Enhanced insights on deep learning systems through explainable AI

wipro holmes

wipro

With the proliferation of AI-based systems in almost every application, the focus has shifted to getting a better insight on their working mechanism.
This can be best explained by the example of an autonomous vehicle (where the decisions on various maneuvers have to be taken in a split second), such as while in the midst of "not so friendly" traffic, it has to find a place to squeeze in and move forward.
In another example, an AI-controlled Clinical Decision Support System (CDSS) analyzes PET/CT images and associated data, and takes a call on the staging of a patient's cancer. The staging, in turn, dictates the therapy regimen that the patent must undergo.

Numerous such examples have forced the European Union to enact a law on the "right for explanation" that was enforced in May 2018. The law requires the providers of AI solutions to explain their reasoning behind outcomes. This can be best explained with an illustration: If the AI is presented with the image, as seen in Figure 1, the concluding solution should be able to reason out why a cat has been classified as a tiger and not as a cat.

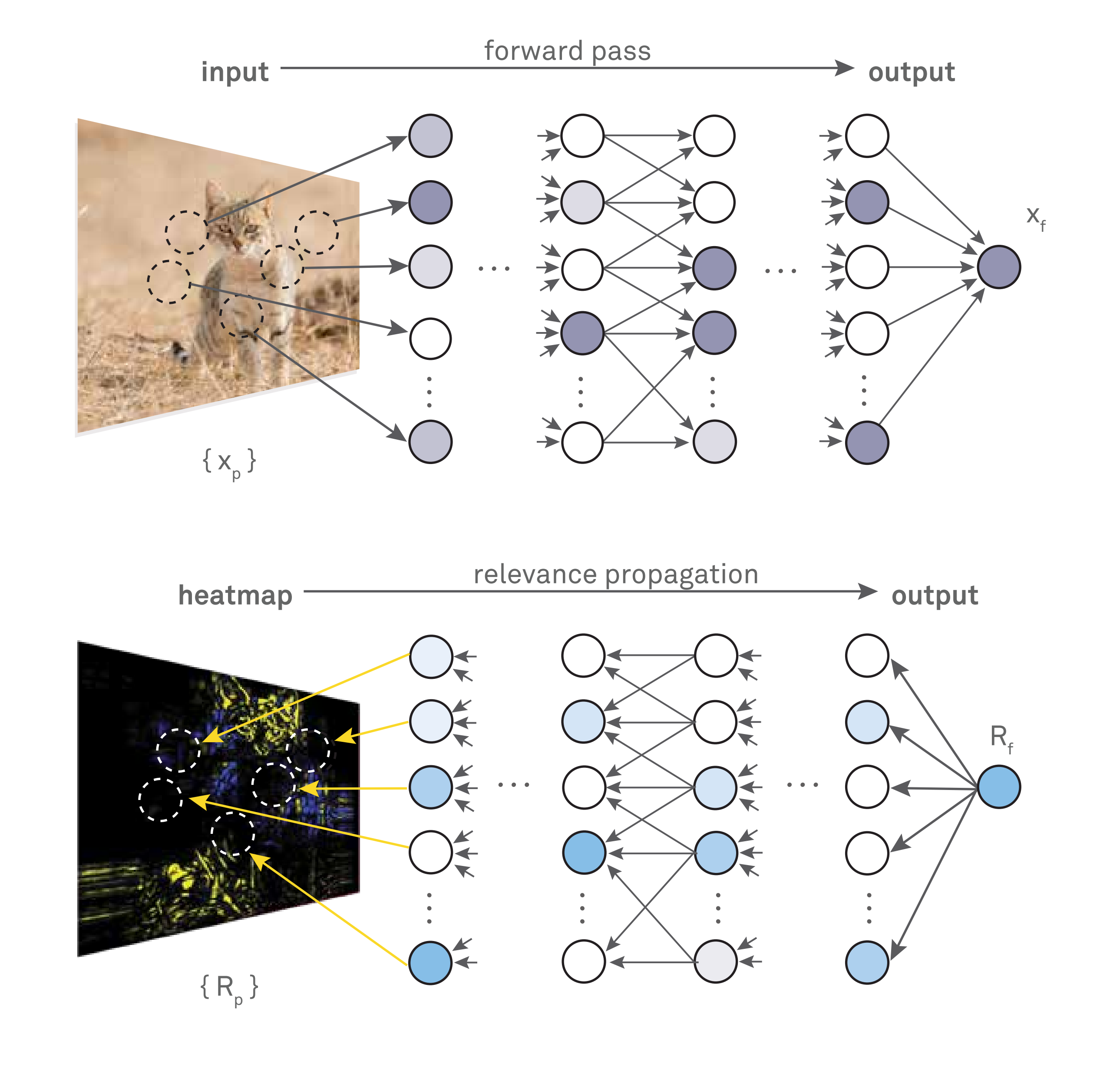

Figure 1: Relevance propagation: a heatmap generation

The explainable AI framework supports image data input as well as text data input with minor changes in the ingestion methodology and the user interface

1. Explainable AI for the image data

In an image data explanation, the explainable AI system has two components: one, to identify the features of the input that results in the said response, and then puts the explanation into words; the second one which opens up the black box model.

Heatmap generation

The heatmap provides visual depiction of the relevance of a feature in the decision making.

In an image classifier, it represents the contribution of the pixels towards a class. The common techniques include:

- **Layer-wise relevance propagation (LRP):** It quantifies the contribution of each pixel towards generating the particular response. It is stable and reveals the proportional contribution of the input regions to the chosen class and is therefore used widely. Figure 2 shows the heatmap of a fighter plane from a classifier, that classifies an input image, as fighter plane or a passenger plane.

