A ADVANCED JOURNAL IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING TRENDS IN BIOLOGICAL SCIENCES
Open Journal

Review

Understanding, Visualizing, and Interpreting Deep Learning Models: A Path to Explanable Artificial Intelligence

Wojciech Samek[1*]; Thomas Wiegand[1,2]; Klaus-Robert Mu¨ller[2,3,4]

[1]*Dept. of Video Coding and Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*
[2]*Dept. of Computer Science, Technische Universita¨t Berlin, 10587 Berlin, Germany*
[3]*Dept. of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea*
[4]*Max Planck Institute for Informatics, Saarbru¨cken 66123, Germany*

[*]***Corresponding author***
Wojciech Samek
Dept. of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany

Article information
Received: January 12th, 2024; **Revised:** January 22nd, 2024; **Accepted:** February 28th, 2024; **Published:** April 18th, 2024

Cite this article
Samek W, Wiegand T, Mu¨ller K-R. Understanding, visualizing, and interpreting deep learning models: A path to explanable artificial intelligence. 2024; 3(1).
doi: https://doi.org/10.70705/ppp.bioai.2024.v03.i01.pp16-21

ABSTRACT

More and more complicated jobs are now being handled by AI systems that perform at least as well as humans, thanks to the availability of huge datasets and recent advancements in deep learning technique. Some fields that have made remarkable strides in this direction include picture categorization, sentiment analysis, voice comprehension, and strategy gaming. Unfortunately, because to their layered non-linear structure, these very effective AI and machine learning models are often implemented in a way that does not reveal how they arrive at their predictions. The development of ways to visualize, explain, and understand deep learning models has lately garnered increased interest due to the fact that this lack of transparency may be a significant negative, for example in medical applications. A need for more interpretability in AI is made in this study, which also summarizes recent advances in the area. Also included are two ways to break down decisions into their component variables, one that calculates the prediction's sensitivity to changes in the input and another that uses deep learning to explain the models' predictions. There are three classification tasks that these approaches are tested on.

Keywords
Artificial intelligence; Deep neural networks; Black box models; Interpretability; Sensitivity analysis; Layer-wise relevance propagation.

INTRODUCTION

Over the last several decades, advancements have been made in the area of artificial intelligence and machine learning. The evolution of deep learning techniques and previous advancements in support vector machines were key motivators for this shift [22]. Important contributing reasons to the success of the project were the availability of big datasets like ImageNet [9] or Sports1M [17], the speed-up advantages achieved with powerful GPU cards, and the great flexibility of software frameworks like Caffe [15] or TensorFlow [1]. Modern AI systems that rely on machine learning are very good at many complicated tasks, such as visual object identification [14], understanding natural languages [8], and processing voice signals [10]. Furthermore, in highly strategic games like Go [34] and Texas hold 'em poker [28], new AI1 systems may even beat expert human players. These remarkable achievements of AI systems, particularly deep learning models, demonstrate the revolutionary nature of this technology. It will not only cause significant changes in enterprises and societies, but will also have far-reaching effects outside the academic sphere.

Despite these models' remarkable prediction accuracy, they are very opaque because to their layered non-linear structure; that is, it is not obvious which input data points are used to generate their judgments. Consequently, these models are usually seen as opaque entities. The lack of transparency in AlphaGo, the artificial intelligence system developed by DeepMind, is highlighted in the 37th move of the second game of the historic Go match between top Go player Lee Sedol and the system. In response to AlphaGo's completely unexpected move, a Go specialist offered the following analysis:
It was the game-winning move for AlphaGo, even though it was unclear throughout the match why the system played it. While AlphaGo's black box nature was irrelevant here, the difficulty of comprehending and verifying an AI system's decision-making process is

a major limitation in many contexts. For example, blindly trusting a black box system's predictions in medical diagnosis would be irresponsible. Rather, a human expert should be able to review and approve any major decisions. Additionally, it is crucial to ensure that the model relies on the proper attributes in self-driving vehicles, since even one wrong prediction might result in significant costs. To provide such a guarantee, AI models must be explainable and interpretable by humans. Expanded analysis of the Section 2 outlines the need for AI that can be explained.

