen Mannschaft. Anfang 2011 wurde er für seine Leistungen zum Weltsportler des Jahres 2010 gewählt.

Nadal hält den Rekord der längsten Siegesserie auf Sand. Zwischen April 2005 und Mai 2007 gewann er auf Sand 81 Spiele in Folge, ehe er im Endspiel des Hamburger Masters-Turniers gegen Roger Federer verlor. Er gewann neben den French Open auch das Masters-Turnier in Monte Carlo und das ATP-500-Turnier in Barcelona je elfmal. Von vielen wird der erfolgreichste Sandplatzspieler der letzten Jahre als der beste Spieler auf diesem Belag in der Geschichte des Tennissports angesehen.[1][2]

Figure 5: An example constituting textual modality in different languages (Top: English, Bottom: German).

1.1.2 *Heterogeneous and Homogeneous Content Applications*

1.1.2.1 *Overview*

Many applications can be build by leveraging heterogeneous and homogeneous content. However, in this thesis, we target only that content which has the **image** and **textual** modalities. Also, we confine ourselves to following tasks:

① **Content Search**: It aims at satisfying information need of an end-user. That is, a user expresses an information need as an input query and the retrieval engine attempts to discover data elements, which are assumed to satisfy that need. However, when the content is heterogeneous (e.g., image and text), input query and the retrieved data elements belong to different modalities, thus making the discovery of such relevant elements hard. Retrieval across modalities can be supported by building representations that share information across modalities and help to improve heterogeneous content-based search significantly.

A similar challenge is also seen when the content is homogeneous, for instance, if the input query and the retrieved data elements belong to different languages, it makes the discovery of relevant elements hard. Retrieval across languages can be supported by building representations that share information across languages and help to improve homogeneous content-based search significantly.

② **Content Categorization**: The goal is to classify content into predefined labels. For the heterogeneous content, label and the given data element belong to different modalities. However, for the homogeneous content, the data element and label belong to the same modality. A classical case for the heterogeneous content is discovering textual labels for the images while for the homogeneous content is learning label information from one language to port that knowledge to another language.

③ **Content Generation**: The goal is to generate one modality from another if the content is heterogeneous while generating one appearance from another if the content is homogeneous. In general, transformation of one form into another is a hard task to achieve.

In the Figure 6, the overall architecture is illustrated. Intuitively, it can be observed that first, we need to find the combination of modalities that we want to handle. Next, depending on the modalities selected, views of the heterogeneous or homogeneous content is determined. Furthermore, using the multi-view content, we produce machine comprehensible representations with two different techniques, i.e., correlation and common space learning to serve our applications.

1.1.2.2 *Challenges*

A significant challenge in the Figure 6 is creating a representation of the heterogeneous and homogeneous content. A useful representation captures the inherent meaning from each view and builds a shared representation from all existing

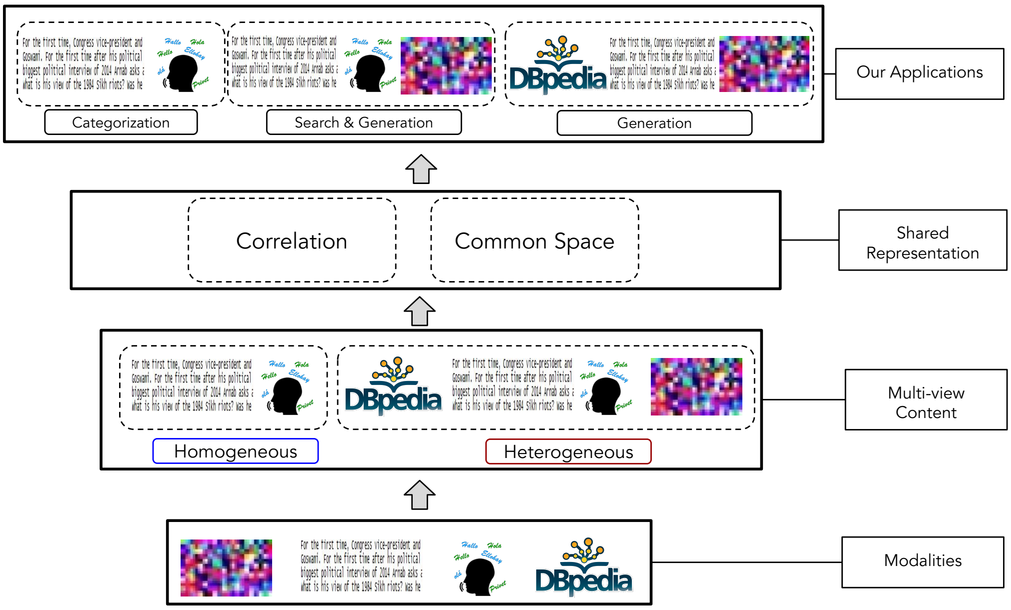


Figure 6: Overall Architecture.

views, which is an essential requirement for the applications mentioned above. Thus, the systems built for each of these applications must address the problem of *how to effectively integrate multiple views in heterogeneous or homogeneous content depicting various languages or modalities into a shared space representation*.

> ϟ **Problem 1: Unifying Multiple Views of Content (i.e., heterogeneous or homogeneous) by Identifying their Correlations**
>
> Build a shared space by identifying correlation among content.

One way to integrate multiple views is by finding their correlation among content. A way to find the correlation is by calculating the joint dimensionality reduction of the homogeneous or heterogeneous content representations. In fact, applications are also depended on the reduced dimensionality of representation which incorporates shared knowledge across content sources. Thus, systems that are built based on the correlated representations allow for a trade-off between result accuracy and computation time. Hence, only shallow representation is built from the multiple views with less computational intensiveness.

> ϟ **Problem 2: Unifying Multiple Views of Content (i.e., heterogeneous or homogeneous) by Building their Common Space**
>
> Build a shared space by mapping content into a common space.

As presented in Section 1.1.2.1, there are several problems in dealing with homogeneous or heterogeneous content. However, in this thesis we are solely concerned with Problem 1 and Problem 2. We also discuss the scope of this thesis in Section 1.3.2.

In the next paragraphs, let us briefly introduce the fields of research concerned with above problems.

1.1.2.3 *Problem 1: Unifying Multiple Views of Content with Correlation*

Unifying multiple views emerging from either homogeneous or heterogeneous content by identifying their *correlation*, is especially crucial for supporting the application of **content search**. Here, the goal is to compute top-ranked query results, given the query and results from either different modalities or languages.

Unifying multiple views of the content target this problem by building a shallow representation of views and then finding the *correlation* across views such that it supports ranking across views. More specifically, these strategies allow retrieval engines to compute ranked results, with the textual modality of variable length. It can lead to significant efficiency across languages for the content-based search – as we will show in Chapter 4.

1.1.2.4 *Problem 2: Unifying Multiple Views of Content with Common Space*

Identifying correlations across multiple views is sometimes not possible due to the large size of datasets representing homogeneous or heterogeneous content. Here, building a *common space* representation is efficient concerning scalability and also crucial for supporting content categorization and generation.

We employ both shallow and deep neural networks to target the Problem 2. They comprise a set of techniques, which allow scaling to larger datasets containing multiple views by providing effective optimization [47]. Neural networks are also found to be effective in representing modalities in the machine comprehensible manner. Recently, different architectures of the neural networks are employed by NLP, CV and semantic web communities for various challenges.

Generally speaking, we consider shallow and deep neural networks regarding two dimensions:

① **Content Categorization**

First, we consider shallow neural networks as an application to the categorization of homogeneous content where views arise from different languages. This way, joint modeling of languages is implemented to create a common space representation – as we will show in Chapter 5.

② **Content Generation**

Second, leveraging shallow neural networks for the generation of heterogeneous content is not adequate. Hence, we employ deep neural networks to generate heterogeneous content. More specifically, we target the problem of generation of textual descriptions for images in multiple languages and also those images with novel visual object categories – as we will show in Chapter 6 and Chapter 7 respectively.

1.2 Heterogeneous and Homogeneous Content Characteristics

In the section mentioned above, challenges that need to be addressed in creating the representations from the heterogeneous and homogeneous content is reviewed. The further analysis shows that the content can have additional characteristics. Based on our overall aim and the research questions that we want to contribute in this thesis, we split the essential characteristics broadly into four different categories. The Figure 7 shows the summary of different characteristics of the heterogeneous and homogeneous content which we encountered in our work.

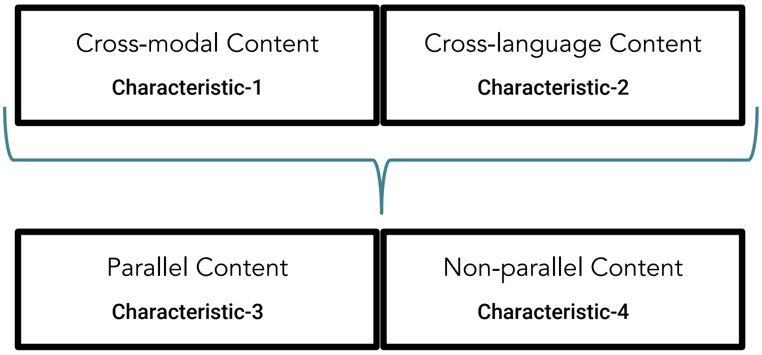


Figure 7: Characteristics of the heterogeneous and homogeneous content used in this thesis. The top row shows cross-modal and cross-language content which can be either parallel or non-parallel as shown in the bottom row.

1.2.1 *Characteristic 1: Cross-modal Content*

A heterogeneous content containing any combination of modalities presented in the Figure 2 is considered *cross-modal*. A sample scenario considering non-aligned image and textual modality is presented in the Figure 4. However, cross-modal content can exist in either parallel or non-parallel setup. There also exist other variations of the cross-modal content, where a view can have more than one example from the second view (e.g., multi-label textual annotation of an image).

Usually, homogeneous content containing is not considered cross-modal as both views emerge from the same modality.

1.2.2 *Characteristic 2: Cross-language Content*

A homogeneous content containing textual modality in different languages is considered cross-language. Similar to the cross-modal content, cross-language content can also exist in either parallel or non-parallel setup.

There is no possibility of cross-modal content being cross-language as views emerge from different modalities. However, cross-language content can still align a modality different than the textual modality to become both cross-language and cross-modal.

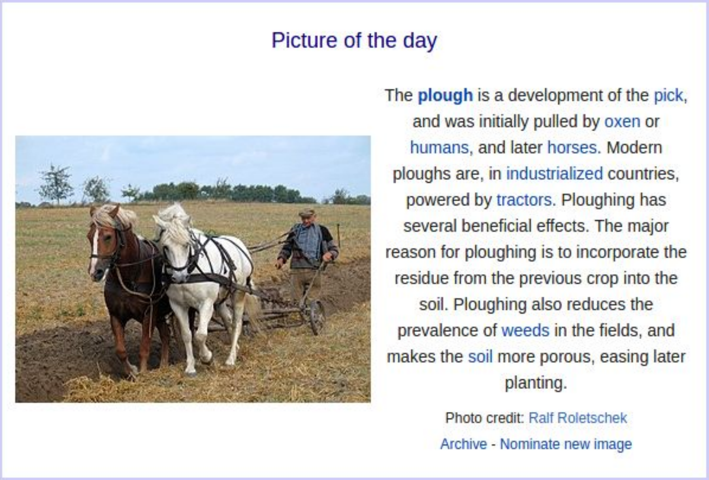
1.2.3 *Characteristic 3: Parallel Content*

The notion of content being *parallel* emerge from the corpus analysis studied as a part of the corpus linguistics [169]. Cross-language or cross-modal content is considered parallel if it satisfies the following definition.

➳ Definition 2: Parallel Content

Any cross-modal or cross-language content that is specially formatted for side-by-side comparison or alignment is referred to as parallel content.

In the Figure 8, we show an example page acquired from Wikipedia picture of the day[12] constituting cross-modal content which is parallel.



Picture of the day

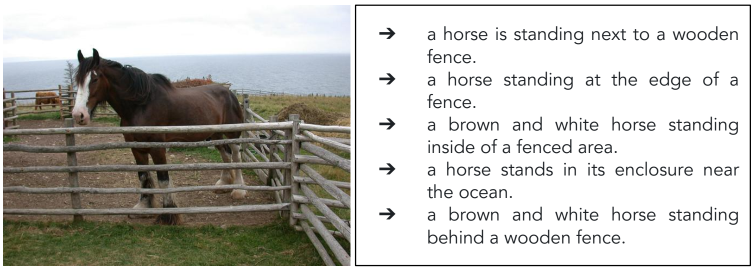
The **plough** is a development of the pick, and was initially pulled by oxen or humans, and later horses. Modern ploughs are, in industrialized countries, powered by tractors. Ploughing has several beneficial effects. The major reason for ploughing is to incorporate the residue from the previous crop into the soil. Ploughing also reduces the prevalence of weeds in the fields, and makes the soil more porous, easing later planting.

Photo credit: Ralf Roletschek

Archive - Nominate new image

Figure 8: An example page constituting parallel cross-modal content.

However, datasets which are annotated by human annotators to create a perfect alignment between either cross-modal or cross-language content is usually a prerequisite for many approaches. A typical example of the annotated cross-modal content that is parallel is image-caption pairs. In the Figure 8, we show an example acquired from the MSCOCO dataset [13] constituting image and its five parallel textual descriptions (i.e., captions).



➔ a horse is standing next to a wooden fence.
➔ a horse standing at the edge of a fence.
➔ a brown and white horse standing inside of a fenced area.
➔ a horse stands in its enclosure near the ocean.
➔ a brown and white horse standing behind a wooden fence.

Figure 9: Image and its parallel captions.

Similarly, a potential scenario where cross-language content is parallel is the precise translation of sentences existing in two different languages. This kind of data plays a critical role in building automatic translation system between two languages.

[12] https://en.wikipedia.org/wiki/Wikipedia:Picture_of_the_day

[13] http://cocodataset.org/#home

1.2.4 *Characteristic 4: Non-parallel Content*

Opposite of the *parallel* content is the *non-parallel* content. Cross-modal or cross-language content is considered non-parallel if it satisfies the following definition.

➾ Definition 3: Non-parallel Content

Any cross-modal or cross-language content that has no direct alignment or side-by-side comparison between intra- or inter-modalities is referred to as the non-parallel content.

However, in some cases, the content can be comparable, i.e., build from bilingual documents which are conceptually aligned. For instance, Wikipedia pages from two different languages describing same concept/topic are considered comparable as shown in the Figure 5. The situation perseveres for cross-modal content as well, where modalities are not the direct translation of each other, but there exists a weak alignment between them. For example, usually on the social media platforms, images are tagged with either hashtags or text with variable length. Sometimes this may not wholly depict the visual content present in an image. However, labels partially describe the visual content.

Heterogeneous and homogeneous content may have several other characteristics, which may motivate additional research questions. However, in this thesis, we concentrate on the above characteristics and present the overall scope of this thesis in the Section 1.3.2.

1.3 RESEARCH QUESTIONS AND SPAN

In this section, we present the span and research questions that will be addressed in the thesis.

1.3.1 *Research Questions*

Based on problems mentioned above, our overall research question is:

✍ Overall Research Question

How to unify subset of text, Entity Relationship graph, and image modality representing languages, relational knowledge, and vision respectively into a shared representation to assist homogeneous or heterogeneous content search, categorization, and generation.

Given the characteristics of heterogeneous and homogeneous content in Section 1.2, the overall question breaks down into several research questions, which we target in Chapter 4, Chapter 5, Chapter 6 and Chapter 7. An overview of ad-

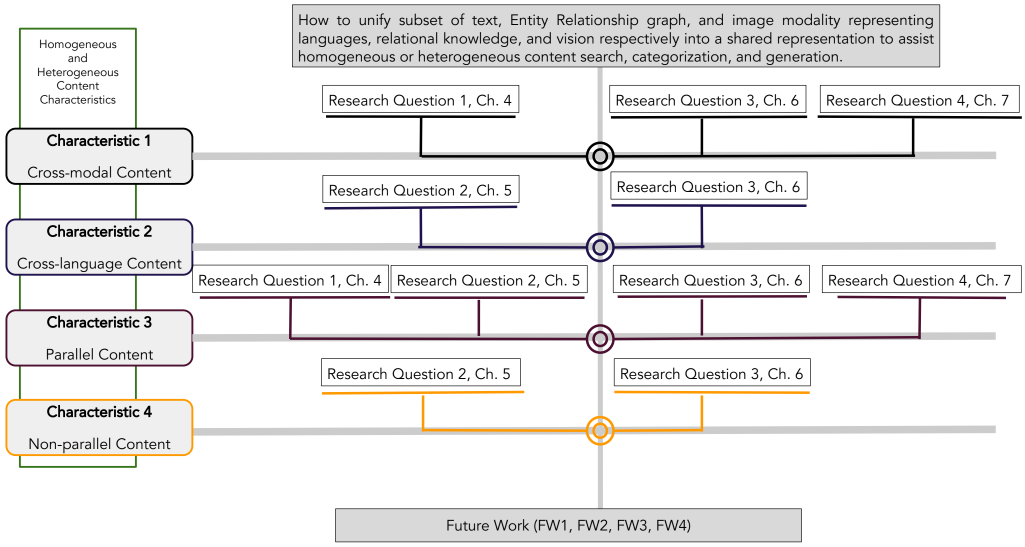

How to unify subset of text, Entity Relationship graph, and image modality representing languages, relational knowledge, and vision respectively into a shared representation to assist homogeneous or heterogeneous content search, categorization, and generation.
Homogeneous and Heterogeneous Content Characteristics
Characteristic 1
Cross-modal Content
Characteristic 2
Cross-language Content
Characteristic 3
Parallel Content
Characteristic 4
Non-parallel Content
Research Question 1, Ch. 4
Research Question 3, Ch. 6
Research Question 4, Ch. 7
Research Question 2, Ch. 5
Research Question 3, Ch. 6
Research Question 1, Ch. 4
Research Question 2, Ch. 5
Research Question 3, Ch. 6
Research Question 4, Ch. 7
Research Question 2, Ch. 5
Research Question 3, Ch. 6
Future Work (FW1, FW2, FW3, FW4)

Figure 10: Overview of heterogeneous and homogeneous content characteristics, and research questions, which are addressed in this thesis.

dressed research questions and heterogeneous and homogeneous content characteristics are depicted in the Figure 10.

> ✍ **Research Question 1**
>
> Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist *search* by finding correlation among their input representations?

Research Question 1 is driven by correlation principles of the heterogeneous content and aims at supporting application of *search* with cross-modal retrieval. Notably, our task is to process *parallel* cross-modal content emerging from two views especially images and different languages text to learn their correlations.

We proposed a novel approach named **correlated centroid space** for this task and extended previous works based on subspace learning to learn correlations across parallel cross-modal content in the Chapter 4.

> ✍ **Research Question 2**
>
> Given two different views of homogeneous content depicting text from different languages, how can we build a shared representation to assist *categorization* by learning a common space by capturing regularities?

Here, we leverage homogeneous content where views emerge from two different languages of the textual modality. We addressed this Research Question 2 by learning a *common space* representation from both parallel or non-parallel cross-language content. The shared representation built bilingual distributed word representations, i.e., embeddings which learned regularities across languages. It has played a crucial role in supporting cross-language textual classification tasks as shown in the Chapter 5.

✍ **Research Question 3**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation of all views if an auxiliary view depicting text in multiple languages is added to assist the *generation* of text from an image?

Adding an auxiliary view to the already existing views of the training instances of heterogeneous or homogeneous content has two inferences from the content prospect.

① Does the auxiliary view provide a novel modality?

② Does it match the modality of existing views?

We address the Research Question 3 by adding an auxiliary view matching the modality of an existing view in the Chapter 6. Our approach proposes to learn a *common space* of all views and further use it to generate text given an image. For this, we leverage deep neural networks in the multi-task learning [37] setting to jointly optimize three different views for delivering consistency among textual descriptions generated across languages.

✍ **Research Question 4**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist the *generation* of text from an image if there are missing views?

Research Question 4 aims to address the learning from those instances which contain missing views (e.g., missing modalities) and is discussed in the Chapter 7. Our approach for handling missing views is achieved with external guidance. To be specific, we leverage deep neural networks augmented with relational knowledge. It is evaluated on the task of textual description generation for images containing visual object categories that are unseen during the training phase.

1.3.2 *Span of this Thesis*

MVRL has been addressed before in many contexts by various studies [156]. They have leveraged modalities representing two different views to learn a *shared representation*. In particular, the recent dissertations concentrated separately on the language and vision [76] problems or addressed only language-specific challenges. There are also few works [299] which summarized MVRL from a theoretical perspective. In contrast to works as mentioned earlier, we will not focus purely on a theoretical perspective or concentrate separately on specific modalities.

Generally speaking, we target the above research questions in Chapter 4, Chapter 5, Chapter 6 and Chapter 7. Concerning those questions, this thesis provides several novel contributions – as we will outline in the next section.

1.4 CONTRIBUTIONS

About the research questions mentioned above, this thesis provides the following contributions:

☞ **Contribution for Research Question 1**

Cross-modal retrieval to assist content search by leveraging correlated centroid space.

Existing work for cross-modal retrieval was built using linear subspace learning approaches such as Canonical Correlation Analysis (CCA) [110], where the correlation between different modalities is captured. Also, they have leveraged only monolingual textual content to built representations which hinders the accessibility of retrieving cross-modal content in multiple languages.

In Chapter 4, we will show how to extend the kernel version of CCA technique (i.e., KCCA) to capture correlated centroid space which is based on our previous publication [1] and target the cross-modal content, i.e., image and textual content in multiple languages.

☞ **Contribution for Research Question 2**

Cross-language text classification to assist content categorization by leveraging Bilingual Paragraph Vectors.

Based on our work in [2], we present a novel bilingual word embeddings learning approach in the Chapter 5. For this, we combine shallow neural networks with the manifold alignment technique. More specifically, we extend existing work of paragraph vectors [144] to build bilingual word representations.

We propose two techniques to handle both parallel and non-parallel content for building bilingual word embeddings. On the one hand, we propose a model that operates on sentence-level parallel content. On the other hand, we propose an extension of the model which works with sentence-level parallel content to work with non-parallel content by leveraging manifold alignment technique.

☞ **Contribution for Research Question 3**

Consistent multi-language image caption generation to assist content generation given auxiliary views by leveraging multi-task attention.

We present an approach in the Chapter 6 which aims to jointly learn from images and their caption pairs in multiple languages. This work helps to reduce divergence across captions generated across languages. To achieve it, we propose deep neural network based attention models that leverage visual features extracted from images and their captions for optimizing a multi-task objective. Learned models are further used to generate captions for images which are highly consistent across languages.

☞ **Contribution for Research Question 4**

Unseen visual object categories caption generation to assist content generation given missing views by leveraging knowledge guided assistance.

Based on our work in [3], we present the knowledge-guided assistance to caption generation for images containing unseen visual object categories in Chapter 7.

Our approach combines ideas from fields such as semantic web and computer vision. First, knowledge graph entities are augmented with their learned embeddings and are also used to annotate images as labels. Furthermore, during training of the image caption generation model, entity embeddings are leveraged to capture attention from an image to calculate the attention weights w.r.t the caption words. Entity labels are also used, however only during the testing phase as a constrained inference. Overall, this approach for caption generation has shown to be compelling enough for scaling to visual object categories that usually lack parallel captions (i.e., missing views) during training.

1.5 OUTLINE

The remainder of this thesis comprises six chapters, which aim at Research Questions 1 - 4 and discuss Contributions 1 - 4.

❷ *Chapter 2 – Foundations*
In the Chapter 2, we provide foundations to our approaches presented in Chapter 4, Chapter 5, Chapter 6 and Chapter 7. Particularly, we introduce fundamentals of MVRL and outline its application for combining heterogeneous or homogeneous content to support several applications such as content search, categorization and generation.

❸ *Chapter 3 – MVRL with Two Views and Correlated Centroid Space*
In Chapter 4, we present a novel approach of MVRL to identify correlations across two-views depicting cross-modal content. For this, we extend the traditional subspace learning technique such as KCCA by adding class specific clustering information such that it correlates semantically similar cross-modal content closer to each other in the shared representation. This work has shown to improve search, especially image retrieval given a textual query and vice versa.

❹ *Chapter 4 – MVRL with Two views and Co-regularization*
We introduce a novel approach to build bilingual embeddings with two views depicting cross-language content in the Chapter 5. For this, we leverage manifold alignment theory and shallow neural networks to project two different languages that are either parallel or non-parallel into a shared representation such that linguistic regularities among them are construable. This work has shown to improve categorization, especially cross-language text classification.

❺ *Chapter 5 – MVRL with Auxiliary Views and Joint Multi-Task Optimization*
In Chapter 6, we introduce a novel approach for image caption generation in multiple languages using more than two views depicting both cross-modal and cross-language content. For this, we leverage multi-task learning and deep neural networks to propose a single model which can generate one modality from another. This work has shown to improve content

generation, especially image caption generation in multiple languages by making them consistent across languages.

❻ *Chapter 6 – MVRL with Missing Views and Knowledge Guided Assistance*
When there are missing views, standard MVRL approaches fail to predict during inference. In Chapter 7, we deal with the missing information in the heterogeneous content, particularly the image-caption parallel content. We introduce a novel approach for generating captions for those images containing visual object categories that are unseen in the training phase. For this, we leveraged entity labels and their embeddings as constraints and external semantic attention respectively for building an image caption generation model. This work has shown to improve content generation, especially scalable to larger visual object categories usually observed on the web.

❼ *Chapter 7 – Conclusion*
Last, we summarize our contributions and results in Chapter 8 and give an outlook on the future work.

FOUNDATIONS

2

FOUNDATIONS

Context of this Chapter. In this chapter, we discuss the preliminaries for remainder of the thesis. First, we present the challenges and advantages of representing homogeneous content with representation learning in Section 2.1 and later introduce the unified representation of heterogeneous and homogeneous content with MVRL in Section 2.2.

2.1 REPRESENTATION LEARNING

The goal of representation learning is to learn useful representations of the content. Efficient representations can identify and extricate the underlying multiple explanatory factors of variation behind the content [25]. Also, representation of the content supports machine learning applications for building useful prediction models. Recently, attaining representations is done with neural network-based approaches. However, Bayesian nonparametric methods [189] and other hierarchical graphical model-based approaches [125] have also shown the ability to learn rich representations of content. In the following, we review approaches used for learning representations and discuss their intricacies.

2.1.1 *Shallow Representation Learning*

We understand the shallow representation learning from the perspective of handcrafted feature extraction. For many machine learning applications such as speech recognition, NLP, and CV, feature extraction plays a key role in building predictive models. Over the past decade, researchers have spent ample amount of time in extracting and selecting relevant shallow features with several feature engineering techniques. However, there are some problems observed with shallow representations such as:

① They can be task specific and hard to generalize to other tasks.

② Cannot capture complex and highly structured dependencies observed in the content.

③ Require some prior knowledge about the content which can be helpful for discarding irrelevant features (e.g., feature selection).

④ They can be very inefficient regarding the number of computational units (e.g., bases, hidden units) [23], and also concerning examples required [24].

Depending on the application domain and data, handcrafted feature extraction methods vary. For example, scale-invariant feature transform (SIFT) [161] use histograms of gradient orientations for extraction of visual data (e.g., images) features. Similarly, shallow features such as term frequency-inverse document frequency (TF-IDF) [224] is used to rank relevant textual documents for a given query to support information retrieval. Natural Language Understanding (NLU) tasks supporting several applications [165] and divided based on syntax, semantics, discourse, and speech are also dependent on feature extraction methods. For example, tasks which fall under the umbrella of "semantics" such as lexical semantics [51], natural language generation [208] and natural language understanding [5] are dependent on the shallow handcrafted feature extraction methods for many years to leverage machine learning methods.

However, for learning useful representations, shallow representation learning usually combines feature extraction with dimensionality reduction techniques for selecting best features. Global or local methods [303] are techniques of dimensionality reduction, where global methods preserve global information of the content and local methods preserve the fundamental structure of high dimensional data in learned representations. In the following, we briefly describe existing dimensionality reduction techniques which are diversified based on local or global methods.

2.1.1.1 *Global Methods of Dimensionality Reduction*

Many global methods are proposed for selection of features and reducing the dimensionality of extracted features. Principal component analysis (PCA) [277] is one such method which makes linear dimensionality reduction by performing an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables. Similarly, other approaches such as Independent Component Analysis (ICA) [117], Linear Discriminant Analysis (LiDA) [213] also extract linear representations using both unsupervised and supervised techniques. However, linear features are insufficient for many real-world complex data scenarios, consequently demanding for more sophisticated techniques to get useful representations.

To assist such scenarios, linear methods were extended with non-linear techniques to include non-linearity in the extracted features. However, these techniques still provide shallow representations. Techniques such as kernel PCA (KPCA) [172] and Generalized Discriminant Analysis (GDA) [18] extends PCA and LiDA respectively for nonlinear dimensionality reduction using the kernel trick [182]. While methods such as Gaussian Process Latent Variable Model (GPLVM) [143] and Gaussian Process Latent Random Field (GPLRF) [302] are designed in a manner to produce low dimensional representations by learning a nonlinear mapping.

Although representations generated by global methods are useful for many tasks, need for local methods which preserve the fundamental structure of high dimensional data are also required.

2.1.1.2 *Local Methods of Dimensionality Reduction*

Local methods that make dimensionality reduction preserve the structure by using local information. Manifold learning [116] methods are significant contributors to the structure-preserving dimensionality reduction techniques by leveraging local information. However, there also exist other techniques such as Locality Preserving Projections (LLP) [99], Marginal Fischer Analysis [287] and Non-negative Matrix Factorization (NMF) [149] which perform linear dimensionality reduction by leveraging local information.

Sometimes, the content exists on the lower dimensional manifold, methods such as ISOMAP [251], local linear embedding (LLE) [218], Laplacian Eigenmap [20] and local tangent space alignment (LTSA) [298] assist in performing nonlinear dimensionality reduction by exploiting the local geometry around each data point. Local methods were proven to be useful for several applications such as hyperspectral image processing [284], face recognition [100], document processing [35] etc. Howbeit, they are still dependent on features which are handcrafted. Hence, this created a necessity for a combined automatic feature extraction and non-linear dimensionality reduction methods for effective representation of complex data.

2.1.2 *Deep Representation Learning*

The goal of deep representation learning is to automatically extract representations of the content by making the learning algorithms less dependent on manual feature engineering. However, these representations are not extracted in a shallow manner, but by stacking multiple nonlinear transformation layers above one another with a final goal of yielding more useful and abstract representations. These representations have achieved state of the art results in several domains such as speech recognition, computer vision, natural language processing and semantic web.

One significant advantage observed when dealing with representations is that they can express many general-purpose priors [25] such as smoothness, sparsity, coherence, and manifolds. However, representation learning is usually achieved with deep architectures which are often difficult to train effectively [70]. Nevertheless, deep architectures enable learning the hierarchy of features which are found to be exponentially more efficient as shown in several theoretical and empirical studies [227]. Also, deep representation learning provides abstract features at higher layers of representations which make them invariant to most local changes of the input. It also makes these representations behave as highly non-linear functions of the raw input [25]. A learned representation is expected to be good if it can distinguish between the related but distinct goals of learning invariant features and extricate explanatory factors for preserving relevant information needed for a specific task. However, it is often difficult to determine a priori which set of features are useful.

Many methods are proposed to learn deep representations. We broadly divide these methods into "*Layerwise pretraining*" and "*Joint training*" based on how new transformations are generated for features at each level of deep architectures.

The fundamental difference between "*Layerwise pretraining*" and "*Joint training*" is that the former learn a hierarchy of features one level at a time, while later train all levels jointly.

2.1.2.1 *Layerwise Pretraining*

The layerwise pretraining is a greedy approach to learn the hierarchy of features one level at a time of a deep architecture and can be applied in the unsupervised or supervised learning setting.

Unsupervised Setting: In this scenario, layerwise pretraining learns deep features which can be used as an initialization to the standard prediction algorithms (e.g., support vector machines (SVM) [101]), to supervised layers of a neural network or a generative model (e.g., deep Boltzmann machine (DBM) [223]). However, stacking of pretrained layers can be approached in several ways. In the following, we discuss some existing methods.

Deep Belief Network (DBN) [107] are built with pretrained Restricted Boltzmann Machine (RBM) [104]. However, a standard belief network (BN) is a directed acyclic graph composed of stochastic variables [192] which has a state of *zero* or *one*. Similarly, RBM consists of stochastic variables and one hidden layer with no connection between hidden units. The major difference between the standard BN and RBM is that the former have directed edges between units, while RBM consists of undirected edges. However, to build DBN, several RBM in a pretraining phase are stacked above one another with a directed connections between the layers and a feed-forward network (e.g., multi -layer perceptron (MLP)) is then used for fine-tuning.

Deep Boltzmann Machine (DBM) introduced by Salakhutdinov et al. [223] is a deep multi-layer Boltzmann machine (BM) where each layer captures higher-order correlations between the activities of hidden features in the layer below. It has undirected connections in contrast with DBN directed connections for the better flow of information [128]. Connections are present only between hidden units in adjacent layers, as well as between visible and the hidden units in the first hidden layer.

DBM has three major advantages.

① Similar to DBNs, the DBMs can learn internal representations by capturing complex structure in the higher layers. Usually, these high-level representations are built with a large amount of unlabeled data and very limited labeled dataset for fine-tuning the model for a specific discriminative task.

② States of variables in DBM can be initialized efficiently by bottom-up pass if DBMs are learned correctly.

③ As opposed to DBNs and other deep feature learning approaches, DBMs approximate inference procedure can incorporate top-down feedback after first bottom-up pass. It allows DBMs to better use higher-level knowledge to resolve uncertainty about intermediate feature representations.

Stacked Denoising Autoencoders (stacked-DAE) [260] are initialized as a deep network using denoising Autoencoders [259] in the same manner as RBMs stacked into DBNs. A denoising autoencoder (DAE) is a variant of autoencoders [23, 199] that are trained to reconstruct a "repaired" input from the corrupted version of it. It is achieved with an architecture that first corrupts the initial input into a different representation using stochastic mapping. Furthermore, the corrupted input is then mapped as in basic autoencoder into a hidden representation from which the reconstruction of it is again obtained. Although the input is corrupted, DAE still minimizes the same reconstruction loss between a uncorrupted input and its reconstruction.

In stacked-DAE, it has to be noted that the corruption of input is only performed for the initial denoising-training of each layer. Once the hidden representation is obtained, it will henceforth be used without corruption and is applied to produce the representation that will serve as clean input for training the next layer. Once the stacked encoder is built, its highest level output representation (i.e., pretrained) can be used as input to a standard supervised learning method (e.g., SVM) or can be leveraged for fine-tuning with a logistic regression layer added on top of it.

In general, the significant difference observed between Autoencoders and RBMs is that the Autoencoders consider the real-valued mean as their hidden representation whereas the stochastic RBMs sample a hidden binary representation from that mean. However, after their initial pretraining, the way layers of RBMs are typically used in practice may vary.

Supervised Setting: Observed in the previous sections show that training each layer of deep architectures is performed in an unsupervised manner. However, a surrogate method for building deep representations is to train in a supervised manner with the greedy and layer-wise approach. In the following, we discuss some of the existing methods.

Deep Network [23] is a multi-layer neural network, where each new hidden layer is trained as the hidden layer of a one-hidden-layer supervised neural network. Further, the output of the last of previously trained layers is considered as input by throwing away the output layer of the supervised neural network. The parameters of the hidden layer of a neural network are used as the pretraining initialization of the new top layer of the deep network.

However, in different settings, it is observed that the purely supervised greedy layer-wise pretraining performs significantly worse than the unsupervised greedy layer-wise pretraining. A possible explanation is that the greedy supervised procedure is too greedy: in the learned hidden units representation, it may discard some of the information about the target, information that cannot be captured easily by a one-hidden-layer neural network but could be captured by composing more hidden layers.

Deep-Structured CRF [293] leverage probabilistic methods [215] to build supervised pretraining by using outputs of the previous layer which can be fed an extra input for the next layer along with the raw input. Deep-structured CRF architecture is a hierarchy of linear-chain conditional random fields (CRFs) [139] that do not use state transition features. The observation sequence at any layer is augmented with both previous layer's observation sequence and marginal posterior probabilities. However, for many applications, the observation sequence of previous layer's and the present layer is usually enough.

Training of deep-structured CRF is achieved by fixing the trained lower-layer CRF parameters and then computing corresponding marginal posterior probabilities so that they can be further fed to the next layer. This process is continued until the model parameters of the highest or final layer of the model are optimized. The inference process is similar, howbeit with slight modifications. Deep-structured CRF make both train and inference on each layer independent which make the computational complexity linear w.r.t the layers used.

Context-Dependent Deep-Neural-Network Hidden Markov Model (CD-DNN-HMM) [52] pretrains in a supervised way all the previously added layers at each step of the iteration. CD-DNN-HMM hybrid architecture contains a context-dependent Hidden Markov Model (HMM) [202] combined with deep neural network (DNN). A generative model such as HMM has the observable features are assumed to be generated from a hidden Markov process that transitions between states. While, DNN is a conventional multi-layer perceptron (MLP) [216].

CD-DNN-HMM adopts a discriminative pretraining approach in contrast with DBNs to reduce inaccuracies in the pretraining. They achieve it with a layerwise backpropagation. Initially, a one-hidden-layer DNN is trained to full convergence using labels discriminatively with backpropagation. Then the softmax layer is replaced by another randomly initialized hidden layer, and again a new softmax layer is stacked on to the top so that it is again discriminatively trained to full convergence. This process is repeated until the required number of hidden layers is reached. However, this approach is similar to Deep Network [23], but differs with it by adding updates only from newly added hidden layers achieving accuracies close to those obtained with DBN pretraining.

Also, CD-DNN-HMM is closer to the standard ANN-HMM [210] architecture which replaces Gaussian mixtures with DNN and computes HMM's state emission likelihoods by converting state posteriors from the DNN to likelihoods.

2.1.2.2 *Joint Training*

In the section mentioned above, we have seen how the layerwise pretraining is leveraged in an unsupervised and supervised setting with different architectures to learn deep representations. The success of layerwise pretraining (unsupervised or supervised) can be attributed to intermediate representations, which provide a more natural way to learn intermediary versions rather than learning everything at once in a single go. Also, especially unsupervised pretraining additionally contributed to the regularization and optimization effect.

On the contrary, joint training of architectures which learn deep representations face challenges such as ill-conditioning (e.g., symmetry breaking in neural networks [180]) and local minima (e.g., optimization difficulty [25]). Few studies [84] also highlighted other problems with joint training, mainly in the context of neural networks and proposed few tricks for improvement. Howbeit, joint training also provide several advantages in a proper arrangement [45, 167, 85, 138, 246]. In the following, we explore advantages of architectures which learn deep representations in one go by overcoming challenges as mentioned earlier.

