Understanding neural networks with neural-symmetric integration

People often conceive of Artificial Intelligence (AI) as a ‘black box’ because they find it difficult to understand the knowledge that is hidden inside it. They also find it difficult to understand the reasoning behind the choices that are made by AI. This understanding is required in many situations involving so-called ‘safety-critical’ tasks, such as autonomous driving. The decisions made by AI may have to be audited for insurance purposes, for greater accountability, or for legal challenges. Moreover, developers and engineers may need to understand those AI decisions so that they can fix them and prevent any potential negative outcomes.

Dr Joe Townsend, Dr Theodoros Kasioumis, and Dr Hiroya Inakoshi from the Artificial Intelligence Research Division at Fujitsu Research of Europe Ltd, explain that there are usually two approaches in response to this challenge. The first is ‘opening’ the black box in order to translate what is inside. The alternative involves training the AI model to be interpretable from the beginning. The researchers have developed solutions for both approaches for a type of machine learning paradigm called the Convolutional Neural Network (CNN). These two methods can also be deployed either separately or in unison.

CONVOLUTIONAL NEURAL NETWORKS

A CNN is a deep neural network that is designed to process structured arrays of data such as images. Artificial neural networks imitate aspects of both the structure and function of the human brain. In particular, CNNs are inspired by the visual cortex – the region of the brain that processes visual information. CNNs are great at picking up on patterns such as lines, gradients, circles, and even eyes and faces. This makes them highly suitable for image recognition tasks, where an image is processed by a series of layers that identify progressively more complex features. In CNNs, however, the information learned is distributed across millions of artificial neurons, making it very difficult to interpret.

CNNs are made up of layers, but the layers are not fully connected. They have filters in the form of sets of cube-shaped weights that are applied throughout the image. These filters are often alternately referred to as ‘kernels’ or ‘feature detectors’. The filters are applied to the original image through parameter-sharing and translation invariance, so the same response is produced regardless of how its input is shifted. The convolutional layers contain most of the network’s user-specified parameters, including the number of filters, the size of the filters and the activation function.

TRUSTED AND EXPLAINABLE AI

Public concern regarding the extent to which AI can be trusted has led to increases in demand for trusted and explainable AI. Trusted AI implies that a decision made by an algorithm can be accounted for, is fair, and will cause no harm. But while AI is concerned with understanding and explaining how models trained through machine learning make their decisions, or how they might be designed or trained to be explainable from the outset.

NEURAL–SYMBOLIC INTEGRATION

The field of Neural-Symbolic Integration concerns explainable AI for artificial neural networks, exploring ways of extracting interpretable, symbolic knowledge from trained networks, injecting such knowledge into those networks, or both. For example, if a neural network is trained to classify animal data, an extracted rule might say ‘if it has wings, it’s a bird’. However, the developer might correct this assumption by injecting the fact ‘bats are mammals but have wings’ into the network.
The researchers demonstrate how such as green traffic signs and cars. Considering the road classification task for example, a category of ‘highway’ images is associated with a small set of filters representing semantic concepts that are present on a highway, such as green traffic signs and cars. The researchers demonstrate how associating each category with an interpretable set of sparse neurons enables them to construct more compact rules that accurately explain the CNN decisions without a significant loss of accuracy. An example of one such extracted rule for the highway category could be ‘the image is a highway because there are green traffic signs and cars and no pedestrians’. Compact rules are more interpretable as there is less information for the reader to digest. These rules are in turn can be manipulated by a logic program and approximate the behaviour of the original CNN.

**ERIC already provides a framework for discovering important symbols that have yet to acquire labels but can distinguish between classes.**

**ELITE BACKPROPAGATION**

The research team have developed another solution, Elite BackPropagation (EBP), that involves training the model to be interpretable from the beginning. EBP does this by enforcing class-wise activation sparsity – that is, by training CNNs to associate each category with a handful of ‘elite’ filters that rarely but strongly activate with respect to that category. Images from a specific class will, therefore, be associated with the same group of elite filters.

Considering the road classification task for example, a category of ‘highway’ images is associated with a small set of filters representing semantic concepts that are present on a highway, such as green traffic signs and cars. The researchers demonstrate how associating each category with an interpretable set of sparse neurons enables them to construct more compact rules that accurately explain the CNN decisions without a significant loss of accuracy. An example of one such extracted rule for the highway category could be ‘the image is a highway because there are green traffic signs and cars and no pedestrians’. Compact rules are more interpretable as there is less information for the reader to digest. These rules are in turn can be manipulated by a logic program and approximate the behaviour of the original CNN.

**Research Objectives**

The Artificial Intelligence Research division at Fujitsu Research of Europe Ltd use Neural-Symbolic Integration to better understand the reasoning of Artificial Intelligence.

**Detail**

Address
Fujitsu Research of Europe Limited
4th Floor, Building 3, Hyde Park Hayes
11 Millington Road
Hayes
UB3 4AZ
United Kingdom

Bio
Fujitsu’s AI research is focused on realizing a sustainable society, developing cutting-edge AI technologies that create new value and help transform society and business. Joe joined Fujitsu Research in 2015 following his PhD on the topic of Neural-Symbolic Integration. Theodoros joined in 2019 and has a PhD in Differential Geometry.

**Funding**

The work is conducted under Fujitsu Research’s Trusted AI Project.

Collaborators
Hiroya Inakoshi

**References**
