EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS

Wojciech Samek[1], *Thomas Wiegand*[1,2], *Klaus-Robert Müller*[2,3,4]

[1]Dept. of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[2]Dept. of Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[3]Dept. of Brain & Cognitive Engineering, Korea University, Seoul 136-713, South Korea
[4]Max Planck Institute for Informatics, Saarbrücken 66123, Germany

ABSTRACT

With the availability of large databases and recent improvements in deep learning methodology, the performance of AI systems is reaching or even exceeding the human level on an increasing number of complex tasks. Impressive examples of this development can be found in domains such as image classification, sentiment analysis, speech understanding or strategic game playing. However, because of their nested non-linear structure, these highly successful machine learning and artificial intelligence models are usually applied in a black box manner, i.e., no information is provided about what exactly makes them arrive at their predictions. Since this lack of transparency can be a major drawback, e.g., in medical applications, the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention. This paper summarizes recent developments in this field and makes a plea for more interpretability in artificial intelligence. Furthermore, it presents two approaches to explaining predictions of deep learning models, one method which computes the sensitivity of the prediction with respect to changes in the input and one approach which meaningfully decomposes the decision in terms of the input variables. These methods are evaluated on three classification tasks.

Index Terms— Artificial intelligence, deep neural networks, black box models, interpretability, sensitivity analysis, layer-wise relevance propagation

1. INTRODUCTION

The field of machine learning and artificial intelligence has progressed over the last decades. A driving force for this development were earlier improvements in support vector machines and more recent improvements in deep learning methodology [22]. Also the availability of large databases such as ImageNet [9] or Sports1M [17], the speed-up gains obtained with powerful GPU cards and the high flexibility of software frameworks such as Caffe [15] or TensorFlow [1] were crucial factors to success. Today's machine learning-based AI systems excel in a number of complex tasks ranging from the detection of objects in images [14] and the understanding of natural languages [8] to the processing of speech signals [10]. On top of that, recent AI[1] systems can even outplay professional human players in difficult strategic games such as Go [34] and Texas hold'em poker [28]. These immense successes of AI systems, especially deep learning models, show the revolutionary character of this technology, which will have a large impact beyond the academic world and will also give rise to disruptive changes in industries and societies.

However, although these models reach impressive prediction accuracies, their nested non-linear structure makes them highly non-transparent, i.e., it is not clear what information in the input data makes them actually arrive at their decisions. Therefore these models are typically regarded as *black boxes*. The 37th move in the second game of the historic Go match between Lee Sedol, a top Go player, and AlphaGo, an artificial intelligence system built by DeepMind, demonstrates the non-transparency of the AI system. AlphaGo played a move which was totally unexpected and which was commented on by a Go expert in the following way:

> *"It's not a human move. I've never seen a human play this move."* (Fan Hui, 2016).

Although during the match it was unclear why the system played this move, it was the deciding move for AlphaGo to win the game. In this case the black box character of the AlphaGo did not matter, but in many applications the impossibility of understanding and validating the decision process of an AI system is a clear drawback. For instance, in medical diagnosis, it would be irresponsible to trust predictions of a black box system by default. Instead every far reaching decision should be made accessible for appropriate validation by a human expert. Also in self-driving cars, where a single incorrect prediction can be very costly, the reliance of the model on the right features must be guaranteed. The use of explainable and human interpretable AI models is a prerequisite for providing such a guarantee. More discussion on the

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC (01IS14013A). We thank Grégore Montavon for his valuable comments on the paper.

[1]The terms artificial intelligence and machine learning are used synonymously.

necessity of explainable AI can be found in Section 2.

Not surprisingly, the development of techniques for "opening" black box models has recently received a lot of attention in the community [6, 35, 39, 5, 33, 25, 23, 30, 40, 11, 27]. This includes the development of methods which help to better understand what the model has learned (i.e., it's representation) [12, 24, 29] as well as techniques for explaining individual predictions [19, 35, 39, 5, 26]. A tutorial on methods from these two categories can be found in [27]. Note that explainability is also important for support vector machines and other advanced machine learning techniques beyond neural networks [20].

The main goal of this paper is to foster awareness for the necessity of explainability in machine learning and artificial intelligence. This is done in Section 2. After that in Section 3, we present two recent techniques, namely sensitivity analysis (SA) [6, 35] and layer-wise relevance propagation (LRP) [5], for explaining the individual predictions of an AI model in terms of input variables. The question of how to objectively evaluate the quality of explanations is addressed in Section 4 and results from image, text and video classification experiments are presented in Section 5. The paper concludes with an outlook on future work in Section 6.

