Improving Explanations with Model Canonization

(Modified-) Backpropagation and rule-based XAI methods are prominent choices to explain neural network predictions. This is due to their speed and efficiency, as the computation of explanations only requires one backward pass through the model. Another important factor for the popularity of backpropagation and rule-based XAI methods is the high quality and faithfulness of their explanations. However, these methods may struggle when being applied to modern model architectures with innovative building blocks or high inter-connectivity.

This is caused by types of neural network layers which have been shown to break implementation invariance, which has been defined as axiom for XAI methods. Specifically, implementation invariance is desirable from a functional perspective, and suggests that explanations computed for two different networks implementing the same mathematical function should always be identical. However, for example in the presence of BatchNorm (BN) layers, implementation invariance is hurt.

What is model canonization?

A simple approach to address the issue is model canonization, which is the process of re-structuring the components of a model ***f*** into a model ***g*** which produces exactly the same output but does not contain problematic components such as BN layers. In practice, BN parameters can simply be merged into neighboring Linear (including Convolutional) layers and then be dropped.

From a mathematical perspective, a BN layer is defined as:

$$BN(x) = w_{BN}^{T}\left(\frac{x-\mu}{\sqrt{\sigma+\epsilon}}\right) + b_{BN}$$

Here, $\boldsymbol{w_{BN}}$ and $\boldsymbol{b_{BN}}$ are learnable weights and the bias term of the BN layer, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the running mean and running variance and $\boldsymbol{\epsilon}$ is a stabilizer. As a BN layer constitutes a linear transformation with constant parameter during test time, these parameters can now be merged into a preceding linear layer with weights $\boldsymbol{w_L}$ and bias $\boldsymbol{b_L}$, resulting into a new linear layer with weights $\boldsymbol{w_{new}}$ and bias $\boldsymbol{b_{new}}$, which can be calculated as follows:

$$w_{new} = \frac{w_{BN}}{\sqrt{\sigma+\epsilon}} w_L \quad \text{and} \quad b_{new} = \frac{w_{BN}}{\sqrt{\sigma+\epsilon}}(b_L - \mu) + b_{BN}$$

We describe the beneficial effects of model canonization on a VGG-16 model trained on the ILSVRC2017 dataset, which can easily be canonized as described above. In Figure 1, we show the resulting attribution heatmaps for Excitation Backprop (EB), Layer-Wise Relevance Propagation (LRP) with α2β1-rule and LRP with ϵ+-rule both with and without model canonization. It can be seen that attribution heatmaps differ after the application of model canonization. However, it is hard to make a judgment regarding the effectiveness of model canonization solely based on the visual inspection of heatmaps.

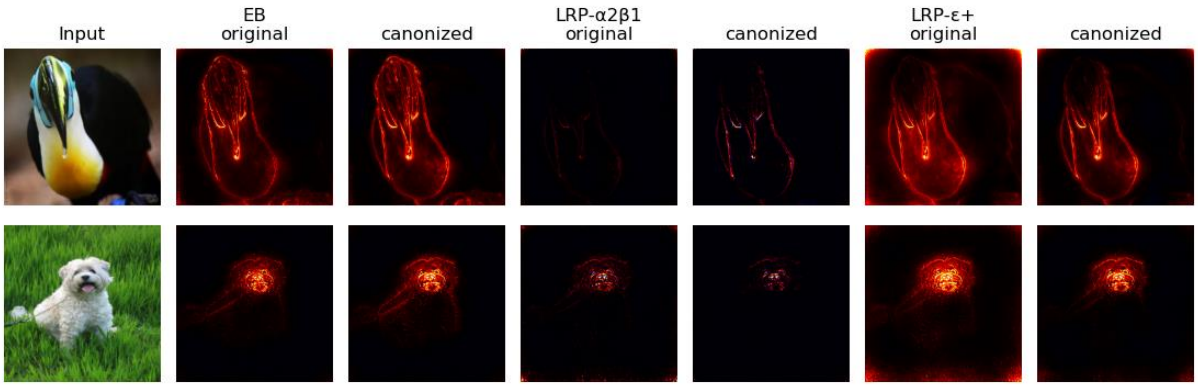

Figure 1: Attribution heatmaps before and after model canonization for different explanation methods

Quantitative Comparison with XAI Evaluation Framework

In order to quantitatively measure the impact of model canonization, we apply the XAI evaluation framework discussed in our previous blog post to evaluate its impact with respect to complexity, faithfulness, localization, randomization and robustness metrics. The results in Tab. 1 indicate that model canonization has a positive impact on the complexity (i.e., the readability for human observers) of explanations. Moreover, except for LRP-α2β1, the faithfulness (i.e., how well the explanation represents the model's reasoning) of explanations increases. The localization capabilities of the applied explainer (i.e., how precisely does the explanation identify the expected object) increase as well when applying model canonization. It has a negative impact on randomization, i.e., explanations look more similar when randomizing the output scores. There is almost no impact on the robustness of explanations, i.e., how sensitive the model is to small perturbations in the input.

Table 1: XAI evaluation results for different explanation methods with and without model canonization. For Complexity, Faithfulness and Localization higher scores are better, and for Randomization and Robustness lower scores are better.

	canonized	**Complexity**	**Faithfulness**	**Localization**	**Randomization**	**Robustness**
EB	no	0.60	0.68	0.76	1.00	**0.01**
	yes	**0.61**	**0.69**	0.76	1.00	0.02
LRP - α2β1	no	0.70	**0.67**	0.66	**0.89**	0.01
	yes	**0.84**	0.65	**0.70**	0.95	0.01
LRP - ϵ+	no	0.51	0.67	0.70	**0.64**	0.02
	yes	**0.62**	**0.68**	**0.73**	0.73	**0.01**

Conclusions

Overall, model canonization has a positive impact on the quality of explanations for most XAI methods, in particular for complexity, faithfulness and localization metrics. Note that while model canonization is straightforward for relatively simple model architectures as VGG-16, it can be harder for more complex and interconnected models, such as DenseNets. Therefore, DenseNet canonization will be demonstrated in a future blog post.

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. Throughout the iToBoS project, innovative state-of-the-art model architectures will be deployed, for which model canonization will be required to optimize the quality of explanations.

Authors

Frederik Pahde, Fraunhofer Heinrich-Hertz-Institute

Galip Ümit Yolcu, Fraunhofer Heinrich-Hertz-Institute

Sebastian Lapuschkin, Fraunhofer Heinrich-Hertz-Institute