A Step Towards Explainable Person Re-identification Rankings

Andreas Specker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
andreas.specker@kit.edu

Abstract

More and more video and image data is available to security authorities that can help solve crimes. Since manual analysis is time-consuming, algorithms are needed that support e.g. re-identification of persons. However, person re-identification approaches solely output image rank lists but do not provide an explanation for the results.

In this work, two concepts are proposed to explain person re-identification rankings and a qualitative evaluation is conducted. Both approaches are based on a multi-task convolutional neural network which outputs feature vectors for person re-identification and simultaneously recognizes a person's semantic attributes. Analyses of the learned weights and the outputs of the attribute classifier are used to generate the explanations.

The results of the conducted experiments indicate that both approaches are suitable to improve the comprehensibility of person re-identification rankings.

1 Introduction

The increased use of surveillance cameras to ensure security in public spaces leads to huge amounts of video data available to law enforcement agencies. On the one hand, this allows the search for specific persons of interest, but on the other hand, it raises the problem of efficient and fast evaluation of the data.

The research field of person re-identification (re-id) addresses this problem by developing approaches that enable automatic searches for persons in a huge image or video database, usually referred to as the gallery. The starting point for a search is typically a so-called query image that shows the target person.

Recent works [4, 11, 20] train a convolutional neural network (CNN) to embed person images into a feature space. This feature space has the characteristics that generated features from images showing the same person are closer together than features from images of different people. Such features, also called embeddings, are represented by vectors with a certain number of elements N. The calculation of the distances between the query embedding vector and all embeddings of images included in the gallery makes it possible to create a ranking of the gallery images sorted by their similarity to the query image.

Task-specific problems that make it difficult to train a CNN for re-id include low image resolution, occlusions, and misaligned person detections. Moreover, large differences between scientific datasets, typically used for developing and training, and real-world data lead to problems. Scientific datasets are only a small excerpt of reality restricted with respect to the variety of persons' visual appearances. They are usually recorded within a short period and at a specific location and scene setup. As a result, many important characteristics of a person's visual appearance, such as different types of clothing in summer and winter or varying lighting conditions, are not included. Hence, the learned feature space is biased and thus imperfect, especially since it is a matter of finding unseen persons in the application who may have unfamiliar characteristics.

Therefore the resulting rankings of the person pictures naturally contain false positive results in the first ranks as well. In this case, the difficulty is that these errors are not necessarily understandable. The reasons for images being ranked at their positions remain unclear.

Since the search is based ond abstract deep feature vectors, it is not possible to intuitively interpret the embeddings. Furthermore, the sole indication of the distance does not solve the problem since such values are not intuitively interpretable either.

To provide meaningful explanations along with ranking lists, re-id embeddings are first thoroughly analyzed in this work. Subsequently, two concrete approaches for explaining rankings are proposed and evaluated. The general concept behind both approaches is to leverage person attributes, such as gender or clothing colors, to add semantics to the re-id model. Thus, the meaning of the feature vectors can be understood to a certain extent via analyzing the relationship between elements of the embedding vector and the attributes.

2 Related work

This technical report is in the realm of two different research fields: person re-identification and explainable artificial intelligence (XAI). As this work can be applied to any person re-id approach, related work regarding re-id will not be discussed further in this section. According to [18], XAI methods can be categorized into three main fields: explaining of inner-workings, counterfactual explanations, and explanation of decisions.

One possible way to explain the **inner-workings** of a CNN is to determine and visualize features that maximize the activation and are thus most relevant [2, 14, 8]. Other recent works focus not on the maximization of activation but instead try to invert neural networks and retrieve explanations based on e.g. parameter gradients [5, 12, 19]. Moreover, some approaches distill the information of deep neural networks into models with better interpretability [17, 10] or aim to characterize hidden features quantitatively [1, 13].