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 doctor 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 important ones:

- **Verification of the system**: As mentioned before, in many applications one must not trust a black box system 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 pneumonia risk of a person arrives at totally wrong conclusions. The application of this model in a black box manner would not reduce but rather increase the number 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 recognize that this can not be true as asthma and hearth problems are factors which negatively affect the prognosis for recovery. However, the AI model does not know anything about asthma or pneumonia, it just infers from data. In this example, the data were systematically biased, because in contrast to healthy persons the majority of asthma and heart patients were under strict medical supervision. Because of that supervision and the increased sensitivity of these patients, this group has a significant lower risk of dying of pneumonia. 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 its weaknesses. Obviously, it is more difficult to perform such weakness analysis on black box models than on models which are interpretable. Also detecting biases in the model or the dataset (as in the pneumonia example) is easier if one understands what the model is doing and why it arrives at its predictions. Furthermore, model interpretability can be helpful when comparing different models or architectures. For instance, the authors 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 identification of the most "appropriate" model requires explainability. 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 systems 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 knowledge from the AI system in order to acquire new insights. 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 strategies to play Go, which certainly now have also been adapted by professional human players. Another domain where information extraction from the model can be crucial are the sciences. To put it simple, physicists, chemists and biologists are rather interested in identifying the hidden laws of nature than just predicting 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 legal aspects, e.g., the assignment of responsibility when the systems makes a wrong decision, have recently received 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. Another example where regulations may become a driving force for more explainability in artificial intelligence are individual rights. Persons immediately affected by decisions of an AI system (e.g., persons rejected 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 decision 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 explaining predictions of deep learning models. The process of explanation 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.

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 evaluated gradient (partial derivative). Mathematically, sensitivity analysis quantifies the importance of each input variable i (e.g., image pixel) as

$$R_i = \Big|\Big|\frac{\partial}{\partial x_i} f(\mathbf{x})\Big|\Big|.$$

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 $f(\mathbf{x})$ itself, but rather a *variation* of it. The following example illustrates why measuring the sensitivity of the function may be suboptimal for explaining predictions of AI systems.

A heatmap computed with sensitivity analysis indicates which pixels need to be changed to make the image look (from the AI system's perspective) more/less like the predicted class. For instance, in the example shown in Fig. 1 these pixels would be the yellow flowers which occlude part of the rooster. Changing these pixels in a specific way would reconstruct the occluded parts of the rooster, which most probably would also increase the classification score, because more of the rooster would be visible in the image. Note that such heatmap would not indicate which pixels are actually pivotal 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 decomposing 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 contrast to sensitivity analysis, this method explains predictions relative to the state of maximum uncertainty, i.e., it identifies pixels which are pivotal for the prediction "rooster". Recent work [26] also shows close relations to Taylor decomposition, 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 decomposition. Mathematically, it redistributes the prediction $f(\mathbf{x})$ backwards using local redistribution rules until it assigns a relevance score R_i 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

$$\sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x}) \quad (1)$$

This property says that at every step of the redistribution process (e.g., at every layer of a deep neural network), the total amount of relevance (i.e., the prediction $f(\mathbf{x})$) is conserved. No relevance is artificially added or removed during redistribution. The relevance scores R_i of each input variable determines how much this variable has contributed to the prediction. Thus, in contrast to sensitivity analysis, LRP truly decomposes the function value $f(\mathbf{x})$.

In the following we describe the LRP redistribution process for feed-forward neural networks, redistribution procedures have also been proposed for other popular models [5, 4, 20].

Let x_j be the neuron activations at layer l, R_k be the relevance scores associated to the neurons at layer $l+1$ and w_{jk} 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:

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k \quad (2)$$

where a small stabilization term ϵ is added to prevent division by zero. Intuitively, this rule redistributes relevance proportionally from layer $l+1$ to each neuron in layer l based on two criteria, namely (i) the neuron activation x_j, i.e., more activated neurons receive a larger share of relevance, and (ii) the strength of the connection w_{jk}, i.e., more relevance flows

