XAI Hyperparameter Optimization

Rule-based eXplainable AI (XAI) methods, such as layer-wise relevance propagation (LRP) and DeepLift, provide large flexibility thanks to configurable rules, allowing AI practitioners to tailor the XAI method to the problem at hand. This comes at the cost of a large number of potential XAI parameterizations, especially for complex models with many layers. However, finding optimal parameters is barely researched and often neglected, which can cause these methods to yield suboptimal explanations.

In this blog post, we demonstrate the hyperparameter optimization for LRP using the XAI evaluation framework presented in an earlier post. Specifically, we want to explain a VGG-11 model with BatchNorm (BN) layers trained on the ILSVRC2017 dataset. We use the LRP γ-rule, in which relevance scores R_j of layer j are computed given scores R_k from the succeeding layer k as:

$$R_j = \sum_k \left(\frac{a_j \cdot \left(w_{jk} + \gamma w_{jk}^{+1}\right)}{\sum_{0,j} \left(a_j \cdot \left(w_{jk} + \gamma w_{jk}^{+1}\right)\right)} \right) \cdot R_k$$

Here, w_{jk} are the weights between layers j and k, w^{+}_{jk} is the positive part of w_{jk} and γ is a configurable parameter controlling the impact of positive and negative contributions. In the extreme case, the LRP γ-rule is identical the $\alpha 1 \beta 0$-rule as $\gamma \rightarrow \infty$, where negative contributions are disregarded. Similarly, the γ-rule is identical to the ε-rule as $\gamma=0$, where negative and positive contributions are treated equally. Note that γ can be defined differently for each layer. To reduce the search space, we group the 11 layers of the VGG-11 (10 convolutional layers and 1 fully-connected layer) into 3 low-level feature layers (Conv 1-3), 4 mid-level feature layers (Conv 4-7), 3 high-level feature layers (Conv 8-10) and the fully connected layer. For each group, we search for the optimal γ with $\gamma \in [0, 0.1, 0.25, 0.5, 1, 10]$. Given these 4 free parameters with 6 options each, we perform a grid search both with and without model canonization. The results for 6 metrics, including localization measured with relevance rank accuracy (RRA) and relevance mass accuracy (RMA), faithfulness, complexity, randomization and robustness are shown in Figure 1.

It can be seen that the impact of the γ-parameter differs by layer-group and metric. For example, the faithfulness is mainly influenced by the γ-value for low-level features, where higher values lead to better faithfulness. The randomization score, instead, is mainly influenced by the γ-parameter for the fully-connected layer in the classifier head and low values lead to better scores.

