Incorporating Domain Knowledge for Learning Interpretable Features

Martin Melchior

Abstract Deep Learning has seen an enormous success in the last years. In several application domains prediction models with remarkable accuracy could be trained, sometimes by using large datasets. Often, these models are configured with huge amounts of parameters and seen by domain experts as hard to understand black boxes and hence of less value or not trustworthy. As a result, we observe a claim for better interpretability in several application domains. This claim can also be seen to arise from the fact that the formulation of the underlying problems is not complete and certain important aspects are disregarded. Interpretability is required particularly in domains where high demands with respect to safety or fairness are posed or, for example, in natural sciences where the application of these techniques aims at knowledge discovery. Alternatively, the gap in problem formulation can be compensated by incorporating a priori domain knowledge into the model. In this article, we highlight the importance to further advance the techniques to support interpretability or the mechanisms to incorporate domain knowledge in the machine learning approaches. When

Martin Melchior
Institute for Data Science, School of Engineering, University of Applied Sciences and Arts of Northwestern Switzerland, CH-5210 Windisch, Switzerland,
✉ martin.melchior@fhnw.ch

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 8, No. 2, 2022

DOI: 10.5445/IR/1000150142
ISSN 2363-9881

transferring these techniques to the application domains, close collaboration with domain experts is indispensable.

1 Success of Machine Learning and Deep Learning

In the last decade, the field of data science has seen an enormous boost in developing techniques to perform tasks in computer vision, text and audio processing. These techniques have entered our everyday lives - ranging from apps for language translation (Popel et al., 2020), recommender systems (Zhang et al., 2019) for music and movies to tools for recognizing and annotating objects in images which improve the search in large image databases (Mellina, 2017).

This development has stimulated many applications in the disciplines of science and engineering. Undoubtedly, in many cases the use of these techniques has proven to be most beneficial because they have made the work of researchers and engineers easier and more efficient. In some application domains where huge amounts of data (Big Data) are involved, techniques for automatic and efficient filtering and extraction of information have become an indispensable part and an enabler of the data exploitation process. This holds true, for example, in natural sciences such as astrophysics (Baron, 2019), solar physics (Bobra and Couvidat, 2015) or particle physics (Albertsson et al., 2018) where huge amounts of data are collected and sophisticated techniques for efficient filtering or for information extraction are needed. But it also applies in commercial applications such as in finance or in medicine where huge amounts of data are available that can be used for better prediction or diagnostics, respectively.

These techniques are based on machine learning approaches which are designed to directly learn from data. In traditional machine learning, suitable features are handcrafted typically by domain experts. These features are then plugged into rather simple models which are trained to accomplish the desired tasks. For not too complex models, the way the input features are combined to produce an output is easy to explain and the way the model derives its decisions can (more) easily be interpreted from a domain perspective. This holds true particularly for rather small decision trees or for linear regression models with a small number of factors.

The deep learning models (LeCun et al., 2015) that have proven to be particularly successful in the computer vision, text or audio processing domains which are characterized by complex, high-dimensional, unstructured data are on the opposite side of the spectrum: These models are extremely flexible, composed of many nested components and come along with a huge number of parameters. During training, they can flexibly be adjusted to capture a large variety seen in the data. The features relevant for the task at hand are learned directly from the data during training and no feature engineering is needed. From the data science perspective, the role of the domain experts is reduced to properly prepare the data, possibly to provide labels and to judge and interpret the result.

In domains with complex, high-dimensional, unstructured data this works because neural networks often seem to provide a structure nicely suited for the problems at hand, also referred to as *inductive bias* (Mitchell, 1980). In typical computer vision applications, this structure decomposes into a succession of layers which leads to deeply nested function calls. The different levels of depth are associated with different levels of abstractions: As can be demonstrated in some computer vision applications, low level features are represented and recognizable in earlier layers, higher level concepts in later layers. A nice example can be found in (Olah et al., 2020) where the authors show the kind of features represented in the first five layers of the Inceptionv1 model (Szegedy et al., 2015). This gives a nice intuition about how and why deep learning models work and why they sometimes almost miraculously generalize well to new data not seen during training.

