*Are attribution methods **useful** in **real-world scenarios**?*

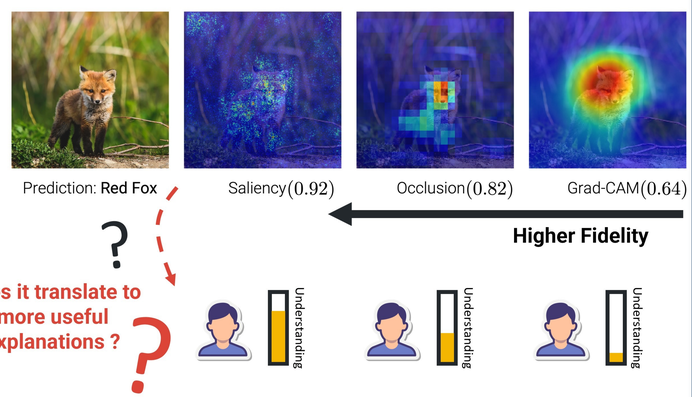
{julien_colin*, thomas_fel*, remi_cadene, thomas_serre}@brown.edu

Summary

- **What do we want to study?**
  - How useful are attribution methods to better understand a model?
  - Do their associated benchmarks predict this utility?
- **How do we do it?**
  - Propose a novel human-centered evaluation framework and run a large-scale evaluation (n=1150) of representative Attribution Method
- **What do we find?**
  - Attribution methods help only in simpler use cases
  - Benchmarks used to score those methods do not correlate with our metric of utility

- **Take away**: Attribution methods help only in simpler scenarios. Complex use cases might require additional "what" information.

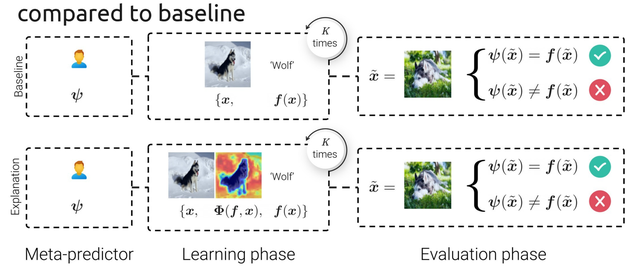
1. Are we moving in the right direction ?



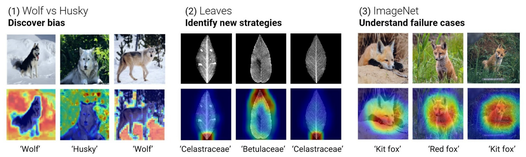
2. Evaluation framework

Evaluate how much attribution maps **help training** participants at getting **better at predicting** a models' decisions on new data

Utility score computed from **relative improvement** in accuracy compared to baseline



3 use cases of XAI
1150 participants
Extensive control & screening



3. Mixed results

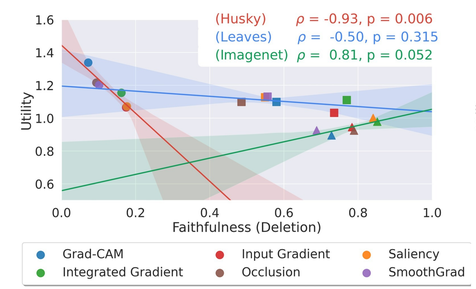
Helpful in **simpler scenarios**, **not** when diagnosing **more challenging** failure cases of the system

Method	*Husky vs. Wolf*	*Leaves*	*ImageNet*
Baseline	1	1	1
Control	0.95	1.02	0.94
Saliency	1.06	**1.13**	1
Integ.-Grad.	1.15	**1.11**	0.98
SmoothGrad	**1.20**	**1.13**	0.93
GradCAM	**1.34**	1.10	0.90
Occlusion	**1.22**	1.10	0.92
Grad.-Input	1.06	1.05	0.95

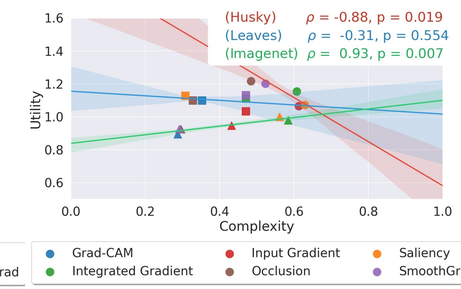
Utility scores

4. Why attribution methods are **not useful** in **more complex scenarios**?

Making attribution methods (1) more faithful or (2) less visually complex not a promising avenue



(1) No correlation

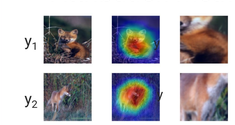


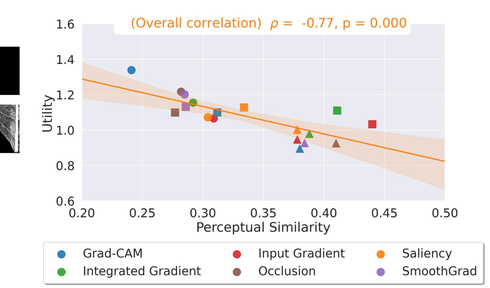
(1) Weak correlation

Need novel (3) "what" information beyond the "where" information contributed by attribution methods (e.g., [3])

*Attribution methods **useful***

*Attribution methods **useless***





(3) Strong correlation

See also:
[1]The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. Nguyen & al. 2021.
[2] HIVE: Evaluating the Human Interpretability of Visual Explanations, Kim & al, 2022
[3] CRAFT: Concept Recursive Activation FacTorization for Explainability, Fel & al, 2022

serre-lab.github.io/Meta-predictor

deel-ai/xplique