Convolutional Neural Network (CNN) [148, 145, 147] are deep architectures which are specifically designed to deal with the topological structure (e.g. domain knowledge of the input) to learn better features. These multilayer networks can learn complex, high-dimensional, nonlinear mappings from a large collection of examples which make them a suitable candidate for joint training. Several recent studies [138, 236, 248, 98, 113] have shown that in a supervised setting with large quantities of labeled data, proper initialization and choice of non-linearity, CNNs can outperform different approaches.

Conceptually, CNN exploits input topological structure and define local receptive fields [115] such that each low-level features are extracted only using a subset of the input (e.g., image patch) by adding topological locality constraint. It provides a gain of having a smaller number of parameters. Computation of local features is expected to be relevant to all positions of the receptive field. Therefore, a stride of such local low-level feature extractor over the subset of input corresponds to the transformation of an input into a similarly shaped feature map [146] leading to *convolution* and sharing of same parameters. Also, local features computed in the neighboring input locations are then summarized through an average [145, 147] or max [118] *pooling* operation supporting invariance to input modifications.

Combination of *convolution* and *pooling* is crucial for the modern CNN architectures which have produced state of the art results in various domains (e.g., object classification [138], semantic segmentation [160] in computer vision, sentence classification [131], sentiment analysis [63] in natural language processing, knowledge base completion [226] in Semantic Web etc.). Few approaches went beyond *convolution* and *pooling* and extended CNNs with several other tricks. For example, usage of residual [98] and dense [113] connections or varied inputs (e.g., character-level representations of the text [297]).

In general, final layers of CNN are fully connected after going through many convolutional and pooling layers. Usually, neurons present in the fully connected layer have connections to all activations in the previous layer, resembling standard neural networks. Finally, training of CNNs is usually achieved with the forward pass over all layers and then backpropagation [220] to update parameters.

Recurrent Neural Network (RNN) [69] was initially proposed to model time series or sequences. The network structure of RNN is similar to that of standard multilayer perceptron, but connections are also allowed between hidden units with a time delay. Hence, this network is capable of storing information emerging from the past and enables it to discover temporal correspondences between events that are distant from each other in the data. Due to this memory efficiency, RNNs have found its application to many tasks which require modeling sequential data such as Language Model (LM) [175], speech recognition [91], Machine Translation (MT) [127] etc.

Although RNNs were found to be very effective in dealing with sequences, howbeit it has been found that they are difficult to train for long-term dependencies due to issues such as exploding and vanishing gradient [21]. Variants of RNN such as Long Short-term Memory (LSTM) [108] and Gated Recurrent Unit (GRU) [44] have mitigated the problems persisting in RNN. LSTM and GRU can handle long sequences with efficient memory management.

2.2 MULTI-VIEW REPRESENTATION LEARNING

Increase in the availability of data containing multiple views such as a combination of image+text, audio+video, and text translations have lead to heterogeneous and homogeneous representations. Different views usually contain information which is complementary, consensus or combination of both. Exploiting multiple views for learning representations is more expressive than separately learning from either of views [186]. Therefore, representation learning with multiple views is very encouraging with broad applicability to many applications in varied domains.

Multi-view representation learning is built on following multi-view data principles [156]:

- *Correlation* – Aims to maximize the correlations among variables between multiple heterogeneous views.
- *Consensus* – Aims to maximize the agreement on the representations learned from multiple heterogeneous views.
- *Complementarity* – Aims to exploit the complementary knowledge contained in multiple views to effectively represent the data.
- *Consensus* and *Complementarity* – In general, combining both of aforementioned principles simultaneously is required for better representations.

In comparison to the representation learning (with single-view), multi-view representation learning acquires different representation (e.g., embedding) for each view and then jointly optimize all representations from multiple views to enhance succeeding learning tasks, such as retrieval, classification, and generation. Howbeit, there is also a possibility of degradation of performance [164], if learning objective cannot correctly capture the properties of the multi-view data. Therefore, careful selection of techniques is required based on characteristics of the multi-view data satisfying aforementioned underlying principles.

Many techniques are proposed to handle multi-view data which leverage fundamentals of Probabilistic Graphical Model (PGM) [136], kernel machines [228] and non-linear neural networks [92]. The fundamental difference between these techniques is whether the architecture of learning models is to be interpreted as a PGM or as a computation graph [25]. This difference has shown an impact when building shallow and deep architectures, where an exact inference of probabilistic models usually becomes intractable, while computation graphs have shown significant impact for learning from large-scale data.

In the following, we explore multi-view representation learning from the perspective of both paradigms as mentioned earlier and also multi-view data principles which they leverage to build learning models. Figure 11 provides the overall view of the sections.

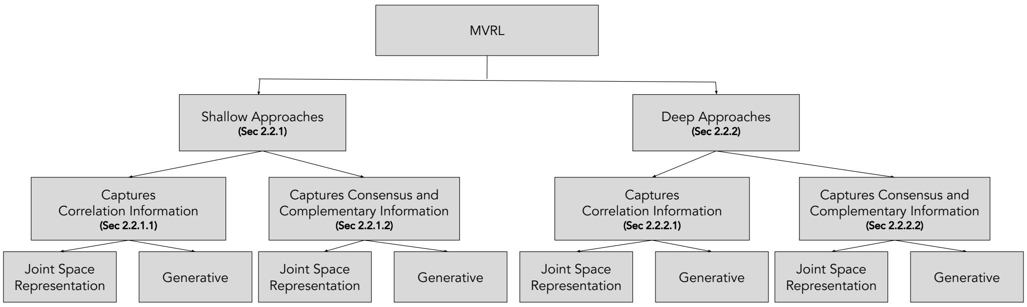

MVRL
Shallow Approaches (Sec 2.2.1)
Deep Approaches (Sec 2.2.2)
Captures Correlation Information (Sec 2.2.1.1)
Captures Consensus and Complementary Information (Sec 2.2.1.2)
Captures Correlation Information (Sec 2.2.2.1)
Captures Consensus and Complementary Information (Sec 2.2.2.2)
Joint Space Representation
Generative
Joint Space Representation
Generative
Joint Space Representation
Generative
Joint Space Representation
Generative

Figure 11: Organization of the Sections. The left part shows the architecture of multi-view representation learning based on shallow approaches which is further divided based on correlation as well as consensus and complementarity principles. While on the right part displays the deep approaches.

2.2.1 *Multi-View Shallow Representation Learning*

Shallow approaches for multi-view representation learning do not leverage any ideas from deep representation learning presented in the Section 2.1.2. However, they are confined to the multi-view data principles and also partially affiliate themselves to the paradigms as mentioned earlier. In the following, we divide the techniques based on the multi-view data principles.

2.2.1.1 *Correlation*

As presented earlier, the goal of correlation principle is to maximize the correlations of variables among multiple heterogeneous views. In the following, those methods are presented which learn shallow representations by finding correlations across views. These approaches are further divided into two categories where the first set of methods learn direct joint representations, while the rest learns a generative model.

Joint Space Representation: Goal of joint space representation approaches is to build a stable deterministic shallow representation from the multiple views of data. In the following, we explore some techniques that are based on the correlation principles and achieve joint space representation.

Canonical Correlation Analysis (CCA) proposed by Hotelling [110] work with two views for finding the linear transformations of each single view such that the correlations between the transformed variables are mutually maximized. Considering a two view data $\{X, Y\} = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $X \in \mathcal{R}^{d_1 \times n}, Y \in \mathcal{R}^{d_2 \times n}$, CCA aims to compute two linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y$ which makes the individual instances in X, Y maximally correlated in the projected space and evaluated using the following correlation coefficient ρ.

$$\rho = \frac{\boldsymbol{W}_x^T C_{xy} \boldsymbol{W}_y}{\sqrt{(\boldsymbol{W}_x^T C_{xx} \boldsymbol{W}_x)(\boldsymbol{W}_y^T C_{yy} \boldsymbol{W}_y)}} \tag{1}$$

where C_{xy} is a cross-covariance matrix given by Equation 2.

$$C_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \tag{2}$$

Here, μ_x, μ_y represent mean of the two views X, Y and C_{xx}, C_{yy} are covariance matrices.

Maximizing linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y$ of CCA is equivalent to solving a pair of generalized eigenvalue problems [95] and optimization is posed as a Lagrangian dual [245]. Finally, correlation between different views is provided by the eigenvector corresponding to the largest eigenvalues. Besides successful application of CCA to multi-view data, it still fails to deal with non-linearity and sometimes overfit.

Kernel Canonical Correlation Analysis (KCCA) [3] provides a non-linear extension of CCA and handle two views. Formalizing kernel CCA in line with CCA, dual representation is leveraged for representing $\boldsymbol{W}_x, \boldsymbol{W}_y$ using $X\boldsymbol{\alpha}, Y\boldsymbol{\beta}$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of size n. Correlation coefficient ρ is now provided by Equation 3.

$$\rho = \frac{\boldsymbol{\alpha}^T X^T X Y^T Y \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T X^T X X^T X \boldsymbol{\alpha} \times \boldsymbol{\beta}^T Y^T Y Y^T Y \boldsymbol{\beta}}} \tag{3}$$

If kernel matrices K_x, K_y of X, Y are provided by $(X^T X, Y^T Y)$ respectively. Then the Equation 3 is rewritten as follows:

$$\rho = \frac{\boldsymbol{\alpha}^T K_x K_y \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T K_x^2 \boldsymbol{\alpha} \times \boldsymbol{\beta}^T K_y^2 \boldsymbol{\beta}}} \tag{4}$$

For maximizing linear projections, in contrast to the linear CCA which works by carrying out an Eigen decomposition of the covariance matrix. The eigenvalue problem for kernel CCA degenerate solutions when either K_x or K_y is invertible [95].

Regularized Canonical Correlation Analysis (regularized CCA) [54] adds regularization to CCA. It can be seen as a way to deal with overfitting and assist to generalize better for the unseen samples. Now if CCA is observed from the perspective of an estimator of a linear system consuming data from two different views X and Y. Then regularized CCA aims to compute two normalized linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y$ which makes the individual instances in X, Y maximally correlated in the projected space and evaluated using the correlation coefficient ρ given by Equation 5 and optimized with the maximum likelihood estimator.

$$\rho = \frac{\boldsymbol{W}_x^T C_{xy} \boldsymbol{W}_y}{\sqrt{(\boldsymbol{W}_x^T C_{xx} \boldsymbol{W}_x + \tau_x \|\boldsymbol{W}_x\|^2)(\boldsymbol{W}_y^T C_{yy} \boldsymbol{W}_y + \tau_y \|\boldsymbol{W}_y\|^2)}} \tag{5}$$

where τ_x, τ_y are the regularization parameters and are bounded between interval [0,1]. Regularized CCA is also extended with kernel variations [233] and is given by Equation 6.

$$\rho = \frac{\boldsymbol{W}_x^T K_x K_y \boldsymbol{W}_y}{\sqrt{(\boldsymbol{W}_x^T (K_x + \tau_x I)^2 \boldsymbol{W}_x)(\boldsymbol{W}_y^T (K_y + \tau_y I)^2 \boldsymbol{W}_y)}} \tag{6}$$

Rates of regularization parameter [81] in the regularized kernel CCA is investigated with theoretical analysis to understand the optimal values.

Cluster Canonical Correlation Analysis (cluster CCA) extends CCA by first projecting multiple views of the data into the lower dimensional subspace before clustering. Chaudhuri et al. [39] assumes that the each view is generated from mixture of Gaussian [253] and are uncorrelated. To ensure a sufficient correlation between views, CCA matrix across the views assumed to be at least $k - 1$, when each view is in isotropic position and the $(k-1)^{th}$ singular value of this matrix to be at least minimum Eigen value. Now the correlation coefficient ρ is provided by Equation 7.

$$\rho = \frac{\mathbb{E}[(\boldsymbol{W_x}.X)(\boldsymbol{W_y}.Y)]}{\sqrt{\mathbb{E}[(\boldsymbol{W_x}.X)^2]\mathbb{E}[(\boldsymbol{W_y}.Y)^2]}} \tag{7}$$

where $\boldsymbol{W_x}, \boldsymbol{W_y}$ are the projections. In the dual formation, minimizing the Equation 7 provide the minimum Eigen value satisfying the aforementioned conditions.

CCA was also combined with clustering algorithms like spectral clustering [27] to realize the correlation spectral clustering. Other variations of cluster CCA also exists such as the one proposed by Rasiwasia et al. [206] where discriminant low dimensional representations are learned to maximize the correlation between the two views.

Sparse Canonical Correlation Analysis (sparse CCA) [94] is designed to reduce the dimensionality of vectors for attaining a stable solution with a sub selection procedure. Sparse CCA also aims to compute pair of linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y$ who maximize the correlation coefficient ρ given by the Equation 1.

Although, this equation is same as CCA. The constraints for sparse CCA [276] are different from those of CCA and are given by Equation 8.

$$\text{s.t.} \quad ||\boldsymbol{W}_x||^2 \leqslant 1, ||\boldsymbol{W}_y||^2 \leqslant 1 \qquad P_x(\boldsymbol{W}_x) \leqslant c_x, P_y(\boldsymbol{W}_y) \leqslant c_y \tag{8}$$

where P_x and P_y are convex penalty functions. For c_x and c_y small, this results in $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ sparse: many of the elements of $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ will exactly equal zero. Advantage of sparse CCA criterion is that it results in unique $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ even when dimensions of each instance in X and Y are greater than total sample size for certain choice of P_x and P_y.

Generalized Canonical Correlation Analysis (generalized CCA)) leverage more than two views of the data as opposed to earlier approaches. Extending to multiple views is achieved with generalized version of CCA [130] by combining all the pairwise correlations among each view through addition operation in the objective function [221].

Given an additional view $\{Z\} = \{z_1, ..., z_n\}$ to the existing two view data $\{X, Y\}$ such that $\{X, Y, Z\} = \{(x_1, y_1, z_1), ..., (x_n, y_n, z_n)\}$, where $X \in \mathcal{R}^{d_1 \times n}, Y \in \mathcal{R}^{d_2 \times n}$ and $Z \in \mathcal{R}^{d_3 \times n}$ form three parallel views. Then generalized CCA aims to compute linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y, \boldsymbol{W}_z$ which makes the individual instances in X, Y, Z maximally correlated in the projected space and evaluated using the following correlation coefficient ρ.

$$\begin{aligned}\rho = \\ &\frac{\boldsymbol{W}_x^T C_{xy} \boldsymbol{W}_y}{\sqrt{(\boldsymbol{W}_x^T C_{xx} \boldsymbol{W}_x)(\boldsymbol{W}_y^T C_{yy} \boldsymbol{W}_y)}} + \frac{\boldsymbol{W}_y^T C_{yz} \boldsymbol{W}_z}{\sqrt{(\boldsymbol{W}_y^T C_{yy} \boldsymbol{W}_z)(\boldsymbol{W}_z^T C_{zz} \boldsymbol{W}_z)}} \\ &+ \frac{\boldsymbol{W}_x^T C_{xz} \boldsymbol{W}_z}{\sqrt{(\boldsymbol{W}_x^T C_{xx} \boldsymbol{W}_x)(\boldsymbol{W}_z^T C_{zz} \boldsymbol{W}_z)}}\end{aligned} \tag{9}$$

Now, Maximizing linear projections $\boldsymbol{W}_x, \boldsymbol{W}_y, \boldsymbol{W}_z$ is equivalent that of two view CCA. Hence, it makes CCA a merely the subset of the generalized CCA [234]. Also, there exists a regularized version of the generalized CCA [252] similar to aforementioned regularized CCA.

Generative Model: Main aim of the generative approaches is to build probabilistic models to learn a compact set of latent random variables that represent a distribution over the observed multi-view data. In the following, we explore some techniques that are based on the correlation principles and build a generative model.

Probabilistic Canonical Correlation Analysis (probabilistic CCA) is built on the principles of CCA and probabilistic generative models [184]. In CCA, the canonical correlation directions given by $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ are obtained by solving generalized eigenvalue problem. While in the probabilistic CCA, latent variable interpretation and building a model is estimated with maximum likelihood estimate (MLE) for computing canonical correlation directions [11].

Given latent variables z, x and y, then the Gaussian prior and conditional distribution is given by $z \sim \mathcal{N}(0, I_d)$ and $x|z \sim \mathcal{N}(\boldsymbol{W}_x z + \mu_x, \theta_x), y|z \sim \mathcal{N}(\boldsymbol{W}_y z + \mu_y, \theta_y)$ respectively. Here, μ_x, μ_y represent mean of the two views X, Y. Projections $\boldsymbol{W}_x, \boldsymbol{W}_y$ are now estimated with MLE and is given by Equation 10 and Equation 11 respectively.

$$\boldsymbol{W}_x = \sum_{xx} \boldsymbol{u}_x P^{1/2} R \tag{10}$$

$$\boldsymbol{W}_y = \sum_{yy} \boldsymbol{u}_y P^{1/2} R \tag{11}$$

where P is a diagonal matrix constituting canonical correlations, R is an arbitrary rotation matrix and $\mathbf{u}_x, \mathbf{u}_y$ are canonical directions. Other variations of probabilistic CCA also exist, where the Gaussian distributions are replaced with Student-t distributions [10], hierarchical Bayesian model is used with variational approximation [271] and noisy data is leveraged for learning using variational Bayesian inference [258].

2.2.1.2 *Consensus and Complementarity*

We explore methods that combine the goal of *consensus* and *complementarity* principles to effectively represent the multi-view data. However, we also have a brief look at those methods that satisfy either of these principles separately.

Joint Space Representation: Goal of joint space representation approaches is to build a stable deterministic shallow representation from the multiple views of data. In the following, we explore some techniques that are based on the consensus/complementarity principles and achieve a joint space representation.

Collective Matrix Factorization (CMF) [237] learns a joint representation from the multi-view data by considering both consensus and complementarity principles. Given multi-view data $\{X_1, X_2, ..., X_I\}$ where $X_i \in \mathcal{R}^{d_i \times n}$ with "I" views, where n denotes the training sample size and d the concatenated dimensions of the data from all "I" views, then CMF factorizes in a way as given by Equation 12.

$$X_i = \mathbf{U}^T \mathbf{V}_i \quad \forall i \in I \tag{12}$$

where $\mathbf{U} \in \mathcal{R}^{k \times d}$ and $\mathbf{V} \in \mathcal{R}^{k \times n}$. It can be observed that the collective factorization has lead the data matrices X_i from multiple views to share a shallow joint representation with the factor matrix $\mathbf{U}$, while each data matrix is factorized into a loading matrix $\mathbf{V}$.

Standard approach for optimization is to minimize Bregman matrix factorization [88] loss or leverage simpler loss such as regularized squared error w.r.t $\mathbf{U}$ and $\{\mathbf{V}_i\}_{i \in I}$ given by Equation 13.

$$\min_{\mathbf{U}, \mathbf{V}_i} \sum_{i \in I} \alpha_{X_i} ||X_i - \mathbf{U}^T \mathbf{V}_i||_F^2 + \alpha_U ||\mathbf{U}||_F^2 + \sum_{i \in I} \alpha_{V_i} ||\mathbf{V}_i||_F^2 \tag{13}$$

where α_{X_i}, α_U and α_{V_i} are regularization parameters. It can be observed that $\mathbf{U}$ is attained by leveraging X_i's, where the consensus ensures the mutual agreement on multiple views of data and the complementarity exploits the exclusive information contained in different views for learning the joint space representation.

Partial Least Squares (PLS) [217] has been successfully applied to model relations between sets of observed variables for building shallow joint representations. Given two view data $\{X, Y\}$. PLS builds a k-dimensional solution with parameter matrices $\mathbf{W}_x \in \mathcal{R}^{d_1 \times k}$ and $\mathbf{W}_y \in \mathcal{R}^{d_2 \times k}$ and are optimized to maximize the the covariance between different sets of variables as given by the Equation 14.

$$\max_{\mathbf{W}_x, \mathbf{W}_y} \operatorname{tr}(\mathbf{W}_x^T C_{xy} \mathbf{W}_y) \quad \text{s.t.} \quad \mathbf{W}_x^T \mathbf{W}_x = I, \mathbf{W}_y^T \mathbf{W}_y = I \tag{14}$$

Close connections are found between PLS and CCA in varied aspects [15]. Nevertheless, CCA finds the directions of maximum correlation, while PLS finds the directions of maximum covariance with consensus principles.

Generative Model: The main aim of the generative approaches is to build probabilistic models for learning a fixed set of latent random variables that represent a distribution over the observed multi-view data. Parameters of these probabilistic models are in general estimated by maximizing the regularized likelihood of the multi-view data. In the following, we explore some techniques that are based on the consensus and complementarity principles and build a generative model.

Probabilistic Multi-View Sparse Coding [120, 159] learns a shallow joint representation by leveraging multi-view data with a set of linear mappings defined as *dictionaries*. Goal of these *dictionaries* is to find a shared representation s^* by selecting the most appropriate bases and eliminating unwanted ones. Ultimately, this leads to a high correlation within the multi-view data and also falls in-line with directed graphical models explaining away effect [193].

Given two view data $\{X, Y\}$, non-probabilistic sparse coding is formulated to learn a representation w.r.t a multi-view data sample and is given by the Equation 15.

$$s^* = \arg\min_{s} ||X - \boldsymbol{W}_x s||_2^2 + ||Y - \boldsymbol{W}_y s||_2^2 + \lambda ||s||_1 \tag{15}$$

where $\boldsymbol{W}_x$, $\boldsymbol{W}_y$ are *dictionaries* and λ denote the regularization constant. Learning the pair of dictionaries is achieved by optimizing the objective w.r.t $\boldsymbol{W}_x$, $\boldsymbol{W}_y$ and is given by the Equation 16.

$$\frac{\partial}{\partial \boldsymbol{W}_x}, \frac{\partial}{\partial \boldsymbol{W}_y} = \sum_{i=1}^{n} (||x_i - \boldsymbol{W}_x s_i^*||_2^2 + ||y_i - \boldsymbol{W}_y s_i^*||_2^2) \tag{16}$$

Usually $\boldsymbol{W}_x$, $\boldsymbol{W}_y$ are regularized by the constraint of having unit-norm columns, while x_i and y_i are input from the different views.

However, if the aforementioned regularized form can be generalized as a probabilistic model. The probabilistic multi-view sparse coding then assume generative distributions with a prior $p(s)$ given by the Equation 17 and its conditional distributions are provided by the Equation 18.

$$p(s) = \prod_{j} \frac{\lambda}{2} \exp(-\lambda |s_j|) \tag{17}$$

$$\begin{aligned} \forall_{i=1}^{n}: \quad & p(x_i|s) = \mathcal{N}(x_i; \boldsymbol{W}_x s + \mu_{x_i}, \sigma_{x_i}^2) \\ & p(y_i|s) = \mathcal{N}(y_i; \boldsymbol{W}_y s + \mu_{y_i}, \sigma_{y_i}^2) \end{aligned} \tag{18}$$

For obtaining a sparse multi-view representation, maximum a posteriori (MAP) value of s i.e., $s^* = \arg\max_{h} p(s|x, y)$ is computed. Further to learn parameters $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$, joint MAP values of s^* as shown in the Equation 19 are leveraged to maximize the likelihood of the data.

$$\arg\max_{\boldsymbol{W}_x, \boldsymbol{W}_y} \prod_{i} p(x_i|s^*) p(y_i|s^*) \tag{19}$$

Alternatively, Expectation Maximization (EM) can be also exploited to learn dictionaries $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ and shared representation s^*.

Multi-View Markov Random Field (Multi-view MRF) [155] is an undirected graphical model leveraged to learn shallow multi-view representation. However, these models are not applied in their direct form but are modified into special cases. For example, Xing et al. [285] leveraged exponential family Harmonium [275], a known special case of Markov Random Field (MRF) to propose a multi-wing harmonium model. It is a Multi-view MRF having an advantage of the faster inference than the directed graphical models [183] due to the conditional independence of the hidden units.

Given two view data $\{X, Y\}$ with the set of hidden units $H = \{h_1, ..., h_n\}$. Multi-wing harmonium take each view and hidden units to construct a complete bipartite graph where units in the same set contain no connections, but are fully connected across sets. Additionally, all the observed (i.e. input) and hidden variables are from exponential family. Furthermore, random variables in the log-domain are coupled with other terms to attain the joint distribution given by Equation 20.

$$\begin{aligned} p(X, Y, H) \propto & \\ & \exp\{\sum_i \alpha_i^T \phi(x_i) + \beta_i^T \psi(y_i) + \gamma_i^T \varphi(h_i) \\ & + \sum_i \phi(x_i)^T W_{ii} \varphi(h_i) \\ & + \sum_i \phi(y_i)^T U_{ii} \varphi(h_i)\} \end{aligned} \tag{20}$$

where $\alpha_i, \beta_i, \gamma_i$ are associated weights of clique potentials $\phi(\cdot), \psi(\cdot), \varphi(\cdot)$, while W_{ii}, U_{ii} are the associated weights of potentials over cliques consisting of pairwise linked nodes $\phi(x_i)\varphi(h_i), \psi(y_i)\varphi(h_i)$.

Training of model parameters is achieved with MLE on the training data. Update rules are obtained by taking partial derivatives of the log-likelihood of the Equation 20 w.r.t to model parameters.

Multi-View Hierarchical Bayesian Model (Multi-view HBM) [16, 29] is seen as an extension to Latent Dirichlet Allocation (LDA) [30] which is a three-level hierarchical Bayesian network that models a sample from a single view (e.g. textual document) as a finite mixture over an underlying set of topics. Multi-view HBM has seen its applications mainly in the joint modeling of two different views emerging from varied modalities (e.g. visual and textual data).

One such multi-view HBM model is "correspondence LDA" proposed by Blei et al. [29]. Correspondence LDA allows simultaneous dimensionality reduction in the joint representation and also models the conditional correspondence between their respectively reduced representations. During the generative process it generates one view after another.

Given two view data $X = \{x_1, ..., x_N\}, Y = \{y_1, ..., y_M\}$ without parallel views, the pair (X_i, Y_i) represents a combination from two varied views. If, $Z = \{z_1, ..., z_N\}$ denote latent variables that generate the view-1 and $W = \{w_1, ..., w_M\}$ be the discrete indexing variables that take values from 1 to N. Then K-factor correspondence LDA model assumes the following generative process for pairs (X, Y):

- Sample $\theta \sim \mathtt{Dirichlet}(\theta|\alpha)$
- For each view X_n, where $n \in \{1, ..., N\}$
  - Sample $Z_n \sim \mathtt{Multivariate}(\theta)$
  - Sample $X_n \sim p(X|Z_n, \mu, \sigma)$ from a multivariate Gaussian distribution conditioned on Z_n
- For each view Y_m, where $m \in \{1, ..., M\}$
  - Sample $W_m \sim \mathtt{Uniform}(1, ..., N)$
  - Sample $Y_m \sim p(Y|W_m, Z, \beta)$ from a multinomial distribution conditioned on the Z_{W_m} factor.

Furthermore, the joint distribution $p(X, Y, \theta, Z, W)$ of the correspondence LDA is given by Equation 21.

$$p(\theta|\alpha)\left(\prod_{n=1}^{N} p(Z_n|\theta)p(X_n|Z_n, \mu, \sigma)\right)\left(\prod_{m=1}^{M} p(W_n|N)p(Y_m|W_m, Z, \beta)\right) \quad (21)$$

Since, exact probabilistic inference for the correspondence LDA is intractable, variational inference method [124] is used to approximate the posterior distribution over the latent variables.

2.2.2 *Multi-View Deep Representation Learning*

The rise of deep representation learning approaches presented in the Section 2.1.2 also influenced shallow multi-view representation learning to build deep representations by capturing the abstract relationship between the multi-view data. In the following, we explore deep multi-view representation methods which leverage multi-view data principles presented in the Section 2.2 from the perspective of both probabilistic and joint space representation.

2.2.2.1 *Correlation*

As discussed in sections mentioned above, the goal of correlation principle is to maximize the correlations of variables among multiple heterogeneous views. In the following, those methods are presented which learn deep representations by finding correlations across views. Furthermore, approaches are divided into two categories where the first set of methods learn direct joint representation, while the rest learns a generative model.

Joint Space Representation: Goal of joint space representation approaches is to build a stable deterministic deep representation from the multiple views of data. In the following, we explore some techniques that are based on the correlation principles and achieve a joint space representation.

Deep CCA [9] learn non-linear mappings with multiple stacked layers between two views which are maximally correlated. This setup resembles the objectives of neural network based CCA-like approaches [19] for capturing high-level associations between data from the multiple views. Earlier, few approaches [140] also investigated a neural network implementation of CCA by maximizing the correlation between the outputs of networks for the different views. Another approach by Hsieh et al. [111] formulated a non-linear CCA method using three feed-forward neural networks. The aim of the first network is to maximize the correlation between canonical variates, while the remaining two networks were aimed to map the canonical variates back to the original two sets of variables.

Coming back to Deep CCA, given two view data $\{X, Y\}$, it first learns deep representations $\mathcal{F}_x(X), \mathcal{F}_y(Y)$ for both views separately with a deep neural network. Then the goal of deep CCA is to jointly learn parameters for both views such that correlation between $(\mathcal{F}_x(X), \mathcal{F}_y(Y))$ is as high as possible. Let Θ_x be the vector constituting all parameters of the first view and similarly Θ_y for the another view. Equation 22 maximizes the correlation.

$$(\Theta_x^*, \Theta_y^*) = \arg\max_{\Theta_x, \Theta_y} \mathrm{Correlation}(\mathcal{F}_x(X; \Theta_x), \mathcal{F}_y(Y; \Theta_y)) \tag{22}$$

Parameters Θ_x^*, Θ_y^* are estimated on the training data by following the gradient of the correlation objective, with stochastic optimization with mini-batches [274].

Deep Canonically Correlated Autoencoder (DCCAE) [273] is a deep neural network based model which consists of two autoencoders and optimizes the combination of canonical correlation between the learned bottleneck representations and the reconstruction errors of the autoencoders. Objective which is optimized to learn correlation between the input deep projections $\mathcal{F}(X), \mathcal{G}(Y)$ for both views is given by the Equation 23.

$$\begin{aligned} \min_{W_f, W_g, W_p, W_q, U, V} -\frac{1}{N}\mathrm{tr}(U^T \mathcal{F}(X) \mathcal{G}(Y)^T V) \\ + \frac{\lambda}{N}\sum_{i=1}^{N}(\|x_i - p(\mathcal{F}(x_i))\|^2 \\ + \|y_i - q(\mathcal{G}(y_i))\|^2) \end{aligned} \tag{23}$$

where $\lambda > 0$, W_f, W_g, W_p, W_q are set of learnable parameters and U, V are canonically correlated directions that project the DNN outputs. This approach is seen as an extension to Deep CCA by adding an autoencoder regularization. For parameter estimation, stochastic optimization [71] is applied to the DCCAE objective. Also, this objective offers a trade-off between the information captured in the mapping within each view, while also finding relationship across views.

2.2.2.2 *Consensus and Complementarity*

We explore methods that combine the goal of consensus and complementarity principles to effectively learn deep representations acquired from the multi-view data. However, we also have a brief look at those methods that satisfy either of these principles separately.

Deep Multi-View Embeddings are inspired from the single view compositional distributional semantics [46] approaches which learn distributed representations of a unit. Especially, it has been extensively applied to modalities such as textual corpora [17] to learn distributed representation of words a.k.a. *embeddings*. Lately, interest has also increased to learn joint representations by leveraging the multi-view input by modifying distributional semantics methods. The joint representation has found its application to several cross-view tasks such as cross-language [219] and cross-modal [33] tasks.

Methods which build upon deep multi-view embeddings are designed by leveraging varied techniques such as learning to rank [158], manifold learning [38], neural language models [22] etc. However, the core goal of all methods remains same, i.e., to map two or more views into a common space representation for supporting varied applications.

Deep Multi-View Autoencoders are good alternatives for learning a joint (or shared) representation between different views due to the flexibility of their objectives. Several autoencoder based approaches are proposed to extract shared representations. In the following, we discuss two such architectures.

– Two-view autoencoder by Ngiam et al. [186] uses two view dataset augmented with additional examples from each of the single view as input. Idea is to use a greedy layer-wise pretraining with an extension to RBMs with sparsity followed by fine-tuning.

– Correspondence autoencoder (Corr-AE) [75] is another autoencoder based approach which construct correlations between hidden representations of two single views with two different deep autoencoders. Corr-AE architecture differs from Ngiam et al. [186] as it consists of two deep autoencoders that are connected by a predefined similarity measure on a specific internal (a.k.a code) layer.

Formally, given two view data $\{X, Y\}$. $\mathcal{F}(x; \boldsymbol{W}_{\mathcal{F}})$ and $\mathcal{G}(y; \boldsymbol{W}_{\mathcal{G}})$ denote mapping of individual views X and Y respectively to the internal (or code) layers with $\boldsymbol{W}_{\mathcal{F}}$ and $\boldsymbol{W}_{\mathcal{G}}$ representing the weights of two separate autoencoders constructed from views X and Y respectively. The similarity measure between pairs of (x_i, y_i) is now calculated with Equation 24 and the loss function used to learn joint representation is given by the Equation 25.

$$Sim(x_i, y_i, \boldsymbol{W}_{\mathcal{F}}, \boldsymbol{W}_{\mathcal{G}}) = ||\mathcal{F}(x_i; \boldsymbol{W}_{\mathcal{F}}) - \mathcal{G}(y_i; \boldsymbol{W}_{\mathcal{G}})||_2^2 \tag{24}$$

$$\begin{aligned} \mathcal{L}(x_i, y_i, \boldsymbol{W}_{\mathcal{F}}, \boldsymbol{W}_{\mathcal{G}}) = \\ & (1-\alpha)(\mathcal{L}_X(x_i, y_i, \boldsymbol{W}_{\mathcal{F}}, \boldsymbol{W}_{\mathcal{G}}) + \mathcal{L}_Y(x_i, y_i, \boldsymbol{W}_{\mathcal{F}}, \boldsymbol{W}_{\mathcal{G}})) \\ & + \alpha \mathcal{L}_J(x_i, y_i, \boldsymbol{W}_{\mathcal{F}}, \boldsymbol{W}_{\mathcal{G}}) \end{aligned} \tag{25}$$

where $\mathcal{L}_X(\cdot) = ||x_i - \hat{x}_i||_2^2$, $\mathcal{L}_Y(\cdot) = ||y_i - \hat{y}_i||_2^2$ and $\mathcal{L}_J(\cdot)$ is given by Equation 24. $\hat{x}_i$ and $\hat{y}_i$ are the reconstructions of x_i and y_i respectively. Usually, $\mathcal{F}$ and $\mathcal{G}$ are chosen as logistic activation functions. $\mathcal{L}_X(\cdot)$ and $\mathcal{L}_Y(\cdot)$ are the losses caused by data reconstruction errors for the given inputs of two separated deep autoencoders. $\mathcal{L}_J(\cdot)$ is the correlation loss and α is trade-off parameter between two groups of objectives.

Multi-view Encoder-Decoder leverage multiple views of the data which are parallel to build representations such that one view can generate another. Previously, autoencoders are the preferred choice for encoding an input into a hidden representation to reconstruct it back. However, with the rise of convolutional and recurrent neural networks, encoder-decoder architectures are now dominated with the combination of them if the views emerge from two different heterogeneous sources. Otherwise, the same type of networks is leveraged in such a way that they support applications spawning across several domains.

For instance, multi-view encoder-decoder architectures are proven to be successful for sequence-to-sequence tasks in the domain of NLP, where the sequence from one view is encoded using an encoder (usually an RNN variant) to generate another sequence emerging from another view (which also is an RNN variant). Some of the tasks are neural machine translation (NMT) [247, 43] and question answering [262]. Encoder-decoder architectures for sequence-to-sequence tasks are improved with the attention mechanism [12] to handle lengthy sequences especially in the case of NMT.

Generative Model: The main aim of the generative approaches is to build probabilistic models for learning a fixed set of latent random variables that represent a distribution over the observed multi-view data using deep architectures. Parameters of these probabilistic models are in general estimated by maximizing the regularized likelihood of the multi-view data. In the following, we explore some techniques that are based on the consensus and complementarity principles and build a generative model.

Deep Multi-view Deep Boltzmann Machine (Deep Multi-view DBM) [243] is an extension to deep Boltzmann machine (DBM) (discussed in the Section 2.1.2.1) to leverage multiple views of the data. Particularly, each data view is modeled using a separate two-layer DBM and then an additional layer of binary hidden units on top of them is added to learn the shared representation.

Given two view data $\{X, Y\}$, the distribution for any sample x in the view-1 using a two-layer DBM with hidden layers $h^{(1)}$,$h^{(2)}$ is provided by the Equation 26 and expanded into the Equation 27.

$$P(x;\theta) = \sum_{h^{(1)},h^{(2)}} P(x;h^{(1)},h^{(2)};\theta) \tag{26}$$

$$
\begin{aligned}
P(x;\theta) = \\
&\frac{1}{\mathcal{Z}(\theta)} \sum_{h^{(1)},h^{(2)}} \exp\Big(- \sum_{i=1}^{d_{v_1}} \frac{(x_i - b_i)^2}{2\sigma_i^2} \\
&+ \sum_{i=1}^{d_{v_1}} \sum_{j=1}^{d_{h_1}} \frac{x_i}{\sigma_i} W_{ij}^{(1)} \\
&+ \sum_{j=1}^{d_{h_1}} \sum_{l=1}^{d_{h_2}} h_j^{(1)} W_{jl}^{(2)} h_l^{(2)} \Big)
\end{aligned}
\tag{27}
$$

where d_{v_1} and d_{h_1}, d_{h_2} represent the dimensions of input and hidden layers respectively. Similar to view-1, two-layer DBM for the second view (i.e., view-2) is leveraged and is defined by combining a replicated softmax model [105] with a binary RBM. Consequently, the deep multi-view DBM has been presented by combining the two-layer DBM of view-1 and view-2 with an additional layer of binary hidden units on top of them. The joint distribution over multiple views is given by the Equation 28.

$$
P(x,y;\theta) = \sum_{h^{(1)},h^{(2)},h^{(3)}} P(h_{v_1}^{(1)}, h_{v_2}^{(2)}, h^{(3)}) \Big(\sum_{h_{v_1}^{(1)}} P(x, h_{v_1}^{(1)}, h_{v_1}^{(2)}) \Big) \Big(\sum_{h_{v_2}^{(1)}} P(y, h_{v_2}^{(1)}, h_{v_2}^{(2)}) \Big)
\tag{28}
$$

where hidden layers are represented accordingly to the view they belong to. For example, $h_{v_1}^{(1)}$ represent hidden layer-1 of the view-1.

