Benchmarking the Attribution Quality of Vision Models

Robin Hesse[1] Simone Schaub-Meyer[1,2] Stefan Roth[1,2]

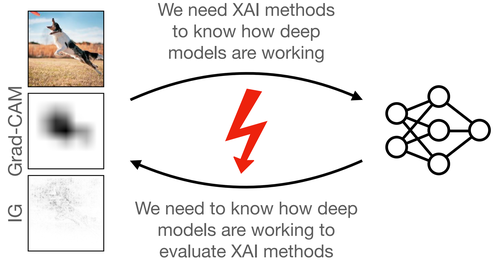
[1] TECHNISCHE UNIVERSITÄT DARMSTADT [2] hessian.AI

Paper & Code

NEURAL INFORMATION PROCESSING SYSTEMS

The chicken and egg problem

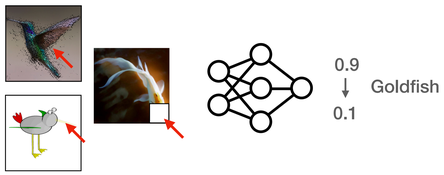
…of evaluating XAI methods



Related work

…and its limitations

Common protocols to evaluate attribution quality



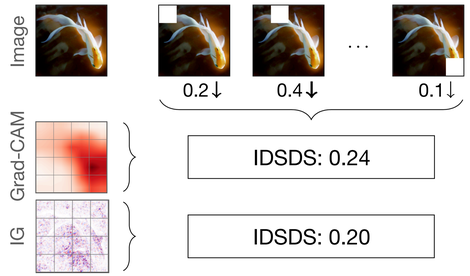
→ Delete patches, pixels, or concepts to measure the effect on the output confidence or accuracy

Protocol	Domain alignment	No inf. leakage	Inter-model comp.	Natural data
Incremental deletion	✗	N/A	✗	✓
Single deletion	✗	N/A	✓	✓
ROAR	✓	✗	✗	✓
FunnyBirds	✓	✓	✓	✗
IDSDS (ours)	✓	✓	✓	✓

In-domain single deletion score (IDSDS)

1. Train the model on images with one patch deleted

2. Measure the rank correlation between output drops and attribution strength for each patch



→ **Aligned train and test domains**
We train and evaluate with corruptions

→ **Provably no class information leakage**
Deletions are independent of the image content

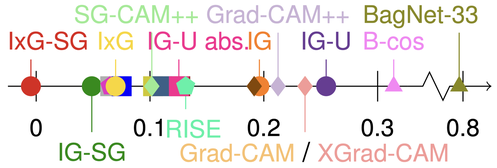
→ **Inter-model comparison**
Rank correlation only improves if the actual task of ranking the patch importances is more effectively solved

→ **Stable under different hyperparameter settings**
Different training seeds and baseline images lead to consistent results

[1] Samek et al. (2017). "Evaluating the visualization of what a deep neural network has learned." In: IEEE Trans. Neural Networks Learn. Syst.
[2] Selvaraju et al. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." In: ICCV
[3] Hooker et al. (2019). "A benchmark for interpretability methods in deep neural networks." In: NeurIPS
[4] Hesse et al. (2023). "FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods." In: ICCV

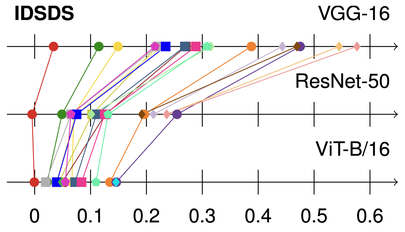
Ranking attribution methods

→ SmoothGrad (SG) impairs performance for all methods
→ Taking the absolute attributions (abs.) impairs performance
→ Intrinsically explainable models (▲) achieve the best results

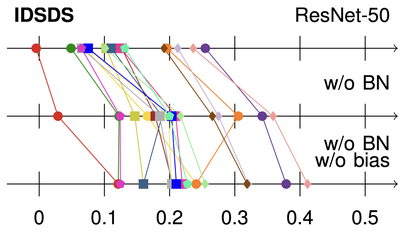


How design choices affect attribution quality

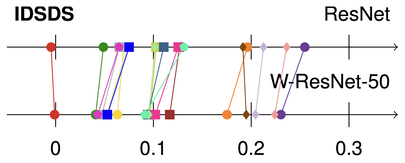
→ Different backbones have significantly different attribution quality



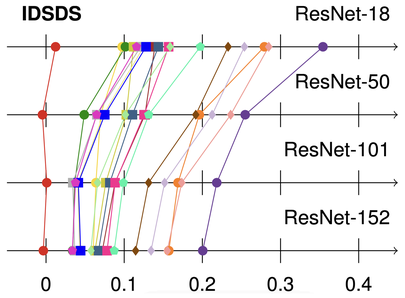
→ Batch norm (BN) and bias terms impair attribution quality



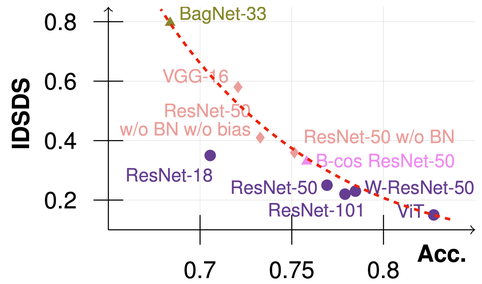
→ Wider models have lower attribution quality



→ Deeper models have lower attribution quality



→ There is an accuracy-attribution quality tradeoff



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). The project has also been supported in part by the State of Hesse through the cluster projects "The Third Wave of Artificial Intelligence (3AI)" and "The Adaptive Mind (TAM)".

erc European Research Council