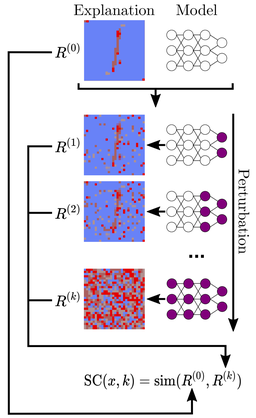
Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek

Some of the presented results have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.

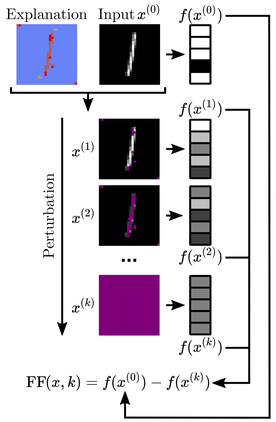
Randomization-based Evaluation of XAI

Model Randomization Tests [1]

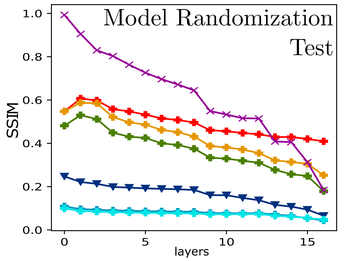


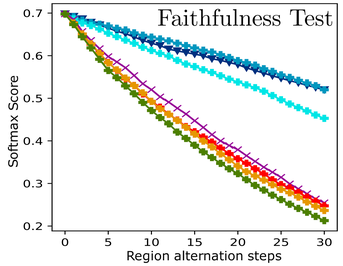
- Randomize **model** towards implausible (top-down)
- Measure similarity between attributions before and after randomization
- Optimally, SC(x, k) should be low, with a **perturbation to model parameters also perturbing explanations**

Faithfulness Tests [2]



- Randomize **data** towards implausible
- Measure difference in prediction scores
- Optimally, FF(x, k) should decrease quickly with perturbation steps k, with the earliest **perturbations based on largest attributions affecting the prediction the most**





Figures for ResNet-50; lower is better

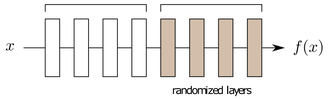
Gradient, GradientxInput, Integrated Gradients, Guided Backprop, LRP-$\alpha1\beta0$, LRP-$\alpha2\beta1$, LRP-Adaptive$\beta3$

Despite their similarities, it seems that **model randomization and faithfulness tests severely disagree** - to the point that they are almost negatively correlated.

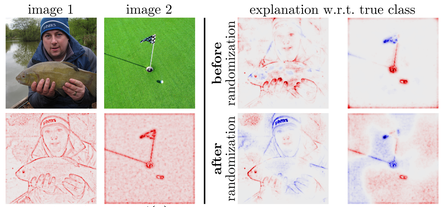
Why?

Model Randomization Preserves some Behavior

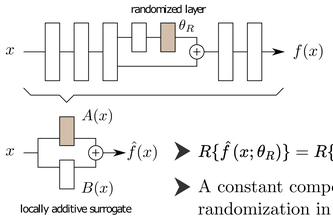
Preservation of Irrelevance



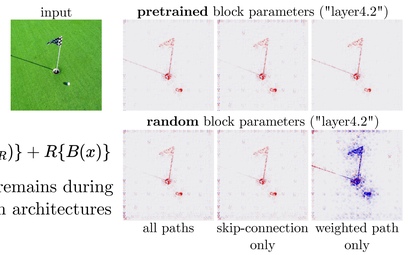
- $\phi(x)$ unresponsive to x_i
  $\Rightarrow \forall g : g \circ \phi(x)$ unresponsive to x_i
  $\Rightarrow \forall g : R_i\{g \circ \phi(x)\}$-small
- Any g based on ϕ shares a similar low-relevance pattern



Preservation of Baseline Explanation

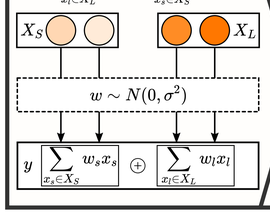


- $R\{\hat{f}(x;\theta_R)\} = R\{A(x;\theta_R)\} + R\{B(x)\}$
- A constant component remains during randomization in certain architectures



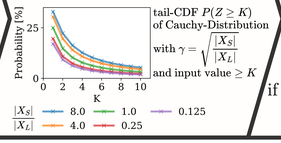
Preservation of Highly Activated Features

Let $\min_{x_l \in X_L} x_l \geq K \max_{x_s \in X_S} x_s > 0$



Large activations contribute highly despite randomization:

$$P\left(0 < \sum_{x_l \in X_L} w_l x_l \leq \sum_{x_s \in X_S} w_s x_s\right) \text{ is low!}$$



Explanations are dominated by large activations in last unrandomized layer:

$$P\left(\sum_{x_l \in X_L} R(w_l x_l) \leq \sum_{x_s \in X_S} R(w_s x_s)\right)$$

is also low if explanation is monotenous

Observation 1

Contrary to intuition, **top-down randomization only alters features to a limited extend**, and we can therefore expect **(faithful) explanations to be altered to a limited extend** as well, especially when the sign of attribution scores is not considered during evaluation, as in [1]. Only unfaithful (e.g. noisy) explanations would enable further changes.

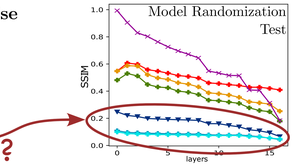
Similarity Metric Minimization is Sensitive to Noise

Shown here for SSIM.

However, the same issue persists for Spearman RC, MSE, ...

$$\mathrm{SSIM}(A,B) = \frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1} \frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}$$

$$\min_{\sigma_{AB} \geq 0} \mathrm{SSIM}(A,B) = \frac{C_2}{\sigma_A^2 + \sigma_B^2 + C_2}$$



- Gradients in ReLU models have statistics resembling **noise processes** [3]
- Effects from model randomization **cannot be disentangled** from noise effects
  - ➡ Gradient-based explanations **can perform well** in model randomization tests **due to noise** ?

Observation 2

The similarity metrics employed by model randomization checks are minimized by attributions drawn from random noise processes. This effect can occur in practice for Gradient-based attribution methods and cannot be disentangled from the effects of the randomization itself. **Using this type of model randomization tests for ranking attribution methods is therefore problematic.**

Conclusion and Outlook

- Model randomization only alters prediction behavior to a certain extent
- Similarity scores used to quantify the effect of model randomization are sensitive to noise in explanations
  - ➡ Due to above issues, model randomization tests as proposed by [1] should **only be used as a binary test** (does model randomization alter explanations or not?), rather than for ranking explanation methods.
  - ➡ **We caution against the use of any singular method as a sole criterion to evaluate and rank explanations**
- Potential Solutions:
  - Measure change only w.r.t. non-irrelevant features (e.g., bottom-up randomization?)
  - Measure change only w.r.t. non-baseline component (removes universality w.r.t architecture)

[1] Adebayo, Julius, et al. "Sanity checks for saliency maps." Advances in neural information processing systems 31, 2018.
[2] Samek, Wojciech, et al. "Evaluating the visualization of what a deep neural network has learned." IEEE transactions on neural networks and learning systems 28.11, 2016.
[3] Balduzzi, David, et al. "The shattered gradients problem: If resnets are the answer, then what is the question?." International Conference on Machine Learning. PMLR, 2017.

Link to our Paper ➡