It is not unexpected that there has been a lot of interest in the development of methods for "opening" black box models lately [6, 35, 39, 5, 33, 25, 23, 30, 40, 11, 27].

Methods for better understanding the model's representation (its learnt information) [12, 24, 29] and approaches for explaining individual predictions [19, 35, 39, 5, 26] are all part of this. There is a guide to these two types of approaches in [27]. Beyond neural networks, other sophisticated machine learning approaches rely on explainability as well, such as support vector machines [20].

This study aims to raise awareness about the importance of explainability in AI and machine learning. Section 2 does this. Section 3 then goes on to provide two modern methods for deriving an explanation for an AI model's prediction from its input variables: sensitivity analysis (SA) [6, 35] and layer-wise relevance propagation (LRP) [5]. In Section 4, we discuss ways to objectively assess explanation quality, and in Section 5, we show the findings of our picture, text, and video categorization experiments. In Section 6, the publication wraps off with a look towards what's to come.

2. WHY DO WE NEED EXPLAINABLE AI ?

The ability to explain the rationale behind one's decisions to other people is an important aspect of human intelligence. It is not only important in social interactions, e.g., a person who never reveals one's intentions and thoughts will be most probably regarded as a "strange fellow", but it is also crucial in educational context, where students aim to comprehend the reasoning of their teachers. Furthermore, the explanation of one's decisions is often a prerequisite for establishing a trust relationship between people, e.g., when a medical doc- tor explains the therapy decision to his patient.

Although these social aspects may be of less importance for technical AI systems, there are many arguments in favor of explainability in artificial intelligence. Here are the most im- portant ones:

Verification of the system: As mentioned before in many applications one must not trust a black box sys- tem by default. For instance, in health care the use of models which can be interpreted and verified by medical experts is an absolute necessity. The authors of [7] show an example from this domain, where an AI system which was trained to predict the pneumo- nia risk of a person arrives at totally wrong conclu- sions. The application of this model in a black box manner would not reduce but rather increase the num- ber of pneumonia-related deaths. In short, the model learns that asthmatic patients with heart problems have a much lower risk of dying of pneumonia than healthy persons. A medical doctor would immediately rec- ognize that this can not be true as asthma and hearth problems are factors which negatively affect the prog- nosis for recovery. However, the AI model does not know anything about asthma or pneumonia, it just in- fers from data. In this example, the data were system- atically biased, because in contrast to healthy persons the majority of asthma and heart patients were under strict medical supervision. Because of that supervi- sion and the increased sensitivity of these patients, this group has a significant lower risk of dying of pneu- monia. However, this correlation does not have causal character and therefore should not be taken as basis for the decision on pneumonia therapy.

Improvement of the system: The first step towards improving an AI system is to understand it's weak- nesses. Obviously, it's more difficult to perform such weakness analysis on black box models than on mod- els which are interpretable. Also detecting biases in the model or the dataset (as in the pneumonia exam- ple) is easier if one understands what the model is do- ing and why it arrives at it's predictions. Furthermore, model interpretability can be helpful when comparing different models or architectures. For instance, the au- thors of [20, 2, 3] observed that models may have the same classification performance, but largely differ in terms of what features they use as the basis for their decisions. These works demonstrate that the identi- fication of the most "appropriate" model requires ex- plainability. One can even claim that the better we understand what our models are doing (and why they sometimes fail), the easier it becomes to improve them.

Learning from the system: Because today's AI sys- tems are trained with Millions of examples, they may observe patterns in the data which are not accessible to humans, who are only capable of learning with a limited number of examples. When using explainable AI systems we can try to extract this distilled knowl- edge from the AI system in order to acquire new in- sights. One example of such knowledge transfer from AI system to human was mentioned by Fan Hui in the quote above. The AI system identifies new strate- gies to play Go, which certainly now have also been adapted by professional human players. Another do- main where information extraction from the model can be crucial are the sciences. To put it simple, physi- cists, chemists and biologists are rather interested in identifying the hidden laws of nature than just predict- ing some quantity with black box models. Thus, only models which are explainable are useful in this domain (c.f., [37, 32]).