Exact maximum likelihood learning in the deep multi-view DBM is intractable, hence approximate learning is implemented using mean-field inference [212] to estimate data-dependent expectations, and an MCMC based stochastic approximation procedure is used to approximate the model expectation.

Multi-view Generative Adversarial Networks [40] are the extension to Generative Adversarial Networks (GAN) [87] where two neural networks compete with each other. A generator neural network emulate the random noise into true distribution of the data in an attempt to fool the discriminator neural network whose goal is to distinguish genuine data from the imitation data created by the generator network. There are several variations[14] of GANs exist. GAN which leverages multi-view data is expected to perform density estimation from multi-view inputs and also can deal with missing views to update its prediction when more views are provided.

Initially, looking into the GAN architecture. Given an input data x, prior $p_z(z)$ over input noise variables is defined along with a differentiable generative function $\mathcal{G}(z;\theta_g)$ and discriminator $\mathcal{D}(x;\theta_d)$ function over input data x to predict a single scalar. $\mathcal{D}$ and $\mathcal{G}$ are now trained to maximize and minimize the label pre-

[14] https://github.com/hindupuravinash/the-gan-zoo

diction and $\log(1-\mathcal{D}(\mathcal{G}(z)))$ respectively with two-player minimax game [254] using a value function $\mathcal{V}(\mathcal{G},\mathcal{D})$ provided by the Equation 29.

$$\min_{\mathcal{G}}\max_{\mathcal{D}} \mathcal{V}(\mathcal{D},\mathcal{G}) = \mathop{\mathbb{E}}_{x\sim p_{data}(x)}[\log(\mathcal{D}(x)] + \mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(1-\mathcal{D}(\mathcal{G}(z))] \tag{29}$$

However, modeling multi-view GANs still require more sophistication than the basic GAN provides. Thus, Bidirectional GANs (BiGANs) [60] are leveraged as they can learn inverse mapping between feature representations and the input noise variables. This helps to get back the learned latent feature representations useful for many auxiliary tasks. The BiGAN introduces additional encoder $\mathcal{E}(x)$ which induces distribution $p_{\mathcal{E}}(z|x)$ along with generator $\mathcal{G}$ that models distribution $p_{\mathcal{G}}(x|z)$. Discriminator $\mathcal{D}$ is modified now to take input from both x, z and aim to comprehend whether the sample is generated from $p_{\mathcal{E}}(z|x)$ or $p_{\mathcal{G}}(x|z)$. Thus the modified training objective is provided by Equation 30 and Equation 31.

$$\begin{aligned}\min_{\mathcal{G},\mathcal{E}}\max_{\mathcal{D}} \mathcal{V}(\mathcal{D},\mathcal{E},\mathcal{G}) = &\\ &\mathop{\mathbb{E}}_{x\sim p_{data}(x)}[\mathop{\mathbb{E}}_{z\sim p_{\mathcal{E}}(.|x)}[\log(\mathcal{D}(x,z)]] \\ &+ \mathop{\mathbb{E}}_{z\sim p_z(z)}[\mathop{\mathbb{E}}_{z\sim p_{\mathcal{G}}(.|z)}[\log(1-\mathcal{D}(x,z)]]\end{aligned} \tag{30}$$

$$\begin{aligned}\min_{\mathcal{E},\mathcal{H}}\max_{\mathcal{D}'} \mathcal{V}(\mathcal{E},\mathcal{H},\mathcal{D}') = &\\ &\mathop{\mathbb{E}}_{\tilde{x}\sim p_{data}(\tilde{x})}[\mathop{\mathbb{E}}_{z\sim p_{\mathcal{E}}(z|x)}[\log(\mathcal{D}'(x,z)]] \\ &+ \mathop{\mathbb{E}}_{\tilde{x}\sim p_{data}(\tilde{x})}[\mathop{\mathbb{E}}_{z\sim p_{\mathcal{H}}(z|x)}[1-\log(\mathcal{D}'(\tilde{x},z)]]\end{aligned} \tag{31}$$

Combining the objective of BiGANs (i.e. $\mathcal{V}(\mathcal{D},\mathcal{G},\mathcal{E})$ with $\mathcal{V}(\mathcal{E},\mathcal{H},\mathcal{D}')$) provide the final objective of single-view BiGAN and easily extended to N different views (assuming all views are available) with the aggregation model provided by:

$$\Psi(\tilde{x}_k) = \sum_{k=1}^{N} \Phi(\tilde{x}_k) \tag{32}$$

where $\Phi(\tilde{x}_k)$ represent the usage of different views from $\tilde{x}$.

STATE OF THE ART

3

STATE OF THE ART

Context of this Chapter. In this chapter, we present those applications of MVRL where their input views emerge from the textual and image modality. In general, these applications can be broadly divided into two tasks: (1) Cross-lingual and (2) Cross-modal. Furthermore, tasks are again divided based on their usage in the real-world scenarios such as content-based retrieval, classification, and generation.

Although, cross-lingual and cross-modal tasks can be approached with varied techniques. We present only those approaches that leverage techniques similar to the ones presented in the Section 2.

3.1 MVRL FOR CROSS-LINGUAL TASKS

MVRL has shown its applicability for those views as well where they emerge from the homogeneous content. However, it has found its actual usage when views emerge from the homogeneous content belonging to different languages text. Applications that are built utilizing such views can inherently support various cross-lingual tasks such as cross-language text retrieval, cross-language text classification, and cross-language text generation. In the following, overall aim and state of the art for each of these tasks are explored separately.

3.1.1 *Retrieval*

Cross-language text retrieval [188] is of interest over past few decades to support those languages that lack sufficient information in the query language. With the advent of representation learning [25], deep architectures were utilized in the Information Retrieval (IR) for tasks such as query-document matching and query expansion [178]. However, for the cross-language text retrieval, representations emerging from different languages depicting multiple views is learned either with joint space or a generative model.

Cross-language text retrieval is also closely aligned with other similar problems that leverage structured knowledge. For instance, cross-language entity linking [235] aims to link mentions written in the non-English documents to entries in the English Wikipedia by comparison of textual clues across languages. A neural network based model is designed by leveraging convolution and tensor networks to train fine-grained similarities and dissimilarities between the query and candidate document from the multiple perspectives.

Problem similar to cross-language text retrieval [268] is question-question similarity re-ranking in the community question answering [168] also leverage MVRL techniques. A cross-language system is trained on the input language is ported to another language given the labeled training data for the first language and only unlabeled data for the second.

Adaptation of the MVRL technique of adversarial training using neural networks [87] for cross-language learning [126] is also explored. High-level features that are discriminative for the primary learning task, and at the same time invariant across the input languages is used as the key component for building the cross-language system.

3.1.2 *Classification*

Interest in the task of Cross-language Text Classification (CLTC) is driven by the availability of human curated labels for rich resource languages and their unavailability for the resource-poor languages. For instance, cross-language sentiment analysis [305] goal is to predict sentiment for those languages which lack abundant training data by leveraging those languages which have it.

Approaches proposed initially for CLTC has only leveraged shallow MVRL generative techniques such as latent topics detection across languages with topic models [28] extended to multilingual setting [176, 80]. In general, extraction of the cross-language latent topics/concepts use either context-insensitive [296] or context-sensitive methods [266] to build word co-occurrence statistics.

However, the rise of distributed representations of words a.k.a word embeddings [22, 173, 195] has shifted focus and created the need for utilizing deep MVRL architectures for learning representations. Furthermore, it is extended to learn from varied views by projecting pair or multiple languages into the shared semantic space to create multilingual [103, 135, 50], bilingual [90, 267, 164] and polylingual [4] word embeddings. Also, representations were extended beyond words to meet the variable-length textual units such as phrases, sentences and documents for both single view [239, 144] and multi-view [197] setting.

The dependency of shared representations has also been extended beyond CLTC and has shown its applicability to other tasks such as cross-language POS tagging [89].

3.1.3 *Generation*

In the sections mentioned above, we have discussed the cross-language tasks such as retrieval and classification. However, cross-language text generation is also an important paradigm where the text given in one language is translated into another language. Hence, cross-language text generation can be otherwise interpreted as the Machine Translation (MT) [238] problem.

The MT approaches initially designed are mostly shallow and are based on statistical NLP [32]. Lately, they are dominated by the neural MT [280] which leverage deep MVRL. These approaches use multi-view encoder-decoder based architectures (refer the Section 2.2.2.2)and have shown tremendous improvement in translating source to target language. Several variations of them are also explored where they either handle only bilingual pairs or cater multiple languages at once [77] by leveraging multi-task learning [62].

3.2 MVRL FOR CROSS-MODAL TASKS

MVRL can provide a significant impact if the view emerges from the heterogeneous views content where each view belongs to different modalities such as text and images. Applications that are built utilizing such views can inherently support cross-modal tasks such as cross-modal retrieval, cross-modal classification, and cross-modal generation. In the following, overall aim and state of the art for each of these tasks are explored separately.

3.2.1 *Retrieval*

The goal of cross-modal retrieval is to retrieve a modality that is different from the query modality. For instance, retrieving images that are similar to the textual query is a plausible case of cross-modal retrieval. Over past decades, many approaches are proposed for cross-modal retrieval using images and textual data (available in variable lengths such as phrases, sentences, and paragraphs) are based on shallow and deep MVRL. Most of the shallow methods such as CCA, Partial least square (PLS) and Bilinear Model (BLM) [54] aim at learning subspaces [31] or a shared space from the cross-modal data, in which the similarity between the modalities is measured using various distance metrics.

However, subspace learning methods are generally susceptible to scaling challenges. To overcome such issues, PGM based generative models are proposed such as correspondence Latent Dirichlet Allocation (Corr-LDA) [29] (refer the Section 2.2.1.2) and others which include topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) [201] and Multi-modal Document Random Field (MDRF) [121]. Howbeit these approaches depend on the exact inference and are intractable. Hence, approximate inference methods such as variational inference [124] is adapted to provide partial solutions.

Deep MVRL methods overcame few challenges observed with shallow ones, by designing robust techniques that can scale to large datasets and also avoid intractable inference problems. Approaches discussed in the Chapter 2 such as deep restricted Boltzmann machine (Deep RBM) [243], deep canonical correlation analysis (DCCA) [9], correspondence autoencoder (Corr -AE) [75] and deep visual-semantic embeddings [79, 133] used multimodal inputs to learn representations of common spaces.

With the availability of large-scale image-caption pair datasets [109, 292, 157] has spawn interest in image to caption retrieval or vice versa either focused purely on visual similarity based approaches [190] or learning of a multimodal space [109]. However, initially only shallow methods were used such as distance metrics to find textual captions when given image as a query.

However, the success of deep representation learning for image recognition and language models has shifted the focus to learn multimodal representations of images and their captions with deep MVRL techniques. For example, a recursive neural network [241] was used to retrieve images for a given caption or vice versa. Although image to caption retrieval produced grammatically correct captions, they failed to generalize to the novel concepts. This drawback has shifted the focus towards cross-modal generation approaches (see Section 3.2.3).

3.2.2 *Classification*

Interest in labeling one modality with an other is seen as the problem of cross-modal classification. For instance, prediction of a textual label as annotation for an image is a plausible case of the cross-modal classification. Previously, labeling images with the textual annotations have been explored by several works from diverse perspectives.

Automatically mining image data from the web [59, 41] and annotating them with textual labels is an unsupervised way of providing classification. Other approaches [281] have focused their efforts on cleaning the data acquired from the web by leveraging pre-trained models built from datasets created with the human supervision (e.g., ImageNet [55]).

Few aimed at directly training a classification model from the web data [301] by automatically discovering the hidden patterns. Other objectives are also perceived, where the label noise is tackled when building models such as by Sukhbaatar et al. [244] and Xiao et al. [283]. They filtered the label noise when learning a image-specific deep representation model. Here, an alternative case is also possible, where instead of directly learning image representation models (e.g., CNNs [236, 98, 112]), leveraging multi-view data (e.g., images and text) is utilized to address the challenge. As learning from CNN with noisy labeled data is still an open problem.

Alternative for clean textual labels are hashtags [15] which are regularly observed in the noisy environments such as social media (e.g., Twitter, Pinterest, Facebook, Instagram) messages. They capture authors perspective on a particular topic. Sometimes messages also accompany images, and hashtags are usually considered as weakly aligned labels for those images. Inspired from the application of deep representation learning (see Section 2.1.2) [58, 86] for modeling messages for prediction and recommendation. Hashtags are also explored for image tagging by Denton et al. [56] and proposed a 3-way multiplicative gating approach, where the image model is conditioned on the user metadata on Facebook[16] dataset. While, Park et al. [191] proposed context sequence memory network (CSMN) model mainly to built a personalized image captioning system to predict hashtags on Instagram[17] dataset. However, hashtags are usually illustrated with n-grams or abbreviations and sometimes difficult to interpret when compared with semantically enriched clean textual labels.

Approaches presented above only operate with the single-label per image, wherein real-world scenarios multiple labels are usually observed per image.

[15] https://en.wikipedia.org/wiki/Hashtag

[16] https://www.facebook.com/

[17] https://www.instagram.com/

3.2.3 *Generation*

The goal of cross-modal generation is to generate one modality from another. For instance, the application which has gained much attention in the recent years is the generation of sentence-level textual descriptions for images. Approaches have leveraged deep MVRL approaches such as encoder-decoder based architectures [43, 12] where image is encoded to decode a textual sequence depicting what is observed in an image. Similarly, vice versa is also investigated, where given a sentence-level textual description, an image is generated [207] by matching critical visual objects depicted in the description.

In the following, overall aim and state of the art for each of these tasks are explored separately.

Image to Text Generation goal is to generate a variable length text conditioned on an image. In general, approaches limit the generated textual descriptions to a sentence. Initially, Kiros et al.[133] explored generation of text conditioned on an image with multimodal neural language model (Neural LM), while Karpathy et al. [129] and Chen et al. [42] used RNN. Mao et al.[166] proposed multimodal variant of RNN that use the image at every time step. Vinyals et al. [262] and Donahue et al. [61] introduced similar architectures but leverages LSTM. Jia et al.[119] used LSTM and extra guidance from the correlated image and textual features obtained using CCA.

Fang et al.[72] slightly deviated from RNN based approaches and used multi-instance learning and maximum-entropy language model. Another noteworthy improvement is seen with encoder-decoder frameworks for caption generation by including semantic [291] and visual [286] attention along with a reviewer module [288]. Lu et al. [163] explored attention mechanism with a visual sentinel, while Anderson et al. [8] used region CNN (R-CNN) [209] visual features to capture visual attention. Few approaches [211] leveraged reinforcement learning to optimize evaluation metric CIDEr [255] along with visual attention mechanism. Another set of approaches [123] which annotate captions to individual regions in an image, while some [102] expanded caption generation to novel objects not seen in an image-sentence parallel corpus.

Approaches mentioned above are designed to handle single images only. However in real-world scenarios, a sequence of images are observed, and they illustrate a story. For tackling such scenarios, the entire order of images is considered for generating descriptions. In general, generated descriptions resemble a paragraph and is referred to as visual storytelling [114].

Text to Image Generation goal is to reverse the process of the image to text generation by leveraging conditional generation, i.e., to generate an image given the variable length text. Initially, Denton et al. [57] synthesized images at multiple resolutions by using a Laplacian pyramid [34] of an adversarial generator and discriminators [87]. This work generated compelling high-resolution images and could also condition on class labels for the controllable generation.

Going beyond the approach of Denton et al. [57], which build a model by conditioning only on the class labels, Reed et al. [207] instead conditions to generate more substantial textual descriptions. It was the first end-to-end differentiable architecture from the character-level to pixel-level. Furthermore, it introduces a manifold interpolation regularizer for the GAN generator that significantly improves the quality of generated samples, including on the held out zero-shot categories.

MVRL WITH TWO VIEWS AND CORRELATED CENTROID SPACE

4

MVRL WITH TWO VIEWS AND CORRELATED CENTROID SPACE

Context of this Chapter. In this chapter, we leverage multi-view shallow representation learning and propose a novel approach for cross-modal retrieval to retrieve images given the textual query in different languages and vice versa. Our approach to retrieve semantically similar documents across modalities in different languages is termed as correlated centroid space unsupervised retrieval (C^2SUR) and consists of two phases. In the first phase, we extract heterogeneous features from a multimodal document and project it to a correlated space using kernel canonical correlation analysis (KCCA). In the second phase, correlated space centroids are obtained using clustering to retrieve cross-modal documents with different similarity measures. Experimental results show that C^2SUR outperforms the existing state-of-the-art English cross-modal retrieval approaches and achieve similar results for other languages.

Our main contributions of this chapter can be broadly summarized as:

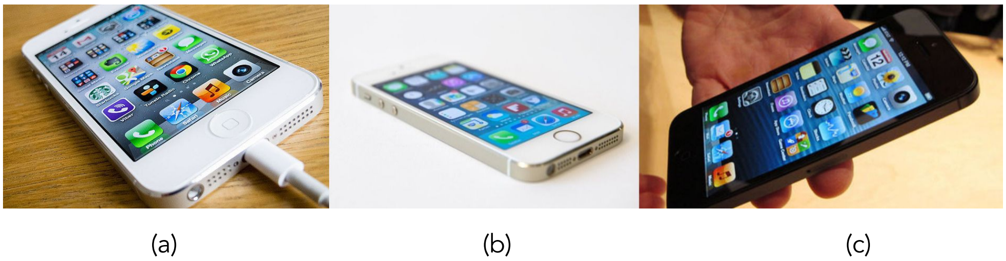
① We designed a novel approach to link text in multiple languages with visual content (i.e. images) and vice versa to facilitate multilingual cross-modal retrieval.

② We extended an existing dataset [18] to multiple languages to facilitate multilingual cross-modal research.

③ We provide empirical evidence to show that C^2SUR outperforms existing state of the art monolingual (English) cross-modal retrieval approaches.

Outline. The remainder of this chapter is organized into following sections. Initially, Section 4.1 presents the motivation in Section 4.1.1 and briefly introduce existing multi-view shallow representation learning approaches in the context of cross-modal retrieval in Section 4.1.2. Next Section 4.2 presents the research question and describes our contribution to cross-modal retrieval. Our approach i.e. C^2SUR is then discussed in the Section 4.3. The dataset and metrics used for evaluation of the approach are described in the Section 4.4. While in the Section 4.4.2 details about the evaluation results are shown, which are further analyzed in the Section 4.4.3. Summary of the chapter is presented in the Section 4.5.

[18] http://www.svcl.ucsd.edu/projects/crossmodal/



(a) (b) (c)

Figure 12: Images (i.e.,(a), (b), (c)) seen in the related news articles written in English, German and Spanish languages respectively.

4.1 introduction

4.1.1 *Motivation*

Web often comprise information which is present in varied modalities such as text, image, video or audio. Sometimes one or more modalities co-exist to represent a multimodal document as found in online news articles. They are either embedded with a video or an image along with the text in different languages. Figure 12 shows images[19] taken from news articles describing the same incident written in English[20], German[21] and Spanish[22] respectively. Similarly, multimodal articles are also found in other web sources such as blogs, social networks, Wikipedia and personal websites.

Mining multimodal documents pose numerous challenges. In the recent years, multimedia and computer vision communities have published considerable research in bridging the gap between modalities to facilitate cross-modal applications [203]. Their research aims to address the problems of automatic image tagging with class labels [181], usage of image queries for text retrieval [177] or vice versa. From an old Chinese proverb and its interpretations [142], we understand that "A picture is worth 10,000 words". Existing multimodal learning approaches [205, 206] has well adopted this for cross-modal retrieval. They combine visual information with text for both image and text-based retrievals. Other cross-modal approaches [171] leverage other modalities such as video and audio. However, most of the work about the text is limited to English.

Similarly, natural language processing(NLP) and information retrieval(IR) communities which work on different cross-lingual applications [196] concentrate only on text and diminish the importance of other modalities present in the multimodal document. Also, some of the cross-language retrieval systems are highly dependent on transliteration or translation tools [231] and support only keyword-based queries.

In this chapter, we aim to tackle this problem of cross-modal retrieval in a multilingual setting, by designing a cross-modal retrieval approach which is invariant to the languages present in a multimodal document. The method similar

[19] Images are of different resolution.
[20] http://bit.ly/1AUcpqG
[21] http://bit.ly/1rA3kCq
[22] https://tinyurl.com/yc24em9k

to our objective is by Wu et al. [279] who propose to identify novelty and redundancy with apparent duplicates in videos using cross-lingual news stories.

4.1.2 *Background on Related Learning Methods*

Our approach for cross-modal retrieval is dependent on the existing multi-view shallow representation learning methods and is comparable to other related approaches. Here, we discuss about existing methods which are already presented in the Section 2.2.1.1 and reiterate them in the context of cross-modal retrieval task.

4.1.2.1 *Text and Image as Input to CCA*

To build a low-dimension correlated space representation of two different modalities i.e text and image using CCA. Two sets of multivariate random variables $\mathbf{t} \in \mathbb{R}^{d_t}$ and $\mathbf{m} \in \mathbb{R}^{d_m}$ representing text and image modality respectively is chosen to find the projections $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{t}$ and $\mathbf{m}$ are highly correlated in the projected space. The transformation can be visualized in the Equation 33.

$$(\mathbf{t_p}, \mathbf{m_p}) \rightarrow (\mathbf{Ut}, \mathbf{Vm}) \tag{33}$$

where $\mathbf{Ut}$ represents the text projection, while $\mathbf{Vm}$ represents an image projection. In order to maximize this correlation ρ, we build an optimization function using Equation 34 with certain constraints as shown in Equation 35. We can observe that the optimization function is invariant to scaling. Also projections are constrained to unit variance [9].

$$\rho = \operatorname*{arg\,max}_{\mathbf{U},\mathbf{V}} \frac{\mathbf{U}^T C_{tm} \mathbf{V}}{\sqrt{\mathbf{U}^T C_{tt} \mathbf{U}} \sqrt{\mathbf{V}^T C_{mm} \mathbf{V}}} \tag{34}$$

$$\rho = \operatorname*{arg\,max}_{\mathbf{U}^T C_{tt} \mathbf{U} = \mathbf{V}^T C_{mm} \mathbf{V} = 1} \mathbf{U}^T C_{tm} \mathbf{V} \tag{35}$$

where C_{tt} represent covariance matrix of the text modality and C_{mm} represent covariance matrix of the image modality; while C_{tm} is a cross-covariance matrix between text and image modalities. Equation 34 is solved with a generalized eigenvalue problem to maximize the correlation by learning projections $\mathbf{U}$,$\mathbf{V}$ and given by the Equation 36 and Equation 37 respectively. Here, λ represent an eigenvalue.

$$C_{tt}^{-1} C_{tm} C_{mm}^{-1} C_{mt} \mathbf{U} = \lambda^2 \mathbf{U} \tag{36}$$

$$C_{mm}^{-1} C_{tm} C_{tt}^{-1} C_{tm} \mathbf{V} = \lambda^2 \mathbf{V} \tag{37}$$

4.1.2.2 *Text and Image as Input to KCCA*

Kernelization of CCA is helpful in finding the correlation between non-linear relationships [95]. Given any two sets of multivariate random variables $\mathbf{t} \in \mathbb{R}^{d_t}$ and $\mathbf{m} \in \mathbb{R}^{d_m}$ representing text and image modalities respectively. We find the kernel functions $K_T = k_T(t_i, t_j)$ and $K_I = k_I(m_i, m_j)$, such that K_T,$K_I \in \mathbb{R}^{n \times n}$ are both positive semi-definite kernel matrices. To find the correlation ρ_{kcca} between the transformed kernel matrices, we follow the similar optimization approach as of CCA given by Equation 38 and Equation 39.

$$\rho_{kcca} = \underset{\mathbf{X},\mathbf{Y}}{\arg\max} \frac{\mathbf{X}^T K_T K_I \mathbf{Y}}{\sqrt{\mathbf{X}^T K_T^2 \mathbf{X}}\sqrt{\mathbf{Y}^T K_I^2 \mathbf{Y}}} \tag{38}$$

$$\rho_{kcca} = \underset{\mathbf{X}^T K_T^2 \mathbf{X} = \mathbf{Y}^T K_I^2 \mathbf{Y} = 1}{\arg\max} \mathbf{X}^T K_T K_I \mathbf{Y} \tag{39}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the projections of $\mathbf{t}$ and $\mathbf{m}$ respectively in the projected correlated space.

4.1.2.3 *Other Related Approaches*

Several approaches have been proposed in bridging modalities with joint dimensionality reduction approaches [205, 206] using extended CCA with semantic class labels. Some approaches formulate an optimization problem [232] where the correlation between modalities is found by separating the classes in their respective feature spaces. As cross-modal data involves heterogeneous features, most of the approaches [294] aim in learning these features implicitly without any external representation. Zhai et al. [295] focus on the joint representation of multiple media types using joint representation learning which incorporates sparse and graph regularization. We use KCCA for maximizing the pair-wise correlation between different media as Blaschko et al. [27] used for correlational spectral clustering.

4.2 RESEARCH QUESTION AND CONTRIBUTIONS

Let us outline the research questions, hypotheses, and contributions, which we target throughout the chapter.

4.2.1 *Research Question and Hypothesis*

As presented in Section 1.3, our overall research question is: How to effectively integrate multiple views of training instances depicting heterogeneous or homogeneous content into a common space representation for supporting applications in different domains? In this chapter, we address the part, where learning correlation among views emerging from different modalities by leveraging their input representations. More specifically, we aim at Research Question 1:

✍ **Research Question 1**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist *search* by finding correlation among their input representations?

For addressing above research question, we verify hypothesis as follows:

□ **Hypothesis 1**

Leveraging shallow multi-view learning approaches such as kernel canonical correlation approaches (Kernel Canonical Correlation Analysis (KCCA)) can effectively learn the correlation between different modalities of the data emerging from heterogeneous sources. However, it also can be extended with more sophisticated approach (see Algorithm 1) for addressing applications such as the image to text retrieval or vice versa in multiple languages.

Intuitively, Hypothesis 1 states that the KCCA can be extended to useful capture correlation among the heterogeneous data. In particular, we expect that the extended KCCA to effectively discriminate the cross-modal data such that similar items are ranked closer while pushing away the dissimilar items. Further, we expect that the extension of KCCA implementation is more straightforward and computationally efficient.

To validate Hypothesis, we present our correlated centroid space unsupervised retrieval (C^{2}SUR) in Section 4.3 to support different languages. Moreover, we implemented the approach and empirically show (see the evaluation in Section 4.4) its effectiveness with state of the art.

4.2.2 *Contributions*

While being naturally appealing, KCCA and its extensions are not studied before in the context of cross-modal retrieval, where textual data can emerge from multiple languages. Aiming at above hypotheses, we provide the following contribution:

- *Contribution for Hypothesis 1*
  Building common space representation from the heterogeneous data using different principles of cross-modal association has been extensively studied from the past decade [153]. However, closest to our work is the use of correlation principles for English text and image retrieval or vice versa [205].

  However, our usage of KCCA and its extended variant in the context of retrieving images for a given query in different languages and vice versa is unique. Also, tackling the challenge thrown by the data, where textual data used has variable length queries, as opposed to short textual queries used in the standard text-based information retrieval. Additionally, representation of complex heterogeneous data requires a more lightweight implementation, existing works either made complex implementations or do not consider other nuances such as different languages.

Facing these characteristics, we propose (C^2SUR). To the best of our knowledge, this is the first work of cross-modal retrieval in the multilingual setting.

We conducted an evaluation using Wikipedia dataset in English, German and Spanish to validate the Hypotheses 1. In these experiments, we could achieve significant performance gains over the state of the art for English cross-modal retrieval, and new scores are reported for German and Spanish. In fact, we could show that our proposed approach (C^2SUR) leads to an improvement in mean average precision (MAP) scores for the image to text and vice versa retrieval.

4.3 CORRELATED CENTROID SPACE APPROACH

In this section, we first formulate the problem and then present the foundation to build correlated spaces using KCCA. Later, our approach C^2SUR is discussed, which is an extension to KCCA.

4.3.1 *Problem Formulation*

As discussed in the motivation Section 4.1.1, multimodal documents on the web are found in the form of pairwise modalities. Sometimes, there can be multiple instances of modalities present in a single document. To reduce the complexity, we assume a multimodal document $D_i = (Text, Media)$ to contain a single media item either an image, video or audio embedded with a textual description. A collection $C_j = \{D_1, D_2 ... D_i ... D_n\}$ of these documents in different languages $L = \{L_{C_1}, L_{C_2} ... L_{C_j} ... L_{C_m}\}$ are spread across web. Formally, our research question is to find a cross-modal semantically similar document across language collections L_{C_o} using unsupervised similarity measures on low-dimension correlation space representation. Figure 13 shows the broad visualization of approach.

4.3.2 *Correlated Space*

Initially, KCCA is utilized to attain correlated low-dimension space of heterogeneous representations. It is then used to find semantically similar cross-modal documents using different unsupervised similarity measures. For instance, various similarity measures such as Cosine Similarity, Normalized Correlation, Minkowski distance, etc. have been well adopted for clustering and other semantic similarity tasks. We leverage five such similarity measures, mainly Cosine, Correlation, Minkowski, Mahalanobis and Chebyshev to build our baseline approach termed as correlated space unsupervised retrieval (CSUR).

4.3.3 *Correlated Centroid Space*

In the correlated centroid space approach, we extend the aforementioned correlated space approach. Correlated low-dimension representation of text and

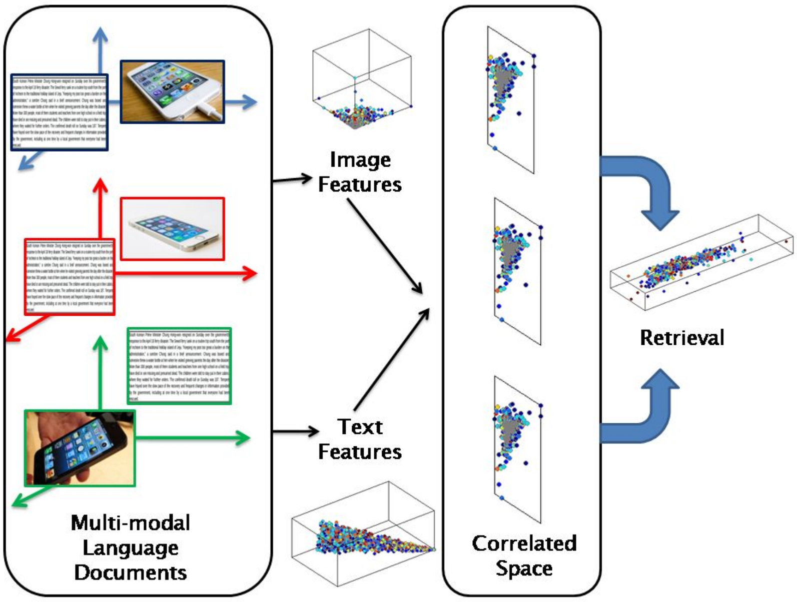


Figure 13: Correlated Space Retrieval

images attained with correlated space approach is replaced with it's closest centroids obtained using k-means clustering [97].

Let $m_T = \{m_{T_1}...m_{T_k}\}$ and $m_I = \{m_{I_1}...m_{I_k}\}$ denote the initial k centroids for the correlated text and image space respectively. Iterating over the samples of the training data, we perform assignment and update steps to obtain final k centroids. The assignment step assigns the each observed sample to its closest mean, while the update step calculates the new means that will be a centroid.

Correlated low-dimension representation of text and image samples of the training data is given by CS_{Tr_T} and CS_{Tr_I} respectively. Choice of k is dependent on number of classes in the training data, while p represents the total training samples. $S^{(t)}_{T_i}$ and $S^{(t)}_{I_i}$ denote new samples of text and image modalities assigned to its closest mean. Algorithm 1 lists the procedure. Now the modified feature space is used for cross-modal retrieval similar to CSUR and termed as C^{2}SUR.

```
Algorithm 1: Correlated Centroid Space
  Require: $CS_{Tr_T} = x_{T_1}...x_{T_p}, CS_{Tr_I} = x_{I_1}...x_{I_p}$
  Ensure: $p > 0$ {Output: Final K-Centroids}
    Assignment Step:
    $S^{(t)}_{T_i} = x_{T_j} : ||x_{T_j} - m_{T_i}|| \leqslant ||x_{T_j} - m_{T_{i^*}}|| \quad \forall i^* = 1...k$
    $S^{(t)}_{I_i} = x_{I_j} : ||x_{I_j} - m_{I_i}|| \leqslant ||x_{I_j} - m_{I_{i^*}}|| \quad \forall i^* = 1...k$
    Update Step:
    $m^{(t+1)}_{T_i} = \frac{\sum_{x_{T_j} \in S^{(t)}_{T_i}} x_{T_j}}{|S^{(t)}_{T_i}|}, m^{(t+1)}_{I_i} = \frac{\sum_{x_{I_j} \in S^{(t)}_{I_i}} x_{I_j}}{|S^{(t)}_{I_i}|}$
```

4.4 EVALUATION

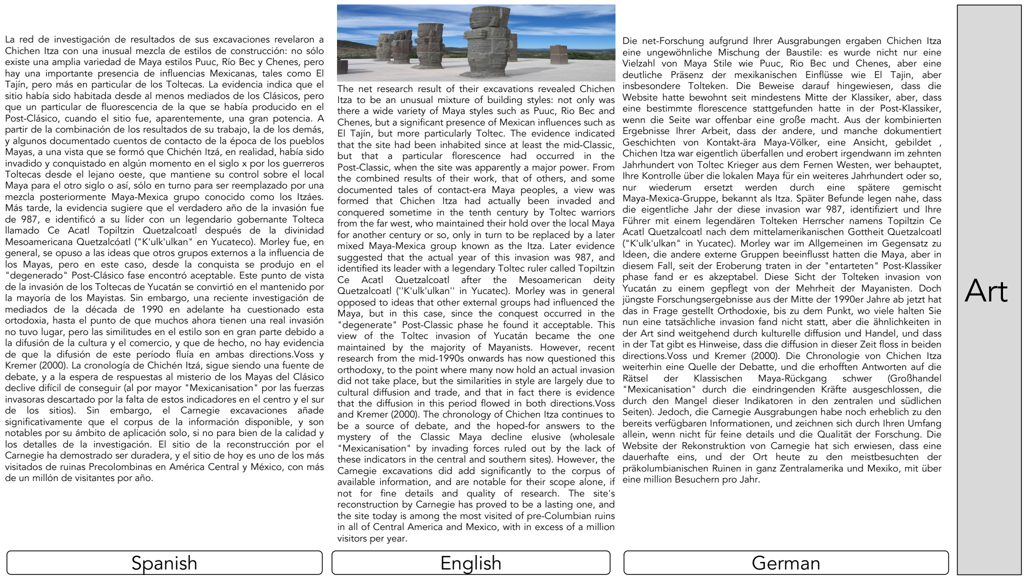
4.4.1 *Evaluation Setup*

In this section, we provide details about the dataset that is used and created to perform the experiments. Also, we describe features that are extracted from text and image modalities to learn a correlated space representation. It is then followed by methods used to evaluate the approach.

4.4.1.1 *Dataset Creation*

We used Wiki dataset[23] created for English texts and images using Wikipedia's featured articles. It has 2866 documents containing selected text paragraph and image pairs belonging to 10 semantic categories taken from art, biology, sport etc. We expanded the dataset into two more languages, mainly German and Spanish, using the Yandex machine translation API[24], while keeping the original images for every language. Thus, the expanded dataset consists of text and image pairs in three different languages. We relied on machine translation, as it is the most efficient way to create such a corpus.[25].

Figure 14 show the sample from the dataset representing category "art".



La red de investigación de resultados de sus excavaciones revelaron a Chichen Itza con una inusual mezcla de estilos de construcción: no sólo existe una amplia variedad de Maya estilos Puuc, Río Bec y Chenes, pero hay una importante presencia de influencias Mexicanas, tales como El Tajín, pero más en particular de los Toltecas. La evidencia indica que el sitio había sido habitada desde al menos mediados de los Clásicos, pero que un particular de fluorescencia de la que se había producido en el Post-Clásico, cuando el sitio fue, aparentemente, una gran potencia. A partir de la combinación de los resultados de su trabajo, la de los demás, y algunos documentado cuentos de contacto de la época de los pueblos Mayas, a una vista que se formó que Chichén Itzá, en realidad, había sido invadido y conquistado en algún momento en el siglo x por los guerreros Toltecas desde el lejano oeste, que mantiene su control sobre el local Maya para el otro siglo o así, sólo en turno para ser reemplazado por una mezcla posteriormente Maya-Mexica grupo conocido como los Itzáes. Más tarde, la evidencia sugiere que el verdadero año de la invasión fue de 987, e identificó a su líder con un legendario gobernante Tolteca llamado Ce Acatl Topiltzin Quetzalcoatl después de la divinidad Mesoamericana Quetzalcóatl ("K'ulk'ulkan" en Yucateco). Morley fue, en general, se opuso a las ideas que otros grupos externos a la influencia de los Mayas, pero en este caso, desde la conquista se produjo en el "degenerado" Post-Clásico fase encontró aceptable. Este punto de vista de la invasión de los Toltecas de Yucatán se convirtió en el mantenido por la mayoría de los Mayistas. Sin embargo, una reciente investigación de mediados de la década de 1990 en adelante ha cuestionado esta ortodoxia, hasta el punto de que muchos ahora tienen una real invasión no tuvo lugar, pero las similitudes en el estilo son en gran parte debido a la difusión de la cultura y el comercio, y que de hecho, no hay evidencia de que la difusión de este período fluía en ambas directions.Voss y Kremer (2000). La cronología de Chichén Itzá, sigue siendo una fuente de debate, y a la espera de respuestas al misterio de los Mayas del Clásico declive difícil de conseguir (al por mayor "Mexicanisation" por las fuerzas invasoras descartado por la falta de estos indicadores en el centro y el sur de los sitios). Sin embargo, el Carnegie excavaciones añade significativamente que el corpus de la información disponible, y son notables por su ámbito de aplicación solo, si no para bien de la calidad y los detalles de la investigación. El sitio de la reconstrucción por el Carnegie ha demostrado ser duradera, y el sitio de hoy es uno de los más visitados de ruinas Precolombinas en América Central y México, con más de un millón de visitantes por año.