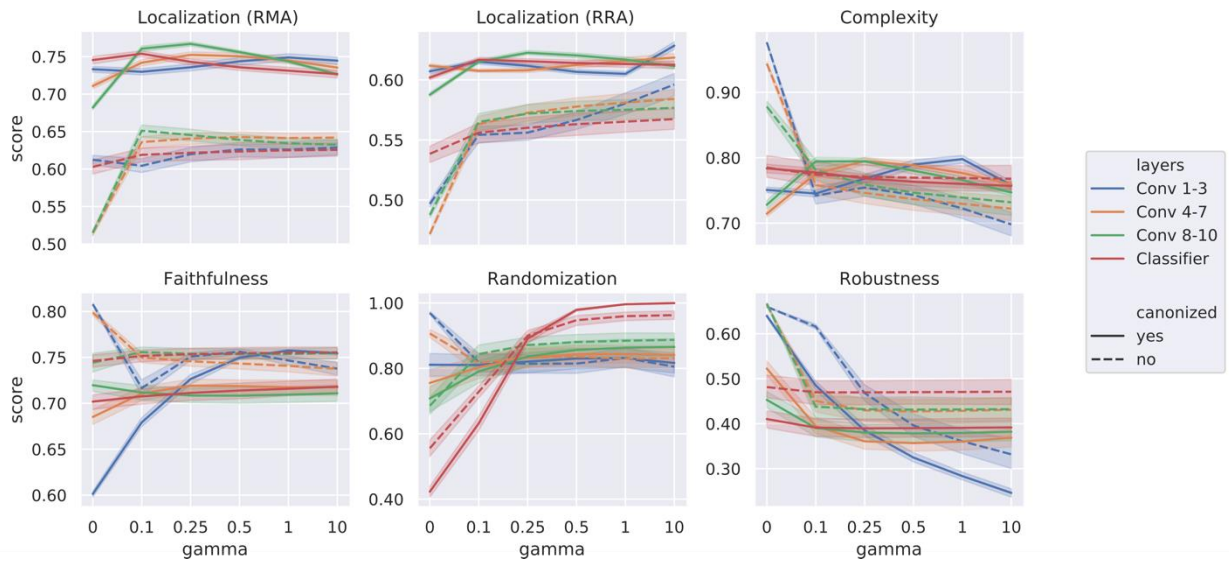

Figure 1: XAI evaluation results for various $\boldsymbol{\gamma}$ *-configurations. We use the metrics RMA, RRA, Complexity, Faithfulness, Randomization and Robustness.*

In Figure 2, we show attribution heatmaps for three random samples using the best and worst parameterizations from our grid search according to the metrics faithfulness, localization and complexity. The heatmaps differ quite significantly, which highlights the importance of the parameter choice for XAI methods.

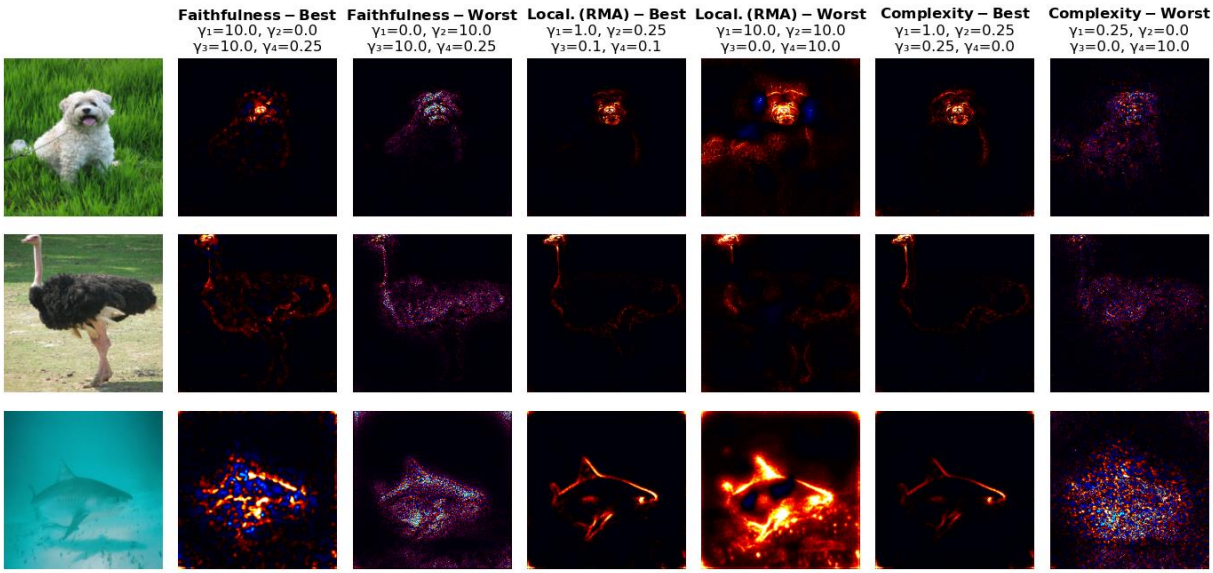

Figure 2: Attribution heatmaps for best and worst parameterization according to metrics faithfulness, localization (RMA) and complexity.

Conclusions

In this blog post, we have demonstrated the application of our XAI evaluation framework to optimize the parameterization of XAI methods. Our results stress the importance of appropriate XAI parameterization according to the problem at hand.

Relevance to IToBoS

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 to optimize the XAI parameterization.

Authors

Frederik Pahde, Fraunhofer Heinrich-Hertz-Institute
Galip Ümit Yolcu, Fraunhofer Heinrich-Hertz-Institute
Sebastian Lapuschkin, Fraunhofer Heinrich-Hertz-Institute