Compliance to legislation: AI systems are affecting more and more areas of our daily life. With that also le- gal aspects, e.g., the assignment of responsibility when the systems makes a wrong decision, have recently re- ceived increased attention. Since it may be impossible to find satisfactory answers for these legal questions when relying on black box models, future AI systems will necessarily have to become more explainable. An- other example where regulations may become a driv- ing force for more explainability in artificial intelli- gence are individual rights. Persons immediately af- fected by decisions of an AI system (e.g., persons re- jected for loan by the bank) may want to know why the systems has decided in this way. Only explainable AI systems will provide this information. These concerns brought the European Union to adapt new regulations which implement a "right to explanation" whereby a user can ask for an explanation of an algorithmic deci- sion that was made about her or him [13].

These examples demonstrate that explainability is not only of important and topical academic interest, but it will play a pivotal role in future AI systems.

3. METHODS FOR VISUALIZING, INTERPRETING AND EXPLAINING DEEP LEARNING MODELS

This section introduces two popular techniques for explain- ing predictions of deep learning models. The process of ex- planation is summarized in Fig. 1. First, the system correctly classifies the input image as "rooster". Then, an explanation method is applied to explain the prediction in terms of input variables. The result of this explanation process is a heatmap visualizing the importance of each pixel for the prediction. In this example the rooster's red comb and wattle are the basis for the AI system's decision. for the prediction "rooster". The presence of yellow flowers is certainly not indicative of the presence of a rooster in the image. Because of this property SA does not perform well in the quantitative evaluation experiments presented in Section

5. More discussion on the drawbacks of sensitivity analysis can be found in [27].

3.2. Layer-Wise Relevance Propagation

In the following we provide a general framework for de- composing predictions of modern AI systems, e.g., feed- forwards neural networks and bag-of-words models [5], long-short term memory (LSTM) networks [4] and Fisher Vector classifiers [20], in terms of input variables. In con- trast to sensitivity analysis, this method explains predictions relative to the state of maximum uncertainty, i.e., it iden- tifies pixels which are pivotal for the prediction "rooster". Recent work [26] also shows close relations to Taylor de- composition, which is a general function analysis tool in mathematics.

A recent technique called Layer-wise relevance propagation (LRP) [5] explains the classifier's decisions by decompo- sition. Mathematically, it redistributes the prediction f (x) backwards using local redistribution rules until it assigns a relevance score Ri to each input variable (e.g., image pixel). The key property of this redistribution process is referred to as relevance conservation and can be summarized as

3.1. Sensitivity Analysis

The first method is known as sensitivity analysis (SA) [6, 35] and explains a prediction based on the model's locally evalu- ated gradient (partial derivative). Mathematically, sensitivity analysis quantifies the importance of each input variable i (e.g., image pixel) as This property says that at every step of the redistribution pro- cess (e.g., at every layer of a deep neural network), the total amount of relevance (i.e., the prediction f (x)) is conserved. No relevance is artificially added or removed during redistri- bution. The relevance scores Ri of each input variable de- termines how much this variable has contributed to the pre- diction. Thus, in contrast to sensitivity analysis, LRP truly decomposes the function value f (x).

In the following we describe the LRP redistribution process for feed-forward neural networks, redistribution procedures

This measure assumes that the most relevant input features are those to which the output is most sensitive. In contrast to the approach presented in the next subsection, sensitivity analysis does not explain the function value f (x) itself, but rather a variation of it. The following example illustrates why measuring the sensitivity of the function may be subop- timal for explaining predictions of AI systems. have also been proposed for other popular models [5, 4, 20].