Counterfactual explanations in the context of XAI describe what has to be changed to the feature vector in order to achieve a prediction of the desired class. For example, such explanations can have the form *"If feature value X would be Y, class C would have been predicted"*. Works that investigate counterfactual explanations to achieve more interpretable machine learning models are [21, 6, 9].

The work in this technical report best fits the research direction of **explaining decisions**. To do this, most approaches rely on the visualization of attribution maps, such as gradient or activation maps, or leverage attention modules to generate valid explanations. Commonly used methods are [16, 3, 7]. In contrast to these methods which primarily focus on the explanation of classification results, person re-id is a retrieval task and does not make any hard decision. Instead abstract feature vectors are compared. To bridge this gap, this work adds an attribute classifier in order to be able to make differences between feature vectors from hidden layers more interpretable.

3 Concepts

The main idea behind the proposed concepts is to use a pre-trained re-id network as a black box and to train an attribute classifier upon it. The attribute classifier takes re-id embeddings as input and outputs classification probabilities of the recognized attributes. The parameters of the re-id network remain frozen while only the weights of the newly added fully connected classification layer are trained. By that, the attribute classifier is forced to interpret the abstract feature vectors and to recognize the attributes based on the information contained therein. The architecture is visualized in Figure 3.1.

Of course, this training procedure does not achieve the best results in terms of attribute recognition accuracy, but it allows the interpretation of the meaning of the elements of the re-id feature vector. The learned weights of the fully connected attribute classification layer enable direct conclusions to be drawn between feature components and their meaning concerning semantic attributes. The weights are understood as a measure of the correlation between the embedding and the attributes.

3.1 Use of classifier outputs

The straightforward way to explain ranking results is to compare attribute predictions of query and gallery images. It would be possible to compute the distances between attribute predictions and to output those for each attribute

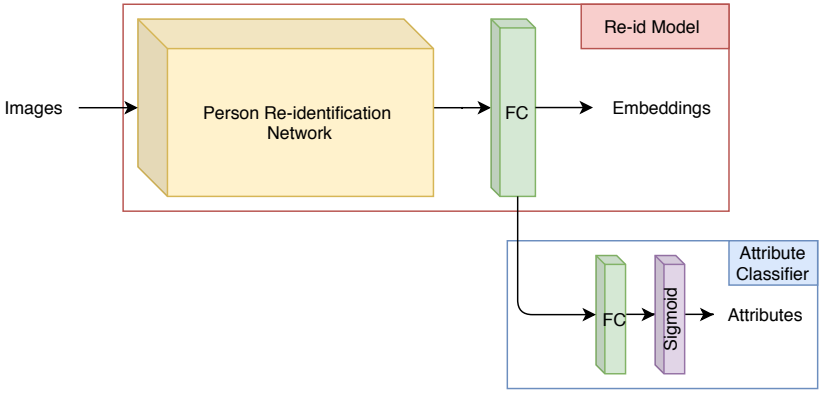


Figure 3.1: Visualization of the multi-task network architecture used in this work. The CNN jointly generates person re-identification embeddings and recognizes semantic attributes.

to explain the matching, but this method suffers from several drawbacks. First, depending on the number of attributes, displaying scores for all attributes could overstrain the system operator because they can not be captured and understood at a glance. Additionally, some attributes might not be visible or relevant to re-identify the person shown in the query image and thus do not need an explanation. Furthermore, absolute errors are hard to interpret without expert knowledge and reference values. As a result, it would be beneficial to have matching scores in percent instead. Building on the identified problems, the following method is proposed to generate meaningful clues on the positions of gallery images. Since the goal is to find occurrences of the person visible in the query image, the first step is to identify the attributes for which the classifier is most certain. Confidently recognized query attributes are determined by computing the distances between the classifier outputs and the attribute decision boundary as a measure of uncertainty. Typically, the decision boundary is 0.5. Afterward, a decision is made based on a threshold t_a. So, with x being the classifier output for attribute a, attribute a is chosen if $|x_a - 0.5| > t_a$ applies. For example, for $t_a = 0.1$ attributes with classification scores below 0.1 or above 0.9 would be selected as suitable candidates to help to explain the ranking results. In the next step, the absolute errors between query and gallery images

are computed for these attributes. As the aim is to provide matching instead of error scores, the error measurements have to be inverted. Last but not least, normalization by the attribute prediction confidence of the query image results in matching scores in percent. The following equation points out the computation formula in detail.

