Minimizing Paralysis of Choice for XAI Methods with multi-metric Evaluation Framework

Due to the ubiquity of AI systems in our society, awareness has been raised for the need of neural networks and their predictions to be transparent and explainable. This has increased the interest in eXplainable AI (XAI) in the machine learning (ML) research community. In recent years, a plethora of XAI methods has been developed. Examples include Saliency, Guided Backprop, Excitation Backprop and Layer-wise Relevance Propagation (LRP). Given an input sample and a ML model, XAI methods compute relevance scores for all input values, e.g., pixels (or voxels) for image classification tasks, which can be presented to the user in the form of heatmaps.

Throughout this post, we use the XAI methods lists above to explain predictions for a VGG-16 model which was trained to classify samples from the ILSVRC2017 dataset. In Fig. 1, explanations for two exemplary input samples are shown. Although the different explanations differ quite a bit, it is hard to tell which heatmap explanation is the "best" one, most accurately describing the model's reasoning when predicting given the input sample on the left.

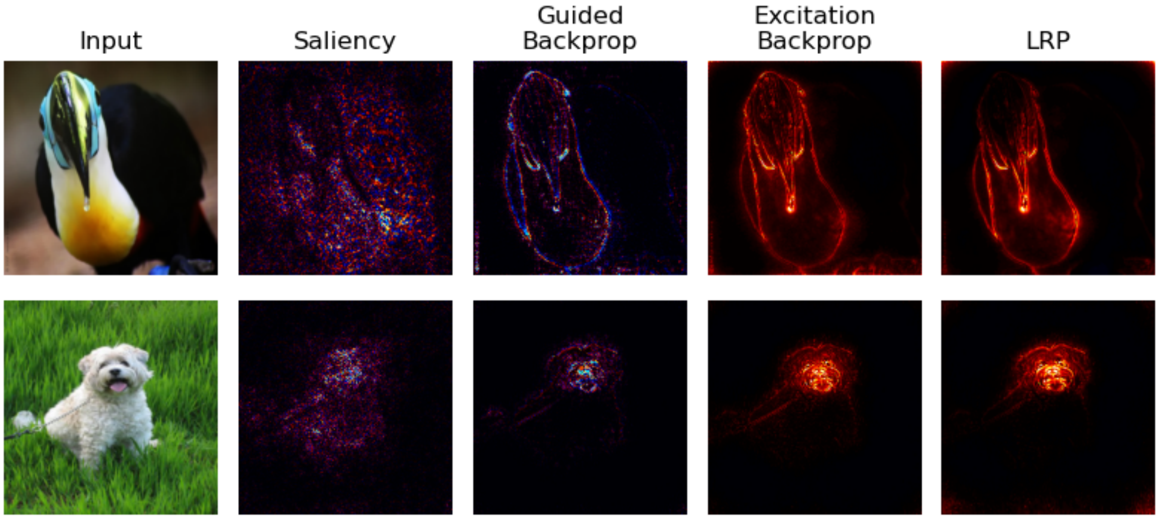

Figure 1: Attribution heatmaps for two input samples for a VGG16 model using different explanation methods

XAI Metrics

To improve the comparability of XAI methods, many evaluation metrics have been proposed, measuring the quality of explanations from different perspectives.

Faithfulness metrics measure whether an explanation truly represents features used by the model. For example, the Pixel Flipping procedure measures the difference in model output scores for the correct class when replacing pixels in descending order of their relevance scores. When plotting these output scores over the percentage of perturbed pixels, this leads to a quickly decreasing curve for accurate explanations. The faithfulness can be measured as area over the perturbation curve.

Localization metrics measure how well an explanation localizes the object of interest for the underlying task. For instance, Relevance Rank Accuracy measures the fraction of high-intensity relevance scores, defined as the top-k values, within the (binary) ground truth mask.

Complexity metrics measure how concise explanations are, e.g., the Sparseness can be evaluated by computing the Gini coefficient of the total attribution vector.

Robustness metrics measure the robustness of explanations towards small changes in the input. A prominent example is Average Sensitivity, which use Monte Carlo sampling to measure the average sensitivity of an explanation for a given XAI method. Note that robust explanations result in a low robustness score.

Randomization metrics measure by how much explanations change when model components and parameters are changed at random. Here, it is expected for the explainer to react to the applied changes. For instance, the random logit test measures the similarity between the original explanation and the explanation with respect to a random other class. Again, a low score is desirable.

All these metrics can easily be computed using Quantus, which is a framework to compute XAI evaluation metrics and has already been introduced in an earlier blog post.

Systematic XAI Evaluation

In the following, we apply these metrics to the XAI methods listed above, i.e., Saliency, Guided Backprop, Excitation Backprop and LRP. The results are summarized Table 1. It can be seen that different metrics prefer different XAI methods. In our example, Guided Backprop leads to the least complex explanations. The most faithful explanations are produced by Excitation Backprop and LRP. The localization of the object of interest is best for Excitation Backprop. In terms of randomization, Saliency leads to the best explanations.

The most robust explanations are produced by Guided Backprop, Excitation Backprop and LRP.

Table 1: XAI Evaluation Results for four different explanation methods. For Complexity, Faithfulness and Localization higher scores are better, and for Randomization and Robustness lower scores are better.

	Complexity	**Faithfulness**	**Localization**	**Randomization**	**Robustness**
Saliency	0.57	0.57	0.65	**0.45**	0.04
Guided Backprop	**0.77**	0.57	0.69	1.00	**0.01**
Excitation Backprop	0.60	**0.68**	**0.76**	1.00	**0.01**
LRP	0.62	**0.68**	0.73	0.73	**0.01**

Conclusions

The choice of XAI method depends on the problem at hand, as explanation metric performs differently under each evaluation metric. In this post, we have run a systematic XAI evaluation using a framework, in which the quality of explanations is measured from different perspectives.

This evaluation framework can be applied to (1) measure the effect of changes to XAI methods or (2) to optimize XAI hyperparameters. Both ideas will be addressed in upcoming blog posts.

Relevance to ITοBoS

In iToBoS, many different AI systems will be trained for specific tasks, which in combination will culminate in an “AI Cognitive Assistant”. All those systems will need to be explained with suitable XAI approaches to elucidate all possible and required aspects of the systems’ decision making. In order to ensure these explanations are correct and of high quality, we will apply the evaluation framework presented in this blog post.

Authors

Frederik Pahde, Fraunhofer Heinrich-Hertz-Institute

Galip Ümit Yolcu, Fraunhofer Heinrich-Hertz-Institute

Sebastian Lapuschkin, Fraunhofer Heinrich-Hertz-Institute