The net research result of their excavations revealed Chichen Itza to be an unusual mixture of building styles: not only was there a wide variety of Maya styles such as Puuc, Rio Bec and Chenes, but a significant presence of Mexican influences such as El Tajín, but more particularly Toltec. The evidence indicated that the site had been inhabited since at least the mid-Classic, but that a particular florescence had occurred in the Post-Classic, when the site was apparently a major power. From the combined results of their work, that of others, and some documented tales of contact-era Maya peoples, a view was formed that Chichen Itza had actually been invaded and conquered sometime in the tenth century by Toltec warriors from the far west, who maintained their hold over the local Maya for another century or so, only in turn to be replaced by a later mixed Maya-Mexica group known as the Itza. Later evidence suggested that the actual year of this invasion was 987, and identified its leader with a legendary Toltec ruler called Topiltzin Ce Acatl Quetzalcoatl after the Mesoamerican deity Quetzalcoatl (''K'ulk'ulkan'' in Yucatec). Morley was in general opposed to ideas that other external groups had influenced the Maya, but in this case, since the conquest occurred in the "degenerate" Post-Classic phase he found it acceptable. This view of the Toltec invasion of Yucatán became the one maintained by the majority of Mayanists. However, recent research from the mid-1990s onwards has now questioned this orthodoxy, to the point where many now hold an actual invasion did not take place, but the similarities in style are largely due to cultural diffusion and trade, and that in fact there is evidence that the diffusion in this period flowed in both directions.Voss and Kremer (2000). The chronology of Chichen Itza continues to be a source of debate, and the hoped-for answers to the mystery of the Classic Maya decline elusive (wholesale "Mexicanisation" by invading forces ruled out by the lack of these indicators in the central and southern sites). However, the Carnegie excavations did add significantly to the corpus of available information, and are notable for their scope alone, if not for fine details and quality of research. The site's reconstruction by Carnegie has proved to be a lasting one, and the site today is among the most visited of pre-Columbian ruins in all of Central America and Mexico, with in excess of a million visitors per year.

Die net-Forschung aufgrund Ihrer Ausgrabungen ergaben Chichen Itza eine ungewöhnliche Mischung der Baustile: es wurde nicht nur eine Vielzahl von Maya Stile wie Puuc, Rio Bec und Chenes, aber eine deutliche Präsenz der mexikanischen Einflüsse wie El Tajin, aber insbesondere Tolteken. Die Beweise darauf hingewiesen, dass die Website hatte bewohnt seit mindestens Mitte der Klassiker, aber, dass eine bestimmte florescence stattgefunden hatte in der Post-Klassiker, wenn die Seite war offenbar eine große macht. Aus der kombinierten Ergebnisse Ihrer Arbeit, dass der andere, und manche dokumentiert Geschichten von Kontakt-ära Maya-Völker, eine Ansicht, gebildet , Chichen Itza war eigentlich überfallen und erobert irgendwann im zehnten Jahrhundert von Toltec Krieger aus dem Fernen Westen, wer behauptet, Ihre Kontrolle über die lokalen Maya für ein weiteres Jahrhundert oder so, nur wiederum ersetzt werden durch eine spätere gemischt Maya-Mexica-Gruppe, bekannt als Itza. Später Befunde legen nahe, dass die eigentliche Jahr der diese invasion war 987, identifiziert und Ihre Führer mit einem legendären Tolteken Herrscher namens Topiltzin Ce Acatl Quetzalcoatl nach dem mittelamerikanischen Gottheit Quetzalcoatl ("K'ulk'ulkan" in Yucatec). Morley war im Allgemeinen im Gegensatz zu Ideen, die andere externe Gruppen beeinflusst hatten die Maya, aber in diesem Fall, seit der Eroberung traten in der "entarteten" Post-Klassiker phase fand er es akzeptabel. Diese Sicht der Tolteken invasion von Yucatán zu einem gepflegt von der Mehrheit der Mayanisten. Doch jüngste Forschungsergebnisse aus der Mitte der 1990er Jahre ab jetzt hat das in Frage gestellt Orthodoxie, bis zu dem Punkt, wo viele halten Sie nun eine tatsächliche invasion fand nicht statt, aber die ähnlichkeiten in der Art sind weitgehend durch kulturelle diffusion und Handel, und dass in der Tat gibt es Hinweise, dass die diffusion in dieser Zeit floss in beiden directions.Voss und Kremer (2000). Die Chronologie von Chichen Itza weiterhin eine Quelle der Debatte, und die erhofften Antworten auf die Rätsel der Klassischen Maya-Rückgang schwer (Großhandel "Mexicanisation" durch die eindringenden Kräfte ausgeschlossen, die durch den Mangel dieser Indikatoren in den zentralen und südlichen Seiten). Jedoch, die Carnegie Ausgrabungen habe noch erheblich zu den bereits verfügbaren Informationen, und zeichnen sich durch Ihren Umfang allein, wenn nicht für feine details und die Qualität der Forschung. Die Website der Rekonstruktion von Carnegie hat sich erwiesen, dass eine dauerhafte eins, und der Ort heute zu den meistbesuchten der präkolumbianischen Ruinen in ganz Zentralamerika und Mexiko, mit über eine million Besuchern pro Jahr.

Figure 14: Example showing the image and its textual description in English, German and Spanish from the semantic category ("art").

4.4.1.2 *Text and Image Representation*

To acquire representations for both text and images, different feature extraction approaches are adopted. For the text, we used polylingual topic models

[23] http://www.svcl.ucsd.edu/projects/crossmodal/

[24] http://api.yandex.com/translate/

[25] Please note, that the approach is invariant to machine translation and capable of *cross-lingual* cross-modal retrieval

(PTM) [176] to extract representation as a distribution of topics in multiple languages. We leveraged the large collections that have interlingual connections like Wikipedia to train the PTM across languages. A trained PTM model on Wikipedia provides the same topic distribution on English, German and Spanish. We have trained PTM model for 10, 100, and 200 topics using the text of around 250k wikipedia articles in each language. The concentration parameter α is initialized to 1T. Using the training and testing parts of our dataset, each text document is represented as 10, 100 and 200 dimension topic distribution vectors. Similarly, each image is represented as 128-dimension SIFT descriptor histograms as used in earlier works [205, 206].

4.4.1.3 *Evaluation Measures*

We evaluated cross-modal retrieval using mean average precision (MAP) [205, 206] and mean reciprocal rank (MRR) [264] scores. Experiments were repeated 10 times with different combinations of training and testing data to reduce selection bias. We used the same split as in Rasiwasia [205] for all languages to create 2173 training documents and 693 testing documents.

4.4.2 *Evaluation Results*

Using the dataset created for different languages, we segregate the tasks and evaluate them separately. First, we attain the MAP and MRR scores obtained for text and image queries using 10 text topics and 128-dimension SIFT descriptor histograms. Then, we show the variation in MAP scores by changing the number of topics.

4.4.2.1 *Text Query - Image Retrieval*

We used the text queries from testing data to find semantically similar images present in testing data. Text from testing data is projected into correlated space of images and text using the projection matrices trained with training data to retrieve images belonging to the same semantic category. Table 1 and Table 2 shows the MAP and MRR results[26] with standard deviation obtained for English, German and Spanish using CCA, Polynomial kernel with degree 2(poly-2) CCA and RBF kernel CCA with CSUR and C^{2}SUR approach respectively.

For the text query, we performed "unpaired t-test" between best performing methods of CSUR and C^{2}SUR for testing statistical significance. The two-tailed P value is less than 0.0001 for all languages, which is considered to be extremely statistically significant.

4.4.2.2 *Image Query - Text Retrieval*

We used image queries from the testing data to find the semantically similar text in the testing data. Image from the testing data is projected into common space of images and text using the projection matrices trained with training data

[26] Tables show only those similarity measures which obtained best results for each of the given kernels.

	Method	MAP	MRR
English	CCA-Mahalanobis	0.224 ± 0.002	0.241 ± 0.001
	(Poly-2)CCA-Correlation	0.233 ± 0.001	0.247 ± 0.002
	(RBF)CCA-Correlation	**0.235** ± 0.005	**0.250** ± 0.003
German	CCA-Cosine	0.219±0.003	0.242 ± 0.002
	(Poly-2)CCA-Chybyshev	**0.256** ± 0.001	**0.308** ± 0.002
	(RBF)CCA-Correlation	0.246 ± 0.003	0.272 ± 0.001
Spanish	CCA-Cosine	0.208 ± 0.002	0.223 ± 0.001
	(Poly-2)CCA-Cosine	**0.249** ± 0.002	**0.283** ± 0.003
	(RBF)CCA-Correlation	0.229 ± 0.002	0.249 ± 0.003

Table 1: Text Query - Image Retrieval **(CSUR)**

	Method	MAP	MRR
English	CCA-Correlation	0.245 ± 0.003	0.273 ± 0.002
	(Poly-2)CCA-Chebyshev	0.245 ± 0.002	0.259 ± 0.001
	(RBF)CCA-Correlation	**0.262** ± 0.003	**0.277** ± 0.001
German	CCA-Correlation	0.215 ± 0.001	0.246 ± 0.002
	(Poly-2)CCA-Correlation	**0.263** ± 0.003	**0.265** ± 0.002
	(RBF)CCA-Chebyshev	0.226 ± 0.002	0.255 ± 0.003
Spanish	CCA-Chebyshev	0.230 ± 0.003	0.255 ± 0.002
	(Poly-2)CCA-Chebyshev	0.259 ± 0.002	0.267 ± 0.001
	(RBF)CCA-Correlation	**0.268** ± 0.002	**0.268** ± 0.002

Table 2: Text Query - Image Retrieval **(C^{2}SUR)**

to retrieve images belonging to the same semantic category. Table 3 and Table 4 shows the MAP and MRR results[27] with standard deviation obtained for English, German and Spanish using CCA, Polynomial kernel with degree 2(poly-2) CCA and RBF kernel CCA using CSUR and C^{2}SUR approach respectively.

For the image query, "unpaired t-test" between best performing methods of CSUR and C^{2}SUR showed that two-tailed P value equals 0.0111 for German and less than 0.0001 for Spanish. Although, there was no significant improvement for English. Topic distribution of text can show influence on the cross-modal retrieval. To apprehend it, we evaluated C^{2}SUR approach on various kernels with different topic distributions. Figure 15, Figure 16 and Figure 17 shows the average of MAP scores obtained for text and image queries using different similarity measures.

[27] Tables only show those similarity measures which obtained best results for each of the given kernels.

	Method	MAP	MRR
English	CCA-Minkowski	0.241 ± 0.002	0.263 ± 0.001
	(Poly-2)CCA-Correlation	0.239 ± 0.002	0.256 ± 0.002
	(RBF)CCA-Mahalanobis	**0.273** ± 0.003	**0.311** ± 0.002
German	CCA-Mahalanobis	0.219 ± 0.001	0.233 ± 0.002
	(Poly-2)CCA-Minkowski	**0.282** ± 0.001	**0.275** ± 0.001
	(RBF)CCA-Mahalanobis	0.248 ± 0.002	0.271 ± 0.001
Spanish	CCA-Chebyshev	0.220 ± 0.002	0.234 ± 0.001
	(Poly-2)CCA-Cosine	**0.238** ± 0.001	**0.257** ± 0.003
	(RBF)CCA-Cosine	0.225 ± 0.004	0.238 ± 0.002

Table 3: Image Query - Text Retrieval **(CSUR)**

	Method	MAP	MRR
English	CCA-Chebyshev	0.253 ± 0.002	0.257 ± 0.003
	(Poly-2)CCA-Chebyshev	**0.273** ± 0.002	**0.293** ± 0.002
	(RBF)CCA-Chebyshev	0.263 ± 0.003	0.287 ± 0.002
German	CCA-Chebyshev	0.226 ± 0.003	0.252 ± 0.002
	(Poly-2)CCA-Minkowski	0.231 ± 0.001	0.241 ± 0.002
	(RBF)CCA-Correlation	**0.284** ± 0.002	**0.274** ± 0.001
Spanish	CCA-Minkowski	**0.250** ± 0.001	**0.284** ± 0.002
	(Poly-2)CCA-Correlation	0.231 ± 0.003	0.258 ± 0.002
	(RBF)CCA-Chebyshev	0.219 ± 0.002	0.244 ± 0.003

Table 4: Image Query - Text Retrieval **(C^{2}SUR)**

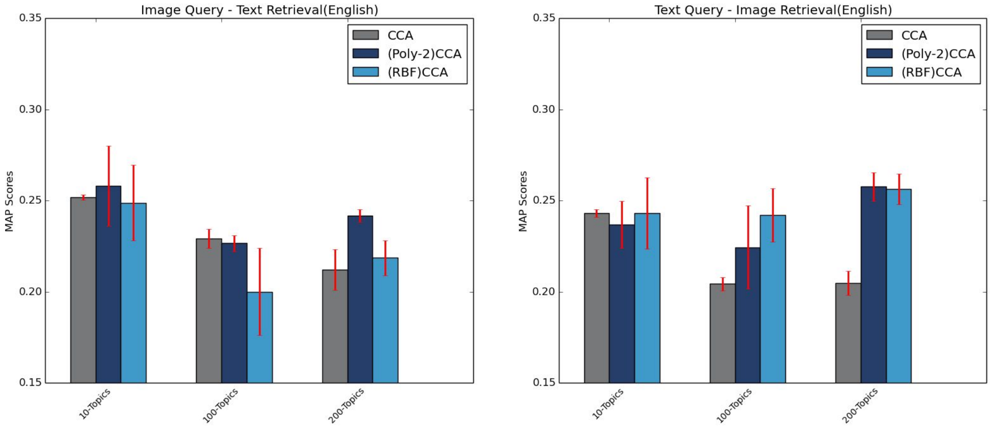


Figure 15: English-C^{2}SUR

4.4.2.3 *Cross-modal Retrieval Comparison*

Most of the earlier works [205, 232, 206] performed cross-modal experiments only on English text with 10-topics and 128-dimension SIFT image features. We compared the best methods of CSUR and C^{2}SUR with the existing approaches[28].

[28] Cluster-CCA [206] and Cluster-KCCA [206] approaches are not directly comparable with ours. They compare the cluster labels of instances, while we compare the original semantic category labels

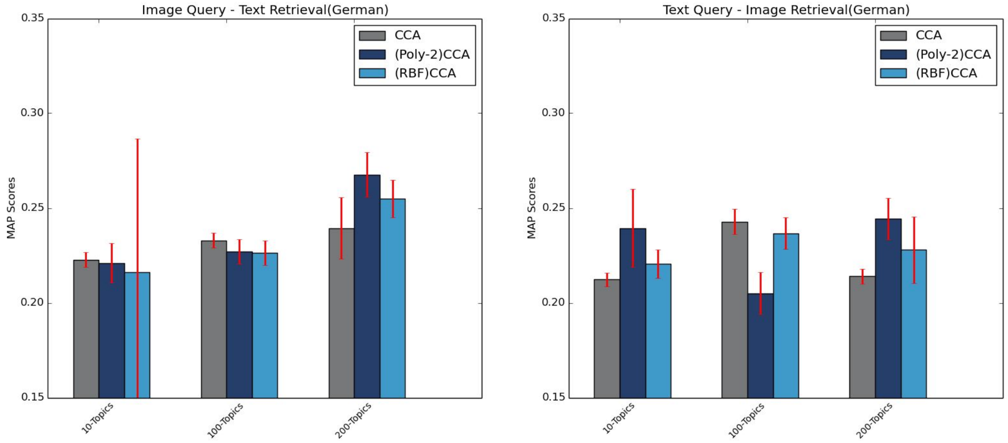


Figure 16: German-C^{2}SUR

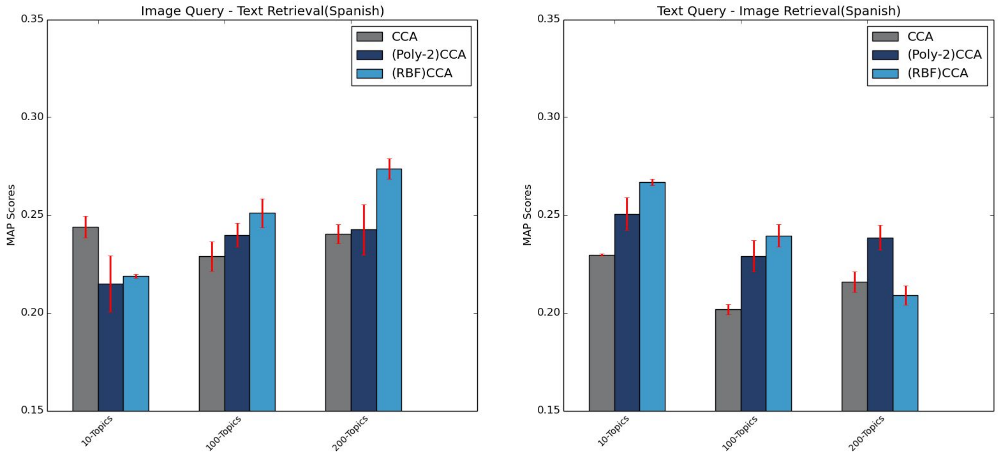


Figure 17: Spanish-C^{2}SUR

Table 5 shows the comparison on text and image queries for English, German and Spanish on the Wiki dataset. We show the best MAP scores for CSUR and C^{2}SUR for German and Spanish with different topic variations. For Example, CSUR-10 represent 10-topics. Please note, that the related work can only be applied to English text.

4.4.3 *Evaluation Results Analyses*

In this section, we analyzed the results obtained using our proposed approaches to perform cross-modal retrieval.

Table 1 and Table 2 shows the results attained using text queries for image retrieval with CSUR and C^{2}SUR approaches respectively. It can be inferred that kernel versions of CCA (KCCA) in both the approaches outperformed baseline CCA on MAP scores. Best performing kCCA used in CSUR and C^{2}SUR approaches had an average improvement of 0.029 and 0.034 respectively over base-

	Method	Image Query	Text Query	Average (MAP)
English	SM [205]	0.225	0.223	0.224
	Mean-CCA [206]	0.246 ± 0.005	0.194 ± 0.005	0.220 ± 0.005
	SCDL [272]	0.252	0.198	0.225
	SliM^2 [306]	0.255	0.202	0.229
	GMLDA [232]	0.272	0.232	0.252
	CSUR-10	**0.273** ± 0.003	0.235 ± 0.005	0.254 ± 0.004
	C^{2}SUR-10	**0.273** ± 0.002	**0.262** ± 0.003	**0.268** ± 0.003
German	CSUR-10	0.282 ± 0.001	0.256 ± 0.001	0.269 ± 0.001
	CSUR-100	0.230 ± 0.002	0.242 ± 0.004	0.236 ± 0.003
	CSUR-200	0.240 ± 0.002	0.243 ± 0.004	0.241 ± 0.003
	C^{2}SUR-10	**0.284** ± 0.002	**0.263** ± 0.003	**0.276** ± 0.003
	C^{2}SUR-100	0.236 ± 0.004	0.250 ± 0.008	0.243 ± 0.006
	C^{2}SUR-200	0.278 ± 0.002	0.253 ± 0.002	0.266 ± 0.002
Spanish	CSUR-10	0.238 ± 0.001	0.249 ± 0.002	0.244 ± 0.002
	CSUR-100	0.254 ± 0.003	0.236 ± 0.003	0.245 ± 0.003
	CSUR-200	0.259 ± 0.002	0.231 ± 0.002	0.245 ± 0.002
	C^{2}SUR-10	0.250 ± 0.001	**0.268** ± 0.002	**0.259** ± 0.002
	C^{2}SUR-100	0.258 ± 0.008	0.243 ± 0.004	0.251 ± 0.006
	C^{2}SUR-200	**0.267** ± 0.003	0.244 ± 0.002	0.256 ± 0.003

Table 5: Text and Image Query Comparison **(Wiki)**

line CCA in all languages. It shows the presence of non-linearity in the data. Also, the best approach in C^{2}SUR achieved an average improvement of 0.017 over the best approach of CSUR in all languages. It exhibits the efficiency of C^{2}SUR in eliminating the noisy information from the correlated space of text and image. A similar analysis can be performed on the image queries.

Table 3 and Table 4 show the results obtained using image queries for text retrieval with CSUR and C^{2}SUR respectively. Similar to text query, best performing kCCA used in CSUR and C^{2}SUR approaches had an average improvement of 0.037 and 0.019 respectively over baseline CCA in all languages. Also, the best performing approach of C^{2}SUR attained an average improvement of 0.007 over the best approach of CSUR in all languages.

Effect of text topic distribution on C^{2}SUR approach is evaluated with different text topic distributions and fixed 128-dimension SIFT image features. It can be observed from the Figure 15 that increase in the number of topics can have an adverse effect. A possible explanation is due to padding of zeros in the correlated space of training data to carry out similarity measures with the testing data. For negating the earlier mentioned behavior, dimensions also have to be increased for image features.

We also compared our best performing approach with the existing approaches based on MAP scores for English cross-modal retrieval. Table 5 shows that C^{2}SUR outperforms existing approaches on the average MAP scores. We assume this is due to the ability of C^{2}SUR to efficiently reduce the error in correlation space by improving the classification of borderline samples. Besides, performance on German and Spanish was comparable to English in finding semantically similar documents across modalities.

4.5 SUMMARY

In this chapter, we addressed the first research question:

> ✍ **Research Question 1**
>
> Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist *search* by finding correlation among their input representations?

For this, we validated Hypothesis 1 by proposing a novel a novel approach C^{2}SUR to perform the cross-modal retrieval in multiple languages. We built a shared space for the heterogeneous representations of a multimodal document using KCCA, which is further modified with K-Means centroids to retrieve similar documents. We found that C^{2}SUR is useful in finding semantically similar multimodal documents across languages.

In the next chapter, we will present an approach to achieving multi-view representation learning with consensus and complementarity principles and support a different application where the heterogeneous data emerge from different languages.

MVRL WITH TWO VIEWS AND CO-REGULARIZATION

5

MVRL WITH TWO VIEWS AND CO-REGULARIZATION

Context of this Chapter. In this chapter, we leverage multi-view shallow representation learning and propose a novel approach for cross-language text classification. In many languages, sparse availability of resources causes numerous challenges for textual analysis tasks. Text classification is one of such standard tasks that hinders due to limited availability of label information in low-resource languages. Transferring knowledge (i.e. label information) from high-resource to low-resource languages might improve text classification as compared to the other approaches like machine translation. We introduce BRAVE (*Bilingual paRAgraph VEctors*), a model to learn bilingual distributed representations (i.e. embeddings) of words without word alignments either from sentence-aligned parallel or label-aligned non-parallel document corpora to support cross-language text classification. The empirical analysis shows that classification models trained with our bilingual embeddings outperform other state-of-the-art systems on three different cross-language text classification tasks.

Our main contributions presented in this chapter can be broadly summarized as follows:

① We jointly train monolingual part of parallel corpora with the improved cross-lingual alignment function that extends beyond bag-of-word models.

② We introduced a novel approach to leverage non-parallel data sets such as label or class aligned documents in different languages for learning bilingual cues.

③ We performed an experimental evaluation on three different CLTC tasks, namely cross-language document classification, multi-label classification and cross-language sentiment classification using learned bilingual word embeddings.

Outline. The remainder of this chapter is organized into following sections. Initially, Section 5.1 presents the motivation in Section 5.1.1 and briefly introduce existing multi-view shallow representation learning approaches in the context of learning representations for variable length text in Section 5.1.2. Next Section 5.2 presents the research question and describes our contribution to cross-language text classification. Our approach i.e. BRAVE and its variations are then discussed in the Section 5.3. The dataset and metrics used for evaluation of the approach are described in the Section 5.4. While in the Section 5.4.2 details about the evaluation results are shown, which are further analyzed in the Section 5.4.3. Summary of the chapter is presented in the Section 5.5.

5.1 introduction

5.1.1 *Motivation*

The availability of language-specific annotated resources is crucial for the efficiency of natural language processing tasks. Still, many languages lack rich annotated resources that support various tasks such as part-of-speech (POS) tagging [265], dependency parsing [187] and text classification [1]. While the growth of multilingual information on the web has provided an opportunity to build these missing annotated resources, but still lots of manual effort is required to achieve high-quality resources for every language separately.

Another possibility is to utilize the unlabeled data present in those languages or transfer knowledge from annotation-rich languages. For the first alternative, recent advancements made in learning monolingual distributed representations of words [173, 195, 151] (i.e., monolingual word embeddings) capturing syntactic and semantic information in an unsupervised manner was useful in numerous NLP tasks [49]. However, this may not be sufficient for several other tasks such as cross-language information retrieval [196], cross-language word semantic similarity [266], cross-language text classification (CLTC, henceforth) [134, 282, 200, 250] and machine translation [300] due to irregularities across languages. In this kind of scenarios, transfer of knowledge can be useful.

Several approaches [103, 225, 90, 50] induced monolingual distributed representations into a language independent space (i.e., bilingual or multilingual word embeddings) by jointly training on a pair of languages. Although the overall goal of these approaches is to capture linguistic regularities in words that share same semantic and syntactic space across languages, they differ in their implementation. One set of methods either performed offline alignment of trained monolingual embeddings or jointly-trained both the monolingual and cross-lingual objectives, while the other set utilized only cross-lingual objective. Jointly-trained or offline alignment methods can be further divided based on the type of parallel corpus (e.g., word-aligned, sentence-aligned) they use for learning the cross-lingual objective. Table 6 summarizes different setups to learn bilingual or multilingual embeddings for the various tasks.

Methods in the Table 6 that use word-aligned parallel corpus as offline alignment [174, 73] assume the single correspondence between the words across languages and ignore polysemy. While the jointly-train methods [134] that use word-alignment parallel corpus and consider polysemy perform a computationally expensive operation of considering all possible interactions between the pairs of words in the vocabulary of two different languages. Methods [103, 225] that overcame the complexity issues of word-aligned models by using sentence-aligned parallel corpora limits themselves to only cross-lingual objective, thus making these approaches unable to explore monolingual corpora. Jointly-trained models [90, 50] overcame the issues of both word-aligned and purely cross-lingual objective models by using monolingual and sentence-aligned parallel

Cross-Language Setups			
Objective	Method	Tasks	Parallel Corpus
Monolingual+ Cross-lingual	klementiev et al. [134]	CLDC	Word-Aligned
	Zou et el. [308]	MT,NER	Word-Aligned
	Mikolov et al. [174]	MT	Word-Aligned
	Faruqi et al. [73]	Word Similarity	Word-Aligned
	Lu et al. [162]	Word Similarity	Word-Aligned
	Gouws et al. [89]	POS,SuS	Word-Aligned
	Gouws et al. [90]	CLDC,MT	Sentence-Aligned
	Coul et al. [50]	CLDC,MT	Sentence-Aligned
Cross-lingual	Hermann et al. [103]	CLDC	Sentence-Aligned
	Lauly et al. [225]	CLDC	Sentence-Aligned
	Luong et al. [164]	Word Similarity, CLDC	Sentence-Aligned
	Pham et al. [197]	CLDC	Sentence-Aligned

Table 6: Summary of bilingual or multilingual embedding methods that support Cross-language Document Classification (CLDC), Machine Translation (MT), Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Super Sense Tagging (SuS).

corpora. Nonetheless, these approaches still have certain drawbacks such as usage of only bag-of-words from the parallel sentences ignoring the order of words. Thus, they are missing to capture the non-compositional meaning of the entire sentence. Also, learned bilingual embeddings were heavily biased towards the sampled sentence-aligned parallel corpora. It is also sometimes hard to acquire sentence-level parallel corpora for every language pair. To subdue this concern, few approaches [204] used pivot languages like English or comparable document-aligned corpora [267] to learn bilingual embeddings specific to only one task.

This major downside can be observed in other methods above also, which are inflexible to handle different types of parallel corpora and have a tight-binding between cross-lingual objectives and the parallel corpora. For example, a method using sentence-level parallel corpora cannot be altered to leverage document-level parallel corpora (if available) that might have better performance for some tasks. Also, none of the approaches do leverage widely available label/class-aligned non-parallel documents (e.g. sentiment labels, multi-class datasets) across languages which share special semantics such as sentiment or correlation between concepts as opposed to parallel texts.

In this chapter, we introduce BRAVE a shallow neural network based multi-view representation learning approach. It is a jointly-trained flexible model to learn bilingual embeddings based on the availability of the type of corpora (e.g. sentence-aligned parallel or label/class-aligned non-parallel document) by just altering the cross-lingual objective. BRAVE leverages paragraph vector embeddings [144] of the monolingual corpora to effectively conceal semantics of the text sequences across languages and build a cross-lingual objective. Method closely related to our approach is by Pham et al. [197] who uses shared context sentence vector across languages to learn multilingual text sequences.

5.1.2 *Background on Variable Length Distributed Representations*

Natural language text can be segmented into many meaningful units such as words, phrases, sentences and paragraphs. For effective natural language understanding, depending on the domain and context, different type of segmentations play a prominent role. Also for several other tasks of natural language processing, building distributed representation [106] for each of these meaningful units has become crucial. In the following, we discuss some related approaches.

5.1.2.1 *Word Distributed Representation*

The distributed representation is learned based on the usage of words. This allows words that are used in similar manner to acquire similar representations, naturally capturing their meaning. This has been supported by theoretical linguistic studies based on distributional hypothesis [96]. Word distributed representation a.k.a word embedding constitute real-valued vector representation and words are usually obtained as fixed vocabulary of the textual corpus. Approaches [22, 173, 195, 151] which are proposed in the past few years mostly leverage shallow neural network architectures and optimize for some task (e.g. document classification or language modeling) or learn in an unsupervised manner.

Out of existing approaches, we present here details about word2vec [173] and discuss about its two different learning models 1) Continuous Bag-of-Words (CBOW) and 2) Continuous Skip-Gram.

CBOW

Given the context, the CBOW model learns the embedding by predicting the current word based on its context.

Skip-Gram

Alternative to CBOW, the continuous skip-gram model learns embedding by predicting the surrounding words given a current word.

5.1.2.2 *Beyond Word Distributed Representation*

Requirement for representations that go beyond words and cater larger pieces of text such as phrases, paragraphs and documents have spawn interest in building shallow and deep neural network architectures. Several approaches [144, 154, 249] are proposed either to optimize for a task (e.g. classification) or learn in an unsupervised manner. Out of existing approaches, we present here details about Paragraph Vectors [144] and discuss about its two different learning models 1) A distributed memory and 2) Distributed bag of words.

Distributed bag of words (PV-DBOW)

This approach is seen similar to the aforementioned CBOW model. This method considers the concatenation of the paragraph vector with the word embeddings to predict the next word in a text window.

Distributed Memory Model (PV-DM)

In this model, paragraph vectors are asked to contribute to the prediction task of the next word given many contexts sampled from the paragraph. The paragraph vector and word embeddings are averaged or concatenated to predict the next word in a context.

5.2 RESEARCH QUESTION AND CONTRIBUTIONS

Let us outline the research questions, hypotheses, and contributions, which we target throughout the chapter.

5.2.1 *Research Question and Hypothesis*

As presented in Section 1.3, our overall research question is: How to effectively integrate multiple views of training instances depicting heterogeneous or homogeneous content into a common space representation for supporting applications in different domains? In this chapter, we address the second part, i.e., learning a common space representation among views emerging from the same modality with consensus and complementarity principles by leveraging their input representations. More specifically, we aim at Research Question 2:

✍ **Research Question 2**

Given two different views of homogeneous content depicting text from different languages, how can we build a shared representation to assist *categorization* by learning a common space by capturing regularities?

For addressing above research question, we verify the hypothesis as follows:

□ **Hypothesis 2**

Leveraging shallow neural network architecture and manifold alignment approach, we can efficiently and effectively learn a common space representation between the data emerging from two different languages and capture their regularities. More specifically, by leveraging co-regularization approach which is built on the ideas of consensus and complementarity principles will benefit to build bilingual distributed word representations, i.e., embeddings. Furthermore, the usefulness of these representations can be optimized based on the cross-language textual classification task.

Intuitively, Hypothesis 2 states that the combination of shallow neural networks with the manifold alignment techniques can effectively capture regularities across different languages text. In particular, we expect to extend the paragraph vector approach with co-regularization to effectively learn common space representation of cross-language data such that similar words are aligned closer to each other in a high dimensional space. Further, we expect that extension of this approach with manifold alignment technique can leverage pseudo-parallel data for learning common space representation.

To validate Hypothesis, we present our BRAVE models in the Section 5.3 to build bilingual distributed word representations. Moreover, we implemented the approach and empirically show (see the evaluation in Section 5.4) its effectiveness with state of the art.

5.2.2 *Contributions*

While being naturally appealing, usage of paragraph vectors and its combination with manifold alignment techniques are not studied before in the context of cross-language text classification, where the textual data can emerge from multiple languages. Aiming at above hypotheses, we provide the following contribution:

- *Contribution for Hypothesis* 2
  Building a common space representation from the data using different correlation or consensus/complementarity principles is studied earlier.

  However, our usage of paragraph vectors and its extension for building common space representations for different languages is unique. Additionally, handling cross-language data which is not parallel and of variable length throws new challenges. Facing these characteristics, we propose BRAVE and its variations. To the best of our knowledge, this is the first work which utilizes manifold alignment techniques for building bilingual distributed word representations that are useful for many cross-language tasks (e.g., cross-language text classification).

  Therefore, we picked one of many cross-language tasks, i.e., cross-language text classification for evaluation and validated the Hypotheses 2. In these experiments, we could achieve performance gains over state of the art for cross-language document classification (CLDC), cross-language sentiment classification (CLSC) and reported comparable results for the multi-label CLDC. In fact, we could show that our proposed approach BRAVE was useful for building bilingual word distributed representations.

5.3 BRAVE MODELS

In this section, we present our BRAVE model along with its variations whose aim is to learn bilingual embeddings that can generalize across different languages.

5.3.1 *Bilingual Paragraph Vectors (BRAVE)*

Most of the NLP tasks require fixed-length representations. Tasks like CLTC also require fixed-length representation to incorporate inherent semantics of sentences or documents. Distributed representation of sentences and documents i.e. paragraph vectors [144] are designed to out-perform certain text classification tasks by overcoming constraints posed by the bag-of-words models.

Here, we leverage paragraph vectors distributed memory model (PV-DM) as the monolingual objective $\mathcal{M}(\cdot)$ and jointly optimize with bilingual regularization function $\varphi(\cdot)$ for learning bilingual embeddings similar to the earlier approaches [90, 50]. Equation 40 shows the formulation of the overall objective function that is minimized.

$$\mathcal{L} = \min_{\theta^{l_1},\theta^{l_2}} \sum_{l\in\{l_1,l_2\}} \sum_{C^l} \mathcal{M}^l(w_t, h; \theta^l) + \frac{\lambda\varphi(\theta^{l_1},\theta^{l_2})}{2} \tag{40}$$

Here, C^l represent the corpus of individual languages (i.e. l_1 or l_2). Given any sequence of words $(w^l_1, w^l_2 ... w^l_T)$ in C^l, w_t is the predicted word in a context h constrained on paragraph p (i.e. sentence or document) and sequence of words.

Formally, the first term (i.e. $\mathcal{M}(\cdot)$) in the Equation 40 maximizes the average log probability based on word vector matrix $\boldsymbol{W}^l$ and a unique paragraph vector matrix $\mathbf{P}^l$. Equation 41 represents the average log probability.

$$\mathcal{M}^l(w_t, h; \theta^l) = \frac{\Sigma_{t=k}^{T-k} y^l_{w_t} - \log(\sum_i e^{y^l_i})}{T} \tag{41}$$

where each y^l_i is log-probability of predicted word i and is given by Equation 42.

$$y^l = b + \mathbf{U}h(w^l_{t-k}....w^l_{t+k}; \boldsymbol{W}^l, \mathbf{P}^l) \tag{42}$$

To optimize for efficiency, hierarchical softmax [179] is used in training with $\mathbf{U}$ and b as parameters. Binary Huffmann tree is utilized to represent hierarchial softmax [173]. Analogous to Pham et al., [197], we also derive h by concatenating paragraph vector from $\mathbf{P}^l$ with the average of word vectors in $\boldsymbol{W}^l$. This helps to fine tune both word and paragraph vectors independently.

Now, to capture the bilingual cues, the regularization function ($\varphi(\cdot)$) is learned in two different ways. In the first approach a sentence-aligned parallel corpora is used, while in the second approach a label-aligned document corpora.

5.3.2 *BRAVE with Sentence-Aligned Parallel corpora (BRAVE-S)*

To compute the bilingual regularization function $\varphi(\cdot)$, we slightly deviate from earlier approaches [90]. Instead of simply performing L2-loss between the mean of word vectors in each sentence pair $(s^{l_1}_j, s^{l_2}_j)$ of the sentence-aligned parallel corpus (PC) at each training step. We use the concept of elastic net regularization [307] and employ linear combination of L2-loss between *sentence paragraph vectors* $\mathbf{sp}^{l_1}_j$ and $\mathbf{sp}^{l_2}_j \in \mathbb{R}^d$ precomputed from the monolingual term $\mathcal{M}(\cdot)$ with L2-loss between the mean of word vectors observed in sentences. This induces a constraint on the usage of monolingual part of parallel training data to learn $\mathcal{M}(\cdot)$. At the same time, it has an advantage of using combination of paragraph and word vectors which combines compositional and non-compositional meanings of sentences.

Also, it eliminates the need for word-alignment and makes an assumption that each word observed in the sentence of language l_1 can potentially find its alignment in the sentence of language l_2. Theoretically, low value of $\varphi(\cdot)$ ensures that words across languages which are similar are embedded closer to each other. Equation 43 shows the regularization term.

$$\alpha \|\mathbf{sp}_j^{l_1} - \mathbf{sp}_j^{l_2}\|^2 + (1-\alpha)\|\frac{1}{m}\sum_{w_i \in s_j^{l_1}}^{m} \boldsymbol{w}_i^{l_1} - \frac{1}{n}\sum_{w_k \in s_j^{l_2}}^{n} \boldsymbol{w}_k^{l_2}\|^2 \tag{43}$$

Where $\boldsymbol{w}_i^{l_1}$ and $\boldsymbol{w}_k^{l_2}$ represent word embeddings obtained for the words w_i and w_k in each sentence (s_j) of length m and n in languages l_1 and l_2 respectively.

5.3.3 *BRAVE with Non-Parallel Document Corpora (BRAVE-D)*

Sometimes it is hard to acquire sentence-aligned parallel corpora for many languages. Availability of non-parallel corpora such as topic-aligned (e.g. Wikipedia) or label/class-aligned document corpora (e.g. sentiment analysis and multi-class classification data sets) in different languages can be leveraged to learn bilingual embeddings for performing CLTC. Earlier approaches like CL-LSI [64] and CL-KCCA [261] were used to learn bilingual document spaces for the tasks comparable to CLTC. Although these approaches provide decent results, they face serious scalability issues and are mostly limited to Wikipedia. Multi-view shallow generative models such as Cross-lingual latent topic extraction models [266] showed promising results for the tasks like word-level or phrase-level translations, but have certain drawbacks for CLTC tasks.