$$s_a = \frac{1 - |x_a^q - x_a^g|}{0.5 + |x_a^q - 0.5|} \tag{3.1}$$

Here, q denotes the query image and $g \in G$ stands for a gallery image from gallery G.

3.2 Attribute-related error

The second concept for explaining the person rank list focuses on the retrieval distance instead of attribute classifications. The goal is to visualize the contribution of each attribute to the distance between q and g. This approach exposes the attributes which contribute most to the distance between the embedding vector. To achieve this, the squared error for each element of the query feature vector f^q and the gallery feature vector f^g is multiplied with the learned attribute classifier weight w_{na}. w_{na} denotes the weight between feature component n and attribute output neuron a. As can be seen in the following Equation 3.2, summing the weighted errors for all feature elements results in an error measurement e_a for attribute a.

$$e_a = \sum_{n=1}^{N} (f_n^q - f_n^g)^2 * w_{na} \tag{3.2}$$

Comparing the errors of the attributes allows the estimation of their contribution to the retrieval distance.

4 Evaluation

This chapter focuses on two main aspects. First, embeddings for person images are analyzed to understand the influence of single feature elements and to examine the correlation with semantic attributes. The more individual vector

elements correlate with single or few attributes, the easier it is to understand and interpret their values and to explain ranking results.

Second, the proposed concepts for explaining the resulting rank list of gallery images are evaluated qualitatively based on some meaningful examples.

4.1 Training and Parameters

All experiments presented in this work are conducted with the well-known Market-1501 [22] dataset as it provides identity labels as well as annotations for 27 attributes. The dataset consists of $32,668$ images of $1,501$ persons, divided into training, query, and test sets. For the experiments, the multi-class attribute *age* is also assumed to be binary, resulting in 30 binary attributes.

For the experiments, the AGW approach [20] is used as the person re-id model. It achieves results comparable to the state-of-the-art with a simple architecture. It is trained using the standard parameters proposed in the original work.

Subsequently, the network parameters of the re-id model remain frozen while a fully-connected classification layer appended to the re-id feature layer is trained. This additional layer consists of 2048×30 weights which equals the number of re-id vector elements times the number of binary attributes in the datasets. Please note that learning a bias is omitted in this layer. Regarding training parameters and procedure, this work orients itself on the findings of [15].

4.2 Embeddings and corresponding attributes

To create Table 4.1, the learned weights between each of the 2048 components of the fully-connected feature layer and the attribute classification layer were examined. For each component, the attribute with the highest weight was determined and summed up with respect to the attributes. For instance, 104 vector components have the greatest weight with attribute *downblue*. The third column refers to the positive ratio of attributes in the training dataset since obviously there is a relationship between positive ratios and the number of top-1 occurrences. The results in the table indicate that the problem of imbalanced or biased data is not only limited to the task of person attribute recognition. Persons

with rarely occurring attributes such as *hat* or *downyellow* are worse represented in feature space and thus the error probability of the resulting ranking increases. This is a particular problem when it comes to ethnically unbalanced training data. Besides, the second factor that is relevant for the number of neurons connected to attributes is the complexity of the attribute. For example, the attribute *bag* occurs in lots of different types, colors, and styles. Thus multiple feature elements are required to represent such a huge variety.

Table 4.1: The number of vector elements that have the greatest weight to the attributes compared to the positive ratios of attributes in the training dataset.