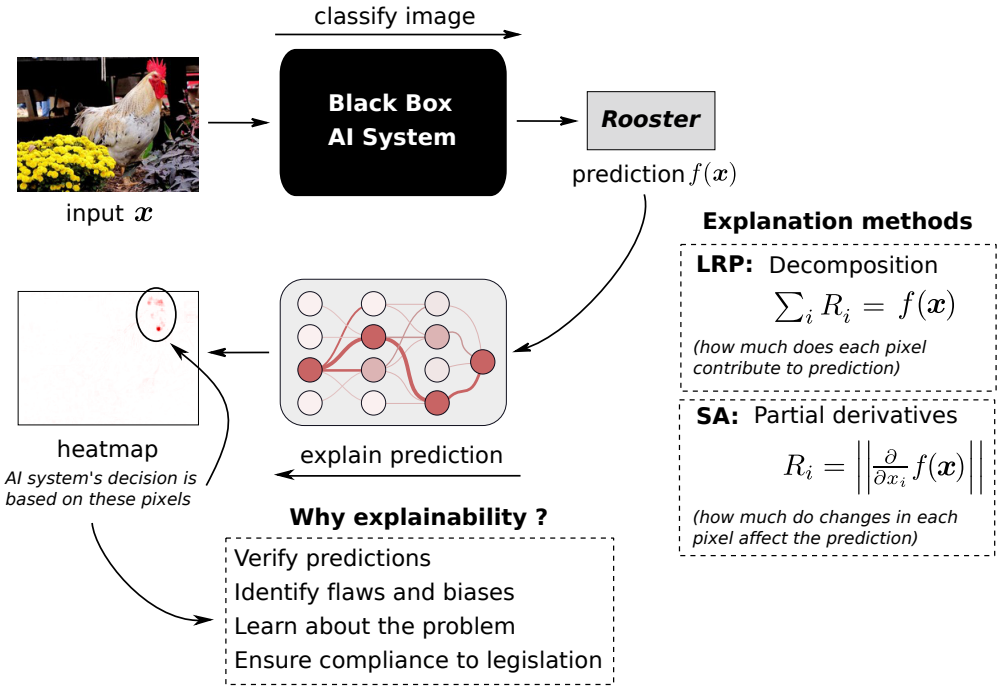

Fig. 1. Explaining predictions of an AI system. The input image is correctly classified as "rooster". In order to understand why the system has arrived at this decision, explanation methods such as SA or LRP are applied. The result of this explanation is an image, the heatmap, which visualizes the importance 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. With the heatmap one can verify that the AI system works as intended.

through more prominent connections. Note that relevance conservation holds for $\epsilon = 0$.

The "alpha-beta" rule is an alternative redistribution rule introduced in [5]:

$$R_j = \sum_k \Big(\alpha \cdot \frac{(x_j w_{jk})^+}{\sum_j (x_j w_{jk})^+} - \beta \cdot \frac{(x_j w_{jk})^-}{\sum_j (x_j w_{jk})^-}\Big) R_k \quad (3)$$

where $()^+$ and $()^-$ denote the positive and negative parts, respectively. The conservation of relevance is enforced by an additional constraint $\alpha - \beta = 1$. For the special case $\alpha = 1$, the authors of [26] showed that this redistribution rule coincides with a "deep Taylor decomposition" of the neural network function when the neural network is composed of ReLU neurons.

3.3. Software

The LRP toolbox [21] provides a python and matlab implementation of the method as well as an integration into popular frameworks such as Caffe and TensorFlow. With this toolbox one can directly apply LRP to other peoples' models. The toolbox code, online demonstrators and further information can be found on `www.explain-ai.org`.

4. EVALUATING THE QUALITY OF EXPLANATIONS

In order to compare heatmaps produced by different explanation methods, e.g., SA and LRP, one needs an objective measure of the quality of explanations. The authors of [31] proposed such a quality measure based on perturbation analysis. The method is based on the following three ideas:

- The perturbation of input variables which are highly important for the prediction leads to a steeper decline of the prediction score than the perturbation of input dimensions which are of lesser importance.
- Explanation methods such as SA and LRP provide a score for every input variable. Thus, the input variables can be sorted according to this relevance score.
- One can iteratively perturb input variables (starting from the most relevant ones) and track the prediction score after every perturbation step. The average decline of the prediction score (or the decline of the prediction accuracy) can be used as an objective measure of explanation quality, because a large decline indicates that the explanation method was successful in identifying the truly relevant input variables.

In the following evaluation we use model-independent perturbations (e.g., replacing the input values by random sample from uniform distribution) in order to avoid biases.

5. EXPERIMENTAL EVALUATION

This section evaluates SA and LRP on three different problems, namely the annotation of images, the classification of 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 objects from the ILSVRC2012 [9] dataset.