Here, we propose a two step approach to build bilingual embeddings with label/class-aligned document corpora.

- In the first step, we perform manifold alignment using Procrustes analysis [269] between sets of documents belonging to same class/label in different languages. This will help to identify the closest alignment of a document in language l_1 with a document in another language l_2.

- In the second step, we use the pair of partially aligned documents belonging to same class or label in different languages to extract bilingual cues similar to the approach mentioned in the Section 5.3.2. Only difference being paragraph vector is learned for the entire document.

Step-1:

Let S^{l_1} and S^{l_2} be the sets containing languages l_1 and l_2 training documents associated to label or a class. Below, we provide the three step procedure to attain partial alignment between the documents present in these sets.

- Learning low-dimensional embeddings of the sets (S^{l_1}, S^{l_2}) is key for alignment. We use document paragraph vectors [144] to learn low-dimensional embeddings of the documents in each language. Let $\mathbf{X}^{l_1}$ and $\mathbf{X}^{l_2}$ be the low-dimensional embeddings of S^{l_1} and S^{l_2} respectively.

- To find the optimal values of transformation, Procrustes superimposition is done by translating, rotating and scaling the objects (i.e. rows of $\mathbf{X}^{l_2}$ is transformed to make it similar to the rows of $\mathbf{X}^{l_1}$). Transformation is achieved by
  - **Translation:** Taking mean of all the members of set to make centroids $(\sum_{i=1}^{|S^{l_1}|} \frac{\mathbf{X}^{l_1}}{|S^{l_1}|}, \sum_{i=1}^{|S^{l_2}|} \frac{\mathbf{X}^{l_2}}{|S^{l_2}|})$ lie at origin.
  - **Scaling and Rotation:** The rotation and scaling that maximizes the alignment is given by orthogonal matrix (Q) and scaling factor (k). They are obtained by minimizing orthogonal Procrustes problem [229] and is provided by Equation 44.

    $$\arg\min_{k,Q} ||\mathbf{X}^{l_1} - \mathbf{X}_*^{l_2}||_F \tag{44}$$

    where $\mathbf{X}_*^{l_2}$ a matrix of transformed $\mathbf{X}^{l_2}$ values given by $k\mathbf{X}^{l_2}Q$ and $||.||_F$ is the Frobenius norm constrained over $Q^TQ = I$.

- If $S_*^{l_2}$ represents the new document set obtained after identifying the close alignment among documents in S^{l_1} and S^{l_2} with cosine similarity between $\mathbf{X}^{l_1}$ and $\mathbf{X}_*^{l_2}$, then the partially aligned corpora $\{S^{l_1}, S_*^{l_2}\}$ contains one-to-one correspondence between the two languages documents that are used to learn bilingual cues in the second step.

From perturbation theory of spectral spaces [137] it can be understood that the difference between low-dimensional embedding subspaces (i.e. $\mathbf{X}^{l_1}$ and $\mathbf{X}_*^{l_2}$) is always bounded, thus the new alignment obtained between document sets $\{S^{l_1}, S_*^{l_2}\}$ is insensitive to perturbations. Which also means that Procrustes analysis has provided best possible document alignments.

Step-2:

Now, document pairs $(d_j^{l_1}, d_j^{l_2})$ of the partially-aligned corpus (PAC) is used to compute bilingual regularization function $\varphi(\cdot)$. At each training step, L2-loss of precomputed *document paragraph vectors* $\mathbf{dp}_j^{l1}$ and $\mathbf{dp}_j^{l2} \in \mathbb{R}^d$ obtained from the monolingual term $\mathcal{M}(\cdot)$ is combined with the L2-loss between vector of words weighted by the probability of their occurrence in a particular label/class of entire **PAC**. Consideration of word probabilities will help to induce label/class specific information. Equation 45 provides the regularization term.

$$\begin{aligned} &\alpha||\mathbf{dp}_j^{l_1} - \mathbf{dp}_j^{l_2}||^2 \\ &+ (1-\alpha)|| \sum_{w_i \in d_j^{l_1}}^{m} \frac{p_{w_i} \boldsymbol{w}_i^{l_1}}{\sum_m p_{w_i}} - \sum_{w_k \in d_j^{l_2}}^{n} \frac{q_{w_k} \boldsymbol{w}_k^{l_2}}{\sum_n q_{w_k}} ||^2 \end{aligned} \tag{45}$$

Where w_i,w_k are words and their embeddings $\mathbf{w}_i^{l_1}$,$\mathbf{w}_k^{l_2}$ observed in each document (d_j) of length m and n in languages l_1 and l_2 respectively. While, p_{w_i} and q_{w_k} represents probability of occurrence of words w_i and w_k in a specific label/class of entire **PAC**. Figure- 18 shows overall goal of both the approaches.

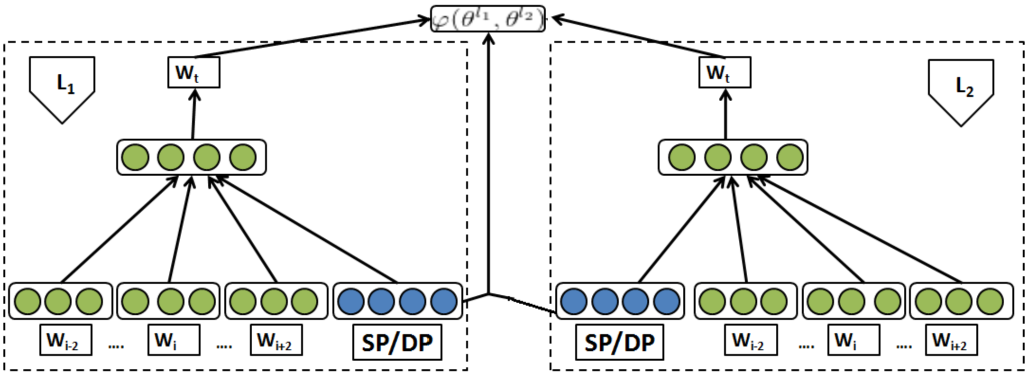


Figure 18: Bilingual word embeddings learned using sentence or document paragraph vectors (SP/DP) along with word vectors.

5.4 EVALUATION

In this section, we report results on three different CLTC tasks to comprehend whether our learned bilingual embeddings are semantically useful across languages. First, cross-language document classification (CLDC) task proposed by Klementiev et al. [134] using the subset of Reuters RCV1/RCV2 corpora [152]. Second, a multi-label CLDC task with more languages using TED corpus[29] of Hermann et al. [103] . Subsequently, a cross-language sentiment classification (CLSC) proposed by Prettenhofer et al., [200] on a multi-domain sentiment dataset.

5.4.1 *Evaluation Setup*

In this section, details about the dataset, implementation and document representation are presented.

5.4.1.1 *Parallel and Non-Parallel Corpora*

For sentence-aligned parallel corpora, Europarl-v7 [30](EP) is used as both monolingual and parallel training data. While for label-aligned non-parallel document corpora, only training and testing collections of the cross-language multi-domain Amazon product reviews(CL-APR) [200] corpus with sentiment labels is used.

5.4.1.2 *Implementation*

Our implementation launches monolingual paragraph vector [144] threads for each language along with bilingual regularization thread. Word and paragraph

[29] http://www.clg.ox.ac.uk/tedcorpus
[30] http://www.statmt.org/europarl/

embeddings matrices are initialized with normal distribution ($\mu = 0$ and $\sigma^2 = 0.1$) for each language and all threads access them asynchronously. Following Pham et al. [197] suggested combination (P=5*W) of paragraph and word embeddings, we chose paragraph embeddings with dimensionality of 200 and 640 when word embeddings are of 40 and 128 dimensions respectively. Asynchronous stochastic gradient descent (ASGD) is used to update parameters (i.e. $\mathbf{P}^l$,$\mathbf{W}^l$,$\mathbf{U}$ and b) and train the model.

For each training pair in parallel or non-parallel corpora, initially monolingual threads sample context h with window size of 8 from a random paragraph (i.e. sentence or document) in each language. Then the bilingual regularization thread along with monolingual threads make update to parameters asynchronously. Learning rate is set to 0.001 which decrease with the increase of epochs, while α is chosen to be 0.6 (can be fine tuned based on empirical analysis) to give more weight to paragraph vectors. All models are trained for 50 epochs.

5.4.1.3 *Document Representation*

Documents are represented with tf-idf weighted sum of word embeddings that are present in them.

5.4.2 *Evaluation Results*

The experimental results for each of the CLTC tasks are presented separately.

5.4.2.1 *Cross-language Document Classification (CLDC) - RCV1/RCV2*

Goal of this task is to classify target language documents with the labeled examples from the source language. To achieve it, we used the subset of Reuters RCV1/RCV2 corpora as the training and evaluation sets and replicated the experimental setting of Klementiev et al. [134]. From the English, German, French and Spanish collection of the dataset, only those documents are selected which was labeled with a single topic (i.e. CCAT, ECAT, GCAT and MCAT). For the classification experiments, 1000 labeled documents from source language are selected to train a multi-class classifier using averaged perceptron [78, 48] and 5000 documents were used as the testing data.

English-German, English-French and English-Spanish portion of **EP** corpora (i.e. each with around 1.9M sentence-pairs) is used both as monolingual and parallel training data with **BRAVE-S** approach to build vocabulary of around 85k English, 144k German, 119k French and 118k Spanish. While training and testing collections belonging to all domains in English-German, English-French languages of **CL-APR** ((i.e. around 12,000 document-pairs)) was used both as monolingual and partially aligned data with **BRAVE-D** approach to build vocabulary of around 21k English, 22k German and 18k French. Further, documents in the training and testing data of RCV1/RCV2 corpora are represented as described in the Section 5.4.1.3 with the vocabulary built. Table 7 and Table 8 shows the comparison of our approaches with the existing systems.

Model	en → de	de → en	en → fr	fr → en	en → es	es → en
Majority class	46.8	46.8	22.5	25.0	15.3	22.2
MT	68.1	67.4	76.3	71.1	52.0	58.4
I-Matrix [134]	77.6	71.1	74.5	61.9	31.3	63.0
BAE-cr [225]	**91.8**	74.2	**84.6**	74.2	49.0	64.4
CVM-Add [103]	86.4	74.7	-	-	-	-
DWA [135]	83.1	75.4	-	-	-	-
BilBOWA [90]	86.5	75	-	-	-	-
UnsupAlign [164]	87.6	77.8	-	-	-	-
Trans-gram [50]	87.8	78.7	-	-	-	-
BRAVE-S$_{(EP)}$	88.1	**78.9**	79.2	**77.8**	**56.9**	**67.6**
BRAVE-D$_{(CL-APR)}$	69.4	67.9	64.1	56.5	-	-

Table 7: CLDC Accuracy with 1000 labeled examples on RCV1/RCV2 Corpus using 40 dimensional embeddings. en/de, en/fr and en/es results of Majority class, MT, I-Matrix and BAE-cr are adopted from Lauly et al. [225]

Model	en → de	de → en	en → fr	fr → en	en → es	es → en
CVM-BI [103]	86.1	79.0	-	-	-	-
UnsupAlign [164]	88.9	77.4	-	-	-	-
BRAVE-S$_{(EP)}$	89.7	**80.1**	82.5	**79.5**	**60.2**	**70.4**
BRAVE-D$_{(CL-APR)}$	70.4	70.6	66.2	57.6	-	-

Table 8: CLDC Accuracy with 1000 labeled examples on RCV1/RCV2 Corpus using 128 dimensional embeddings.

5.4.2.2 *Multi-label CLDC - TED Corpus*

To understand the applicability of our approaches to wider range of languages[31] and class labels, we perform experiments with the subset of TED corpus [103]. Aim of this task is same as CLDC in Section 5.4.2.1, but experiments were conducted with larger variety of languages and class labels. TED Corpus contains English transcriptions and their sentence-aligned translations for 12 languages from the TED conference. Entire corpus is further classified into 15 topics (i.e. class labels) based on the most frequent keywords appearing in them.

To conduct our experiments, we follow the *single* mode setting of Hermann et al. [103] (i.e. embeddings are learned only from a single language pair). Entire language pair (i.e. en→L2) training data of the TED corpus is used both as monolingual and parallel training data to learn bilingual word embeddings with dimensionality of 128 using **BRAVE-S** approach. Bilingual word embeddings of 128 dimensions learned with **EP** and **CL-APR** are also used for comparison. Documents in the training and testing data of TED corpus are represented as described in the Section 5.4.1.3 using each of these embeddings. A multi-class

[31] Our goal is not to evaluate shared multilingual semantic representation.

classifier using averaged perceptron is built using training documents in source language to be applied on target language testing data for predicting the class labels. Table 9 and Table 10 shows the cumulative F1-scores.

Method	de	es	fr	it	nl
			en → L2		
MT-*Baseline*	0.465	**0.518**	**0.526**	**0.514**	**0.505**
DOC/ADD	0.424	0.383	0.476	0.485	0.264
DOC/BI	0.428	0.416	0.445	0.473	0.219
BRAVE-S(TED)	**0.484**	0.436	0.456	0.507	0.328
BRAVE-S(EP)	0.418	0.365	0.387	0.418	0.284
BRAVE-D(CL-APR)	0.385	-	0.212	-	-
			L2 → en		
MT-*Baseline*	0.469	0.486	0.358	**0.481**	**0.463**
DOC/ADD	0.476	0.422	0.464	0.461	0.251
DOC/BI	0.442	0.365	**0.479**	0.460	0.235
BRAVE-S(TED)	**0.492**	**0.495**	0.465	0.475	0.384
BRAVE-S(EP)	0.458	0.404	0.437	0.443	0.338
BRAVE-D(CL-APR)	0.366	-	0.278	-	-

Table 9: Cumulative F1-scores on TED Corpus using training data in English language and evaluation on other languages (i.e. German (de), Spanish (es), French (fr), Italian (it), Dutch (nl)) and vice versa. MT-*Baseline*, DOC/ADD, DOC/BI represents single language pair of Hermann et al., [103] as document features. Underline shows the best results amongst embedding models.

5.4.2.3 *Cross-language Sentiment Classification (CLSC)*

The objective of the third CLTC task is to identify sentiment polarity (e.g., positive or negative) of the data in target language by exploiting the labeled data in source language. We chose subset of publicly available Amazon product reviews (CL-APR) [200] dataset mainly English(E), German(G) and French(F) languages belonging to three different product categories (books(B), dvds(D) and music(M)) to conduct our experiments. For each language-category pair, corpus consists of training, testing sets comprising 1000 positive and 1000 negative reviews each with an additional unlabeled reviews varying from 9,000 to 170,000.

We constructed 12 different CLSC tasks using different languages (i.e. E, G and F) for three categories (i.e. B, D and M). For example, EFM refers English music reviews as source language and French music reviews as target language. Bilingual word embeddings with dimensionality of 128 learned with **BRAVE-S** and **BRAVE-D** are used to represent each review as described in the Section 5.4.1.3. To have fair comparison with earlier approaches, sentiment classification model

Method	pt	po	ro	ru	tr
			en → L2		
MT-*Baseline*	0.470	0.445	**0.493**	0.432	0.409
DOC/ADD	0.354	0.402	0.418	0.448	0.452
DOC/BI	0.400	0.403	0.467	0.421	0.457
BRAVE-S(TED)	**0.506**	**0.453**	$\underline{0.488}$	**0.456**	**0.491**
BRAVE-S(EP)	0.454	0.412	0.424	-	-
BRAVE-D(CL-APR)	-	-	-	-	-
			L2 → en		
MT-*Baseline*	0.374	**0.460**	**0.486**	0.404	0.441
DOC/ADD	0.338	0.400	0.407	0.471	0.435
DOC/BI	0.380	0.393	0.426	**0.467**	0.477
BRAVE-S(TED)	**0.388**	$\underline{0.442}$	$\underline{0.464}$	0.457	**0.484**
BRAVE-S(EP)	0.312	0.374	0.418	-	-
BRAVE-D(CL-APR)	-	-	-	-	-

Table 10: Cumulative F1-scores on TED Corpus using training data in English language and evaluation on other languages (i.e. Portuguese (pt), Polish (po), Romanian (ro), Russian (ru) and Turkish (tr)) and vice versa. MT-*Baseline*, DOC/ADD, DOC/BI represents single language pair of Hermann et al., [103] as document features. Underline shows the best results amongst embedding models.

is then trained with libsvm[32] default parameter settings using source language training reviews[33] to classify target language test reviews. Table 11 shows the accuracy and standard deviation results after we randomly chose subset of target language testing documents and repeated the experiment for 10 times for all CLSC tasks.

5.4.3 *Evaluation Results Analyses*

First CLTC task (i.e., CLDC) results presented in the Table 7 and Table 8 shows that BRAVE-S was able to outperform most of the existing systems. The success of BRAVE-S can be attributed to its ability to incorporate both non-compositional and compositional meaning observed in an entire sentence and the individual words respectively. Thus making it different from other models which use only bag-of -words [90] or bi-grams [103].

Similarly, second CLTC task (i.e. multi-label CLDC) results presented in the Table 9 and Table 10 shows that BRAVE-S learned with the training data of TED corpus outperformed *single mode* DOC/* embedding models [103], BRAVE-S learned with **EP** and BRAVE-D. The BRAVE-S(TED) was able to capture bet-

[32] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[33] We do not use 100 labeled target language reviews in model training, as it was shown by earlier approaches that 100 labeled target language reviews does not have much impact.

Cross-Language Sentiment Classification (en→L2 and Vice versa)					
Task	CL-SCL	CL-SSMC	CL-SLF	BRAVE-S (EP)	BRAVE-D (CL-APR)
EFB	79.86±0.22	83.05±0.26	82.61±0.25	72.24±0.31	82.57±0.33
EFD	78.80±0.25	82.70±0.20	82.70±0.45	74.95±0.25	**82.90±0.35**
EFM	75.95±0.31	80.46±0.20	80.19±0.40	72.80±0.20	**80.70±0.45**
FEB	77.26±0.22	80.05±0.26	80.48±0.33	75.45±0.38	80.28±0.21
FED	76.57±0.20	79.40±0.28	78.76±0.38	73.75±0.26	**79.80±0.15**
FEM	76.76±0.25	78.82±0.17	79.18±0.33	73.66±0.17	78.56±0.33
EGB	77.77±0.28	81.88±0.42	79.91±0.47	75.95±0.16	**81.75±0.45**
EGD	79.93±0.23	82.25±0.20	81.86±0.31	78.30±0.42	81.56±0.26
EGM	73.95±0.30	81.30±0.20	79.59±0.42	75.95±0.33	81.20±0.17
GEB	77.85±0.27	79.06±0.23	78.61±0.34	72.25±0.20	**80.23±0.17**
GED	77.83±0.33	80.89±0.16	80.27±0.35	73.28±0.23	**80.78±0.20**
GEM	77.37±0.34	79.85±0.17	79.80±0.26	74.41±0.22	**79.77±0.36**

Table 11: Average classification accuracies and standard deviations for 12 CLSC tasks. Results of other baselines are adopted from CL-SCL [200], CL-SSMC [284], CL-SLF [304]

Top-3 Nearest Neighbors (Euclidean Distance)			
English Words	Models	German	French
great	BRAVE-S	wachstum	éminent
		super	maintenus
		spielen	m'efforcerai
	BRAVE-D	schärfe	festival
		mögen	interressante
		kraftvolle	attachant
bored	BRAVE-S	boykottiert	ennuyé
		leere	précédera
		ausgehen	compromettent
	BRAVE-D	ableben	réserve
		lichtblick	intensité
		traurigen	consterné

Table 12: Nearest Neighbors for English Words in German and French.

ter linguistic regularities across languages that are more specific to the corpus, than the general purpose bilingual embeddings learned with **EP**. Though in some cases, all our embedding models could not outperform machine translation baseline. It can be due to the asymmetry between languages induced by the language-specific words which could not find its equivalents in English.

Also, it can be apprehended from both CLDC and multi-label CLDC that BRAVE-D results are not as expected. Though being a general approach like BRAVE-S which can capture both non-compositional and compositional meaning from larger pieces of texts, a minimal overlap of vocabulary learned with BRAVE-D using cross-language sentiment label-aligned corpora with other domains (i.e., Reuters and TED) produce unfavorable results. Thus, we understand that the choice of label/class-aligned corpora is crucial.

Final CLTC task (i.e., CLSC) results presented in the Table 11 shows that BRAVE-D outperforms other baseline approaches in most of the cases. As BRAVE-D learns bilingual word embeddings using **CL-APR**, it was able to inherently encompass sentiment label information effectively like previous approaches [250, 305] than the general purpose embeddings learned using BRAVE-S with **EP** and similar approaches [170]. Thus making it more suitable for sentiment classification task. Also, unlike CL-SSMC [284] and CL-SLF [304], BRAVE-D is not highly parameter dependent where the results of the former approaches show significant variance based on the parameter settings. To visualize the difference in embeddings learned with BRAVE-S and BRAVE-D, we selected sentiment words and identified cross-language nearest neighbors in Table 12. It can be observed that BRAVE-D was able to identify better sentiment (either positive or negative) word neighbors than BRAVE-S.

5.5 SUMMARY

In this chapter, we addressed the second research question:

> ✍ **Research Question 2**
>
> Given two different views of homogeneous content depicting text from different languages, how can we build a shared representation to assist *categorization* by learning a common space by capturing regularities?

For this, we validated Hypothesis 2 by proposing an approach that leverages paragraph vectors and manifold alignment technique to learn bilingual word embeddings with sentence-aligned parallel and label-aligned non-parallel corpora. Empirical analysis exhibited that embeddings learned from both of these types of corpora have shown the remarkable impact on CLTC tasks.

In the next chapter, we will present an approach for achieving multi-view deep representation learning with consensus and complementarity principles and support an application where the heterogeneous data emerge from three views, i.e., different languages and image.

MVRL WITH AUXILIARY VIEWS AND JOINT MULTI-TTASK OPTIMIZATION

6

MVRL WITH AUXILIARY VIEWS AND JOINT MULTI-TASK OPTIMIZATION

Context of this Chapter. In this chapter, we leverage multi-view deep representation learning along with the multi-task learning to propose a novel approach for multi-language consistent image caption generation. Lately, generation of natural language descriptions at sentence-level for an image has received significant attention. Most of the earlier proposed approaches accomplish this task only with datasets that align images with English sentences. At present, descriptions are already available in more than one language. Porting models that are built for English to other languages can lead to the generation of descriptions which are very different from English and also irrelevant to an image. A possible solution to minimize such issues is by controlling the diversity of the other language caption by leveraging correspondences between languages when building the image caption generation model. To realize this, we introduce a multi-task learning based image caption generation model that helps to control cross-language diversity and incorporates correspondences between different language captions aligned to an image. The empirical analyses show that the proposed model can effectively control the diversity of closely related languages.

Our main contributions presented in this chapter can be broadly summarized as follows:

① We proposed a framework to leverage *inter-language correspondences* as the initial input to an image caption model for controlling diversity and making semantically similar descriptions across languages.

② We explored two different architecture variations of proposed model that leverage multi-task learning.

③ We showed using two different datasets that it is less complicated to control the diversity of generated captions for closely related languages in contrast with distantly related languages.

Outline. The remainder of this chapter is organized into following sections. Initially, Section 6.1 presents the motivation in Section 6.1.1 and briefly introduce existing multilingual multimodal representations methods in Section 6.1.2. Next Section 6.2 presents the research question and describes our contribution to multi-language image caption generation. Our approach i.e. multi-task attention (MTA) and its variations are then discussed in the Section 6.3. The dataset and metrics used for evaluation of the approach are described in the Section 6.4. While in the Section 6.4.2 details about the evaluation results are shown, which are further analyzed in the Section 6.4.3. Summary of the chapter is presented in the Section 6.5.

6.1 INTRODUCTION

6.1.1 *Motivation*

Generation of natural language text from the input data is of interest over past few decades [208]. This data-to-text generation is an instance of Natural Language Generation (NLG) and has leveraged different types of input data, i.e, linguistic or non-linguistic to develop systems such as weather and financial report generation [198], summaries of patient information in clinical contexts [13], generation of paraphrases of input sentences [14] and many more. According to Gatt et al. [82], NLG system target subproblems such as:

① Content determination

- Decides what information has to be included in the generated text.

② Text structuring

- Determines the order of information that needs to be presented in the text.

③ Sentence aggregation

- If more than one sentence is generated, it determines the distribution of information in the sentences.

④ Lexicalization

- Identification of the words and phrases to express information.

⑤ Referring expression generation

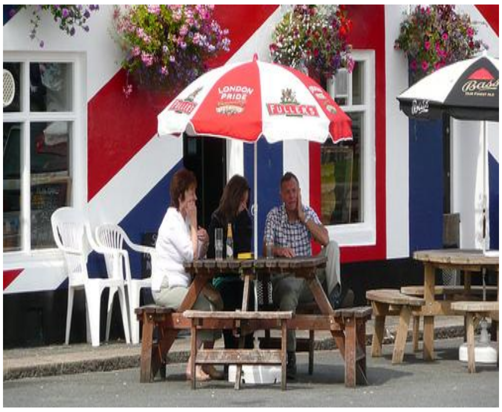
- Those words and phrases are selected which represent the domain objects.

⑥ Linguistic realization

- Words and phrases are combined such that they are well-formed sentences.

However in the past few years, there is a significant interest in designing systems where input data emerge from the visual information. These systems are also expected to face similar subproblems as the traditional data-to-text generation systems. One such application of vision-to-text generation is image description generation at the sentence-level [129, 286]. Howbeit, most of the proposed approaches are fine-tuned to corpora such as Flickr8K [109], Flickr30K [292] and MSCOCO [157] that contain only English descriptions.

Generating descriptions in languages other than English requires translation of the generated English descriptions. Observation from the previous research [290]



STA(En): A group of people are sitting on the ground
STA(De): Eine gruppe junger leute auf einer treppe (A group of young people sitting on stairs)
MTA (En): A group of people are sitting on a bench
MTA (De): Eine Gruppe von Menschen auf einer Bank (A group of people on a bench)

Figure 19: Example image from the Multi30K dataset with generated descriptions in English (En) and German (De) using single-task (STA) and multi-task (MTA) models.

show that the translation of English descriptions propagates errors into the target language. This scenario motivates us to create separate models and datasets for each language. Corpora's such as IAPR-TC12 [93], Multi30K [67], and STAIR [290] are extended using English description datasets to capture image descriptions in multiple languages and facilitate the creation of separate caption generation models. Howbeit, there are circumstances where models trained separately per language can lead to inconsistent caption generation across languages.

To visualize the challenge, Figure 19 shows an example image with generated descriptions across languages. It can be observed that the generated descriptions are very diverse and vary across languages. This is counter-intuitive, as we expect the same model trained on two different language descriptions of an image to generate semantically similar and consistent captions across languages.

This behavior exhibits that the description generation across languages is non-trivial and demand additional anticipation from the description models. They are currently confined to generating concise descriptions by covering all possible interactions between objects/attributes present in an image. Consequently, we append a supplemental criterion of generating less diverse and semantically similar descriptions for models that aim to generate descriptions in multiple languages.[34]

Existing methods for image description generation can be broadly divided into two categories (1) template-based and (2) encoder-decoder framework based. We further divide the encoder-decoder framework based methods into those who use attention mechanisms [12] and those who do not. Table 13 summarizes the different setups for generating image captions either in one or more languages using the encoder-decoder framework.

It can be observed in the Table 13 that most of the earlier proposed methods do not generate image descriptions in more than one language. Two potential reasons for that could be:

[34] The expectation here is distinct from other objectives such as multimodal machine translation or cross-lingual caption generation [68].

Approach	LM	Dataset	Language
MLBL [133]	LBL	4	En
m-RNN [166]	RNN	1,2,3,4	En
Minds Eye [41]	RNN+MELM	1,2,3	En
BRNN [129]	biRNN	1,2,3	En
NIC [262]	LSTM	1,2,3	En
LRCN [61]	LSTM	2,3	En
Guided LSTM [119]	guided-LSTM	1,2,3	En
Multilingual Multimodal LM [66]	LSTM	4	En,De
Deep Bidirectional LSTM [270]	biLSTM	1,2,3	En
Regional Visual Attributes [278]	LSTM	1,2,3	En
Japanese-Generator [290]	LSTM	6	Ja
Visual Attention [286]	Att-LSTM	1,2,3	En
Region-based Attention [122]	SF-LSTM	1,2,3	En
Attribute Attention [291]	Att-LSTM	2,3	En
Review Attention [288]	Review-Att-LSTM	3	En
Adaptive Attention [163]	Sentinel-LSTM	2,3	En
Self-Critical Attention [211]	Att-LSTM	3	En
Areas of Attention [194]	Att-RNN	3	En
Contrastive Adaptive Attention [53]	CL-Sentinel-LSTM	3	En
Up-Down Attention [8]	Att-LSTM+LSTM	3	En

Table 13: Summary of Encoder-Decoder based Image caption generation methods using different datasets: (1) Flickr8K (2) Flickr30K (3) MSCOCO (4) IAPR-TC12 (5) Multi30K (6) STAIR and language models (LM). Att-LSTM → Attention based LSTM, biLSTM → Bidirectional LSTM and rest are other variations.

① Unavailability of the corpora containing captions in more than one language.

② Inflexibility of the approaches to leverage more than one language caption to build a single model that could generate semantically similar and consistent captions across languages.

We intend to address the later with the help of corpora that provide captions in more than one language. Given such a challenge, we understand that for consistent caption generation across languages, it is essential to jointly consider different language captions of an image while building the description generation models. Recently, multi-task learning has been leveraged to address similar problems for varied tasks such as neural machine translation [77], discourse representation and identification [141] and sequence to sequence learning [164].

In this chapter, we introduce a novel multi-task attention-based image caption model for generating consistent descriptions across languages. The underlying assumption of our proposed framework is that different languages that describe an image may differ lexically, but share same semantics. Hence, we explore such correspondences across languages and realize it with multi-task learning for providing transfer which introduces inductive bias [219]. In this way, our proposed model makes full use of different language captions given for each image. Thus, making the model generalize better by preferring hypotheses that explain more than one task. Furthermore, we only require a single model for multiple languages.

6.1.2 *Background on Multilingual Multimodal Representations*

Recently, learning a multilingual multimodal space has shown to achieve image caption generation [204] with parallel corpora and multimodal machine translation [242, 68]. Later, image caption retrieval [83, 36] was also explored in multiple languages with multilingual multimodal embeddings. However, such approaches restrict corpora to be parallel and perform only caption retrieval.

6.2 research question and contributions

Let us outline the research questions, hypotheses, and contributions, which we target throughout the chapter.

6.2.1 *Research Question and Hypothesis*

As presented in Section 1.3, our overall research question is: How to effectively integrate multiple views of training instances depicting heterogeneous or homogeneous content into a common space representation for supporting applications in different domains? In this chapter, we address the third part, i.e., learning a common space representation among views emerging from the same modality with consensus and complementarity principles by leveraging their input representations. More specifically, we aim at Research Question 3:

✍ **Research Question 3**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation of all views if an auxiliary view depicting text in multiple languages is added to assist the *generation* of text from an image?

For addressing above research question, we verify the hypothesis as follows:

□ **Hypothesis 3**

Leveraging deep neural network architectures and multi-task learning, we can effectively learn a common space representation of all views emerging from heterogeneous data to generate one modality from another, i.e., especially generating text from an image. Multi-task learning has proven to be effective in capturing knowledge across shared tasks. We can leverage its potential and design a shared layer in deep neural network architecture with a multi-task loss for generating consistent caption text in multiple languages for a given image.

Intuitively, Hypothesis 3 states that the combination of deep neural networks with multi-task learning can effectively capture common space representation across different languages text and image to generate consistent text in multiple languages for a given image. Mainly, we expect to extend the image to caption (i.e., text) generation model with our proposed LSTM shared layer for effectively capturing the common space representation of the different languages text and an image. Further, we expect to use a multi-task loss for learning parameters.

To validate Hypothesis, we present our multi-task attention model in the Section 6.3 to build a consistent multi-language image caption generation model. Moreover, we implemented the approach and empirically show (see the evaluation in Section 6.4) its effectiveness in contrast with other state of the art.

6.2.2 *Contributions*

While being naturally appealing, usage of deep neural networks and its combination with multi-task learning is not studied before in the context of image caption generation, where the caption data can emerge from multiple languages. Aiming at above hypotheses, we provide the following contribution:

- *Contribution for Hypothesis 3*
  Building a common space representation from the heterogeneous data depicting only two views using either correlation or consensus/complementarity principles is studied earlier. However, our usage of deep encoder-decoder architecture and its extension in the multi-task learning setting for building common space representation for more than two views is unique. Additionally, designing a LSTM shared layer for leveraging shared information from two deep encoder-decoder architectures throws diverse challenges. Facing these characteristics, we propose a multi-task attention model and its variations. To the best of our knowledge, this is the first work which utilizes multi-task learning for image caption generation across languages which are highly consistent.

  We conducted an evaluation using different language image-caption datasets and validated the Hypotheses 3. In these experiments, we could achieve performance comparable to the English state of the art caption generation models and reported new results for other languages. In fact, we could show that our proposed approach was useful for building multilingual multimodal representation.

6.3 consistent caption generation in multiple languages

In this section, we propose our models for transforming the source image into captions in many languages.

6.3.1 *Objective*

Let $\{I_n, S_n\}_{n=1}^{N}$ be our dataset containing an image with more than one language caption, where $S_n \subseteq \mathcal{S}$ and $\mathcal{S} \triangleq \{1, 2, ..., K\}$ is set of existing language captions. Each image have same number of captions $I_n = |S_n|$. Our goal now is to build a image caption joint-model for multiple languages.

6.3.2 *Single-task Attention-Based Image Description Model*

The aim of single-task approach is to build image caption model separately for each language. Given an image I, its global visual features $\mathbf{I}_v \in \mathbb{R}^V$ represent the encoding of full image and $\mathbf{a}_v = \{\mathbf{a}_{v_1}, ..., \mathbf{a}_{v_L}\}$, $\mathbf{a}_{v_j} \in \mathbb{R}^D$ the spatial attention features set. Similar to previous works [163, 8], our proposed image description model also leverages soft attention mechanism to weigh each spatial attention feature during description generation using the partial output sequence as context.

Specifically, our image description generation model is built with two-layers i.e. layer-1 (L-1) and layer-2 (L-2) using long short-term memory (LSTM) [108] with no peepholes. At any given time step t, LSTM obtain input $\boldsymbol{w}_t$, the previous hidden state $\mathbf{h}_{t-1}$ and the memory cell $\mathbf{c}_{t-1}$ that store previous state information for updating the input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$ given as follows:

$$\mathbf{i}_t = \sigma(W_{wi}\boldsymbol{w}_t + W_{hi}\mathbf{h}_{t-1}) \quad (46)$$

$$\mathbf{f}_t = \sigma(W_{wf}\boldsymbol{w}_t + W_{hf}\mathbf{h}_{t-1}) \quad (47)$$

$$\mathbf{o}_t = \sigma(W_{wo}\boldsymbol{w}_t + W_{ho}\mathbf{h}_{t-1}) \quad (48)$$

$$\tilde{\mathbf{c}}_t = \tanh(W_{wc}\boldsymbol{w}_t + W_{hc}\mathbf{h}_{t-1}) \quad (49)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad (50)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (51)$$

Here, $W_{w,h(i,f,o,c)}$ represent the LSTM weights. σ represents the sigmoid activation function and ⊙ represents the element-wise multiplication.

Initially, L-1 of the model receives input from the textual sequence, where each word ($\boldsymbol{w}_t \in \mathbb{R}^T$) at time step t in the textual sequence is initialized with the pretrained word embeddings. Now, visual context provided by global visual features $\mathbf{I}_v$ can be either provided along with the input words $\boldsymbol{w}_t$ or can be leveraged during prediction of the next word in sequence.

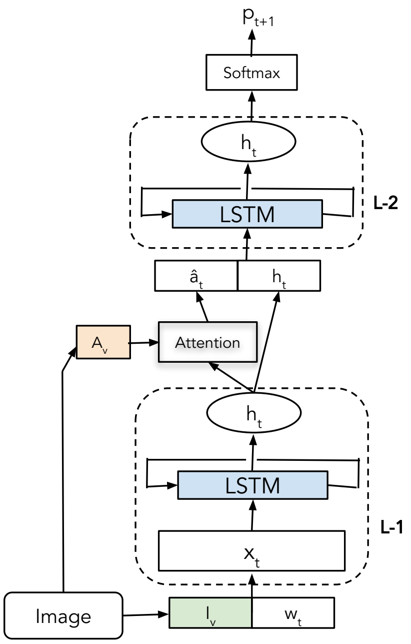


Figure 20: Illustration of Single-task architecture → STA-GVC-I.

If ($\mathbf{h}_t^1 \in \mathbb{R}^{H_1}$) and ($\mathbf{h}_t^2 \in \mathcal{R}^{H_2}$) represent forward hidden vectors at the time step t of L-1 and L-2 respectively. In the following, we present two scenarios where global visual context can be incorporated and later provide details about the inclusion of spatial attention features followed by our final model. For convenience and to reduce many parameter names, we use Θ as the reference for the parameters of LSTM.

6.3.2.1 *Global Visual Context at Input (GVC-I)*

In the first scenario, $\mathbf{w}_t$ is concatenated with the global visual features $\mathbf{I}_v$ at each time step t to provide as an input to the L-1 for generating hidden vectors encoded as follows:

$$\mathbf{x}_t = \mathbf{I}_v \oplus \mathbf{w}_t \tag{52}$$

$$\mathbf{h}_t^1 = \text{L-1}(\mathbf{x}_t, \mathbf{h}_{t-1}^1; \Theta) \tag{53}$$

where ⊕ represents concatenation. Figure 20 illustrates the GVC-I based single-task attention-based image caption generation model.