Attribute	#Top1	Positive Ratio
bag	122	24.63
age2	104	75.77
downblue	104	16.38
backpack	103	26.50
downgray	101	16.38
...	...	...
age4	39	1.07
downgreen	33	1.86
downyellow	32	1.33
downpurple	15	0.27
hat	8	2.66

4.3 Attributes and corresponding features

Next, it is examined what the individual elements of the embedding represent. Since there are connections between all feature elements and attributes, single elements likely represent combinations of several attributes rather than single attributes. Figure 4.1 shows rank lists of gallery images for meaningful and representative feature vector elements sorted by their values. The first column contains the three attributes with the highest and lowest weights, respectively. It

is noteworthy that since weights can have negative values, the attributes with low weights are inhibited if the respective vector element has a high value.

Figure 4.1: Rank lists of gallery images for selected vector elements. The images are sorted in descending order by the value of the corresponding embedding element.

The rank lists indicate that single embedding components stand for combinations of attributes. For example, all persons with high values for the element wear yellow shirts and belong to age group 2 (*upyellow, up, age2*) in Figure 4.1(b). Besides, if the color of the upper-body clothing is yellow, it is reasonable to inhibit the prediction probability of the attribute *upblack*. Another interesting finding is that many components not only stand for an attribute combination but are almost able to identify specific persons, like in Figure 4.1(a). Possible explanations are that the dataset consists of fewer identities than elements in the embedding vectors or that the last layer of a re-id network is already very focused on different persons and not concepts similar to attributes. The second explanation is in line with the results of [15]. The authors achieved the best results if the last network layer is not shared between the re-id and attribute recognition tasks in a multitask network. The results showed that there is interference due to different training goals. Features from the last layer of a re-id network are too focused on different identities, even with the same set of semantic attributes.

4.4 Towards understanding rankings

This section provides examples of the proposed concepts to explain rankings based on semantic attributes. First, Figure 4.2 shows ranked gallery images with certain query attributes and corresponding matching scores for images in the top-10 ranks. Second, the error composition for the last sample is discussed based on Figure 4.3.

The first example in Figure 4.2(a) shows a frontal view of the person of interest in good quality. As a result, many attributes are reliably recognized. These attributes cover all types of semantic attributes ranging from global ones like *gender* and *age* over accessories to clothing styles and colors. At first glance, early ranks only contain persons that share a very similar visual appearance. However, there are also cases of error. For example, when looking at rank 8, one can observe that the image shows a woman although a man is visible in the query image. Concerning the proposed matching scores, it gets clear why this image occurs in an early rank. The matching scores for the attributes are high even for the *gender* attribute (! denotes that an attribute is not present. In this case, *gender* means female while *!gender* stands for male). It can be concluded that the CNN was not able to recognize correctly that the image on rank 8 shows a woman in contrast to the query image. This is indeed a difficult case because there are few clues. For instance, the long hair of the person is hardly visible from the frontal view and could also be a part of the background.

In contrast to the first example, the query image of the second example (see Figure 4.2(b)) is blurry and shows heavy occlusions. As a result, the re-id network is unsure about most of the attributes. For example, it is noticeable that irrelevant features such as the concealing bicycle are re-identified instead of the person. Together with the fact that even early ranks only have matching scores significantly below 75%, it indicates that the result is not very reliable. In practical application, the system operator should use another query image with better quality and fewer occlusions in such a case.