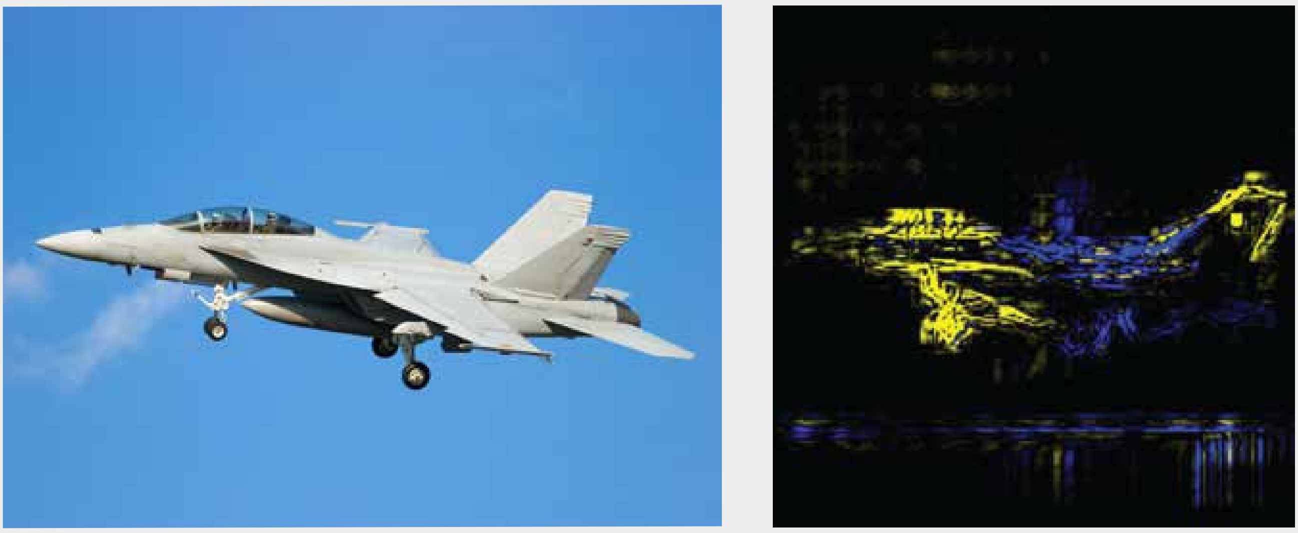

Figure 2: Fighter plane image and the heatmap from the second classifier

- **Sensitivity analysis:** This method provides a change in the response when the weightage of the pixels are changed. It represents the impact made by the deletion of a certain feature in the decision making process. It does not indicate relative importance of the features of the input. It fails to capture the importance or relevance of each input feature or the relation between them.

- **Deconvolution:** This technique provides the heatmap of the matching input pattern for the classified object. It is limited to convolutional neural network.

- **Saliency map:** It provides the heat map for important regions of the input. It is limited to paying attention of the parts contributing to the class, and does not speak about the other areas of input.

Visual explanation generator

It generates the explanation text through LSTM with relevant features derived from the heatmap. The attributes of the features get mapped on to the words in the explanation. In one implementation, the penultimate layer output is taken as the feature and is trained with the explanation.

Visualization of the model

A classifier model in general, is a black box. The tools Deepvis, RNNvis and LSTMvis help to visualize and interactively plot the activation relevance produced in each layer of the classifier. It also depicts the learnings of each neuron.

2. Explainable AI for text data

Explaining predictions made by models in the textual domain is tricky as textual input is transferred to the models in the form of embeddings. These embeddings are obtained by projecting textual units in a vector space of low dimension. The concepts of relevance and heatmap for images have been explained previously. The same is true for text as well. The difference is that, in the case of text, the unit used for mapping the relevance is a textual unit which, in most cases, is a word.

For obtaining wordwise relevance in a chunk of input text for predictions made by the text classification model, the LRP and sensitivity analysis algorithms, as explained above, are primarily used. However, the usage is in the context of words in the text.

Conclusion

Explainable AI paradigm provides better insights into the AI system and helps to open up the black box of deep learning architecture. The framework can provide visual explanation through a heatmap and descriptive explanation through the text.

The explanation provided by the framework can be used for a variety of applications such as debugging, test case designs and optimal designing of the system.

About the authors

Tapati Bandopadhyay
General Manager and Global Practice Head,
Wipro HOLMES™ AI & Automation
Ecosystem, Wipro Ltd.

Tapati Bandopadhyay drives strategy and thought leadership for AI innovations, top line impact metrics, new themes and use-cases, AI for a cause, market making, positioning, ecosystem strategy, use of design thinking in AI solution architecture, ethical AI design principles, etc.

She is a gold medalist in engineering from Jadavpur University, a DFID scholar at the University of Strathclyde, and a PhD in AI.

Vinutha B N
Consulting Partner, CTO Office, Wipro Ltd.

Vinutha heads the CTO research team focusing mainly on the cognitive space. She has more than 27 years' experience in the software industry, out of which she has spent two decades at Wipro. She has worked extensively in embedded domain areas ranging from device drivers, networking, storage and printers, in various customer engagements.

Her main areas of focus are Artificial Intelligence, machine learning and deep learning. She was one of the key members that pioneered Wipro HOLMES™ – Wipro's AI platform.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035, India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**