6.3.2.2 *Global Visual Context at Output (GVC-O)*

Global visual features can also be used to provide visual context before prediction of next word in the sequence. Hence, $\mathbf{I}_v$ is concatenated with the hidden vectors ($\mathbf{h}_t^2$) of L-2 at any time step t before passing on to the final softmax layer for the next word prediction.

$$\mathbf{h}_t' = \mathbf{I}_v \oplus \mathbf{h}_t^2 \tag{54}$$

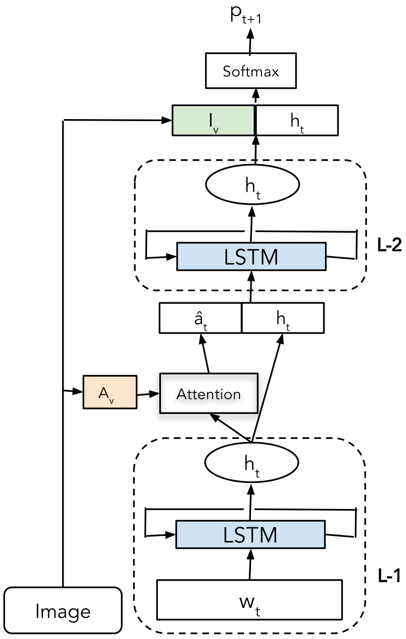


Figure 21: Illustration of Single-task architecture → STA-GVC-O.

$$p_{t+1} = \text{softmax}(W_{vocab}\mathbf{h}'_t) \tag{55}$$

where $W_{vocab} \in \mathbb{R}^{vocab \times (V+H_2)}$ and $vocab$ refers to vocabulary of the caption dataset. Figure 21 illustrates the GVC-O based single-task attention-based image caption generation model.

6.3.2.3 *Spatial Attention Features*

Formerly, we only presented the utilization of $\mathbf{I}_v$. To leverage spatial attention features set $\mathbf{a}_v$, hidden sequences $\mathbf{h}^1_t$ at each time step t is used to generate a normalized attention weight $\boldsymbol{\alpha}_t$ for each of the spatial attention features $(\mathbf{a}_{v_j})$ given as follows:

$$\alpha_{tj} = \frac{\exp(\mathbf{e}_{tj})}{\sum_{k=1}^{L} \exp(\mathbf{e}_{tk})} \tag{56}$$

$$\mathbf{e}_{tj} = \tanh(W_{ae}\mathbf{a}_{v_j} + W_{he}\mathbf{h}^1_t) \tag{57}$$

where L represent cardinality of set a_v, $W_{ae} \in \mathbb{R}^{M \times D}$, $W_{he} \in \mathbb{R}^{M \times H_1}$ are learned parameters. The attended spatial features $(\hat{\mathbf{a}}_t)$ which are used as input along with $\mathbf{h}^1_t$ to the L-2 at every time step t is calculated as:

$$\hat{\mathbf{a}}_t = \sum_{j=1}^{L} \boldsymbol{\alpha}_{tj}\, a_{v_j} \tag{58}$$

6.3.2.4 *Final Models*

Our final models include two variations due to the usage of global visual context at varied locations as presented in the Section 6.3.2.1 and Section 6.3.2.2. However, usage of spatial attention features $\mathbf{a}_v$ remains equivalent for both.

STA-GVC-I

It uses $\mathbf{x}_t$ as input to L-1, while $\mathbf{h}_t^1$ given by Equation 64 and $\hat{\mathbf{a}}_t$ is concatenated using Equation 59 is provided as input to L-2 to generate $\mathbf{h}_t^2$ at any time step t given by Equation 60. Further, $\mathbf{h}_t^2$ is used to predict next words in the sequence with Equation 61.

$$x_t' = \hat{a}_t + h_t^1 \tag{59}$$

$$h_t^2 = \text{L-2}(x_t', h_{t-1}^2; \Theta) \tag{60}$$

$$p_{t+1} = \text{softmax}(W_{vocab} h_t^2) \tag{61}$$

STA-GVC-O

It uses $\mathbf{w}_t$ as input to L-1, while $\mathbf{h}_t^1$ given in Equation 62 and $\hat{\mathbf{a}}_t$ is concatenated using Equation 59 is provided as input to L-2 to generate $\mathbf{h}_t^2$ at any time step t provided by Equation 60. Further, $\mathbf{h}_t^2$ is modified with Equation 54 and is used to predict next words in the sequence with Equation 55.

$$h_t^1 = \text{L-1}(w_t, h_{t-1}^1; \Theta) \tag{62}$$

6.3.3 *Multi-task Attention-Based Image Description Model*

The models presented in the Section 6.3.2 are built separately for each language. Howbeit, for consistent description across languages it is essential to simultaneously capture intrinsic relatedness between the generated descriptions across languages. Therefore, we propose a joint model by integrating the aforementioned models into multi-task learning (MTL) framework [37].

In the joint model, we introduce a shared LSTM layer after **L-1** to enhance the interaction between task-specific layers of different languages. Figure 22 and Figure 23 illustrates the two variations of the proposed multi-task approach leveraged over single-task models GVC-I and GVC-O respectively.

6.3.3.1 *Shared LSTM Layer*

Aim of the shared LSTM layer is to receive input from multiple tasks in the joint model to capture their shared information. If $\mathbf{x}_t'^{(l_1)}$ and $\mathbf{x}_t'^{(l_2)}$ denote the output

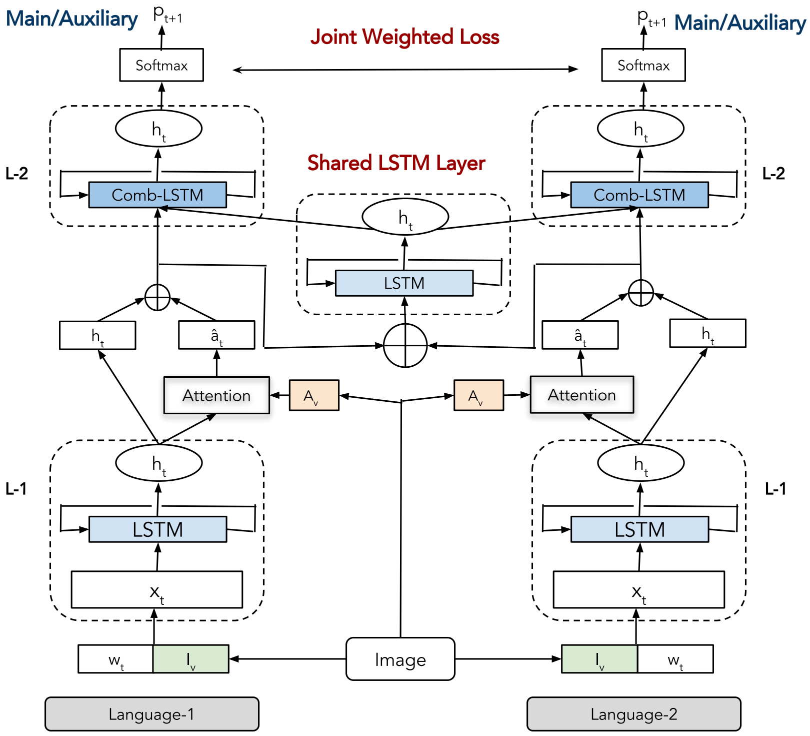


Figure 22: Illustration of Multi-task architecture → MTA-GVC-I.

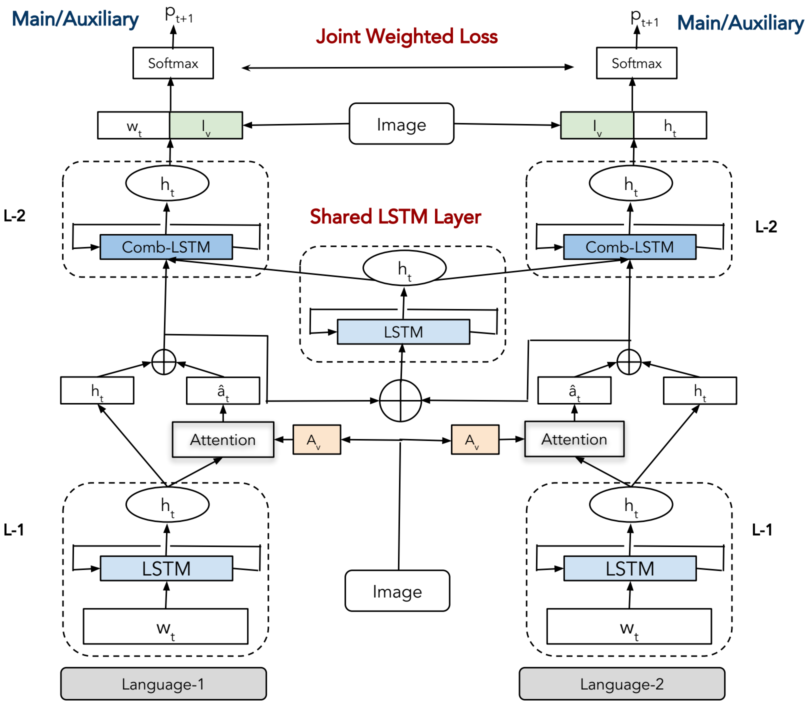


Figure 23: Illustration of Multi-task architecture → MTA-GVC-O.

of task-specific (henceforth, language-1 (l_1) and language-2 (l_2)) L-1 layer and their spatial attentions. Input ($\mathbf{In}_t \in \mathbb{R}^N$) to the shared LSTM layer and its hidden layer output ($\mathbf{h}_t^{(s)} \in \mathbb{R}^{H_s}$) is given by:

$$\mathbf{In}_t = W_{h^{(l_1)}in} x_t^{\prime(l_1)} + W_{h^{(l_2)}in} \mathbf{x}_t^{\prime(l_2)} \tag{63}$$

$$\mathbf{h}_t^{(s)} = \mathrm{LSTM}(\mathbf{In}_t, \mathbf{h}_{t-1}^{(s)}; \Theta) \tag{64}$$

where $W_{h^{(l_1)}in} \in \mathbb{R}^{N \times (H_1 + D)}$, $W_{h^{(l_2)}in} \in \mathbb{R}^{N \times (H_1 + D)}$.

6.3.3.2 *Combination LSTM (Comb-LSTM)*

Output from the L-1 layer of language-1 and language-2 is combined with the hidden layer output ($\mathbf{h}_t^{(s)}$) of the shared LSTM layer using combination LSTM (Comb-LSTM). The Comb-LSTM achieves this at the input stage where the output of the L-1 layer ($\mathbf{x}_t'$) is merged with the hidden layer output ($\mathbf{h}_t^{(s)}$) of the shared LSTM layer using Equation 65 for language-1.

$$\mathbf{Input}_t = W_{h^{(l_1)}x'} \mathbf{x'}_t^{(l_1)} + W_{h^{(l_1)}s} \mathbf{h}_t^{s} \tag{65}$$

Rest of the settings are same as standard LSTM. Hidden layer output $\mathbf{h}_t^{(l_1)^2}$ of the Comb-LSTM specific to each language is further fed into softmax layer provided by the Equation 66.

$$p_{t+1}^{(l_1)} = \mathrm{softmax}(W_{vocab} \mathbf{h}_t^{2(l_1)}) \tag{66}$$

Similar interpretation can be made for language-2.

6.3.4 *Training and Inference*

6.3.4.1 *Training*

Parameters of our multi-task models with parameters θ are trained to optimize the cost function ($\mathcal{C}$) which minimizes the weighted cross-entropy loss of appropriate ground truth word (y_t^*) at each time step t of each individual task.

$$\begin{aligned} \mathcal{C}(\theta) = \\ & -\frac{1}{N} \sum_{n=1}^{N} (\lambda^{(l_1)} \sum_{t=0}^{T^{(n)}} \log p_\theta (y_t^{*(l_1)}) \\ & + \lambda^{(l_2)} \sum_{t=0}^{T^{(n)}} \log p_\theta (y_t^{*(l_2)})) \end{aligned} \tag{67}$$

Where $(\lambda^{(l_1)}, \lambda^{(l_2)}) \in (0, 1]$ is weight hyper-parameter, $T^{(n)}$ represents the length of sentence at n-th training sample and N denote the number of samples used for training. In contrast with multi-task models, single-task models use seperate cost function for each task which minimizes the cross-entropy loss of appropriate ground truth word (y_t^*) at each time step t.

6.3.4.2 *Inference*

As in earlier approaches [129] we also leverage beam search for decoding. However, our multi-task model uses l_1 and l_2 simultaneously in the training phrase. Leveraging beam search still require a prior understanding of the target direction (i.e. l_1 or l_2) it should take to decode a given image. Hence, although we decode l_1 and l_2 simultaneously. We prior decide the target decoder that will be considered (i.e. l_1 or l_2) and decode it until the end-of-sentence (eos) symbol is generated. This means, if there are two tasks we generate twice. The size of beam is set to 5 in our experiments.

6.4 evaluation

6.4.1 *Evaluation Setup*

In this section, we present the datasets and measures used for performing experiments.

6.4.1.1 *Datasets*

We leverage those datasets that provide image captions in more than one language and are of different sizes.

Multi30K

A multilingual multimodal dataset created to serve tasks such as image description and multimodal machine translation. Multi30K extends Flickr30K dataset with German sentences. We use the dataset from task of Cross-lingual image description[35], where any English-German pair of descriptions for a given image is considered a comparable translation pair.

STAIR

It constitute Japanese captions for the MSCOCO images. For each image in MSCOCO, five Japanese captions are created in similar manner as English captions. Overall, 820,310 captions were created for 164,062 images. However in contrast with MSCOCO annotation format, STAIR annotation provide an additional field of "tokenized_caption" where Japanese words are tokenized with spaces. Here, we assume any English-Japanese pair of descriptions for a given image as comparable translation pair. For the experimental evaluation, we use the splits of Karpathy et al. [129].

Table 14 summarizes the training, validation and test splits of all datasets, while Figure 24 shows the sample image and its descriptions in English and Japanese taken from MSCOCO and STAIR respectively.

Cross-Language Out-of-Domain Resources

[35] http://www.statmt.org/wmt16/multimodal-task.html

	Multi30K	MSCOCO&STAIR
Languages	En,De	En,Ja
Sentences	5	5
Training	29,000	113,287
Validation	1014	5000
Test	1000	5000
Sentence-Length	12.3,9.6	11.3,12.54
Vocabulary	7471,8514	9989, 12534

Table 14: Statistics of the datasets

MSCOCO (English)

- ➔ A long restaurant table with rattan rounded back chairs
- ➔ A long table with a plant on top of it surrounded with wooden chairs
- ➔ A long table with a flower arrangement in the middle for meetings
- ➔ A table is adorned with wooden chairs with blue accents
- ➔ A restaurant has modern wooden tables and chairs

STAIR (Japanese)

- ➔ 長い 机 と 椅子 が 、並ん で 置い て ある
- ➔ テーブル の 上 に 、植物 が 置い て ある
- ➔ 背もたれ が 丸い 椅子 が たくさん 並べ られ て いる テーブル の 真ん中 に 大きな ガラス の 花瓶 に 活け られ た 花 が 飾ら れ て いる
- ➔ 長い テーブル に お 揃い の イス が 置か れ て いる
- ➔ 10 人 掛け の テーブル の 上 に は 分厚い 本 や フラワー アレンジメント が 置か れ て いる

Figure 24: An example image and its five descriptions. Please note that English and Japenese descriptions are non-parallel.

To check consistency across languages, we use those out-of-domain resources that contain parallel data such as machine translation [36] for learning cross-language word embeddings. Only those corpora's are selected which contains at-least one European and Asian language in addition to English. Table 15 presents the available corpora belonging to diverse domains and genres. Selecting different corpora will help us to adequately asses the knowledge transfer provided by such domains to the caption generation in different languages.

Dataset	Parallel-Sentences	Languages
OpenSubtitles2018	22512639, 2083600	En-De, En-Ja

Table 15: Out-of-Domain Parallel Sentence Data.

[36] http://www.statmt.org/wmt17/translation-task.html

Open subtitles contain parallel sentence-data from the movie subtitles. They cover various genres and time periods and combine features from spoken language corpora and narrative texts including many dialogs, idiomatic expressions, dialectal expressions and slang.

6.4.1.2 *Evaluation Measures*

The goal of evaluation measures is to analyze the two essential expectations from CBL models, i.e., the efficaciousness of generated caption and its consistency across languages.

① To evaluate the effectiveness, we use evaluation measures such as BLEU, METEOR, CIDEr and SPICE[37] as in earlier approaches [8] and calculate them using extended Microsoft evaluation server[38] by adapting it to multilingual caption datasets.

② For measuring consistency, we measure average cosine similarity (CosSim) to report variance in the generated captions. Zero indicates that semantics of the sentences across languages are wholly independent and one denotes their semantic uniformity.

6.4.2 *Evaluation Results*

6.4.2.1 *Implementation*

In the following, we present the implementation details of each component utilized in the caption generation model.

Spatial Attention Features

Set a_v is extracted in two different ways. (1) Extracted from images using the Faster R-CNN [209] in conjunction with the ResNet-101 [98] trained on visual genome data by Anderson et al. [8]. Only top 36 image region features are selected with each region feature a_{v_j} of dimension 2048. We refer this set to Att→RCNN (2) Spatial feature outputs of the last convolutional layer of ResNet-101 pretrained on ImageNet is used, which have a dimension of $2048 \times 7 \times 7$. This means 49 image region features are selected with each region feature a_{v_j} of dimension 2048. We refer this set to Att→Spatial.

Global Visual Features

I_v of dimension 2048 is extracted using the average pooling of the aforementioned image region features.

Description Generation Model

It is initialized with 512 dimensions word embeddings w_t pre-trained using

[37] Only for English, as the approach is tightly coupled with English parser.

[38] https://github.com/peteanderson80/coco-caption.git

Glove [195] with the image-caption training corpora of English and other languages separately. The dimensions of hidden units $\mathbf{h}_t^1, \mathbf{h}_t^2$ in **L-1** and **L-2** of models are set to 512. Also, hidden units of shared layer $\mathbf{h}_t^{(s)}$ is set to 512. All models are then trained with Adam optimizer [132] with gradient clipping having maximum norm of 1.0 and mini-batch size of 50 for 25 epochs. Initially, learning is set to 0.001 and is reduced by factor of 10 if there is no improvement in the validation loss for 3 continuous epochs. For the multi-task models, $\lambda^{(l_1)}$ and $\lambda^{(l_2)}$ are set to 0.5.

6.4.2.2 *Baselines*

We compared our approaches with existing methods that provide open source implementations.

Visual Attention

Proposed by Xu et al. [286], visual attention model (*Visual-Att*) is used for training separate English, German and Japanese captions models. We regenerated the test descriptions using publicly available code[39]. Furthermore, machine translation is used to translate English captions to other languages using Google translate [40] and is denoted with *MT-Visual-Att*.

Japanese Generator (Ja-Gen)

Japanese descriptions dataset of MSCOCO images proposed by Yoshikawa et al [290] is also used as another baseline.

6.4.2.3 *Quantitative Results*

We compared our proposed models with aforementioned baselines in the Section 6.4.2.2. Results attained are shown in the Table 16 and Table 17. It can be observed that the multi-task models were comparable to the state of the art English generation model results while generating consistent caption across languages.

6.4.2.4 *Qualitative Results*

We performed qualitative analysis by finding the overlap of frequent starting bigrams, and also verbs of generated captions across languages.

Frequent Words

To understand the language differences, we examine the generated descriptions by our best model (i.e., MTA-GVC-O) across languages by extracting Top-5 highly frequent starting bigrams in different languages as shown in the Table 18. We can observe that there is a considerable overlap of English and German usage in the Multi30K dataset.

[39] https://github.com/kelvinxu/arctic-captions

[40] https://translate.google.com

	Multi30K			
	BLEU-4	CIDEr	SPICE	CosSim
Model	En,De	En,De	En	
Ja-Gen	-,-	-,-	-	-
Visual-Att	16.9,9.4	35.7,22.3	11.3	0.525
MT-Visual-Att	16.9,8.5	35.7,16.9	11.3	0.553
MTA-GVC-I				
+Att→Spatial	17.8,10.4	36.7,27.6	12.0	0.614
+Att→RCNN	-,-	-,-	-	-
MTA-GVC-O				
+Att→Spatial	**17.9,10.5**	**37.2,27.7**	**12.1**	**0.628**
+Att→RCNN	-,-	-,-	-	-

Table 16: Results achieved with our models in comparison with baseline approaches. For future comparisons, our MTA-GVCO+Att→Spatial model METEOR score for Multi30k English captions is 18.0.

	MSCOCO&STAIR			
	BLEU-4	CIDEr	SPICE	CosSim
Model	En,Ja	En,Ja	En	
Ja-Gen	-,38.5	-,83.3	-	-
Visual-Att	25.0,29.0	89.2,77.4	17.2	0.453
MT-Visual-Att	25.0,25.8	89.2,58.0	17.2	0.489
MTA-GVC-I				
+Att→Spatial	30.5,35.4	92.4, 83.1	17.6	0.490
+Att→RCNN	31.3,36.6	94.9,88.9	18.1	0.522
MTA-GVC-O				
+Att→Spatial	30.6,35.6	92.8,83.8	17.8	0.498
+Att→RCNN	**31.6,37.4**	**95.3,90.3**	**18.2**	**0.525**

Table 17: Results achieved with our models in comparison with baseline approaches. For future comparisons, our MTA-GVCO+Att→RCNN model METEOR score for MSCOCO English captions is 30.1.

Multi30K			
English	Count	German	Count
A man	487	Ein Mann (A man)	458
A group	178	Eine Frau (A woman)	118
A woman	81	Ein paar (A few)	53
A little	49	Eine gruppe (A group)	34
A young	30	Ein hund (A dog)	32

Table 18: Frequent bigrams as starting tokens used in the generated captions along with their translations. Number of captions in each dataset is: Multi30K → 1000

This shows that a jointly trained model can make closely related languages (e.g., West Germanic languages) generated captions closer to each other resulting in semantically similar captions.

Frequent Part-of-Speech (POS)

We also analyzed the POS tags mainly verbs generated across languages in the Table 19. Analyzing verbs will help to understand the actions that are captured in the generated captions across languages. It can be observed that there is an overlap of verbs (considering root verbs and removing conjugation in German) showing that image descriptions use similar verbs across languages.

Out-of-Vocabulary (OOV)

To induce embeddings for the words present in the generated captions to evaluate cross-language semantic similarity. We leverage the embeddings of the out-of-domain dataset vocabulary. Percentage of OOV words for each in-domain dataset is presented in Table 20.

6.4.3 *Evaluation Results Analyses*

We first start our analysis with the effect of image features on the models. Observing Table 16 and Table 17 shows that the models trained with Att→RCNN features were comparatively better than the models trained with spatial image features. We also comprehend from the results that the position (i.e., Input or Output) at which global image features are provided in the model also play a crucial for consistent caption generation across languages.

Furthermore, when results across languages are compared for effectiveness, we observed that the high BLEU-4 scores are obtained for German, Japanese

Verbs			
English	Count	German	Count
is	525	sitzt (sitting)	159
are	331	steht (standing)	112
playing	153	stehen (stand)	67
sitting	147	spielt (play)	65
standing	110	sitzen (sit)	61
walking	96	springt (jump)	52
riding	59	spielen (play)	33
running	36	tanzen (dancing)	28
holding	34	gehen (walk)	23
dancing	32	läuft (running)	20

Table 19: Frequent Top-10 verbs observed in generated captions of Multi30K

Dataset	Language	%
MSCOCO	English	0.499
STAIR	Japanese	5.190
Multi30K	English	0.184
Multi30K	German	2.581

Table 20: Out-of-Vocabulary (OOV) Percentange.

when compared against English. We attribute this outcome mostly to the difference in length of descriptions that are generated. In most cases, English had shorter descriptions than German and Japanese. Since BLEU weighs recall over precision, it shows that English descriptions are more coherent than German and Japanese.

The variance among generated captions across languages is also examined with the cosine similarity (CosSim). It is also observed that an improvement in the similarity assessment is achieved with our best MTA model when compared with the *Visual-Att* and *MT-Visual-Att*. It shows that MTA models were able to generate semantically closer captions across languages. Also, it is perceived that the machine translation of English captions degraded the performance measures. It conveys that the machine translation (MT) can induce errors and not a right approach for generating multi-language image captions.

6.5 SUMMARY

In this chapter, we addressed the third research question:

> ✍ **Research Question 3**
>
> Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation of all views if an auxiliary view depicting text in multiple languages is added to assist the *generation* of text from an image?

For this, we validated Hypothesis 3 by proposing models to generate consistent image captions across languages. We built these models by jointly optimizing two different language captions of an image by leveraging multi-task learning. Empirical analyses exhibited that single-task models generate different captions for a given image and this can be mitigated with joint learned models and knowledge sharing from different languages.

In the next chapter, we will present an approach to achieving multi-view deep representation learning with consensus and complementarity principles and support an application where the heterogeneous data has missing views.

MVRL WITH MISSING VIEWS AND KNOWLEDGE GUIDED ASSISTANCE

7

MVRL WITH MISSING VIEWS AND KNOWLEDGE GUIDED ASSISTANCE

Context of this Chapter. In this chapter, we leverage multi-view deep representation learning along with knowledge guided assistance to propose a novel approach for unseen visual object categories caption generation. More specifically, our approach for unseen or novel image caption generation is guided by an external resource such as knowledge graph (KG). Entities in KG are leveraged to identify the critical regions of an image for attention mechanism and also serve as image labels for caption generation during training and inference. Moreover, KG entities as image labels are also used while inference to constrain visual object categories. In particular, this work will allow us to scale the image caption generation to unseen or novel objects categories present on the web.

Our main contributions presented in this chapter can be broadly summarized as follows:

① We designed a novel approach, called Knowledge Guided Assistance (KGA), to improve the task of generating captions for images which contain visual objects that are not seen in the training data.

② We created a image classifier for linking the depicted visual objects to KG entities. Based on that, we introduce the first mechanism that exploits the relational structure of entities in KGs for guiding the attention of a caption generator towards picking the correct KG entity to mention in its descriptions.

③ We conducted an extensive experimental evaluation showing the effectiveness of our KGA method. Both, regarding generating effectual captions and also scaling it to more than 600 visual objects.

Outline. The remainder of this chapter is organized into following sections. Initially, Section 7.1 presents the motivation in Section 7.1.1 and briefly introduce existing unseen or novel image caption generation methods in Section 7.1.2. Next Section 7.2 presents the research question and describes our contribution to unseen visual object categories caption generation. Our approach knowledge guided assitance is then discussed in the Section 7.3. The dataset and metrics used for evaluation of the approach are described in the Section 7.4. While in the Section 7.4.2 details about the evaluation results are shown, which are further analyzed in the Section 7.4.3. Summary of the chapter is presented in the Section 7.5.

7.1 INTRODUCTION

7.1.1 *Motivation*

Content on the Web is highly heterogeneous and consists mostly of visual and textual information. In most cases, these different modalities complement each other, which complicates the capturing of the full meaning of automated knowledge extraction techniques. An approach for making information in all modalities accessible to automated processing is linking the information represented in the different modalities (e.g., images and text) into a shared conceptualization, like entities in a Knowledge Graph (KG). However, obtaining a robust formal representation of textual and visual content has remained a research challenge for many years.

Recently, a different approach has shown impressive results, namely the transformation of one unstructured representation into another. Specifically, the task of generating natural language descriptions of images or videos [256, 263] has gained much attention. While such approaches are not relying on formal conceptualizations of the domain to cover, the systems that have been proposed so far are limited by a tiny number of objects that they can describe (less than 100). Such methods – as they need to be trained on manually crafted image-caption parallel data – do not scale to real-world applications, and can't be applied to the cross-domain web-scale content.

In contrast, visual object classification techniques have improved considerably, and they are now scaling to thousands of objects more than the ones covered by caption training data [55]. Also, KGs have grown to cover all of those objects plus millions more accompanied by billions of facts describing relations between those objects. Thus, it appears that those information sources are the missing link to make existing image captioning models scale to a more significant number of objects without having to create additional image-caption training pairs with those missing objects.

In this chapter, we investigate the hypothesis, which conceptual relations of entities – as represented in KGs – can provide information to enable caption generation models to generalize to objects that they have not seen during training in the image-caption parallel data. While there are existing methods that are tackling this task, none of them have exploited any form of conceptual knowledge so far. In our model, we use KG entity embeddings to guide the attention of caption generator to the correct (unseen) object that is depicted in the image.

The contribution of this work on a broader scope is its progress towards the integration of the visual (and textual) information available on the Web with KGs.

7.1.2 *Background on Describing Images with Unseen Objects*

Existing methods such as Deep Compositional Captioning (DCC) [102], Novel object Captioner (NOC) [257], Constrained Beam Search (CBS) [7] and LSTM-C [289] address the challenge by transferring information between seen and unseen objects either before inference (i.e. before testing) or by keeping constraints on the generation of caption words during inference (i.e. during testing). Figure 25 provides a broad overview of those approaches.

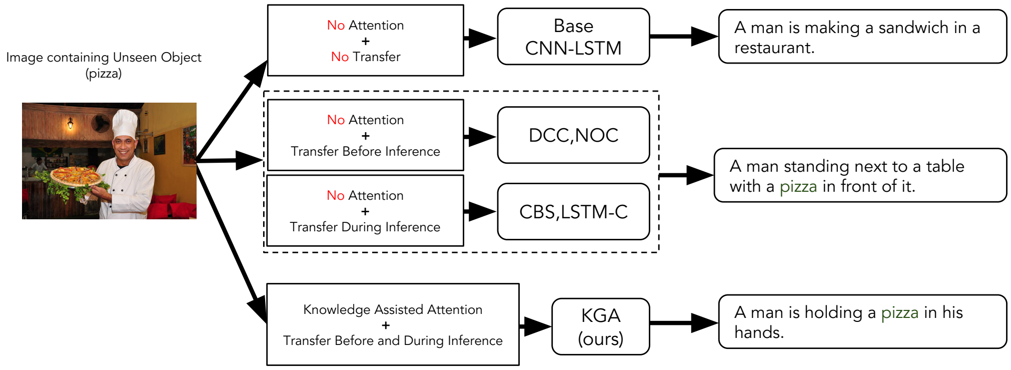


Image containing Unseen Object (pizza)
No Attention + No Transfer
Base CNN-LSTM
A man is making a sandwich in a restaurant.
No Attention + Transfer Before Inference
DCC,NOC
No Attention + Transfer During Inference
CBS,LSTM-C
A man standing next to a table with a pizza in front of it.
Knowledge Assisted Attention + Transfer Before and During Inference
KGA (ours)
A man is holding a pizza in his hands.

Figure 25: KGA goal is to describe images containing unseen objects by building on the existing methods i.e. DCC [102], NOC [257], CBS [7] and LSTM-C [289] and going beyond them by adding relational knowledge assistance. Base refers to our base description generation model built with CNN [236] - LSTM [108].

In DCC, an approach which performs information transfer only before inference, the training of the caption generation model is solely dependent on the corpus constituting words which may appear in the similar context as of unseen objects. Hence, explicit transfer of learned parameters is required between seen and unseen object categories before inference which limits DCC from scaling to a wide variety of unseen objects. NOC tries to overcame such issues by adopting a end-to-end trainable framework which incorporates auxiliary training objectives during training and detaching the need for explicit transfer of parameters between seen and unseen objects before inference. However, NOC training can result in sub-optimal solutions as the additional training attempts to optimize three different loss functions simultaneously. CBS, leverages an approximate search algorithm to guarantee the inclusion of selected words during inference of a caption generation model. These words are however only constrained on the image tags produced by a image classifier. And the vocabulary used to find similar words as candidates for replacement during inference is usually kept very large, hence adding extra computational complexity. LSTM-C avoids the limitation of finding similar words during inference by adding a copying mechanism into caption training. This assists the model during inference to decide whether a word is to be generated or copied from a dictionary. However, LSTM-C suffers from confusion problems since probabilities during word generation tend to get very low.

In general, aforementioned approaches also have the following limitations: (1) The image classifiers used cannot predict abstract meaning, like “hope”, as observed in many web images. (2) Visual features extracted from images are confined to the probability of occurrence of a fixed set of labels (i.e. nouns, verbs and adjectives) observed in a restricted dataset and cannot be easily extended to varied categories for large-scale experiments. (3) Since an attention mechanism is missing, important regions in an image are never attended. While, the attention mechanism in our model helps to scale down all possible identified concepts to the relevant concepts during caption generation. For large-scale applications, this plays a crucial role.

We introduce a new model called Knowledge Guided Assistance (KGA) that exploits conceptual knowledge provided by a knowledge graph (KG) [150] as external semantic attention throughout training and also to aid as a dynamic constraint before and during inference. Hence, it augments an auxiliary view as done in multi-view learning scenarios. Usage of KGs has already shown improvements in other tasks, such as in question answering over structured data, language modeling [2], and generation of factoid questions [230].

7.2 RESEARCH QUESTIONS AND CONTRIBUTIONS

Let us outline the research question, hypothesis, and contributions, which we target throughout the chapter.

7.2.1 *Research Question and Hypothesis*

As presented in Section 1.3, our overall research question is: How to effectively integrate multiple views of training instances depicting heterogeneous or homogeneous content into a common space representation for supporting applications in different domains? In this chapter, we address the fourth part, i.e., learning a common space representation among views emerging from the same modality with consensus and complementarity principles by leveraging their input representations. More specifically, we aim at Research Question 4:

> ✍ **Research Question 4**
>
> Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist the *generation* of text from an image if there are missing views?

For addressing above research question, we verify the hypothesis as follows:

> □ **Hypothesis 4**
>
> Leveraging deep neural network architecture with knowledge guided assistance, we can effectively learn a common space representation emerging from the heterogeneous data to generate one modality from another, i.e., especially generating text from an image containing unseen visual object categories. Knowledge guidance can be used in two ways. First, as an attention mechanism to identify import entities observed in the image modality. Second during inference to guide the generation of caption text.

Intuitively, Hypothesis 4 states that the combination of deep neural networks with knowledge graph embeddings can be used to adequately capture common space representation between entities observed in a knowledge graph and the spatial observed visual content. Mainly, we expect to provide an explicit knowledge graph grounding of entities observed in the visual content. Further, we expect to use knowledge graph entity embeddings and labels for image caption generation in the form of attention mechanism and as a constraint during inference respectively.

To validate Hypothesis, we present our KGA caption generation model in the Section 7.3 to generate captions for the unseen visual object categories. Moreover, we implemented the approach and empirically show (see the evaluation in Section 7.4) its effectiveness with state of the art.

7.2.2 *Contributions*

While being naturally appealing, usage of deep neural networks and its combination with knowledge graph entity embeddings is not studied before in the context of image caption generation, where the visual object categories are unseen before in the training data. Aiming at above hypotheses, we provide the following contribution:

- *Contribution for Hypothesis 4*
  Building a common space representation from the heterogeneous data depicting two views using either correlation or consensus/complementarity principles is studied earlier.

  However, our usage of deep encoder-decoder architecture and its extension with knowledge guided assistance for building a common space representation for those samples which has missing views in the training data is unique. Additionally, designing a caption generation approach by leveraging knowledge graph entity annotation on images and knowledge graph entity embeddings has new challenges. Facing these characteristics, we propose a knowledge-guided assistance caption generation model which uses entity embeddings to calculate attention score, while entity labels are used as constraints for guiding caption generation during inference. To the best of our knowledge, this is the first work which utilizes knowledge guided assistance for caption generation for images containing unseen visual object categories.

  We conducted an evaluation using out-of-domain image-caption dataset and ImageNet images to validate the Hypotheses 4. In these experiments, we could achieve performance comparable to state of the art for caption generation for images containing unseen visual object categories. In fact, we could show that our proposed approach progress towards the integration of the visual (and textual) information available on the Web with KGs.

7.3 DESCRIBING IMAGES WITH UNSEEN OBJECTS USING KGA

In this section, we present our caption generation model to generate captions for unseen visual object categories with knowledge guided assistance (KGA). Core goal of KGA is to introduce external semantic attention (ESA) into the learning and also work as a constraint before and during inference for transferring information between seen words and unseen visual object categories.

7.3.1 *Caption Generation Model*

Our image caption generation model (henceforth, KGA-CGM) combines three important components: a language model pretrained on unpaired textual corpora, external semantic attention (ESA) and image features with a textual (T), semantic (S) and visual (V) layer (i.e. TSV layer) for predicting the next word in the sequence when learned using image-caption pairs. In the following, we present each of these components separately while Figure 26 presents the overall architecture of KGA-CGM.

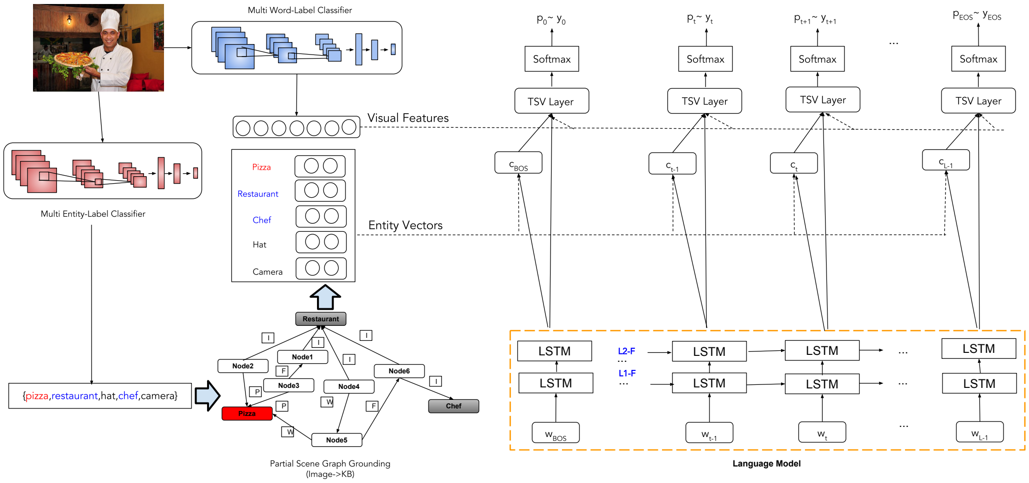


Figure 26: KGA-CGM is built with three components. A language model implemented with a 2-layer forward LSTM where L1-F and L2-F represents layer-1 and layer-2 respectively, a multi-word-label classifier to generate image visual features and a multi-entity-label classifier that generates entity-labels linked to a KG serving as a partial image specific scene graph. This information is further leveraged to acquire entity vectors for supporting ESA. w_t represents the input caption word, c_t the semantic attention, p_t the output of probability distribution over all words and y_t the predicted word at each time step t. BOS and EOS represent the special beginning and end of sentence tokens respectively.

7.3.1.1 *Language Model*

This component is crucial to transfer the sentence structure for unseen visual object categories. Language model is implemented with two long short-term memory (LSTM) [108] layers to predict the next word given previous words in a sentence. If $\overrightarrow{w_{1:L}}$ represent the input to the forward LSTM of layer-1 for capturing forward input sequences into hidden sequence vectors ($\overrightarrow{\mathbf{h}^1_{1:L}} \in \mathbb{R}^H$), where L is the final time step. Then encoding of input word sequences into hidden layer-1 and then into layer-2 at each time step t is achieved as follows:

$$\overrightarrow{\mathbf{h}^1_t} = \text{L1-F}(\overrightarrow{w_t}; \Theta) \tag{68}$$

$$\overrightarrow{\mathbf{h}^2_t} = \text{L2-F}(\overrightarrow{\mathbf{h}^1_t}; \Theta) \tag{69}$$

where Θ represent hidden layer parameters. The encoded final hidden sequence ($\overrightarrow{\mathbf{h}_t^2} \in \mathbb{R}^H$) at time step t is then used for predicting the probability distribution of the next word given by $\mathbf{p}_{t+1} = \mathtt{softmax}(\mathbf{h}_t^2)$. The softmax layer is only used while training with unpaired textual corpora and not used when learned with image captions.

