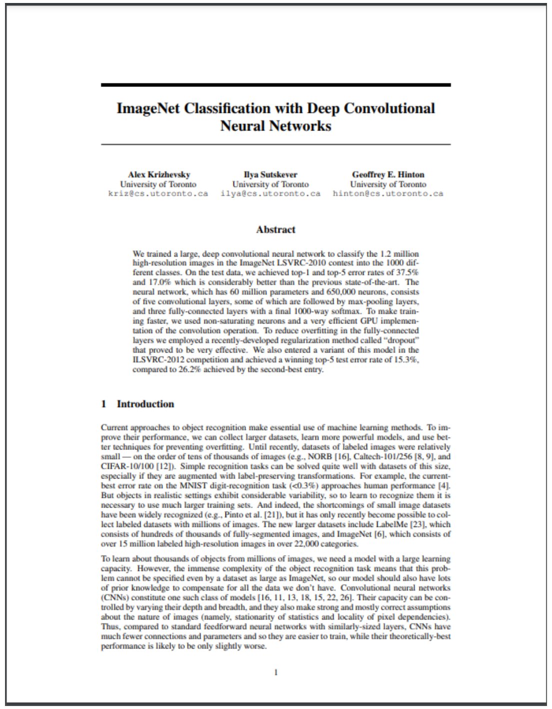
University of Antwerp
Applied Data Mining
Research Group Engineering Management

Explainable Techniques to Contribute to a more Fair and Transparent Tax AI

Dieter Brughmans

The complexity of AI

AlexNet



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

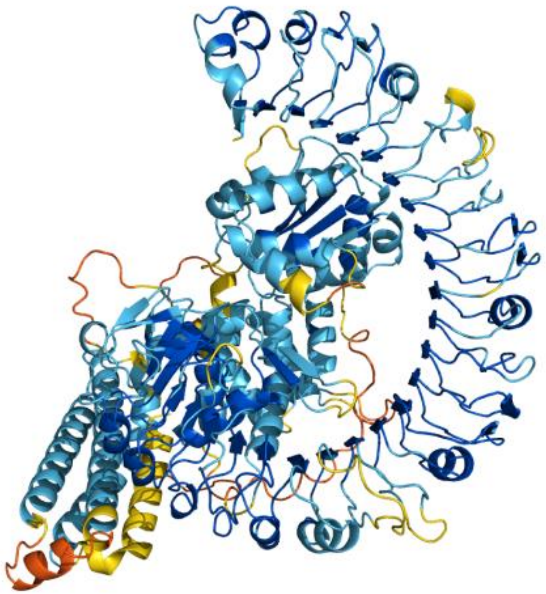
Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