Fig. 2 (A) shows two images from this dataset, which have been correctly classified as "volcano" and "coffee cup", respectively. The heatmaps visualize the explanations obtained with SA and LRP. The LRP heatmap of the coffee cup image 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 R_i are assigned to regions consisting of pure background, e.g., the sky, although these pixels are not really indicative for image category "volcano". In contrast to LRP, SA does not indicate how much every pixel contributes to the prediction, but it rather measures 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 perturbation 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 random 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

In this experiment, a word-embedding based convolutional neural network was trained to classify text documents from the 20Newsgroup dataset[2].

Fig. 2 (B) shows SA and LRP heatmaps (e.g., a relevance score R_i is assigned to every word) overlayed on top of a document, which was classified as topic "sci.med", i.e., the text is assumed to be about a medical topic. Both explanation methods, SA and LRP, indicate that words such as "sickness", "body" or "discomfort" are the basis for this classification decision. In contrast to sensitivity analysis, LRP distinguishes between positive (red) and negative (blue) words, i.e., words which support the classification decision "sci.med" and words which are in contradiction, i.e., speak for another category (e.g.,"sci.space"). Obviously, words such as "ride", "astronaut" and "Shuttle" strongly speak for the topic space, but not necessarily for the topic medicine. With the LRP heatmap, we can see that although the classifier decides for the correct "sci.med" class, there is evidence in the text which contradicts this decision. The SA method does not distinguish between positive and negative evidence.

The lower part of the figure shows the result of the quantitative evaluation. The y-axis displays the relative decrease of the prediction accuracy over 4154 documents of the 20Newsgroup dataset. At every perturbation step, the most important words (according to SA or LRP score) are deleted by setting the corresponding input values to 0. Also this result confirms quantitatively that LRP provides more informative heatmaps than SA, because these heatmaps lead to a larger decrease in classification accuracy compared to SA heatmaps.

More discussion on this text document classification experiment can be found in [3].

5.3. Human Action Recognition in Videos

The last example demonstrates the explanation of a Fisher Vector / SVM classifier [16], which was trained for predicting human actions from compressed videos. In order to reduce 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 visualize the relevant locations of the action within a video frame (i.e., *where* the relevant action happens), but they also identify the most relevant time points within a video sequence (i.e., *when* the relevant action happens).

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

6. CONCLUSION

This paper approached the problem of explainability in artificial intelligence. It was discussed why black box models are not acceptable for certain applications, e.g., in the medical domain where wrong decisions of the system can be very