7.3.1.2 *External Semantic Attention (ESA)*

Our objective in ESA is to extract semantic attention from an image by leveraging relational knowledge in KG as entity-labels obtained using a multi-entity-label image classifier (presented in the Section 7.4.1.2). Here, entity-labels are analogous to patches or attributes of an image. In formal terms, if $\boldsymbol{ea}_i$ is an entity-label and $\boldsymbol{e}_i \in \mathbb{R}^E$ the entity-label vector among set of entity-label vectors ($i = 1, .., L$) and $\boldsymbol{\beta}_i$ the attention weight of $\boldsymbol{e}_i$ then β_i is calculated at each time step t using Equation 70.

$$\boldsymbol{\beta}_{ti} = \frac{\exp(\mathbf{O}_{ti})}{\sum_{j=1}^{L} \exp(\mathbf{O}_{tj})} \tag{70}$$

where $\mathbf{O}_{ti} = f(\boldsymbol{e}_i, \mathbf{h}_t^2)$ represent scoring function which conditions on the layer-2 hidden state ($\mathbf{h}_t^2$) of a caption language model. It can be observed that the scoring function $f(\boldsymbol{e}_i, \mathbf{h}_t^2)$ is crucial for deciding attention weights. Also, relevance of the hidden state with each entity-label is calculated using Equation 71.

$$f(\boldsymbol{e}_i, \mathbf{h}_t^2) = \tanh((\mathbf{h}_t^2)^T W_{he} \boldsymbol{e}_i) \tag{71}$$

where $W_{he} \in \mathbb{R}^{H \times E}$ is a bilinear parameter matrix. Once the attention weights are calculated, the soft attention weighted vector of the context $\mathbf{c}$, which is a dynamic representation of the caption at time step t is given by Equation 72

$$\mathbf{c}_t = \sum_{i=1}^{L} \beta_{ti} \boldsymbol{e}_i \tag{72}$$

Here, $\mathbf{c}_t \in \mathbb{R}^E$ and L represent the cardinality of entity-labels per image-caption pair instance.

7.3.1.3 *Image Features & TSV Layer & Next Word Prediction*

Visual features for an image are extracted using multi-word-label image classifier (discussed in the Section 7.4.1.2). To be consistent with other approaches [102, 257] and for a fair comparison, our visual features (I) also have objects that we aim to describe outside of the caption datasets besides having word-labels observed in paired image-caption data.

Once the output from all components is acquired, the TSV layer is employed to integrate their features i.e. textual (T), semantic (S) and visual (V) yielded by language model, ESA and images respectively. Thus, TSV acts as a transformation layer for molding three different feature spaces into a single common space for prediction of next word in the sequence.

If $\mathbf{h}_t^2 \in \mathbb{R}^H$, $\mathbf{c}_t \in \mathbb{R}^E$ and $\mathbf{I}_t \in \mathbb{R}^I$ represent vectors acquired at each time step t from language model, ESA and images respectively. Then the integration at TSV layer of KGA-CGM is provided by Equation 73.

$$\mathbf{TSV}_t = W_{\mathbf{h}_t^2}\mathbf{h}_t^2 + W_{\mathbf{c}_t}\mathbf{c}_t + W_{\mathbf{I}_t}\mathbf{I}_t \tag{73}$$

where $W_{\mathbf{h}_t^2} \in \mathbb{R}^{\nu s \times H}, W_{\mathbf{c}_t} \in \mathbb{R}^{\nu s \times E}$ and $W_{\mathbf{I}_t} \in \mathbb{R}^{\nu s \times I}$ are linear conversion matrices and $\boldsymbol{\nu s}$ is the image-caption pair training dataset vocabulary size.

The output from the TSV layer at each time step t is further used for predicting the next word in the sequence using a softmax layer given by $p_{t+1} = \mathtt{softmax}(\mathbf{TSV}_t)$.

7.3.2 *KGA-CGM Training*

To learn parameters of KGA-CGM, first we freeze the parameters of the language model trained using unpaired textual corpora. Thus, enabling only those parameters to be learned with image-caption pairs emerging from ESA and TSV layer such as $W_{he}, W_{\mathbf{h}_t^2}, W_{\mathbf{c}_t}$ and $W_{\mathbf{I}_t}$. KGA-CGM is now trained to optimize the cost function that minimizes the sum of the negative log likelihood of the appropriate word at each time step given by Equation 74.

$$\min_{\theta} -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{L^{(n)}} \log(\mathbf{p}(y_t^{(n)})) \tag{74}$$

Where $L^{(n)}$ represent the length of sentence (i.e. caption) with beginning of sentence (BOS), end of sentence (EOS) tokens at n-th training sample and N as a number of samples used for training.

7.3.3 *KGA-CGM Constrained Inference*

Inference in KGA-CGM refer to the generation of descriptions for test images. Here, inference is not straightforward as in the standard image caption generation approaches [263] because unseen visual object categories have no parallel captions throughout training. Hence they will never be generated in a caption. Thus, unseen visual object categories require guidance either before or during inference from similar seen words that appear in the paired image-caption dataset and likely also from image labels. In our case, we achieve the guidance both before and during inference with varied techniques.

Guidance before Inference

We first identify the seen words in the paired image-caption dataset similar to the visual object categories unseen in image-caption dataset by estimating the semantic similarity using their Glove embeddings [195] learned using unpaired textual corpora (more details in Section 7.4.1.1). Furthermore, we utilize this information to perform dynamic transfer between seen words visual features ($W_{\mathbf{I}}$), language model ($W_{\mathbf{h}_t^2}$) and external semantic attention ($W_{\mathbf{c}_t}$)

weights and unseen visual object categories. To illustrate, if (v_{unseen}, i_{unseen}) and $(v_{closest}, i_{closest})$ denote the indexes of unseen visual object category “zebra” and its semantically similar known word “giraffe” in a vocabulary (v_s) and visual features (i_s) respectively. Then to describe images with “zebra” in the similar manner as of “giraffe”, the transfer of weights is performed between them by assigning $W_{\mathbf{c}_t}[v_{unseen},:]$, $W_{\mathbf{h}_t^2}[v_{unseen},:]$ and $W_{\mathbf{I}_t}[v_{unseen},:]$ to $W_{\mathbf{c}_t}[v_{closest},:]$, $W_{\mathbf{h}_t^2}[v_{closest},:]$ and $W_{\mathbf{I}_t}[v_{closest},:]$ respectively.

Furthermore, $W_{\mathbf{I}_t}[i_{unseen}, i_{closest}]$, $W_{\mathbf{I}_t}[i_{closest}, i_{unseen}]$ is set to zero for removing mutual dependencies of seen and unseen words presence in an image. Hence, aforementioned procedure will update the KGA-CGM trained model before inference to assist the generation of unseen visual object categories during inference as given by Algorithm 2.

Algorithm 2: Constrained Inference Overview (Before)

```
Input: M={$W_{he}$, $W_{\mathbf{h}_t^2}$, $W_{\mathbf{c}_t}$, $W_{\mathbf{I}_t}$}
Output: $M_{new}$
Initialize List(closest) = cosine_distance(List(unseen),vocabulary) ;
Initialize $W_{\mathbf{c}_t}[v_{unseen},:]$, $W_{\mathbf{h}_t^2}[v_{unseen},:]$, $W_{\mathbf{I}_t}[v_{unseen},:]$ = 0 ;
Function Before Inference
    forall items T in closest and Z in unseen do
        if T and Z is vocabulary then
            $W_{\mathbf{c}_t}[v_Z,:] = W_{\mathbf{c}_t}[v_T,:]$ ;
            $W_{\mathbf{h}_t^2}[v_Z,:] = W_{\mathbf{h}_t^2}[v_T,:]$ ;
            $W_{\mathbf{I}_t}[v_Z,:] = W_{\mathbf{I}_t}[v_T,:]$ ;
        end
        if $i_T$ and $i_Z$ in visual features then
            $W_{\mathbf{I}_t}[i_Z,i_T]$=0 ;
            $W_{\mathbf{I}_t}[i_T,i_Z]$=0 ;
        end
    end
    $M_{new}$ = M ;
    return $M_{new}$ ;
end
```

Guidance during Inference

The updated KGA-CGM model is used for generating descriptions of unseen visual object categories. However, in the before-inference procedure, the closest words to unseen visual object categories are identified using embeddings that are learned only using textual corpora and are never constrained on images. This obstructs the view from an image leading to spurious results. We resolve such nuances during inference by constraining the beam search used for description generation with image entity-labels ($\mathbf{ea}$). In general, beam search is used to consider the best k sentences at time t to identify the sentence at the next time step. Our modification to beam search is achieved by adding a extra constraint to check if a generated unseen visual object category is part of the entity-labels. If it's not, unseen visual object categories are never replaced with their closest seen words. Algorithm 3 presents the overview of KGA-CGM guidance during inference.

Algorithm 3: Constrained Inference Overview (During)

```
Input: $M_{new}$, $Im_{labels}$, beam-size $k$, word $w$
Output: best $k$ successors
Initialize $Im_{labels}$ = Top-5 ($ea$) ;
Initialize beam-size $k$ ;
Initialize word $w$=null ;
Function During Inference
    forall State $st$ of $k$ do
        $w$=$st$ ;
        if closest[$w$] in $ea$ then
            $st$ = closest[$w$];
        end
        else
            $st$ = $w$ ;
        end
    end
    return best $k$ successors ;
end
```

7.4 Evaluation

7.4.1 *Evaluation Setup*

7.4.1.1 *Resources and Datasets*

Our approach is dependent on several resources and datasets.

Knowledge Graphs (KGs) and Unpaired Textual Corpora

There are several openly available KGs such as DBpedia[41], Wikidata[42], and YAGO[43] which provide relational knowledge encapsulated in entities and their relationships. We choose DBpedia as our KG for entity annotation, as it is one of the extensively used resource for semantic annotation and disambiguation [150][44].

For learning weights of the language model and also Glove word embeddings, we have explored different unpaired textual corpora from out-of-domain sources (i.e. out of image-caption parallel corpora) such as the British National Corpus

[41] http://wiki.dbpedia.org/

[42] https://www.wikidata.org/wiki/Wikidata:Main_Page

[43] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

[44] we presume other KGs also have high quality information and do not distinguish them based on qualitative measures. DBpedia is chosen for convenience.

(BNC)[45], Wikipedia (Wiki) and subset of SBU1M[46] caption text containing 947 categories of ILSVRC12 dataset [222]. NLTK[47] sentence tokenizer is used to extract tokenizations and around 70k+ words vocabulary is extracted with Glove embeddings.

Unseen Objects Description (Out-of-Domain MSCOCO & ImageNet)

To evaluate KGA-CGM, we use the subset of MSCOCO dataset [157] proposed by Hendricks et al. [102]. The dataset is obtained by clustering 80 image object category labels into 8 clusters and then selecting one object from each cluster to be held out from the training set. Now the training set does not contain the images and sentences of those 8 objects represented by bottle, bus, couch, microwave, pizza, racket, suitcase and zebra. Thus making the MSCOCO training dataset to constitute 70,194 image-caption pairs. While validation set of 40504 image-caption pairs are again divided into 20252 each for testing and validation. Now, the goal of KGA-CGM is to generate caption for those test images which contain these 8 unseen object categories. Henceforth, we refer this dataset as "out-of-domain MSCOCO".

To evaluate KGA-CGM on a more challenging task, we attempt to describe images that contain wide variety of objects as observed on the web. To imitate such a scenario, we collected images from collections containing images with wide variety of objects. First, we used same set of images as earlier approaches [257, 289] which are subset of ImageNet [55] constituting 642 object categories used in Hendricks et al. [102] who do not occur in MSCOCO. However, 120 out of those 642 object categories are part of ILSVRC12.

7.4.1.2 *Multi-Label Image Classifiers*

The important constituents that influence KGA-CGM are the image entity-labels and visual features. Identified objects/actions etc. in an image are embodied in visual features, while entity-labels capture the relational knowledge in an image grounded in KG. In this section, we present the approach to extract both visual features and entity-labels.

Multi-Word-label Image Classifier

To extract visual features of out-of-domain MSCOCO images, emulating Hendricks et al. [102] a multi-word-label classifier is built using the captions aligned to an image by extracting part-of-speech (POS) tags such as nouns, verbs and adjectives attained for each word in the entire MSCOCO dataset. For example, the caption "A young child brushes his teeth at the sink" contains word-labels such as "young (JJ)", "child (NN)", "teeth (NN)" etc., that represent concepts in an image. An image classifier is trained now with 471 word-labels using a sigmoid cross-entropy loss by fine-tuning VGG-16 [236] pre-trained on the training part of the ILSVRC12. The visual features extracted for a new image represent the

[45] http://www.natcorp.ox.ac.uk/

[46] http://vision.cs.stonybrook.edu/~vicente/sbucaptions/

[47] http://www.nltk.org/

probabilities of 471 image labels observed in that image. For extracting visual features from ImageNet images, we replace the multi-word-label classifier with the lexical classifier [102] learned with 642 ImageNet object categories.

Multi-Entity-label Image Classifier (MSCOCO)

To extract relational knowledge for out-of-domain MSCOCO images analogous to the word-labels, a multi-entity-label classifier is build with entity-labels attained from a knowledge graph annotation tool such as DBpedia spotlight[48] on training set of MSCOCO constituting 82,783 training image-caption pairs. In total around 812 unique labels are extracted with an average of 3.2 labels annotated per image. To illustrate, considering the caption presented in the aforementioned section, entity labels extracted are "Brush[49]" and "Tooth[50]". An image classifier is now trained with multiple entity-labels using sigmoid cross-entropy loss by fine-tuning VGG-16 [236] pre-trained on the training part of the ILSVRC12.

Fine-tuning with entity labels is explored with varied feature representation of images extracted using three different layers such as pool5, fc6 and fc7 of VGG-16 [236] pre-trained on ILSVRC12. Furthermore, we analyzed image classifiers built separately using pool5, fc6 and fc7. Our analysis revealed that pool5 features overfit even with regularization.

To address this challenge, we trained a classifier with Caffe[51] by fine-tuning the layers above fc6 and fc7 which gave us an improvement in the accuracy as observed in the Table 21.

The classifier fine-tuned on fc6 features constitute two fully connected layers of dimensions 4096 and an output layer comprising a sigmoid activation with 812 dimensions. Similarly, the classifier fine-tuned with fc7 features have an output layer of 812 dimensions comprising a sigmoid activation. The loss function used during training is sigmoid cross-entropy, while only sigmoid is used during prediction for exhibiting the presence of label probabilities.

Figure 27 shows the predictions on the test dataset. It can be observed that fc6 gave the best result with an accuracy around 70% for top-12 and 74.4% top-16 label predictions. Table 22 shows sample entity-label predictions on MSCOCO test images with our multi entity-label image classifier fine-tuned on VGG-16 fc6 layer. Visual object categories used are subset of MSCOCO objects (i.e. book, bed, carrot, elephant, spoon, toilet, truck and umbrella) mainly used by NOC [257] as alternate set of "out of domain" MSCOCO objects.

Multi-Entity-label Image Classifier (ImageNet)

For extracting entity-labels from ImageNet images, we again leveraged lexical classifier [102] learned with 642 ImageNet object categories. However, as all 642 categories denote WordNet synsets, we build a connection between these categories and DBpedia by leveraging BabelNet [185] for multi-entity-label classifier.

[48] https://github.com/dbpedia-spotlight/

[49] http://dbpedia.org/resource/Brush

[50] http://dbpedia.org/resource/Tooth

[51] http://caffe.berkeleyvision.org/

	Model		
Hyper Parameters	pool5	fc6	fc7
weight_decay	0.05	0.03	0.01
base_lr	0.001	0.0003	0.003
gamma	0.5	0.5	0.33
stepsize	7.5K	10K	8K
maxiter	60K	50K	40K
momentum	0.9	0.9	0.9
batch_size	256	256	256
	Results		
Validation Loss	11.0035	**10.1152**	10.3372
Accuracy@12	0.6572	**0.7018**	0.6868
Accuracy@K	0.4526	**0.4892**	0.4778

Table 21: Validation results of different VGG-16 layers. Hyper parameters are used to fine-tune Caffe VGG-16 model. Accuracy@K is calculated by predicting as many labels as in ground truth for each image.

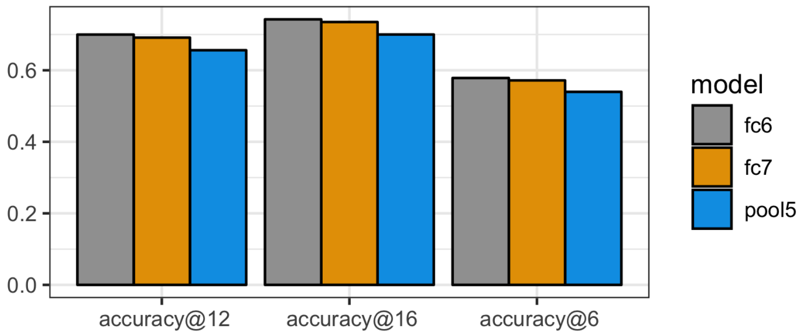


Figure 27: Accuracy of the predicted labels on the test set by Multi Entity-Label Classifier.

To illustrate, for visual object category "wombat" (wordnetid: *n1883070*) in ImageNet can be linked to DBpedia Wombat[52]. Hence, this makes our method very modular for building new image classifiers to incorporate relational knowledge.

7.4.1.3 *Entity-Label Embeddings*

We presented earlier that the acquisition of entity-labels for training multi-entity-label classifiers were obtained using DBpedia spotlight entity annotation and disambiguation tool. Hence, entity-labels are expected to encapsulate relational knowledge grounded in KB. Approaches [240] earlier have transformed such entities in a KB into embeddings to capture their relational information for tasks such as knowledge base completion. In our work, we see the efficacy of these embeddings for caption generation. We leverage entity-label embeddings for

[52] http://dbpedia.org/page/Wombat

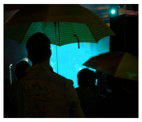
Image	MSCOCO Object	Labels
	book	Plant, Television, Coffee_table_book (Original) [Furniture, Television,Couch] (Predicted)
	bed	[Bedding, Wood, Textile Canopy_(biology)] (Original) [Pillow, Canopy, Hanging] (Predicted)
	carrot	Knife, Vegetable, Meat, Wine, Potato, Fork, Glass, Carrot (Original) [Vegetable, Meat, Carrot] (Predicted)
	elephant	Poaceae, Elephant, Grass (Original) [Elephant, Enclosure, Poaceae] (Predicted)
	spoon	Grape, Milk, Spoon, Fruit (Original) [Fruit, Spoon, Apple] (Predicted)
	toilet	Light, Cabinetry, Pedestal, Medicine, Toilet (Original) [Toilet, Mirror, Hanging] (Predicted)
	truck	Truck, Straw (Original) [Truck, Poaceae, Bus] (Predicted)
	umbrella	Light, Umbrella (Original) [Light, Umbrella, Light_fixture] (Predicted)

Table 22: Sample predictions of Multi entity-label classifier (MSCOCO).

computing semantic attention observed in an image with respect to the caption as observed from KB. To obtain entity-label embeddings, we adopted the RDF2Vec [214] approach and generated 500 dimensional vector representations for 812 and 642 entity-labels to describe out-of-domain MSCOCO and ImageNet images respectively.

Furthermore, we qualitatively evaluate the entity-label embeddings. There are total 812 entity-labels in total used to represent images in the entire MSCOCO. Most of these images are represented with more than one entity-label, thus providing multi-label information for each image. However, directly using their embeddings for ESA can affect caption generation if the label embeddings are not closely related. To check for their closely relatedness, we perform entity similarity. Table 23 shows the results of unseen or novel mscoco objects.

Unseen Object	Top-5 Closely Related Entities
Bottle	Wine_bottle, Wine_glass, Table_setting Nap_(textile), Tablecloth
Bus	Truck, Double-decker_bus, Transit_bus Cargo, Tram
Couch	Pillow, Cupboard, Bathtub Hair_dryer, Living_room
Microwave	Blender, Oven, Paper_bag Dishwasher, Refrigerator
Pizza	Pasta, Pepperoni, Salad Sauce, Grilling
Racket	Ball, Flying_disc, Snowboard Glove, Cricket_ball
Suitcase	Baggage, Backpack, Hair_dryer Apron, Bathtub
Zebra	Giraffe, Elephant, Horn_(anatomy) Calf, Ox

Table 23: Top-5 closely related entities of unseen MSCOCO Objects

It can be perceived from the Table 23 that most of the closely related entities always co-occur in an image as shown with few examples in the paper. Thus enhancing the caption generation model with ESA proven to be effective. We also performed t-SNE visualization of all entity-labels to check how they cluster together. It can be seen from the Figure 28 that some of the closely related objects that occur in the same context cluster close to each other.

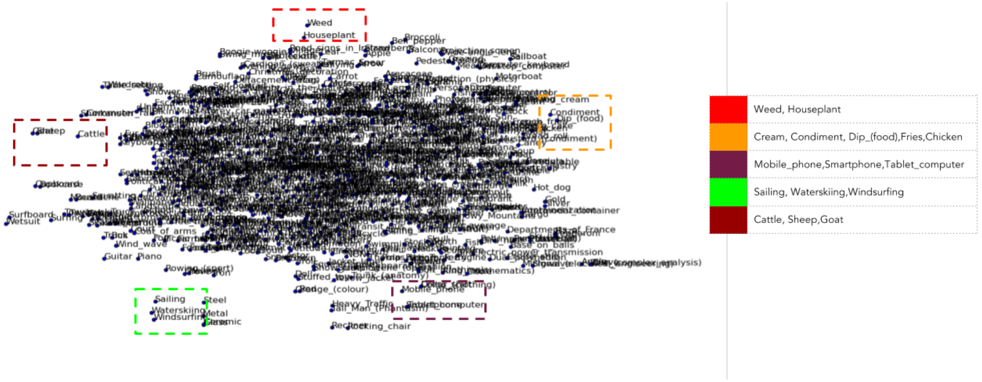


Figure 28: t-SNE visualization of the entity-label embeddings.

7.4.1.4 *Evaluation Measures*

To evaluate generated descriptions for the unseen MSCOCO visual object categories, we use similar evaluation metrics as earlier approaches [102, 257, 289] such as METEOR and also SPICE [6]. However, CIDEr [255] metric is not used as it is required to calculate the inverse document frequency used by this metric across the entire test set and not just unseen object subsets. F1 score is also calculated to measure the presence of unseen objects in the generated captions when compared against reference captions. Furthermore, to evaluate ImageNet object categories description generation: we leveraged F1 and also other metrics such as Unseen and Accuracy scores [257, 289]. The Unseen score measures the percentage of all novel objects mentioned in generated descriptions, while accuracy measure percentage of image descriptions correctly addressed the unseen objects.

7.4.2 *Evaluation Results*

The experiments are conducted to evaluate the efficacy of KGA-CGM model for describing out-of-domain MSCOCO and ImageNet images.

7.4.2.1 *Implementation*

KGA-CGM model constitutes three important components i.e. language model, visual features and entity-labels. Before learning KGA-CGM model with image-caption pairs, we first learn the weights of language model and keep it fixed during the training of KGA-CGM model. To learn language model, we leverage unpaired textual corpora and provide input word embeddings representing 256 dimensions pre-trained with Glove [195] on the same unpaired textual corpora. However, different hidden layer dimensions are explored to see their consequences on caption generation. KGM-CGM model is then trained using image-caption pairs with Adam optimizer [132] with gradient clipping having maximum norm of 1.0 for about 15~50 epochs. Validation data is used for fine tuning parameters and model selection.

7.4.2.2 *Describing Out-of-Domain MSCOCO Images*

In this section, we evaluate KGA-CGM using out-of-domain MSCOCO dataset described in the Section 7.4.1.1.

Quantitative Analysis

We compared our complete KGA-CGM model with the other existing models that generated image descriptions on out-of-domain MSCOCO. To have a fair comparison, only those results are compared that used VGG-16 to generate image features. Table 24 and Table 25 shows the comparison of individual and average scores based on METEOR, SPICE and F1 on all 8 unseen visual object categories with beam size 1.

		F1			
Model	Beam	microwave	racket	bottle	zebra
DCC [102]	1	28.1	52.2	4.6	79.9
NOC [257]	>1	24.7	55.3	17.7	89.0
CBS(T4) [7]	>1	29.7	57.1	16.3	85.7
LSTM-C [289]	>1	27.8	70.2	29.6	91.4
KGA-CGM	1	**50.0**	**75.3**	**29.9**	**92.1**
		METEOR			
DCC [102]	1	22.1	20.3	18.1	22.3
NOC [257]	>1	21.5	24.6	21.2	21.8
LSTM-C [289]	>1	-	-	-	-
CBS(T4) [7]	>1	-	-	-	-
KGA-CGM	1	**22.6**	**25.1**	**21.5**	**22.8**
		SPICE			
DCC [102]	>1	-	-	-	-
CBS(T4) [7]	>1	-	-	-	-
KGA-CGM	1	13.3	16.8	13.1	19.6

Table 24: Individual measures for four unseen objects. Best results are highlighted, while underline shows second best.

It can be noticed that KGA-CGM with beam size 1 was comparable to other approaches even though it used fixed vocabulary from image-caption pairs. For example, CBS [7] used expanded vocabulary of 21,689 when compared to 8802 by us. Also, our word-labels per image are fixed, while CBS uses a varying size of predicted image tags (T1-4). This makes it non-deterministic and can increase uncertainty, as varying tags will either increase or decrease the performance. Furthermore, we also evaluated KGA-CGM for the rest of seen visual object categories in the Table 26. It can be observed that our KGA-CGM outperforms existing approaches as it did not undermine the in-domain description generation, although it was tuned for out-of-domain description generation.

7.4.2.3 *Ablation Study*

To understand how different components of KGA-CGM influence the unseen visual object categories caption generation, we perform ablation study by removing different components of KGA-CGM. Table 27 present the results obtained. All reported scores are average of 8 unseen visual object categories. It can be noticed that *None*, which refers to our **base CNN-LSTM** model which did not use either ESA or constrained inference (CI) in the KGA-CGM model has F1 measure of zero. Enabling ESA into our base CNN-LSTM model (i.e. Attention + No CI) has shown an increase in the METEOR and SPICE as observed in *Only ESA*.

			F1			
Model	Beam	pizza	couch	bus	suitcase	Average
DCC [102]	1	64.6	45.9	29.8	13.2	39.7
NOC [257]	>1	69.3	25.5	68.7	39.8	48.8
CBS(T4) [7]	>1	**77.2**	**48.2**	67.8	**49.9**	54.0
LSTM-C [289]	>1	68.1	38.7	**74.4**	44.7	**55.6**
KGA-CGM	1	70.6	42.1	54.2	25.6	55.0
			METEOR			
DCC [102]	1	**22.2**	**23.1**	**21.6**	18.3	21.0
NOC [257]	>1	21.8	21.4	20.4	18.0	21.3
LSTM-C [289]	>1	-	-	-	-	23.0
CBS(T4) [7]	>1	-	-	-	-	23.3
KGA-CGM	1	21.4	23.0	20.3	**18.7**	22.0
			SPICE			
DCC [102]	>1	-	-	-	-	13.4
CBS(T4) [7]	>1	-	-	-	-	**15.9**
KGA-CGM	1	13.2	14.9	12.6	10.6	14.3

Table 25: Individual and Average measures for all 8 unseen objects. Best results are highlighted, while underline shows second best.

		Seen Objects	
Model	Beam	METEOR	SPICE
DCC [102]	1	23.0	15.9
CBS(T4) [7]	>1	24.5	18.0
KGA-CGM	1	24.1	17.2
KGA-CGM	>1	**25.1**	**18.2**

Table 26: Average measures of MSCOCO seen objects.

Model	Beam	METEOR	SPICE	F1
None	1	19.7	11.7	0
Only ESA	1	20.5	12.8	0
Only CI	1	20.1	12.3	39.8
ESA+CI	1	22.0	14.3	55.0

Table 27: KGA-CGM Ablation Study

However, the F1 measure has remained zero due to no transfer of information between seen words and unseen visual object categories. Alternatively, enabling CI showed a jump in F1 measure as seen in *Only CI*. However, both METEOR and SPICE are lower than *Only ESA* due to missing attention from ESA. Enabling both ESA and CI make our complete KGA-CGM model equipped with both external semantic attention from the image as well as the constrained transfer of information between seen words and unseen visual object categories providing highest METEOR and SPICE scores of 22.0 and 14.3 respectively as observed in *ESA+CI*. Also, it has an increased F1 measure when compared to *Only CI*. This shows that the coherent and accurately generated caption is important for presence of an object in the caption.

7.4.2.4 *Language Model Hidden Layers Influence*

The language model in KGA-CGM is a 2-layer forward LSTM. For learning KGA-CGM with image-caption pairs, input caption word embeddings are chosen to be 256 dimensions, while the LSTM hidden layer dimensions for both layer-1 and layer-2 is selected as 512. However, varying hidden layer dimensions can show an influence on the caption generation results. In this section, we vary the hidden layer dimensions and analyze the consequences. Table 28 shows the METEOR, SPICE and F1 average measures on 8 unseen MSCOCO visual object categories.

Layer-1	Layer-2	Beam	METEOR	SPICE	F1-score
256	256	1	20.9	13.5	50.8
256	512	1	21.1	13.6	48.2
512	512	1	**22.0**	**14.3**	**55.0**
256	256	>1	20.2	13.2	42.9
256	512	>1	20.2	13.1	41.8
512	512	>1	21.5	13.9	48.9

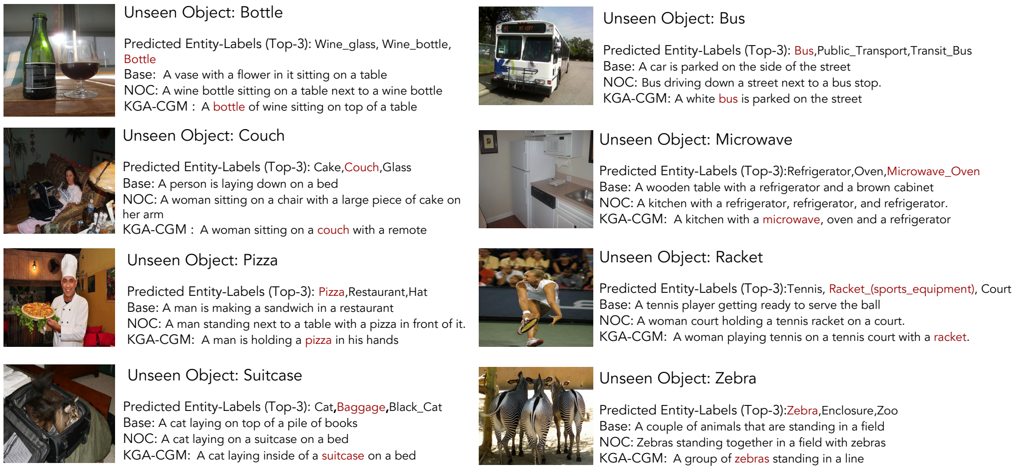
Table 28: Effect on KGA-CGM with varying LSTM hidden layer dimensions in Language model.

7.4.2.5 *Qualitative Analysis*

In Figure 29, sample predictions of our best KGA-CGM model is presented. It can be observed that entity-labels has shown an influence for caption generation. Since, entities as image labels are already disambiguated, it attained high similarity in the prediction of a word thus adding useful semantics. Figure 29 presents the example unseen visual objects descriptions.

7.4.2.6 *Describing ImageNet Images*

ImageNet images do not contain any ground-truth captions and contain exactly one unseen visual object category per image. Initially, we first retrain different language models using unpaired textual data (Section 7.4.1.1) and also the entire



Unseen Object: Bottle

Predicted Entity-Labels (Top-3): Wine_glass, Wine_bottle, Bottle
Base: A vase with a flower in it sitting on a table
NOC: A wine bottle sitting on a table next to a wine bottle
KGA-CGM : A bottle of wine sitting on top of a table

Unseen Object: Bus

Predicted Entity-Labels (Top-3): Bus,Public_Transport,Transit_Bus
Base: A car is parked on the side of the street
NOC: Bus driving down a street next to a bus stop.
KGA-CGM: A white bus is parked on the street

Unseen Object: Couch

Predicted Entity-Labels (Top-3): Cake,Couch,Glass
Base: A person is laying down on a bed
NOC: A woman sitting on a chair with a large piece of cake on her arm
KGA-CGM : A woman sitting on a couch with a remote

Unseen Object: Microwave

Predicted Entity-Labels (Top-3):Refrigerator,Oven,Microwave_Oven
Base: A wooden table with a refrigerator and a brown cabinet
NOC: A kitchen with a refrigerator, refrigerator, and refrigerator.
KGA-CGM: A kitchen with a microwave, oven and a refrigerator

Unseen Object: Pizza

Predicted Entity-Labels (Top-3): Pizza,Restaurant,Hat
Base: A man is making a sandwich in a restaurant
NOC: A man standing next to a table with a pizza in front of it.
KGA-CGM: A man is holding a pizza in his hands

Unseen Object: Racket

Predicted Entity-Labels (Top-3):Tennis, Racket_(sports_equipment), Court
Base: A tennis player getting ready to serve the ball
NOC: A woman court holding a tennis racket on a court.
KGA-CGM: A woman playing tennis on a tennis court with a racket.

Unseen Object: Suitcase

Predicted Entity-Labels (Top-3): Cat,Baggage,Black_Cat
Base: A cat laying on top of a pile of books
NOC: A cat laying on a suitcase on a bed
KGA-CGM: A cat laying inside of a suitcase on a bed

Unseen Object: Zebra

Predicted Entity-Labels (Top-3):Zebra,Enclosure,Zoo
Base: A couple of animals that are standing in a field
NOC: Zebras standing together in a field with zebras
KGA-CGM: A group of zebras standing in a line

Figure 29: Sample predictions of KGA-CGM on out-of-domain MSCOCO Images with Beam Size 1 when compared against base model and NOC [257]

MSCOCO training set. Furthermore, the KGA-CGM model is rebuilt for each one of them separately. To describe ImageNet images, image classifiers presented in the Section 7.4.1.2 are leveraged. Table 29 summarizes the experimental results attained on 634 categories (i.e. not all 642) to have fair comparison with other approaches. By adopting only MSCOCO training data for language model, our KGA-CGM makes the relative improvement over NOC and LSTM-C in all categories i.e. unseen, F1 and accuracy. Figure 30 shows few sample descriptions.

Model	Unpaired Text	Unseen	F1	Accuracy
NOC [257]	MSCOCO	69.1	15.6	10.0
	BNC&Wiki	87.7	31.2	22.0
LSTM-C [289]	MSCOCO	72.1	16.4	11.8
	BNC&Wiki	89.1	33.6	31.1
KGA-CGM	MSCOCO	74.1	17.4	12.2
	BNC&Wiki	90.2	34.4	33.1
	BNC&Wiki&SBU1M	**90.8**	**35.8**	**34.2**

Table 29: Describing ImageNet Images with Beam size 1. Results of NOC and LSTM-C (with Glove) are adopted from Yao et al. [289]

7.4.2.7 *KGA-CGM More Qualitative Results*

Earlier, we presented caption generation qualitative results only with beam size 1. In this section, more results of unseen/novel MSCOCO objects is presented in Table 30 and Table 31 with both beam 1 and > 1. Also Table 32 and Table 33 demonstrates some failure instances of generated captions which lack either semantics, grammar or unseen objects.

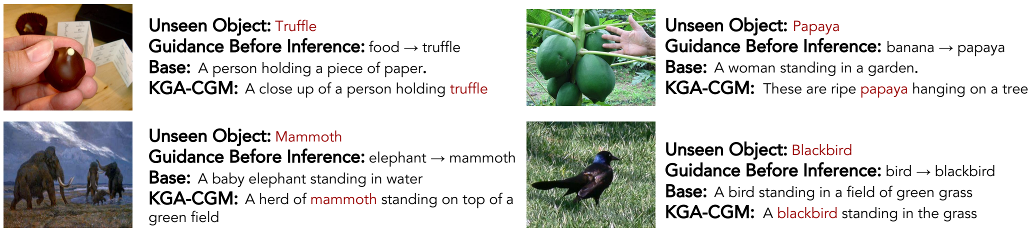

Unseen Object: Truffle
Guidance Before Inference: food → truffle
Base: A person holding a piece of paper.
KGA-CGM: A close up of a person holding truffle
Unseen Object: Papaya
Guidance Before Inference: banana → papaya
Base: A woman standing in a garden.
KGA-CGM: These are ripe papaya hanging on a tree
Unseen Object: Mammoth
Guidance Before Inference: elephant → mammoth
Base: A baby elephant standing in water
KGA-CGM: A herd of mammoth standing on top of a green field
Unseen Object: Blackbird
Guidance Before Inference: bird → blackbird
Base: A bird standing in a field of green grass
KGA-CGM: A blackbird standing in the grass

Figure 30: ImageNet images with best KGA-CGM model from Table 29. Guided before inference shows which words are used for transfer between seen and unseen.

7.4.3 *Evaluation Results Analyses*

The critical observations of our research are:

① The ablation study conducted to understand the influence of different components in KGA-CGM has shown that using external semantic attention and constrained inference has superior performance when compared to using only either of them. Also, increasing the beam size during inference has shown a drop in all measures. It primarily adheres to the influence of multiple words on unseen objects.

② Observations show that the performance advantage becomes more explicit if the domain of unseen objects is broadened. In other words: KGA-CGM improves explicitly over state of the art in settings that are larger and less controlled. At this moment, KGA-CGM scales to one order of magnitude more unseen objects with moderate performance decreases.

③ The influence of the closest seen words (i.e., observed in image-caption pairs) and the unseen visual object categories played a prominent role in generating descriptions. For example in out-of-domain MSCOCO, words such as "suitcase"/"bag", "bottle"/"glass" and "bus/truck" are semantically similar and are also used similarly in a sentence added excellent value. However, some words usually cooccur such as "racket"/"court" and "pizza"/"plate" played different roles in sentences and led to few grammatical errors.

④ The decrease in performance has a high correlation with the discrepancy between the domain where seen and unseen objects come.