1

61 000 000

AlphaFold



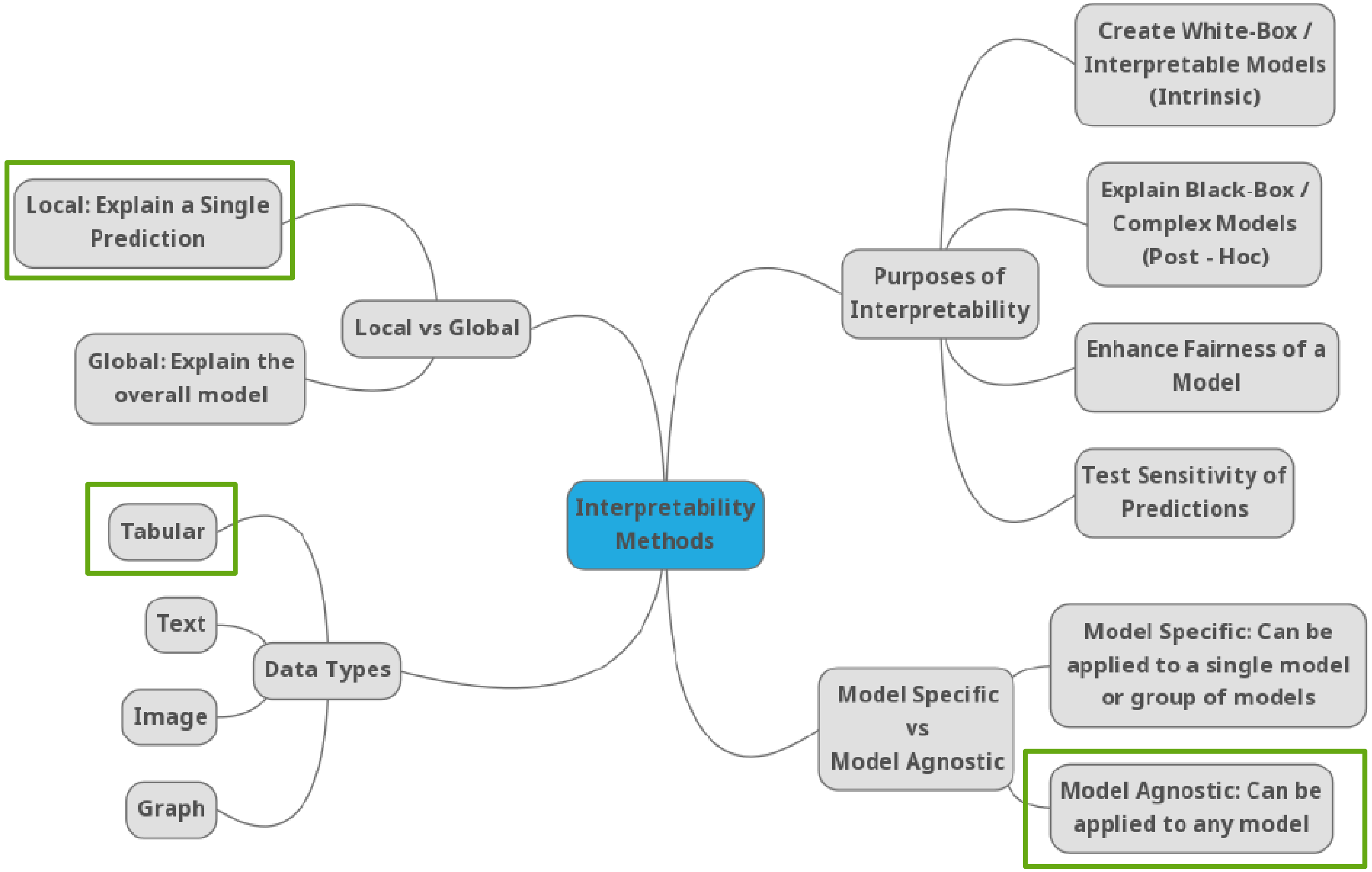
21 000 000

ChatGPT

GPT-3

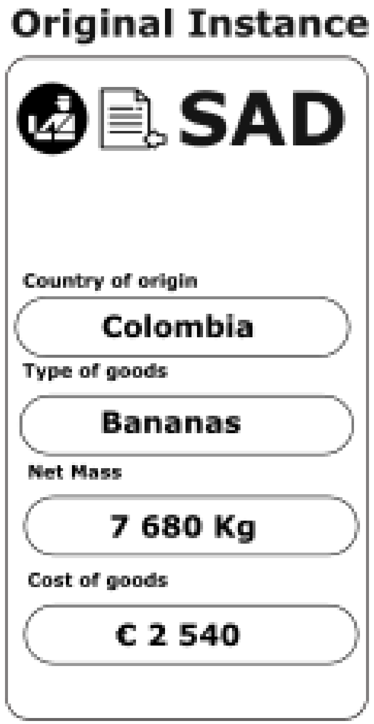
175 000 000 000

XAI Characteristics



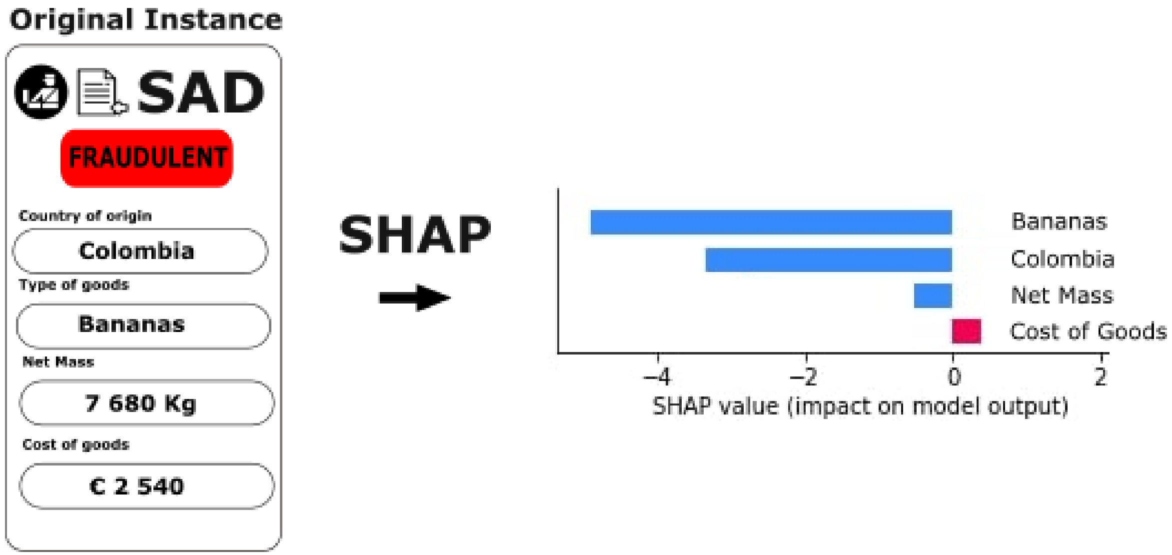
Source*: Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis.* "Explainable ai: A review of machine learning interpretability methods." *Entropy* 23.1 (2020): 18.

