RESEARCH NEWS

Making artificial intelligence explainable

A look inside neural networks

Artificial intelligence (AI) is already firmly embedded in our everyday lives and is conquering more and more territory. For example, voice assistants are already an everyday item in many people's smartphones, cars and homes. Progress in the field of AI is based primarily on the use of neural networks. Mimicking the functionality of the human brain, neural networks link mathematically defined units with one another. But in the past it was not known just how a neural network makes decisions. Researchers at the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, HHI and Technische Universität Berlin have developed a technology that reveals the criteria AI systems use when making decisions. The innovative Spectral Relevance Analysis (SpRAy) method based on Layer-wise Relevance Propagation technology provides a first peek inside the "black box".

Today it's almost impossible to find an area in which artificial intelligence is irrelevant, whether in manufacturing, advertising or communications. Many companies use learning and networked AI systems, for example to generate precise demand forecasts and to exactly predict customer behavior. This approach can also be used to adjust regional logistics processes. Healthcare also uses specific AI activities, such as prognosis generation on the basis of structured data. This plays a role for example in image recognition: X-ray images are input into an AI system which then outputs a diagnosis. Proper detection of image content is also crucial to autonomous driving, where traffic signs, trees, pedestrians and cyclists have to be identified with complete accuracy. And this is the crux of the matter: AI systems have to provide absolutely reliable problem-solving strategies in sensitive application areas such as medicinal diagnostics and in security-critical areas. However, in the past is hasn't been entirely clear how AI systems make decisions. Furthermore, the predictions depend on the quality of the input data. Researchers at the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, HHI and Technische Universität Berlin have now developed a technology, Layer-wise Relevance Propagation (LRP), which renders the AI forecasts explainable and in doing so reveals unreliable problem solution strategies. A further development of LRP technology, referred to as Spectral Relevance Analysis (SpRAy), identifies and quantifies a broad spectrum of learned decision-making behaviors and thus identifies undesirable decisions even in enormous datasets.

Transparent AI

In practice the technology identifies the individual input elements which have been used to make a prediction. Thus for example when an image of a tissue sample is input into an AI system, the influence of each individual pixel is quantified in the classification results. In other words, as well as predicting how “malignant” or “benign” the imaged tissue is, the system also provides information on the basis for this classification. “Not only is the result supposed to be correct, the solution strategy is as well. In the past, AI systems have been treated as black boxes. The systems were trusted to do the right things. With our open-source software, which uses Layer-Wise Relevance Propagation, we’ve succeeded in rendering the solution-finding process of AI systems transparent,” says Dr. Wojciech Samek, head of the "Machine Learning” research group at Fraunhofer HHI. “We’re using LRP to visualize and interpret neural networks and other machine learning models. We use LRP to measure the influence of every input variable in the overall prediction and parse the decisions made by the classifiers,” adds Dr. Klaus-Robert Müller, Professor for Machine Learning at TU Berlin.

Unreliable solution strategies

Trusting the results of neural networks necessarily means understanding how they work. According to the research team’s tests, AI systems don’t always apply the best strategies to reach a solution. For example, one well-known AI system classifies images based on context. It allocated photographs to the category ‘Ship’ when a large amount of water was visible in the picture. It wasn’t solving the actual task of recognizing images of ships, even if in the majority of cases it picked out the right photos. “Many AI algorithms use unreliable strategies and arrive at highly impractical solutions,” says Samek, summarizing the results of the investigations.

Watching neural networks think

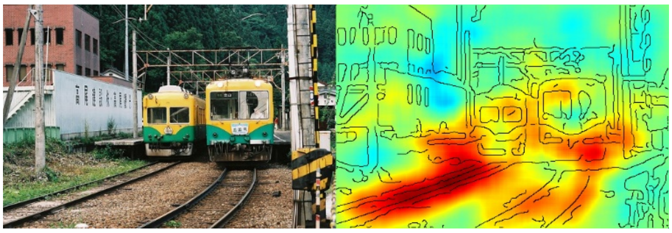
The LRP technology decodes the functionality of neural networks and finds out which characteristic features are used, for example to identify a horse as a horse and not as a donkey or a cow. It identifies the information flowing through the system at each node of the network. This makes it possible to investigate even very deep neural networks.

The Fraunhofer HHI and TU Berlin research teams are currently formulating new algorithms for the investigation of further questions in order to make AI systems even more reliable and robust. The project partners have published their research results in the journal Nature Communications (see link below).