2 Need for Explainable Models

In the data science community, the development in the field has strived mostly for high accuracy. This has led to huge models with a huge number of trainable parameters. Recent models such as GPT-3 of OpenAI (Brown et al., 2020) or Switch-C of Google (Fedus et al., 2022) have 100s of billions of trained parameters. The capabilities of these models are impressive but come at the price that they are considered as black-box systems, hard to understand and interpret. Indeed, in some domains, there is a high demand for understanding the way a specific model operates and the mechanisms underlying the decisions made by the model.

This is of particular importance in critical applications or when using these techniques e.g. in the natural sciences where a primary goal is to discover new knowledge. In critical applications, users or adopters want to understand and have transparency about how a model infers its predictions. From a user perspective, there are often other criteria than accuracy that also play an important role. Examples are safety in collision warning systems or non-discrimination in applications of the criminal justice system (for examples see Doshi-Velez and Kim (2017)). Criteria such as safety or fairness are more difficult to assess, though. For guaranteeing safety, a complete list of possible failover scenarios must be checked. Similarly, for proving fairness all features in the underlying data needs to be assessed so that fair and non-discriminative predictions are inferred - eliminating the most obvious features (such as race or sex) is often not sufficient.

The reason for this difficulty in assessing other criteria is, as argued in (Doshi-Velez and Kim, 2017), related to the fact that the problem is not completely formalized which results in an *unquantified bias*. One way to compensate for that is to claim to have sound explanations for how the reasoning of the system works under the hood and these explanations are believed to provide trust in that the system works properly (or not).

In fields of application such as in various science disciplines, machine learning is adopted for optimizing or even producing scientific outcomes, it often plays an important role in the knowledge discovery process and can help to reveal patterns hidden in huge data. The scientists strive for understanding the whole knowledge discovery chain with all its processing, filtering and information acquisition steps. Hence a sound understanding of how the possibly applied machine learning tools work is an imminent part of a scientifically sound approach. When observing new patterns in the data, they need to be associated with and expressed by key quantities of the considered field of science, in the following referred to as domain quantities.

Generally, the demand for explanations for how data-driven models work aims at acquiring more trust into and acceptance of these tools in the user communities. Users or adopters want to have answers to questions such as why a model makes a certain decision, what drives the model predictions or how model predictions can be trusted (Sarkar, 2018). Explanations will even be required by law in some cases. Recent European Union regulations, released in 2018, require to provide explanations for algorithms that make decisions based on user-level predictors which significantly affect users - they have a *right to explanation* (see e.g. Goodman and Flaxman (2017) for a discussion on the implications).

The level of explanation and the depth of understanding is, however, rather elusive. It depends on the task to be solved and on the application domain or on how well the system has been tested in real-world scenarios. Furthermore, the request for explanations or the acceptance without depends also on the availability of competing alternatives. Deep learning applications that are proven to be much more accurate than any other existing tool and that significantly exceed human performance might be less exposed to claims for explainability than systems that provide a performance comparable to humans.

3 Explainable AI (XAI)

In the last few years a lot of research has flown into turning deep learning models into more transparent, better explainable, or interpretable approaches. These activities are summarized under the notion of Explainable AI (XAI). Here, the terms *transparency*, *explainability* and *interpretability* are important notions but seem rather elusive. For instance, interpretability and explainability are used inconsistently in the community. For example, (Miller, 2019) treats them as identical whereas (Montavon et al., 2018) makes a distinction. We follow the terminology defined in the latter.

(Montavon et al., 2018) defines, *interpretability* as the capability of providing a mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of. In contrast, an *explanation* is defined as the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression). Therefore, explainability as used in the given context rather focuses on explaining by example. It is a post-hoc consideration while interpretability works rather at the conceptual level. Finally, by following (Roscher et al., 2020), an approach is *transparent* if the process of extracting model parameters from training data and generating labels from test data can be described and motivated by the designer of the approach. Hence, while a transparent approach can explain the mechanics, interpretability establishes the links and associations to the domain context.

Several techniques have been developed to elucidate how a network draws its decisions and contributing to explainability. Several tools have been designed particularly for neural networks. Many of them provide visually assisted explanations. For example, saliency maps (or its improved alternatives) highlight the pixels that are relevant for a certain image classification by a neural network (see e.g. Zeiler and Fergus (2014), Simonyan et al. (2014) or Samek et al. (2019) for an overview). Such tools are very helpful to explore what types of patterns in the input data the model responds to. This allows to develop an intuition that may also be helpful in obtaining interpretability.