The Counterfactual Explanation



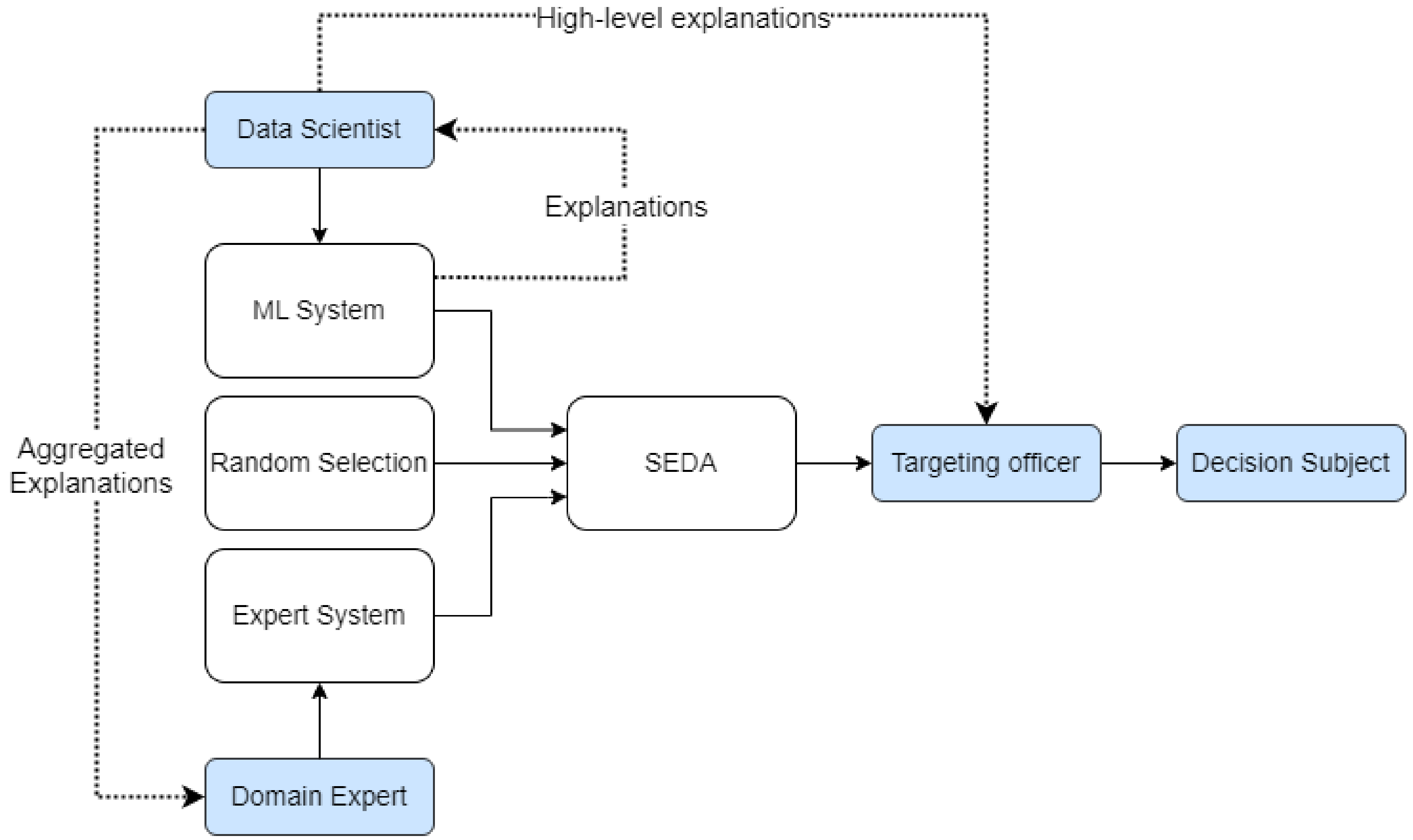
- Why was a declaration predicted as fraudulent?
- The minimal change in input that changes the output.
- "If the country of origin would change from **Colombia** to **USA** and the type of goods from **bananas** to **avocados**, the predicted class would change from **fraudulent** to **compliant**."
- No limitations on complexity of the model

SHAP



- What is the influence of each feature on the prediction score?
- Based on Game Theory
- Explains score, not predicted class
- Can be biased

Counterfactual Explanations for Customs Fraud Detection



Data Scientist

- Model improvement
- Identify Trends

Targeting Officer

- High-level explanations
- Improve investigation time

Domain Expert

- Aggregated explanations
- New insights from data

Decision Subject

- No explanations

Open Issues and Challenges XAI

- Which explanation method to use?
- How to choose among explanations: moral hazard
- What explanations do users want?
- Who should get access to explanations