Let xj be the neuron activations at layer l, Rk be the rele- vance scores associated to the neurons at layer l + 1 and wjk be the weight connecting neuron j to neuron k. The simple LRP rule redistributes relevance from layer l + 1 to layer l in the following way:

4. EXPERIMENTAL EVALUATION

This section evaluates SA and LRP on three different prob- lems, namely the annotation of images, the classificationof text documents and the recognition of human actions in videos.

5.1. Image Classification

In the first experiment we use the GoogleNet model [38], a state-of-the art deep neural network, to classify general ob- jects from the ILSVRC2012 [9] dataset.

Fig. 2 (A) shows two images from this dataset, which have been correctly classified as "volcano" and "coffee cup", re- spectively. The heatmaps visualize the explanations obtained with SA and LRP. The LRP heatmap of the coffee cup im- age shows that the model has identified the ellipsoidal shape of the cup to be a relevant feature for this image category. In the other example, the particular shape of the mountain is regarded as evidence for the presence of a volcano in the image. The SA heatmaps are much noisier than the ones computed with LRP and large values Ri are assigned to re- gions consisting of pure background, e.g., the sky, although these pixels are not really indicative for image category "vol- cano". In contrast to LRP, SA does not indicate how much every pixel contributes to the prediction, but it rather mea- sures the sensitivity of the classifier to changes in the input. Therefore, LRP produces subjectively better explanations of the model's predictions than SA.

The lower part of Fig. 2 (A) displays the results of the per- turbation analysis introduced in Section 4. The y-axis shows the relative decrease of the prediction score average over the first 5040 images of the ILSVRC2012 dataset, i.e., a value of

0.8 means that the original scores decreased on average by 20%. At every perturbation step a 9x9 patch of the image (selected according to SA or LRP scores) is replaced by ran- dom values sampled from an uniform distribution. Since the prediction score decrease is much faster when perturbing the images using LRP heatmaps than when using SA heatmaps, LRP also objectively provides better explanations than SA.

More discussion on this image classification experiment can be found in [31].

5.2. Text Document Classification

To categorize text articles from the 20Newsgroup dataset2, this experiment trained a convolutional neural network based on word embedding.

The document, labeled as "sci.med" (meaning it is presumed to be about a medical issue), is superimposed on top of SA and LRP heatmaps (e.g., a relevance score Ri is given to every word) in Fig. 2 (B).

The SA and LRP explanation approaches both point to terms like "sickness," "body," and "discomfort" as the foundation for this categorization. Unlike sensitivity analysis, LRP differentiates between positive (red) and negative (blue) terms. For example, "sci.med" would be words that support the classification conclusion, while "sci. space" would be words that call for a different category. Words, of course Here is the link to the newsgroups: 2http://qwone.com/□ -jason/20Newsgroups.

Space-related terms, such "ride," "astronaut," and "Shuttle," are more indicative of the subject space than the topic medicine. Even if the classifier chooses the right "sci.med" class, the LRP heatmap shows that there is textual evidence that goes against this judgment. There is no way to distinguish positive evidence from negative evidence using the SA approach.

The quantitative evaluation's outcome is shown in the figure's bottom half. A reduction in the relative accuracy of predictions across 4154 documents in the 20News-group dataset is shown on the y-axis. At each perturbation step, the input values corresponding to the most significant words (as determined by SA or LRP score) are set to 0. Because these heatmaps produce a bigger drop in classification accuracy than SA heatmaps, this finding also quantitatively shows that LRP gives more informative heatmaps than SA.

In [3], we find more explanation of this text document categorization experiment

5.3. Human Action Recognition in Videos

The last examples demonstrates the explanation of a Fisher Vector / SVM classifier [16], which was trained for predict- ing human actions from compressed videos. In order to re- duce computational costs, the classifier was trained on block- wise motion vectors (not individual pixels). The evaluation is performed on the HMDB51 dataset [18].

Fig. 2 (C) shows LRP heatmaps overlayed onto five exemplar frames of a video sample. The video was correctly classified as showing the action "sit-up". One can see that the model mainly focuses on the blocks surrounding the upper body of the person. This makes perfectly sense, as this part of the video frame shows motion which is indicative of the action "sit-up", namely upward and downward movements of the body.