As pointed out in (Rudin, 2019), such post-hoc explanations are not capable of providing perfect fidelity for the original model. They provide incomplete and possibly inaccurate insights into or representations of the original model. If we cannot know for certain whether our explanation is correct we cannot know whether to trust either the explanation or the model. This leads (Rudin, 2019) to the statement that black-box models that provide some explainability but no interpretability should be avoided. This position is justified in critical applications with strong safety or ethics requirements. In other domains such as in the fundamental sciences where machine learning is an important tool for knowledge discovery a more moderate claim may be justifiable by requesting only partial interpretability. An example with partial interpretabilty would be the situation where a latent representation is learned by some dimensionality reduction technique and where only part of the variables have been interpreted. By exploring and analysing the set of uninterpreted variables, domain experts may learn more about hidden patterns in the data and gain new knowledge (see also 4.1 below).

Explainability and partial interpretability can only be successfully reached when accompanied by an intense and close discourse of the data scientists with domain experts. The data scientists know the details of the models and their underlying mechanics and know the tools to extract information or visualize patterns - but only the domain experts can provide a sound judgement when it comes to relate the findings of the model back to the raw data or the domain context in general. Only the domain experts can establish a formulation of the outcomes and their explanations by using the terminology of the domain. Furthermore, the domain experts will be in the best position to judge whether a given application scenario works given the premises under which the model has been designed, trained, and tested.

4 Role of Domain Knowledge

Traditionally, domain knowledge has entered the work of the data scientists in the form of labeling the data or when exploring the raw data and making it ready for machine learning. Labels are expensive to collect and, therefore, labeling is not always feasible. For sure, domain knowledge will continue to play an important role when exploring the raw data which provides important insights into what factors of variation exist, may give hints on how these could best be mapped into a model and what parts need to be corrected at the preprocessing stage. Domain knowledge can also be incorporated when designing and evaluating model architectures. It may provide guidance in the form of established domain theories that need to be respected. Examples are constraints to be fulfilled (such as preserved quantities), symmetries to be respected or mapping equations governing the dynamics of a system with parts consisting of components learnable from the data. See (von Rueden et al., 2021) for an overview.

Learning models that are consistent with domain theories help improving the acceptance in the application domain and help to fill the incompleteness in the problem formulation and the associated unquantified bias as mentioned in Section 2. Incorporating domain knowledge manifestly in the machine learning models compensates some of the need for interpretability.

Generally, it is expected that the connection between domain knowledge and machine learning models will further strengthen in the years to come. This will happen at the level of designing suitable model architectures that allow for incorporating constraints in the models and it will also consist in learning schemes that will allow for correlating the features learned from the data with domain knowledge. Overall, this will lead to better interpretability and acceptance in the application domains. Note that, in addition to better interpretability, the possibility to incorporate domain theories in models that learn from data may give the scientists a tool to more easily test and compare conflicting theories.

Various approaches developed in the last few years seem very promising to accommodate domain knowledge - depending on the specific domains, the typical datasets available there and the specific tasks to be solved. In the following, we sketch a few of these approaches - well aware that the presented selection is highly subjective and rather reflects some recent architectures the author is working with.

4.1 Learning Lower-Dimensional Latent Representations

When starting with high-dimensional unstructured data, an important step is to reduce the dimensionality by extracting the informative features which can then be used to solve downstream tasks such as a prediction or classification problem. A common approach to achieve this is by using models of the autoencoder family. In the last years, this family has evolved to a rich set of generative models that learn the underlying data distribution and that allow to incorporate different nice aspects in the lower dimensional features. Of particular interest are models that learn disentangled representations, with maximal information content while providing excellent reconstruction quality. Different architectures and loss functions help shaping the features in latent space in one or another direction - depending on the needs for the downstream tasks. The corpus of literature here is very rich - an overview can be found in (Voloshynovskiy et al., 2020) where many of these modeling approaches are put in a common, comprehensive framework.