7.5 SUMMARY

In this chapter, we addressed the fourth research question:

✍ **Research Question 4**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist the *generation* of text from an image if there are missing views?

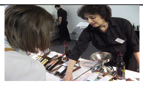
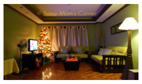
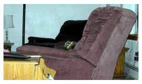
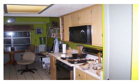
MSCOCO Unseen Object	Images and Predicted Captions		
bottle	 **Beam(1):** A bottle of wine sitting on a table next to a bottle of wine **Beam(>1):** A bottle of wine sitting on top of a table	 **Beam(1):** A woman is sitting at a table with a bottle of wine **Beam(>1):** A woman sitting at a table with a bottle of wine	 **Beam(1):** A bottle of beer and a beer are sitting on a counter **Beam(>1):** A bottle of beer next to a bottle of beer
bus	 **Beam(1):** A white bus is parked in a lot **Beam(>1):** A white buses parked in a parking lot	 **Beam(1):** A large bus is parked on the side of the street **Beam(>1):** A large bus is parked on the street	 **Beam(1):** A red bus driving down a street next to a building **Beam(>1):** A red bus driving down a street next to buildings
couch	 **Beam(1):** A room with a tv and a couch **Beam(>1):** A living room with a tv and a couch	 **Beam(1):** A cat sitting on a couch in a room **Beam(>1):** A cat sitting on a couch in a room	 **Beam(1):** A man sitting on a couch using a laptop computer **Beam(>1):** A man sitting on a couch using a laptop
microwave	 **Beam(1):** A kitchen with a microwave oven and a microwave **Beam(>1):** A kitchen with a microwave and a refrigerator	 **Beam(1):** A kitchen with a microwave oven and a black microwave **Beam(>1):** A kitchen with a microwave oven and a black microwave	 **Beam(1):** A kitchen with a microwave oven and a sink **Beam(>1):** A kitchen with a microwave oven and a sink

Table 30: Positive predictions of KGA-CGM

For this, we validated Hypothesis 4 by proposing an approach to generate captions for images that lack parallel captions during training with the assistance of knowledge encapsulated in KGs.

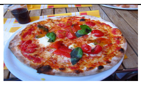
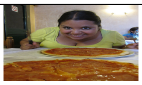
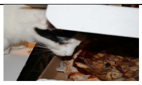
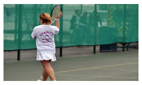
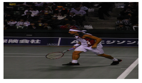
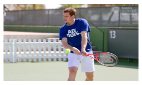
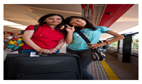
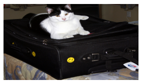
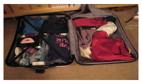
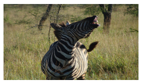
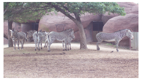
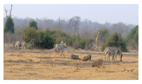
MSCOCO Unseen Object	Images and Predicted Captions		
pizza	 **Beam(1):** A pizza covered in cheese and tomatoes on top of a table **Beam(>1):** A close up of a pizza on a table	 **Beam(1):** A woman sitting at a table with a pizza in front of her **Beam(>1):** A woman sitting at a table in front of a pizza	 **Beam(1):** A cat is sitting on a white and black pizza **Beam(>1):** A cat sitting on top of a white pizza
racket	 **Beam(1):** A woman is playing tennis on a court with a racket **Beam(>1):** A woman is playing tennis with a racket	 **Beam(1):** A man playing tennis on a tennis court with a racket **Beam(>1):** A man playing tennis with a tennis rackets	 **Beam(1):** A tennis player is hitting the ball on the court with a racket **Beam(>1):** A tennis player hitting a tennis ball with a rackets
suitcase	 **Beam(1):** A woman holding a luggage **Beam(>1):** A woman holding a luggage	 **Beam(1):** A black cat laying on top of a suitcase **Beam(>1):** A black and white cat laying on top of a luggage	 **Beam(1):** A suitcase **Beam(>1):** A luggage and bags
zebra	 **Beam(1):** A zebra standing in a field of grass **Beam(>1):** A zebra standing in a field of grass	 **Beam(1):** A group of zebras standing in front of a wall **Beam(>1):** A group of zebras standing in front of a building	 **Beam(1):** A herd of zebra walking across a dirt field **Beam(>1):** A herd of zebra walking across a field

Table 31: More positive predictions of KGA-CGM

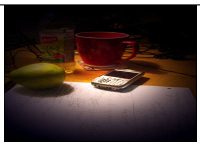
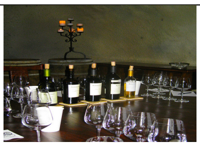
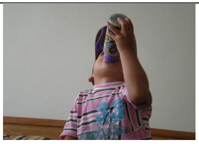
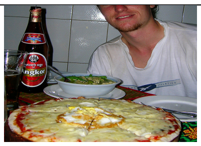
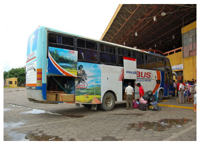
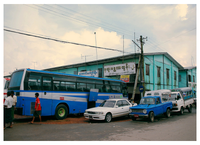
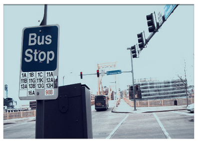
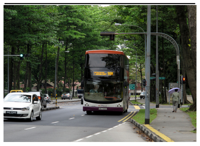
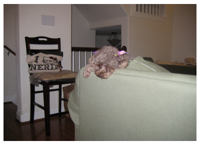
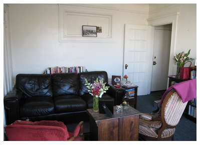
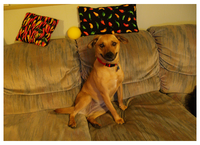
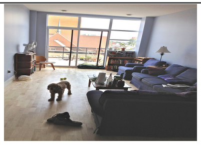
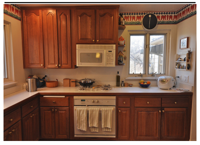
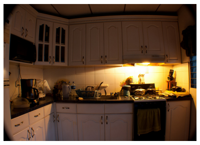
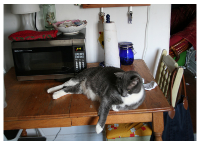
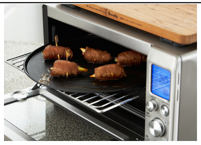
Hallucination	Grammar	No object	Semantics
			
Beam(1): a table with a laptop and a bottle of water	**Beam(1):** A table with many bottles of wine bottles	**Beam(1):** A child is laying down on a bed	**Beam(1):** A person is holding a pizza in a bottle
			
Beam(1): A food bus parked in front of a building	**Beam(1):** A blue and white buses parked in front of a blue building	**Beam(1):** A street sign that is on a pole	**Beam(1):** A bus driving down a street with cars driving down it
			
Beam(1): A cat sitting on a couch next to a person	**Beam(1):** A room with a bed couch and a couch	**Beam(1):** A dog laying on a bed	**Beam(1):** A dog is sitting in the living room couch
			
Beam(1): A kitchen with a sink and a microwave	**Beam(1):** A kitchen with a microwave and a microwave	**Beam(1):** A cat is standing on a table in a kitchen	**Beam(1):** A pan of food that is on a microwave

Table 32: Failure instances of our best KGA-CGM Model (Beam(1)) with failures underlined of objects Bottle, Bus, Couch and Microwave

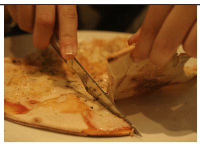
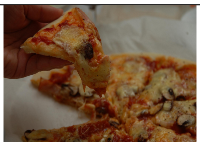
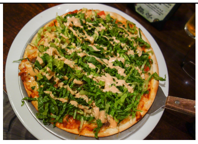
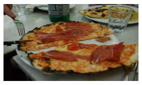
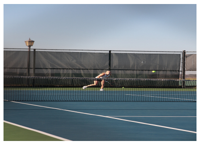
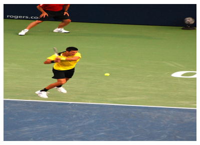
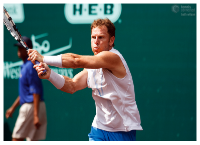
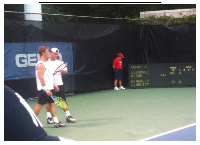
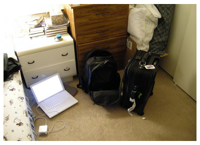
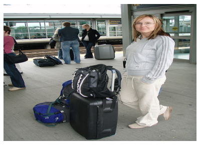
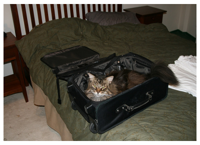
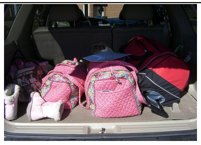
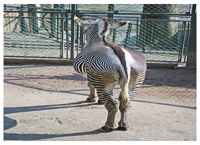
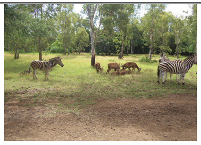
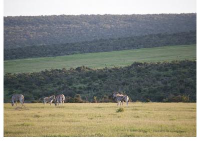
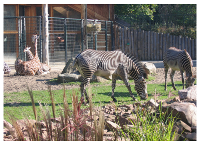
Hallucination	Grammar	No object	Semantics
			
Beam(1): A close up of a pizza on a pizza	**Beam(1):** A close up of a pizza on a pizza covered in cheese	**Beam(1):** A salad and a salad on a white plate	**Beam(1):** A pizza with meat and cheese on a pizza
			
Beam(1): A man is standing on a tennis court with a racket	**Beam(1):** A man playing tennis on a court with a racket with a crowd	**Beam(1):** A man is playing tennis	**Beam(1):** A man standing on a tennis court with a racket holding a tennis
			
Beam(1): A cat laying on a bed next to a luggage	**Beam(1):** a woman standing next to a man in a suit and a luggage	**Beam(1):** A cat laying on top of a bed	**Beam(1):** A woman laying in a pink suitcase with a suitcase
			
Beam(1): Two zebras stand together in the dirt near a fence	**Beam(1):** Two zebras stand in the water near some water	**Beam(1):** A couple of animals that are standing in the grass	**Beam(1):** Two zebras are standing in a fenced in area

Table 33: Failure instances of our best KGA-CGM Model (Beam(1)) with failures underlined of objects Pizza, Racket, Suitcase and Zebra

CONCLUSION

8

CONCLUSION

8.1 SUMMARY

In this thesis, we addressed the following research question:

✍ **Overall Research Question**

How to unify subset of text, Entity Relationship graph, and image modality representing languages, relational knowledge, and vision respectively into a shared representation to assist homogeneous or heterogeneous content search, categorization, and generation.

We identified content characteristics, which are crucial for the above research question: different modalities, different languages, parallel and non-parallel.

Based on these content characteristics, we split the overall research question into four subquestions, which we addressed in Chapter 4, Chapter 5,Chapter 6 and Chapter 7.

✍ **Research Question 1**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist *search* by finding correlation among their input representations?

In Chapter 4, we aimed at Research Question 1 – targeting an approach for learning shallow common space representation of two different heterogeneous sources mainly text and images by leveraging correlation principle of multi-view representation learning. From the data perspective it satisfies two heterogeneous content characteristics, mainly characteristic 1 and characteristic 3, see Figure 10).

For this, we provided Contribution 1.

☞ **Contribution for Research Question 1**

Cross-modal retrieval to assist content search by leveraging correlated centroid space.

Our unsupervised cross-modal retrieval approach, correlated centroid space (C^{2}SUR) builds on KCCA to effectively capture correlation among the heterogeneous data. In particular, C^{2}SUR can effectively discriminate the cross-modal data such that similar items are ranked closer while pushing away the dissimilar items. Also, its implementation is more straightforward and computationally efficient than other methods.

Moreover, C^{2}SUR has been proven to be useful if the text emerges from different languages for cross-modal retrieval. It is a significant advantage about the growth of non-English content and their need of applications.

✍ **Research Question 2**

Given two different views of homogeneous content depicting text from different languages, how can we build a shared representation to assist *categorization* by learning a common space by capturing regularities?

In Chapter 5, we proposed a shallow neural network approach combined with manifold alignment techniques for the above Research Question 2. More specifically, we provided a tailored solution for tackling different cross-language data which is either parallel or non-parallel (Characteristic 2, Characteristic 3, and Characteristic 4, see Figure 10). Employing such a approach, we build bilingual word embeddings for supporting cross-lingual task such as cross-language text classification.

☞ **Contribution for Research Question 2**

Cross-language text classification to assist content categorization by leveraging Bilingual Paragraph Vectors.

We proposed the BRAVE approach for Research Question 2. Here, we extended technique called Paragraph Vectors to compactly model textual content with multiple views emerging from different languages. In contrast to previous works, our BRAVE approach was explored for both parallel and non-parallel cross-language content.

Moreover, for the non-parallel content, one of the manifold alignment technique called Procrustes analysis was leveraged to create pseudo-parallel content. Furthermore, both parallel and pseudo-parallel content is used to build bilingual word embeddings by capturing regularities across languages. Uniformly capturing these inter-language dependencies was shown to be essential for cross-language text classification.

✍ **Research Question 3**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation of all views if an auxiliary view depicting text in multiple languages is added to assist the *generation* of text from an image?

In Chapter 6, we targeted an approach for multi-language image caption generation using parallel image and language content (Characteristic 1 and Characteristic-3 in Figure 10). This way, a single caption model could be built for different languages and an image.

☞ **Contribution for Research Question 3**

Consistent multi-language image caption generation to assist content generation given auxiliary views by leveraging multi-task attention.

Concerning Research Question 3, we proposed a multi-task attention model by leveraging deep neural network architectures and multi-task learning. We learned a common space representation of all views emerging from both homogeneous and heterogeneous data to generate one modality from another, i.e., especially generating text from an image.

In particular, multi-task learning was used to share knowledge across languages, such that one language guides another when building a caption model. Shared layer in CNN-LSTM architecture having a multi-task loss is designed to achieve sharing. Furthermore, LSTM is extended with combination LSTM to capture information from the shared layer.

✍ **Research Question 4**

Given two different views of heterogeneous content depicting text and image modality, how can we build a shared representation to assist the *generation* of text from an image if there are missing views?

Last, we addressed Research Question 4 in Chapter 7. Here, we are concerned with generating captions for those images which contain unseen visual object categories. More specifically, images which are observed in the testing phase contain visual object categories that are unseen during training. For this, we exploited non-parallel knowledge graph entity data and the image-caption parallel data along with unpaired textual resource data (Characteristic 1, Characteristic 3, and Characteristic-4 in Figure 10).

☞ **Contribution for Research Question 4**

Unseen visual object categories caption generation to assist content generation given missing views by leveraging knowledge guided assistance.

We introduced a novel knowledge guided assistance approach for the above Research Question 4. Within our approach, conceptual knowledge provided by a knowledge graph is utilized as external semantic attention throughout training and also to aid as a dynamic constraint before and during inference. Hence, it augments an auxiliary view as done in multi-view learning scenarios.

Mainly, this explicit knowledge graph grounding of entities observed in the visual content and usage of knowledge graph entity embeddings and labels has shown to be useful for image caption generation where image contain unseen object categories. The contribution of this work on a broader scope is also seen as a sign of progress towards the integration of the visual (and textual) information available on the Web with KGs.

8.2 future work

In the following, we will briefly outline relevant future work concerning our overall research question.

✲ Future Work – FW1

Joint correlation analysis of different languages along with an image for multilingual cross-modal retrieval.

We proposed a MVRL approach for performing cross-modal retrieval using shallow representations of language content from multiple languages and an image separately in the Chapter 4. Intuitively speaking, we built separate models for each language.

However, for adequate representation, jointly exploiting correlation across languages and an image is necessary. It is achieved by building a shared representation of multiple views with joint analysis.

It can also offer other intriguing possibilities:

① For example, it can support machine translation by translating a word/phrase never seen in the parallel data by seeking help from an image, provided that the representations be learned from both language corpora and limited image-text parallel corpora.

② It can provide a possibility to create multilingual multimodal embeddings for supporting tasks beyond cross-modal retrieval.

In our future work, we concentrate on expanding correlated centroid space approach to cater more than two views at once. Also, instead of utilizing shallow image and language representations, we build on the work done in Chapter 6 and Chapter 7 and leverage deep representations. It will be an exciting direction to explore where we evaluate and see how deep representations from heterogeneous sources correlate and contribute.

✲ Future Work – FW2

Extension of bilingual to multilingual embeddings.

In the Chapter 5, we proposed a MVRL approach based on neural network and manifold alignment technique for performing cross-language text classification using language content from multiple languages. Intuitively speaking, we built an approach which can leverage only two languages at once.

However, for adequate representation, jointly exploiting regularities across many languages is necessary. It is achieved by building a shared multilingual representation of views emerging from multiple languages with joint analysis.

It can also offer other intriguing possibilities:

① Multilingual embeddings for words, entities, and concepts built by combining many languages into a shared space representation can support natural language processing applications.

② Massive monolingual corpora can be leveraged easily along with limited parallel corpora which usually exists pairwise. It will help to build a single model for transfer learning, in which model can be fine-tuned per language.

In our future work, we concentrate on expanding our approach which currently leverages only two views to more with a simple extension such as summing up individual bilingual objectives.

✲ Future Work – FW3

Improvement of image caption generation in multiple languages by incorporating annotators translation preferences.

In the Chapter 6, we proposed a MVRL approach based on deep neural networks for generation of descriptions (i.e., caption) for an image in multiple languages such that they are consistent across languages. At present, our work leverage only parallel data without understanding the intricacies involved in the creation of such data.

However, for an adequate language generation, the background preferences of annotators who may belong to diverse communities and the choices they make in the translation of captions from one language to other has to be taken into consideration. Also, choice of vocabulary they use in the creation of caption in multiple languages.

It can also offer other intriguing possibilities:

① Based on personal preferences of the audience, a personalized image caption generation can be attained for each language.

② Analysis of the independently collected annotations for the image in a new language, when compared against the English translation can provide new insights about crowdsourcing. It will explore different cultures and shared bodies of knowledge among them.

In our future work, we concentrate on understanding the shared knowledge observed in the form of concrete entities and objects across different language descriptions when building image caption generation models.

✲ Future Work – FW4

Extension of caption generation to images found in the wild.

In the Chapter 7, we proposed a MVRL approach based on deep neural networks for generation of descriptions (i.e. caption) for images containing novel objects. Currently, we built an approach which can only comprehend limited visual objects that are observed in hand curated datasets (e.g., MSCOCO and ImageNet).

However, to assist in generating descriptions for images in large-scale, we need robust generative models which can detect visual objects in the wild.

It can also offer other intriguing possibilities:

① Large-scale object recognition has been a long-standing goal of computer vision and related fields. With advancements of multimodal language processing, generation of captions for images at large-scale is also sought-after by leveraging external knowledge.

② Machine learning models usually fail to predict for those examples it has never seen before. Zero-shot or one-shot prediction of visual objects in images is of interest to improve the caption models so that they can work with less image-caption parallel data.

In our future work, we concentrate on leveraging external knowledge for improving caption generation models especially for those visual objects which lack clean and annotated training data.

REFERENCES

[1] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

[2] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. A neural knowledge language model. *arXiv preprint arXiv:1608.00318*, 2016.

[3] Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

[4] Rami Al-Rfou, Perozzi Bryan, and Skiena Steven. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of CoNLL*, pages 183–192. ACL, 2013.

[5] James Allen. *Natural language understanding*. Pearson, 1995.

[6] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[7] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[9] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[10] Cédric Archambeau, Nicolas Delannay, and Michel Verleysen. Robust probabilistic projections. In *Proceedings of the 23rd International conference on machine learning*, pages 33–40. ACM, 2006.

[11] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[13] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Towards nlg for physiological data monitoringwith body area networks. In *14th European Workshop on Natural Language Generation, Sofia, Bulgaria, August 8-9, 2013*, pages 193–197, 2013.

[14] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics, 2005.

[15] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of chemometrics*, 17(3):166–173, 2003.

[16] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.

[17] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721, 2010.

[18] Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[19] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161, 1992.

[20] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[21] Yoshua Bengio, Paolo Frasconi, and Patrice Simard. The problem of learning long-term dependencies in recurrent networks. In *Neural Networks, 1993., IEEE International Conference on*, pages 1183–1188. IEEE, 1993.

[22] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3 (Feb):1137–1155, 2003.

[23] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[24] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

[25] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[26] Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In *The 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

[27] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[28] D. M. Blei. Probabilistic topic models. *Communications of the ACM.*, 55(4): 77–84, 2012.

[29] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

[30] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[31] Thierry Bouwmans. Subspace learning for background modeling: A survey. *Recent Patents on Computer Science*, 2(3):223–234, 2009.

[32] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2): 79–85, 1990.

[33] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

[34] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in Computer Vision*, pages 671–679. Elsevier, 1987.

[35] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 911–920. ACM, 2008.

[36] Iacer Calixto, Qun Liu, and Nick Campbell. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*, 2017.

[37] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[38] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[39] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[40] Mickaël Chen and Ludovic Denoyer. Multi-view generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 175–188. Springer, 2017.

[41] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[42] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.

[43] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[44] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[45] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220, 2010.

[46] Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140. Oxford, 2008.

[47] A Cochocki and Rolf Unbehauen. *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc., 1993.

[48] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *ACL-EMNLP.*, pages 1–8. 2002.

[49] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[50] J. Coulmance, J. M. Marty, G. Wenzek, and A. Benhalloum. Trans-gram, fast cross-lingual word-embeddings reyes- mannde= reginait- femmefr. In *EMNLP.* 2015.

[51] D Alan Cruse. *Lexical semantics*. Cambridge University Press, 1986.

[52] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[53] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017.

[54] Tijl De Bie and Bart De Moor. On the regularization of canonical correlation analysis. *Int. Sympos. ICA and BSS*, pages 785–790, 2003.

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[56] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1731–1740. ACM, 2015.

[57] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[58] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. Tweet2vec: Character-based distributed representations for social media. *arXiv preprint arXiv:1605.03481*, 2016.

[59] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[60] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[61] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[62] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732, 2015.

[63] Cicero dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

[64] S. T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval.* 1997.

[65] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*, 2016.

[66] Desmond Elliott, Stella Frank, and Eva Hasler. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709*, 2015.

[67] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[68] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*, 2017.

[69] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[70] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pages 153–160, 2009.

[71] Yu M Ermoliev and RJ-B Wets. *Numerical techniques for stochastic optimization*. Springer-Verlag, 1988.

[72] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[73] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *ACL*. 2014.

[74] Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[75] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.

[76] Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. A survey of current datasets for vision and language research. *arXiv preprint arXiv:1506.06833*, 2015.

[77] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016.

[78] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *The Journal of Machine Learning Research.*, 37(3):277–296, 1999.

[79] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[80] K. Fukumasu, Koji Eguchi, and Eric P. Xing. Symmetric correspondence topic models for multilingual text analysis. In *NIPS*, pages 1295–1303. 2012.

[81] Kenji Fukumizu, Francis R Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383, 2007.

[82] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[83] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*, 2017.

[84] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[85] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[86] Yuyun Gong and Qi Zhang. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, pages 2782–2788, 2016.

[87] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[88] Geoffrey J Gordon. Generalized^2 linear^2 models. In *Advances in neural information processing systems*, pages 593–600, 2003.

[89] S. Gouws and A. SÃžgaard. Simple task-specific bilingual word embeddings. In *NAACL-HLT.*, pages 1386–1390. 2015.

[90] S. Gouws, Y. Bengio, and G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML.* 2015.

[91] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[92] Stephen Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks*, 1(1):17–61, 1988.

[93] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 5, page 10, 2006.

[94] David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[95] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[96] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[97] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[98] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[99] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in neural information processing systems*, pages 153–160, 2004.

[100] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.

[101] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[102] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, pages 1–10, 2016.

[103] K. M. Hermann and P. Blunsom. Multilingual models for compositional distributed semantics. In *ACL*. 2014.

[104] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[105] Geoffrey E Hinton and Ruslan R Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614, 2009.

[106] Geoffrey E Hinton, James L McClelland, David E Rumelhart, et al. Distributed representations. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(3):77–109, 1986.

[107] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[108] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[109] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[110] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28 (3/4):321–377, 1936.

[111] William W Hsieh. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13(10):1095–1105, 2000.

[112] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[113] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.

[114] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.

[115] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

[116] Xiaoming Huo, Xuelei Ni, and Andrew K Smith. A survey of manifold-based learning methods. *Recent advances in data mining of enterprise data*, pages 691–745, 2007.

[117] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *What is independent component analysis?* Wiley Online Library, 2001.

[118] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009.

[119] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415, 2015.

[120] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems*, pages 982–990, 2010.

[121] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407–2414. IEEE, 2011.

[122] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*, 2015.

[123] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[124] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[125] Michael Irwin Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

[126] Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, 2017.

[127] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

[128] Juha Karhunen, Tapani Raiko, and KyungHyun Cho. Unsupervised deep learning: A short review. In *Advances in Independent Component Analysis and Learning Machines*, pages 125–142. Elsevier, 2015.

[129] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[130] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[131] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[132] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[133] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603, 2014.

[134] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*. 2012.

[135] T. Kočiský, K. M. Hermann, and P. Blunsom. Learning bilingual word representations by marginalizing alignments. In *ACL*, pages 224–229. 2014.

[136] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[137] V. Kostrykin, K. Makarov, and A. Motovilov. On a subspace perturbation problem. *American Mathematical Society.*, pages 3469–3476, 2003.

[138] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[139] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[140] Pei Ling Lai, Colin Fyfe, et al. Canonical correlation analysis using artificial neural networks. In *ESANN*, pages 363–368. Citeseer, 1998.

[141] Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319, 2017.

[142] Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[143] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pages 329–336, 2004.

[144] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196. 2014.

[145] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4):541–551, 1989.

[146] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[147] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[148] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, pages 143–155, 1989.

[149] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[150] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2015.

[151] O. Levy and Y. Goldberg. Dependency based word embeddings. In *ACL.*, pages 302–308. 2014.

[152] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research.*, 5:361–397, 2004.

[153] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 604–611. ACM, 2003.

[154] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

[155] Stan Z Li. Markov random field models in computer vision. In *European conference on computer vision*, pages 361–370. Springer, 1994.

[156] Yingming Li, Ming Yang, and Zhongfei Zhang. Multi-view representation learning: A survey from shallow methods to deep methods. *arXiv preprint arXiv:1610.01206*, 2016.

[157] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[158] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[159] Weifeng Liu, Dacheng Tao, Jun Cheng, and Yuanyan Tang. Multiview hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding*, 118:50–60, 2014.

[160] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[161] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[162] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep multilingual correlation for improved word embeddings. In *NAACL-HLT*. 2015.

[163] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016.

[164] Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

[165] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[166] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[167] James Martens. Deep learning via hessian-free optimization. In *ICML*, volume 27, pages 735–742, 2010.

[168] Giovanni Da San Martino, Salvatore Romeo, Alberto Barrón-Cedeno, Shafiq Joty, Lluis Marquez, Alessandro Moschitti, and Preslav Nakov. Cross-language question re-ranking. *arXiv preprint arXiv:1710.01487*, 2017.

[169] Anthony M McEnery and Anita Wilson. *Corpus linguistics: an introduction.* Edinburgh University Press, 2001.

[170] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. Cross-lingual mixture model for sentiment classification. In *ACL.*, pages 572–581. 2012.

[171] Florian Metze, Duo Ding, Ehsan Younessian, and Alexander Hauptmann. Beyond audio and video retrieval: topic-oriented multimedia summarization. *International Journal of Multimedia Information Retrieval*, 2(2):131–144, 2013.

[172] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[173] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781.* 2013.

[174] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. In *arXiv preprint arXiv:1309.4168.* 2013.

[175] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[176] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of EMNLP*, pages 880–889. ACL, 2009.

[177] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3040–3047. IEEE, 2013.

[178] Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval (to appear). Google Scholar*, 2017.

[179] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In *NIPS.*, pages 1081–1088. 2009.

[180] Rémi Monasson and Dominic O'Kane. Domains of solutions and replica symmetry breaking in multilayer neural networks. *EPL (Europhysics Letters)*, 27(2):85, 1994.

[181] Sean Moran and Victor Lavrenko. Sparse kernel learning for image annotation. In *Proceedings of International Conference on Multimedia Retrieval*, page 113. ACM, 2014.

[182] K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181–201, 2001.

[183] Kevin P Murphy. Directed graphical models. 2006.

[184] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.

[185] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[186] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[187] Joakim Nivre and Mario Scholz. Deterministic dependency parsing of english text. In *Proceedings of the 20th international conference on Computational Linguistics*, page 64. Association for Computational Linguistics, 2004.

[188] Douglas W Oard and Anne R Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33: 223–56, 1998.

[189] Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2011.

[190] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011.

[191] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. *arXiv preprint arXiv:1704.06485*, 2017.

[192] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

[193] Judea Pearl. Bayesian networks. 2011.

[194] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*, 2016.

[195] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[196] Carol Peters, Martin Braschler, and Paul Clough. Cross-language information retrieval. In *Multilingual Information Retrieval*, pages 57–84. Springer, 2012.

[197] H. Pham, M. T. Luong, and C. D. Manning. Learning distributed representations for multilingual text sequences. In *NAACL-HLT*, pages 88–94. 2015.

[198] Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L Leidner, Dezhao Song, and Frank Schilder. Interacting with financial data using natural language. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1121–1124. ACM, 2016.

[199] Christopher Poultney, Sumit Chopra, Yann L Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2007.

[200] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *ACL.*, pages 1118–1127. 2010.

[201] Duangmanee Putthividhy, Hagai T Attias, and Srikantan S Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415. IEEE, 2010.

[202] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[203] Dimitrios Rafailidis, Stavroula Manolopoulou, and Petros Daras. A unified framework for multimodal retrieval. *Pattern Recognition*, 46(12):3358–3370, 2013.

[204] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. In *arXiv preprint arXiv:1510.03519*. 2015.

[205] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[206] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 823–831, 2014.

[207] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[208] Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

[209] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[210] Steve Renals, Nelson Morgan, Hervé Bourlard, Michael Cohen, and Horacio Franco. Connectionist probability estimators in hmm speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):161–174, 1994.

[211] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[212] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[213] Robert Harry Riffenburgh. *Linear discriminant analysis*. PhD thesis, Virginia Polytechnic Institute, 1957.

[214] Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer, 2016.

[215] Christian Robert. Machine learning, a probabilistic perspective, 2014.

[216] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961.

[217] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34–51. Springer, 2006.

[218] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[219] Sebastian Ruder. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*, 2017.

[220] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[221] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010.

[222] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[223] Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 693–700, 2010.

[224] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[225] A. P. Sarath Chandar, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. In *NIPS.*, pages 1853–1861. 2014.

[226] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.

[227] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[228] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[229] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika.*, 31(1):1–10, 1966.

[230] Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016.

[231] Azadeh Shakery and ChengXiang Zhai. Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. *Information retrieval*, 16(1):1–29, 2013.

[232] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012.

[233] John Shawe-Taylor and S Sun. Kernel methods and support vector machines. *Lecture Notes*, 2009.

[234] Cencheng Shen, Ming Sun, Minh Tang, and Carey E Priebe. Generalized canonical correlation analysis for classification. *Journal of Multivariate Analysis*, 130:310–322, 2014.

[235] Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. Neural cross-lingual entity linking. *arXiv preprint arXiv:1712.01813*, 2017.

[236] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[237] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658. ACM, 2008.

[238] Jonathan Slocum. A survey of machine translation: its history, current status, and future prospects. *Computational linguistics*, 11(1):1–17, 1985.

[239] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP*, pages 1201–1211. 2012.

[240] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.

[241] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014.

[242] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553, 2016.

[243] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[244] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[245] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[246] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[247] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[248] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[249] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[250] X. Tang and X. Wan. Learning bilingual embedding model for cross-language sentiment classification. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences.*, volume 2, pages 134–141. 2014.

[251] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500): 2319–2323, 2000.

[252] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257, 2011.

[253] Volker Tresp. Mixtures of gaussian processes. In *Advances in neural information processing systems*, pages 654–660, 2001.

[254] Thomas R Truscott. *Techniques used in minimax game-playing programs*. PhD thesis, Duke University, 1981.

[255] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[256] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[257] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

[258] Jaakko Viinikanoja, Arto Klami, and Samuel Kaski. Variational bayesian mixture of robust cca models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 370–385. Springer, 2010.

[259] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[260] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[261] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS.*, pages 1497–1504. 2003.

[262] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[263] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017.

[264] Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.

[265] Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[266] I. Vulić and M. F. Moens. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *EMNLP*. 2014.

[267] I. Vulić and M. F. Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL*. 2015.

[268] Ivan Vulić and Marie-Francine Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372. ACM, 2015.

[269] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of ICML*, pages 1120–1127. ACM, 2008.

[270] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 988–997. ACM, 2016.

[271] Chong Wang. Variational bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

[272] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2216–2223. IEEE, 2012.

[273] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[274] Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 688–695. IEEE, 2015.

[275] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in neural information processing systems*, pages 1481–1488, 2005.

[276] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[277] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[278] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[279] Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 168–177. ACM, 2007.

[280] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[281] Yan Xia, Xudong Cao, Fang Wen, and Jian Sun. Well begun is half done: Generating high-quality seeds for automatic image dataset construction from web. In *European Conference on Computer Vision*, pages 387–400. Springer, 2014.

[282] M. Xiao and Y. Guo. Semi-supervised representation learning for cross-lingual text classification. In *EMNLP.*, pages 1465–1475. 2013.

[283] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.

[284] Zhiyong Xiao. Non-negative matrix factorization with local preservation for hyperspectral image dimensionality reduction. *Remote sensing letters*, 5 (9):793–802, 2014.

[285] Eric P Xing, Rong Yan, and Alexander G Hauptmann. Mining associated text and images with dual-wing harmoniums. *arXiv preprint arXiv:1207.1423*, 2012.

[286] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[287] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.

[288] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pages 2361–2369, 2016.

[289] Ting Yao, Pan Yingwei, Li Yehao, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

[290] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017.

[291] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.

[292] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[293] Dong Yu, Li Deng, and Shizhen Wang. Learning in the deep-structured conditional random fields. In *Proc. NIPS Workshop*, pages 1–8, 2009.

[294] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In *International Conference on Multimedia Modeling*, pages 312–322. Springer, 2012.

[295] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2014.

[296] D. Zhang, Qiaozhu Mei, and ChengXiang Zhai. Cross-lingual latent topic extraction. In *ACL*, pages 1128–1137. 2010.

[297] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[298] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.

[299] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[300] K. Zhao, H. Hassan, and M. Auli. Learning translation models from monolingual continuous representations. In *NAACL-HLT*. 2015.

[301] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: building a web-scale landmark recognition engine. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, pages 1085–1092. IEEE, 2009.

[302] Guoqiang Zhong, Wu-Jun Li, Dit-Yan Yeung, Xinwen Hou, Cheng-Lin Liu, et al. Gaussian process latent random field. In *AAAI*, 2010.

[303] Xiao Zhong and David Enke. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, 67:126–139, 2017.

[304] G. Zhou, T. He, J. Zhao, and W. Wu. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *IJCAI*. 2015.

[305] H. Zhou, L. Chen, F. Shi, and D. Huang. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *ACL*, pages 430–440. 2015.

[306] Yueting Zhuang, Yanfei Wang, Fei Wu, Yin Zhang, and Weiming Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.

[307] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *The Journal of the Royal Statistical Society: Series B (Statistical Methodology).*, 67(2):301–320, 2005.

[308] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP.*, pages 1393–1398. 2013.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

IR	Information Retrieval
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
CV	Computer Vision
MT	Machine Translation
LM	Language Model
PGM	Probabilistic Graphical Model
MVRL	Multi-view Representation Learning
CCA	Canonical Correlation Analysis
KCCA	Kernel Canonical Correlation Analysis
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-term Memory
GRU	Gated Recurrent Unit
GAN	Generative Adversarial Networks
LDA	Latent Dirichlet Allocation
RBM	Restricted Boltzmann Machine
DBM	Deep Boltzmann Machine
HMM	Hidden Markov Model
MRF	Markov Random Field
EM	Expectation Maximization
CLTC	Cross-language Text Classification

MATHEMATICAL NOTATION

In this thesis, author would like to make mathematical notations consistent. However, in some places they may look different from how they are generally used in the literature.

Parameters and Variables

Usually, lower-case Greek letters are used to denote hyper parameters and variables. For example, λ is used for the regularization constant and η is used for learning rate. Parameters of different shallow and deep architectures are defined by Θ.

Other variables like a vector used is always denoted by a bold lower-case Roman letter such as $\mathbf{x}$. While, matrix by a bold upper-case Roman letter such as $\mathbf{X}$. A component of a vector is denoted by a lower-case non-bold letter with the index of the component as a subscript. Similarly, an element of a matrix is denoted by a lower-case non-bold letter with a pair of the indices of the component as a subscript. For example, x_i and X_{ij} indicate the i-th component of $\mathbf{x}$ and the element of $\mathbf{X}$ on its i-th row and j-th column, respectively.

Subscripts and Superscripts

In data-driven learning, a set of training examples are given. Assuming the training size as N, usually each sample in the training set is denoted by its index in the superscript such that $\mathbf{x}^{(n)}$ represent the n-th training sample. Howbeit, it should be understood that the order of elements in the set can be arbitrary.

Sub- or superscripts are also used when designing different layers of deep architectures. For example, $\mathbf{h}^{(l)}$ and $\mathbf{W}^{(l)}$ respectively denote the vector of hidden units and matrix of weight parameters in the l-th layer.

Functions

All functions are denoted with a upper-case letter. Similar to the vector notation, a subscript is used to denote a component of a function such that $\mathcal{F}_i(\mathbf{x})$ is the i-th component of a function $\mathcal{F}$. For instance, in neural networks, commonly used functions for non-linear activations such as sigmoid (σ), hyperbolic tangent ($tanh$) etc., are commonly represented with either ψ, ϕ or φ. Similarly, for representing clique potentials in the graphical models ψ, ϕ or φ are again leveraged.

Data Distribution

Type of distribution that will be discussed in this thesis is mostly about data distribution. Training samples are sampled with i.i.d assumption and the heterogeneous data belongs to either text or images. The data distribution is denoted by $p(\cdot)$ and its distribution is selected based on application.

GLOSSARY

① **View**

A single view is a modality represented by either image, text, video or audio.

② **Multi-view**

Combination of different views.

③ **Representation Learning**

Representation to identify and extricate the underlying multiple explanatory factors of variation behind the content.

④ **Data instance**

A single sample from the content.

⑤ **Varied form**

Different types of modalities.

⑥ **Heterogeneous Content**

Data instances containing views from different modalities.

⑦ **Homogeneous Content**

Data instance containing views from the same modality.