Publication in Nature Communications:
https://www.nature.com/articles/s41467-019-08987-4

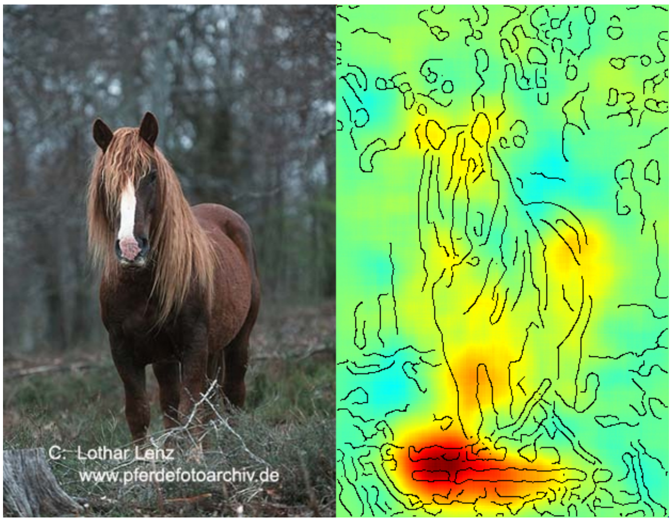
AI, machine learning and more

Artificial intelligence is concerned with the development of systems that can independently solve problems and act analogously to patterns of human thought and behavior. At present the greatest progress is being made in the area of machine learning, a subfield of AI. Machine learning deals with methods of extracting knowledge from data and independently learning contexts contained in the data. The progress is a result of using artificial neural networks based on connections between mathematical calculation units that in principle imitate the neural structure of the human brain. A subfield of machine learning, deep learning, covers a class of new procedures that make it possible to teach and train complex artificial neural networks. These networks consist of a large number of levels which are linked with one another in many-layered structures.



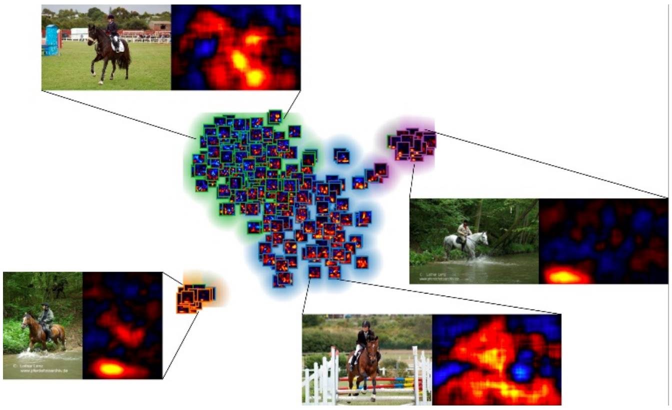
Picture 1: Figure. 1: Here the AI system classifies an image as a train because tracks are present.

© Fraunhofer HHI



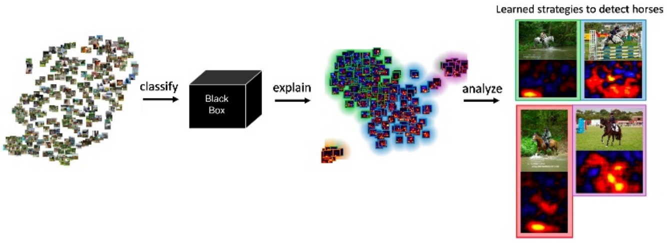
Picture 2: Here the AI system allocates the image to the correct category based on the copyright banner. Nevertheless, the solution strategy is defective.

© Fraunhofer HHI



Picture 3: The new Spectral-wise Relevance Analysis technology renders visible the criteria used by AI systems when making decisions.

© Fraunhofer HHI



Picture 4: Layer-wise Relevance Propagation provides a look inside the "black box".© Fraunhofer HHI

The **Fraunhofer-Gesellschaft** is the leading organization for applied research in Europe. Its research activities are conducted by 72 institutes and research units at locations throughout Germany. The Fraunhofer-Gesellschaft employs a staff of more than 26,600, who work with an annual research budget totaling 2.6 billion euros. Of this sum, 2.2 billion euros is generated through contract research. Around 70 percent of the Fraunhofer-Gesellschaft's contract research revenue is derived from contracts with industry and from publicly financed research projects. International collaborations with excellent research partners and innovative companies around the world ensure direct access to regions of the greatest importance to present and future scientific progress and economic development.