Regarding the example in Figure 4.2(c), clothing colors are certainly predicted, but global attributes like *gender* can not be surely determined. As a result, early ranks contain persons with the same combination of short, red upper-body clothing and black trousers with high matching scores. Furthermore, top-10

age2	99.94%	99.74%	99.55%	97.70%	99.29%	99.69%	96.87%	99.02%	97.41%	96.77%
backpack	97.99%	97.44%	93.78%	97.45%	95.37%	97.03%	95.05%	96.52%	92.61%	98.71%
!bag	96.10%	76.44%	99.40%	90.47%	68.91%	88.77%	99.55%	96.24%	97.58%	96.50%
clothes	99.36%	99.20%	99.35%	98.82%	97.98%	98.38%	98.03%	95.57%	93.49%	98.27%
down	96.96%	99.77%	96.07%	95.69%	92.65%	94.80%	96.82%	96.66%	97.80%	81.16%
downgray	98.97%	99.86%	98.12%	100.00%	99.09%	99.61%	99.16%	93.70%	98.55%	90.94%
!gender	99.01%	95.86%	96.52%	97.50%	98.64%	97.34%	92.97%	98.54%	97.23%	79.84%
!hair	97.19%	95.12%	97.03%	96.96%	93.12%	96.65%	98.75%	94.02%	96.36%	90.69%
upwhite	99.56%	99.11%	97.56%	98.45%	97.91%	98.60%	98.52%	97.86%	94.31%	97.54%

(a)

age2	96.87%	95.88%	91.07%	93.52%	85.64%	67.28%	86.51%	96.37%	95.92%	63.54%
upwhite	99.58%	97.75%	74.79%	97.43%	99.16%	89.60%	97.35%	98.24%	98.27%	98.13%

(b)

clothes	95.45%	98.48%	98.76%	91.01%	91.23%	98.37%	95.09%	98.65%	80.96%	99.48%
downblack	99.47%	99.53%	99.17%	99.72%	99.76%	98.96%	98.81%	99.34%	95.99%	99.09%
upred	96.31%	99.11%	99.96%	98.21%	96.84%	89.67%	96.38%	97.64%	98.40%	97.39%

(c)

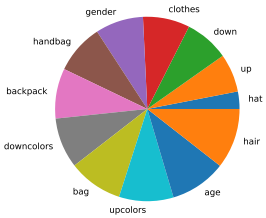
down	98.88%	98.37%	98.12%	99.53%	98.69%	98.29%	96.61%	98.65%	97.36%	99.79%
downgray	99.43%	98.14%	93.00%	98.09%	97.65%	95.31%	91.44%	95.13%	93.61%	99.58%
gender	99.92%	99.69%	99.89%	99.52%	99.46%	99.14%	99.72%	99.51%	99.38%	99.52%
hair	99.86%	99.48%	98.22%	99.36%	99.21%	99.76%	98.76%	97.49%	88.84%	98.20%
!hat	99.18%	97.68%	99.05%	94.70%	94.28%	96.45%	98.78%	96.18%	97.01%	95.97%

(d)

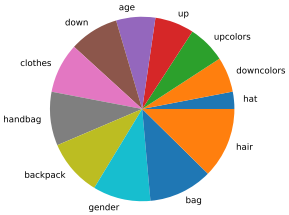
Figure 4.2: Examples for the proposed matching scores based on semantic attributes. ! before an attribute indicates that the attribute is not present in the image.

ranks include both men and women, since gender is not clear from the query image.

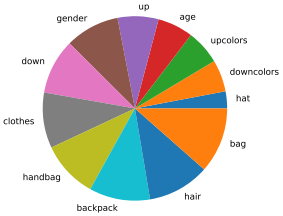
The last example again contains only gallery images with high matching scores for certain query attributes. Except for the last picture, the top-10 images show the person of interest. However, the tenth rank visually differs significantly from the target person. The woman wears a yellow shirt in contrast to the dress the query person is wearing. This example is used to look into detail regarding the error or distance composition as explained in Section 3. Figure 4.3 visualizes error composition for images on rank 3, 9, and 10 as pie charts.



(a) Rank 3



(b) Rank 9



(c) Rank 10

Figure 4.3: Error composition for three different rank images for the rank list shown in 4.2(d)

It attracts attention that the upper-body color greatly contributes to the retrieval distance for rank 3, but less for ranks 9 and 10. The reason for this is that backpacks cover large parts of the upper body clothing when a person is visible from behind. This fact explains the early position of the image on the tenth rank. Matching is not done by the actual color of the upper-body clothing. Instead, the network focuses on the color of the backpack and does not notice the yellow shirt. As a result, the proposed concept allows the understanding of re-id rankings and enable easier identification of weaknesses of the re-id approach used.