[2] http://qwone.com/~jason/20Newsgroups

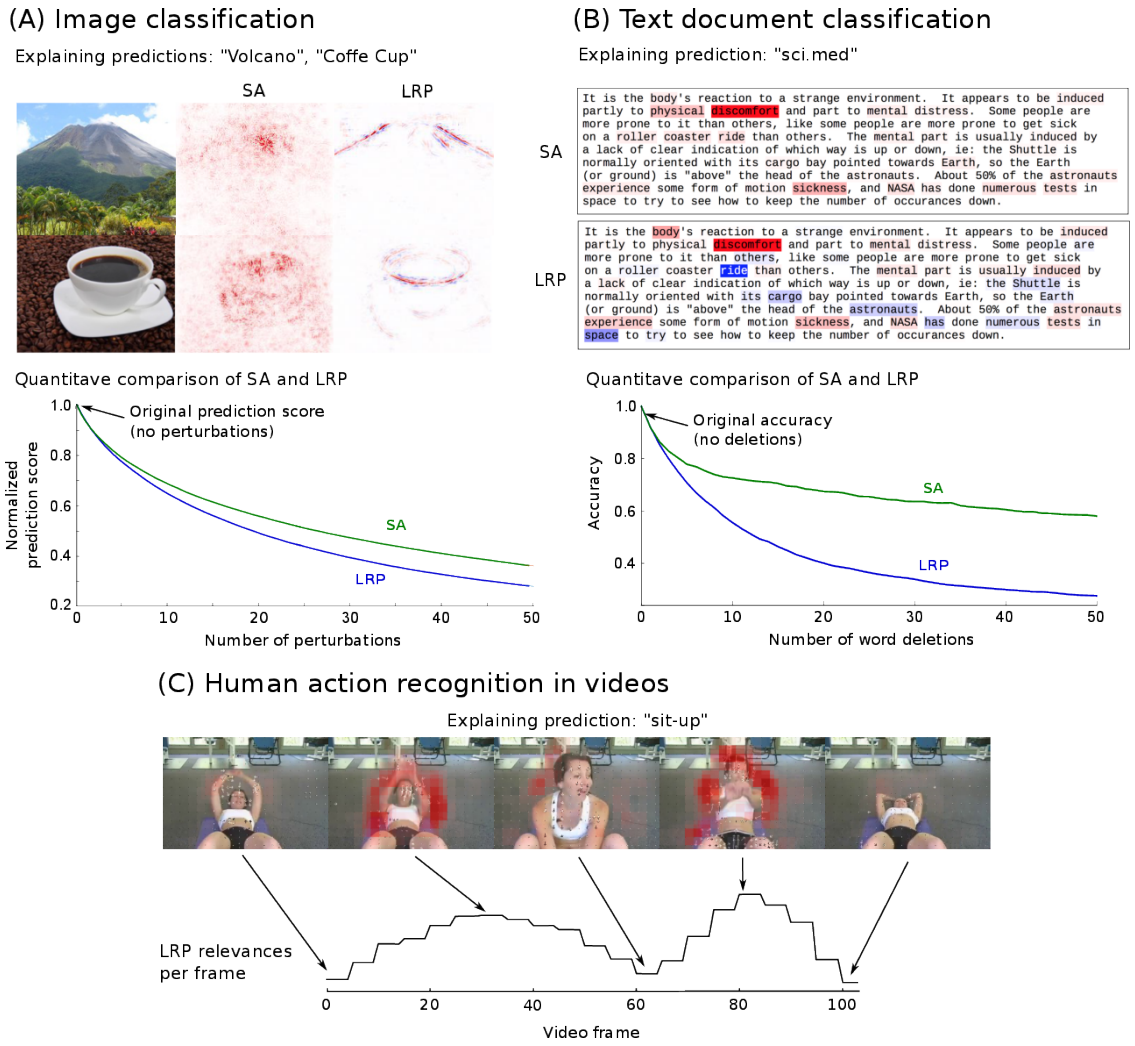

Fig. 2. Explaining predictions of AI systems. (A) shows the application of explainable methods to image classification. The SA heatmaps are noisy and difficult to interpret, whereas LRP heatmaps match human intuition. (B) shows the application of explainable methods to text document classification. The 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.

harmful. Furthermore, explainability was presented as prerequisite for solving legal questions which are arising with the increased usage of AI systems, e.g., how to assign responsibility in case of system failure. Since the "right to explanation" has become part of the European law, it can be expected that it will also greatly foster explainability in AI systems.

Besides being a gateway between AI and society, explainability is also a powerful tool for detecting flaws in the model and biases in the data, for verifying predictions, for improving models, and finally for gaining new insights into the problem at hand (e.g., in the sciences).

In future work we will investigate the theoretical foundations of explainability, in particular the connection between post-hoc explainability, i.e., a trained model is given and the goal is to explain it's predictions, and explainability which is incorporated directly into the structure of the model. Furthermore, we will study new ways to better understand the learned representation, especially the relation between generalizability, compactness and explainability. Finally, we will apply explaining methods such as LRP to new domains, e.g., communications, and search for applications of these methods beyond the ones described in this paper.

7. REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. Explaining predictions of non-linear classifiers in nlp. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7. ACL, 2016.

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

[4] L. Arras, G. Montavon, K.-R. Müller, and W. Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, pages 159–168, 2017.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015.

[6] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[10] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8599–8603, 2013.

[11] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[12] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, 2009.

[13] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[16] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2593–2600, 2014.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014.

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563. IEEE, 2011.

[19] W. Landecker, M. D. Thomure, L. M. A. Bettencourt, M. Mitchell, G. T. Kenyon, and S. P. Brumby. Interpreting individual classifications of hierarchical networks. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 32–38, 2013.

[20] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2912–2920, 2016.

[21] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. The layer-wise relevance propagation toolbox for artificial neural networks. *Journal of Machine Learning Research*, 17(114):1–5, 2016.

[22] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[23] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[24] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196, 2015.

[25] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[26] G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[27] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *arXiv preprint arXiv:1706.07979*, 2017.

[28] M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, et al. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[29] A. Nguyen, J. Yosinski, and J. Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[31] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. in press.

[32] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017.

[33] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[34] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[36] V. Srinivasan, S. Lapuschkin, C. Hellge, K.-R. Müller, and W. Samek. Interpretable human action recognition in compressed domain. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1692–1696, 2017.

[37] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[39] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference Computer Vision - ECCV 2014*, pages 818–833, 2014.

[40] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.