The curve at the bottom of Fig. 2 (C) displays the distribution of relevance over (four consecutive) frames. One can see that the relevance scores are larger for frames in which the person is performing an upwards and downwards movement. Thus, LRP heatmaps not only visualizes the relevant locations of the action within a video frame (i.e., where relevant action happens), but it also identifies the most relevant time points within a video sequence (i.e., when relevant action happens).

More discussion on this experiment can be found in [36].

5. CONCLUSION

The challenge of AI explainability was the focus of this work. We spoke about how black box models couldn't be used in the medical field, for example, since the system's mistakes might have serious consequences. Legal concerns, such as how to allocate blame in the event of a system failure, are emerging with the widespread use of AI systems, and explainability was posited as a necessary condition for their resolution. European law now includes the "right to ex-planation," therefore it may be

Fig. 2. Explaining predictions of AI systems. (A) shows the application of explainable methods to image classification. The SA heatmaps are noisy and difficult to interpret, whereas LRP heatmaps match human intuition. (B) shows the application of explainable methods to text document classification. The SA and LRP heatmaps identify words such as "discomfort", "body" and "sickness" as the relevant ones for explaining the prediction "sci.med". In contrast to sensitivity analysis, LRP distinguishes between positive (red) and negative (blue) relevances. (C) shows explanations for a human action recognition classifier based on motion vector features. The LRP heatmaps of a video which was classified as "sit-up" show increased relevance on frames in which the person is performing an upwards and downwards movement. anticipated that it would also significantly enhance AI systems' explainability.

Aside from serving as a bridge between AI and society, explain-ability is a potent tool for discovering model errors and data biases, validating predictions, improving models, and ultimately, obtaining fresh perspectives on the current issue (for instance, in the scientific realm).

Our next steps will involve delving into the theoretical underpinnings of explainability, specifically exploring the relationship between explainability built into the model's structure and post-hoc explainability, wherein a trained model is provided with the objective of deducing more information about its predictions. Finally, we will investigate novel approaches to elucidating the learnt representation, with a focus on the connection among gener-alizability, compactness, and explainability. Lastly, we're going to use LRP and other explanatory techniques and apply them to other domains, including communications, and look for further uses for these approaches beyond what's in this article.

REFERENCES

The authors of the cited work are [1] Abadi, Agarwal, Barham, Brevdo, Chen, Citro, and others. Tensorflow is a distributed system for large-scale machine learning. arXiv preprint 2016–03–4467-8, published online.

G. Montavon, K.-R. Mueller, L. Arras, and F. Horn Samek, W. Interpreting the results of non-linear classification in natural language processing. Part of the 1st Workshop on Representation Learning for Natural Language Processing, pages 1–7. 2016 ACL.

[3] L. Arras, F. Horn, G. Montavon, K.-R. Voller, and "What is important in a text document?": An approach to interpretable machine learning by W. Samek. 2017. PLoS ONE, 12(8): e0181142.

[4] L. Arras, G. Montavon, K.-R. Mu¨ller, and W. Samek. Determining the meaning of sentiment analysis with recurrent neural network predictions. Publication: 2017 EMNLP Workshop on

Computational Approaches to Subjectiv-ity, Sentiment & Social Media Analysis (WASSA), pages 1–10.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Mu¨ller, and W. Samek. Concerning the layer-wise relevance propagation explanations for pixel-wise non-linear classifier judgments. Publishing in 2015, the PLoS ONE article has the DOI: 10.717/e0130140.

[6] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawan- abe, K. Hansen, and K.-R. Mu¨ller. Ways to clarify the reasoning behind certain categorization choices. The 2010 issue of the Journal of Machine Learning Research, volume 11, pages 1803–1831, is here.

Y. Lou, J. Gehrke, P. Koch, M. Sturm, and R. Caruana

The use of intelligent models in healthcare for the prediction of pneumonia risk and 30-day readmission rates was discussed by N. Elhadad. Presented at the 2015 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, this book covers the years 1721–1730.