5 Conclusion

This work presented concepts to explain and understand rank lists of person re-id system. For this, an attribute classifier is trained with the goal of adding interpretable semantics. Qualitative evaluations show that the proposed concepts work and provide a solid basis for explainable person re-id. It seems to be worth it to conduct further investigations in future work. Interesting research directions include the development of methods for quantitative evaluation as well as generative adversarial networks (GAN). GANs would allow the manipulation of certain aspects of a person's visual appearance and thereby an examination of the effects and influences on ranking results.

References

[1] Chirag Agarwal, Peijie Chen, and Anh Nguyen. "Intriguing generalization and simplicity of adversarially trained neural networks". In: *arXiv preprint arXiv:2006.09373* (2020).

[2] Santiago A Cadena et al. "Diverse feature visualizations reveal invariances in early layers of deep neural networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 217–232.

[3] Aditya Chattopadhay et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 839–847.

[4] T. Chen et al. “Abd-net: Attentive but diverse person re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 8351–8361.

[5] Alexey Dosovitskiy and Thomas Brox. “Inverting visual representations with convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4829–4837.

[6] Yash Goyal et al. “Counterfactual visual explanations”. In: *arXiv preprint arXiv:1904.07451* (2019).

[7] Dong Huk Park et al. “Multimodal explanations: Justifying decisions and pointing to the evidence”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8779–8788.

[8] Qi Li et al. “Improving sample diversity of a pre-trained, class-conditional GAN by changing its class embeddings”. In: *arXiv:1910.04760* (2019).

[9] Shusen Liu et al. “Generative counterfactual introspection for explainable deep learning”. In: *arXiv preprint arXiv:1907.03077* (2019).

[10] Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Improving the interpretability of deep neural networks with knowledge distillation”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2018, pp. 905–912.

[11] Hao Luo et al. “Bag of tricks and a strong baseline for deep person re-identification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[12] Aravindh Mahendran and Andrea Vedaldi. “Understanding deep image representations by inverting them”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5188–5196.

[13] Uday Singh Saini and Evangelos E Papalexakis. “A Peek Into the Hidden Layers of a Convolutional Neural Network Through a Factorization Lens”. In: *arXiv preprint arXiv:1806.02012* (2018).

[14] Shibani Santurkar et al. “Computer Vision with a Single (Robust) Classifier.” In: (2019).

[15] Andreas Specker, Arne Schumann, and Jürgen Beyerer. “A multitask model for person re-identification and attribute recognition using semantic regions”. In: *Artificial Intelligence and Machine Learning in Defense Applications II*. Ed. by Judith Dijk. Vol. 11543. International Society for Optics and Photonics. SPIE, 2020, pp. 98–110. DOI: 10.1117/12.2573981. URL: https://doi.org/10.1117/12.2573981.

[16] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806* (2014).

[17] Sarah Tan et al. “Distill-and-compare: Auditing black-box models using transparent model distillation”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 303–310.

[18] Giulia Vilone and Luca Longo. *Explainable Artificial Intelligence: a Systematic Review*. 2020. arXiv: 2006.00093 [cs.AI].

[19] Eric Wong and J Zico Kolter. “Neural network inversion beyond gradient descent”. In: ().

[20] Mang Ye et al. “Deep Learning for Person Re-identification: A Survey and Outlook”. In: *arXiv preprint arXiv:2001.04193* (2020).

[21] Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. “Interpreting neural network judgments via minimal, stable, and symbolic corrections”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 4874–4885.

[22] Liang Zheng et al. “Scalable Person Re-Identification: A Benchmark”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.