These representations are, though lower-dimensional, not interpretable per se. Furthermore, the dimensionality is typically still rather high - rather in the 100s which is far beyond of what is manageable for the human. Note that Miller (Miller (1956), or also Lage et al. (2018)) argued in 1956 that humans can hold about seven items simultaneously in working memory and suggested that human-interpretable explanations should obey some kind of capacity limit. Here, visual tools such as those provided with the dimensionality reduction tool kit (e.g. t-SNE (van der Maaten and Hinton, 2008) or U-Map (McInnes et al., 2018)) in combination with possibly only a few labeled samples may help to explore whether the features are arranged in an interpretable structure. Although this procedure may work in some cases, it does in general not guarantee that the learned features can directly be associated with domain quantities. To achieve interpretability, close communication with the domain experts is indispensable.

4.2 Suitable Architectures for Incorporating Domain Knowledge

The association of input or learned features (such as obtained in 4.1) with domain quantities could be facilitated by using invertible flow networks (see e.g. Kobyzev et al. (2020) for a review). Here, a non-linear mapping between the feature and target space is learned. Some of the dimensions in the target space can be constrained to known domain quantities that characterize the original samples (see e.g. Ardizzone et al., 2019)). This domain information may come in form of meta data provided in addition to the training data set. Such an approach has successfully been adopted in (Lanusse et al., 2021) to images of galaxies.

However, not all the target dimensions can directly be associated with domain quantities. As pointed out above, the number of latent dimensions obtained in the previous stage is still too large for all features to be directly interpretable. Some of the remaining dimensions will be associated with noise hence uninformative and irrelevant. Other dimensions are related to features that are not of interest for the domain expert but may be interpretable. For a task that is known to be invariant with respect to a symmetry group, the information that allows to identify the element of the symmetry group for a specific input sample (e.g. a rotation angle) must not have predictive power for solving the problem at hand. An example is the task of classifying the type of galaxy from images of galaxies (see e.g. Lanusse et al., 2021). The type of galaxy is independent of its orientation angle. Thus, the information in the latent space that encodes the orientation angle is not relevant for the task at hand (it may, however, be of interest in another context).

Hence, only part of the features not yet mapped to domain quantities may contain promising candidates for knowledge discovery and interpretation. To identify them, the different dimensions may be explored by projecting selected values back to the input space which then can be carefully analyzed. In an encoder-decoder architecture (such as Kingma and Welling, 2014), the mapping from the latent space to the input space is obtained by the decoder. When analyzing the footprint of the latent space dimensions in the input space, domain knowledge is crucial for identifying and interpreting variables of interest or filtering out variables that can be ignored such as uninformative noise or quantities associated with symmetry transformations.

When learning tasks that are known to be invariant with respect to symmetry transformations an alternative approach is to choose a model architecture that is intrinsically invariant with respect to these symmetry transformations. This is the goal of geometric deep learning (Bronstein et al., 2021), a very promising research direction that has gained a lot of attention in the last few years. When using a network architecture that is invariant with respect to symmetry transformations the lower dimensional representation of the input obeys the same invariance properties. Hence, a reduced latent space is learned that encodes the relevant information in a representation that is invariant to the symmetry transformations. Here, the domain expert is needed to provide the knowledge about symmetry groups to be respected.

In conclusion, here again, close interaction with domain experts will be key to create sensible and interpretable output for the domain.

4.3 Graph Neural Networks for Mapping Domain Knowledge

Another promising research direction are graph neural networks (GNN). These provide rich structures that allow to map relations between entities and rules for composing them. This is also an aspect that can be found in symbolic AI and, not surprisingly and as pointed out in (Battaglia et al., 2018), GNNs provide a fusion between symbolic and connectionist AI. In the given context, GNNs allow for incorporating constraints or a priori knowledge provided by domain theories. An example are GNNs that allow to incorporate domain knowledge graphs into an image classification pipeline (see von Rueden et al. (2021) and Marino et al. (2017)). In the latter, it is also shown how the incorporation of the knowledge graph into the GNN model allows to improve the performance of the classification task.

GNNs also allow to map causal relations which represent structural knowledge about the data generating process and which are intimately related to domain knowledge. Once properly included, this leads to (more) transparency and interpretability. Nevertheless, one central problem that needs to be solved on the machine learning part also remains here: Efficient techniques need to be designed and implemented that allow to discover high-level interpretable candidates for causal variables from low-level observations (Schölkopf et al., 2021). The key research areas and research questions are identified and the community is very actively searching for solutions.