This sentence is paraphrased from an article by K. Cho, B. Van Merrie¨nboer, C. Gulcehre, D. Bah-danau, F. Bougares, H. Schwenk, and Y. Bengio. Building statistical machine translation models for learning phrase representations using RNN encoder-decoder. The preprint may be found at arXiv:1406.1078, 2014.

It was written by J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and

A large-scale hierarchical image database: Imagenet, by L. Fei-Fei. Volume 23, Issue 5, 2009, Pages 248–255, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

G. Hinton, B. Kingsbury, and L. Deng [10]. An outline of new kinds of deep neural networks trained for voice recognition and related tasks. Volume 8599, Issue 803, 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing.

[11] B. Kim and F. Doshi-Velez. In the direction of an authoritative study of interpretable ML. 2017, arXiv preprint arXiv:1702.08608, from arXiv.

[12] P. Vincent, D. Erhan, Y. Bengio, and A. Courville. Vi- sualizing higher-layer aspects of a deep network. Report No. 1341, Technical, University of Montreal, 2009.

Goodman, B., and Flaxman, S. (2013). "Right to explanation" and algorithmic decision-making regulations enacted by the European Union. Publication on arXiv:1606.08813 in 2016.

Authors: K. He, X. Zhang, S. Ren, and J. Sun [14]. Advanced RL for picture identification. Volume 7, Issue 7, Pages 770–778 of the 2016 IEEE Conference on Computer Vision and Pattern Recognition proceedings. Junahue, J., Jia, Y., Karayev, S., and Long, J. [15]

Tarell, R., Guadarrama, S., and Girshick, R. A convolutional architecture for rapid feature embedding is Caffe. Pages 675-678 of the 2014 Proceedings of the 22nd Annual ACM International Conference on Multimedia.

I. Laptev and V. Kantorov. Accurate action recognition via efficient feature extraction, encoding, and categorization. This article is part of the 2014 IEEE Computer Vision and Pattern Recognition Conference Proceedings, volume 25, pages 2593–2600.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Suk-thankar, and L. Fei-Fei were the authors of the article [17]. Implementing convolutional neural networks for large-scale video classification? Pages 1725–1732 in the 2014 IEEE Computer Vision and Pattern Recognition (CVPR) conference proceedings.

P. Poggio, H. Kuehne, E. Garrote, H. Jhuang, and [18]

Serre, T. Hmdb is a massive video database that can recognize human movements. This is published in the ICCV proceedings, which cover pages 2556–2563. 2011. Published by IEEE.

[19] B. M. A. Bettencourt, W. Landecker, and M. D. Thomure

S. P. Brumby, G. T. Kenyon, and M. Mitchell respectively. Deciphering distinct categorizations of network hierarchies. Published in 2013 in the proceedings of the IEEE Conference on Information and Data Mining (CIDM), pages 32–38.

[20] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Mu¨ller, and W. Samek. Examining classifiers: deep neural networks and Fisher vectors. The 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) proceedings, pp 2912-2920.

[21] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Mu¨ller, and W. Samek. The toolset for artificial neural networks that propagates relevance layer-wise. Volume 17, Issue 11, Pages 1–5, 2016: Journal of Machine Learning Research.

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Muller were the authors of the provided article. Backprop that works well. On pages 9–48 of Neural Networks: Insider Tips. Press, 2012.

[23] The legend around the interpretability of models by Z. C. Lipton. publication on arXiv:1606.03490, 2016.

[24] Published by A. Mahendran and A. Vedaldi. Gaining insight into implicit representations using inversion. Volume 5, Issue 5, Pages 5188–5196, 2015, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

For example, [25] A. Mahendran and A. Vedaldi. Deep convolutional neural networks shown using natural pre-images. The 2016 edition of the International Journal of Computer Vision, volume 120, issue 3, pages 233–255.

A. Binder, W. Samek, G. Montavon, S. Bach, and K. Using deep taylor decomposition to explain nonlinear classification judgments, R. Mueller wrote. Recognition of Patterns, 65: 211–222, 2017.