Definitely, the approaches described above could be extended. Typically, these approaches need to be adjusted and refined when applied to domain problems. Further work is needed on developing procedures and techniques that help to better incorporate domain knowledge either in form of identifying interpretable features learned with suitable architectures from the data or including a priori knowledge directly in the model architecture or in the learning process.

5 Conclusions

We have highlighted a high demand for improving the interpretability of deep learning models in several application domains and we have argued that this demand is related to an incomplete formulation of the underlying problem where not all relevant aspects are considered. For closing this incompleteness gap, we have suggested the application of suitable architectures that give better support for learning and identifying interpretable features or that allow to include a priori information in the model architecture or the learning process. We have indicated a few research directions that seem promising for learning and defining interpretable features that can then be used for better interpretable downstream prediction tasks. As examples, we have mentioned models from the autoencoder family for learning suitable lower dimensional representations of the original data, invertible flow networks to learn a nonlinear mapping to domain quantities, hence constraining to domain quantities or, graph neural networks that allow to map causal relations or incorporate a priori information directly in form of a knowledge graph.

References

Albertsson K, Altoe P, Anderson D, Andrews M, Espinosa JPA, Aurisano A, Basara L, Bevan A, Bhimji W, Bonacorsi D, Calafiura P, Campanelli M, Capps L, Carminati F, Carrazza S, Childers T, Coniavitis E, Cranmer K, David C, Davis D, Duarte J, Erdmann M, Eschle J, Farbin A, Feickert M, Castro NF, Fitzpatrick C, Floris M, Forti A, Garra-Tico J, Gemmler J, Girone M, Glaysher P, Gleyzer S, Gligorov V, Golling T, Graw J, Gray L, Greenwood D, Hacker T, Harvey J, Hegner B, Heinrich L, Hooberman B, Junggeburth J, Kagan M, Kane M, Kanishchev K, Karpiński P, Kassabov Z, Kaul G, Kcira D, Keck T, Klimentov A, Kowalkowski J, Kreczko L, Kurepin A, Kutschke R, Kuznetsov V, Köhler N, Lakomov I, Lannon K, Lassnig M, Limosani A, Louppe G, Mangu A, Mato P, Meinhard H, Menasce D, Moneta L, Moortgat S, Narain M, Neubauer M, Newman H, Pabst H, Paganini M, Paulini M, Perdue G, Perez U, Picazio A, Pivarski J, Prosper H, Psihas F, Radovic A, Reece R, Rinkevicius A, Rodrigues E, Rorie J, Rousseau D, Sauers A, Schramm S, Schwartzman A, Severini H, Seyfert P, Siroky F, Skazytkin K, Sokoloff M, Stewart G, Stienen B, Stockdale I, Strong G, Thais S, Tomko K, Upfal E, Usai E, Ustyuzhanin A, Vala M, Vallecorsa S, Vasel J, Verzetti M, Vilasís-Cardona X, Vlimant JR, Vukotic I, Wang SJ, Watts G, Williams M, Wu W, Wunsch S, Zapata O (2018) Machine Learning in High Energy Physics Community White Paper. Journal of Physics: Conference Series 1085:022008, IOP Publishing. DOI: 10.1088/1742-6596/1085/2/022008.

Ardizzone L, Kruse J, Rother C, Köthe U (2019) Analyzing Inverse Problems with Invertible Neural Networks. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=rJed6j0cKX.

Baron D (2019) Machine Learning in Astronomy: a practical overview. arXiv. DOI: 10.48550/ARXIV.1904.07248.

Battaglia P, Hamrick JBC, Bapst V, Sanchez A, Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R, Gulcehre C, Song F, Ballard A, Gilmer J, Dahl GE, Vaswani A, Allen K, Nash C, Langston VJ, Dyer C, Heess N, Wierstra D, Kohli P, Botvinick M, Vinyals O, Li Y, Pascanu R (2018) Relational inductive biases, deep learning, and graph networks. arXiv. URL: https://arxiv.org/pdf/1806.01261.pdf.

Bobra MG, Couvidat S (2015) Solar Flare Prediction Using SDO-HMI Vector Magnetic Field Data With a Machine-Learning Algorithm. The Astrophysical Journal 798(2):135, American Astronomical Society. DOI: 10.1088/0004-637x/798/2/135.

Bronstein MM, Bruna J, Cohen T, Velickovic P (2021) Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. CoRR abs/2104.13478. URL: https://arxiv.org/abs/2104.13478.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., Vol. 33, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Doshi-Velez F, Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning. arXiv e-prints.

Fedus W, Zoph B, Shazeer N (2022) Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research 23(120):1–39. URL: http://jmlr.org/papers/v23/21-0998.html.

Goodman B, Flaxman S (2017) European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine 38(3):50–57. DOI: 10.1609/aimag.v38i3.2741.

Kingma DP, Welling M (2014) Auto-Encoding Variational Bayes. CoRR abs/1312.6114.

Kobyzev I, Prince S, Brubaker M (2020) Normalizing Flows: An Introduction and Review of Current Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence PP:1–1. DOI: 10.1109/TPAMI.2020.2992934.

Lage I, Chen E, He J, Narayanan M, Gershman S, Kim B, Doshi-Velez F (2018) An Evaluation of the Human-Interpretability of Explanation. In: 2018 Conference on Neural Information Processing Systems (NeurIPS).

Lanusse F, Mandelbaum R, Ravanbakhsh S, Li CL, Freeman P, Paczos B (2021) Deep generative models for galaxy image simulations. Monthly Notices of the Royal Astronomical Society 504(4):5543–5555. DOI: 10.1093/mnras/stab1214.

LeCun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521:436–44. DOI: 10.1038/nature14539.

van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research 9(86):2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Marino K, Salakhutdinov R, Gupta A (2017) The More You Know: Using Knowledge Graphs for Image Classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20–28. DOI: 10.1109/CVPR.2017.10.

McInnes L, Healy J, Saul N, Grossberger L (2018) UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 3(29):861, The Open Journal. DOI: 10.21105/joss.00861.

Mellina C (2017) Introducing Similarity Search at Flickr. URL: https://code.flickr.net/2017/03/07/introducing-similarity-search-at-flickr.

Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63(2):81–97, American Psychological Association (APA).

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267:1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007.

Mitchell TM (1980) The Need for Biases in Learning Generalizations. Tech. Rep., Rutgers University, New Brunswick, NJ. URL: http://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73:1–15. DOI: https://doi.org/10.1016/j.dsp.2017.10.011.

Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) An Overview of Early Vision in InceptionV1. Distill. DOI: 10.23915/distill.00024.002. Https://distill.pub/2020/circuits/early-vision.

Popel M, Tomkova M, Tomek J, Kaiser L, Uszkoreit J, Bojar O, Žabokrtský Z (2020) Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nature Communications 11(1):4381. DOI: 10.1038/s41467-020-18073-9.

Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8:42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5):206–215. DOI: 10.1038/s42256-019-0048-x.

von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J (2021) Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. IEEE Transactions on Knowledge and Data Engineering, pp. 1–1. DOI: 10.1109/TKDE.2021.3079836.

Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (eds.) (2019) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing. DOI: 10.1007/978-3-030-28954-6.

Sarkar D (2018) The importance of human interpretable machine learning. URL: https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476.

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021) Toward Causal Representation Learning. Proceedings of the IEEE 109(5):612–634. DOI: 10.1109/JPROC.2021.3058954.

Simonyan K, Vedaldi A, Zisserman A (2014) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Bengio Y, LeCun Y (eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings. URL: http://arxiv.org/abs/1312.6034.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. DOI: 10.1109/cvpr.2015.7298594.

Voloshynovskiy S, Taran O, Kondah M, Holotyak T, Rezende D (2020) Variational Information Bottleneck for Semi-Supervised Classification. Entropy 22(9). DOI: 10.3390/e22090943.

Zeiler MD, Fergus R (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, pp. 818–833. ISBN: 978-3-319105-90-1.

Zhang S, Yao L, Sun A, Tay Y (2019) Deep Learning Based Recommender System: A Survey and New Perspectives. ACM Comput. Surv. 52(1), Association for Computing Machinery, New York, NY, USA. DOI: 10.1145/3285029.