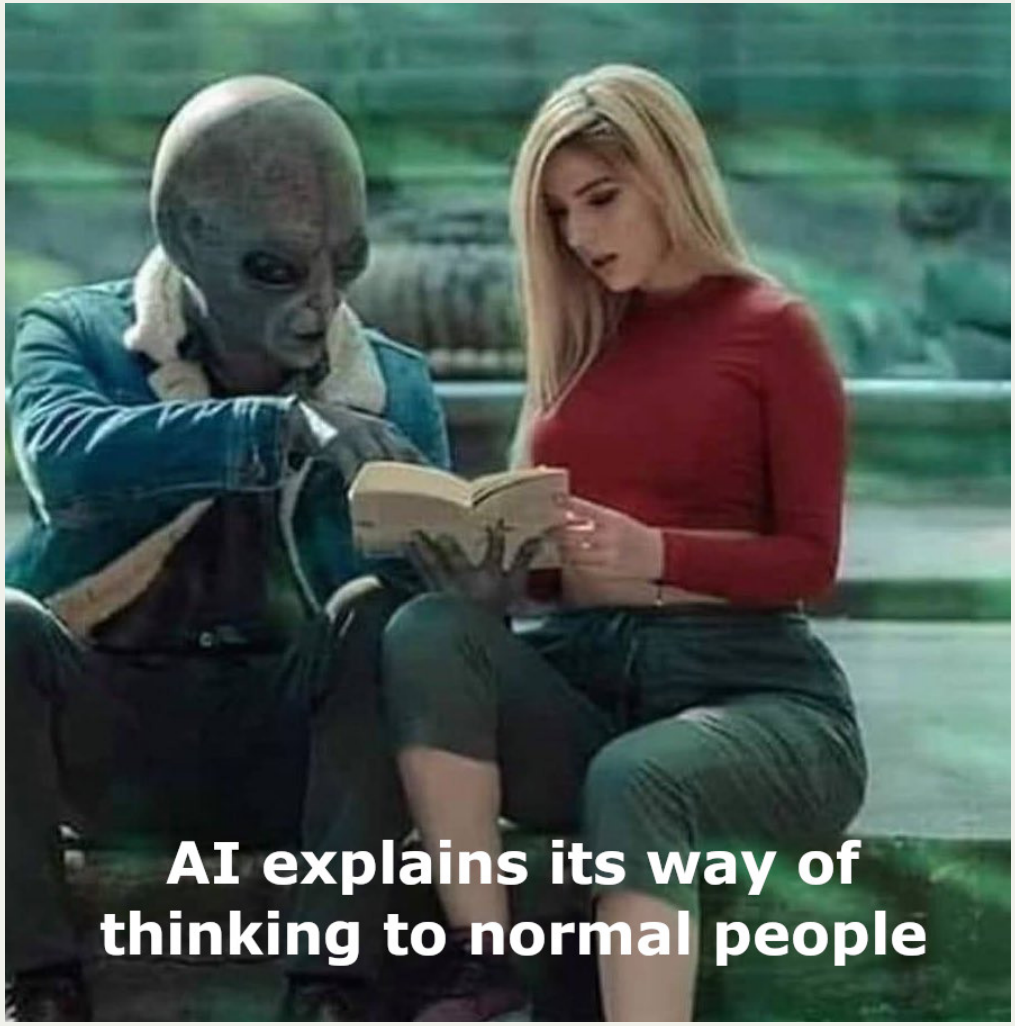
Attribution methods in interpretability of CNNs

Author: Kemal Erdem

What is a problem?

- We don't know how to compare XAI methods
- No one is checking the metrics on real data



Source: "cell" on Twitter

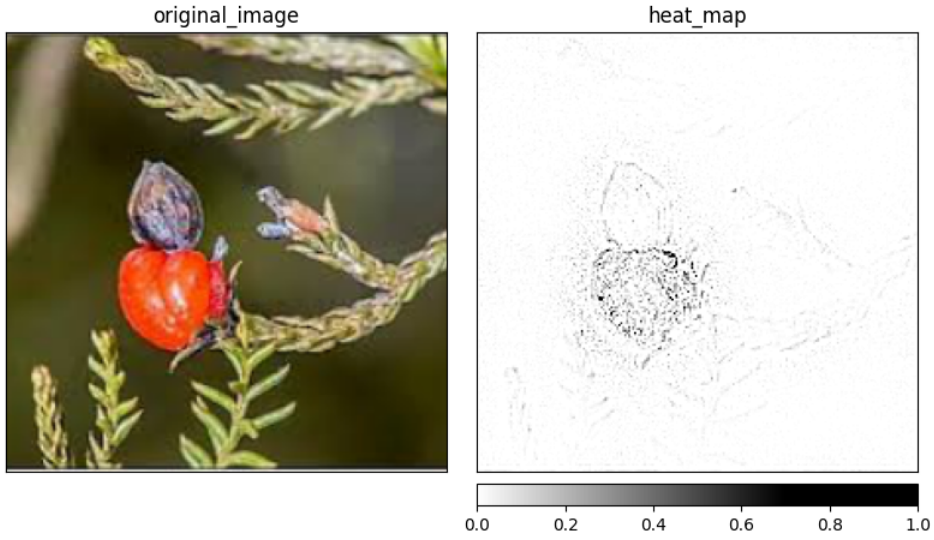
The goal?

The goal?

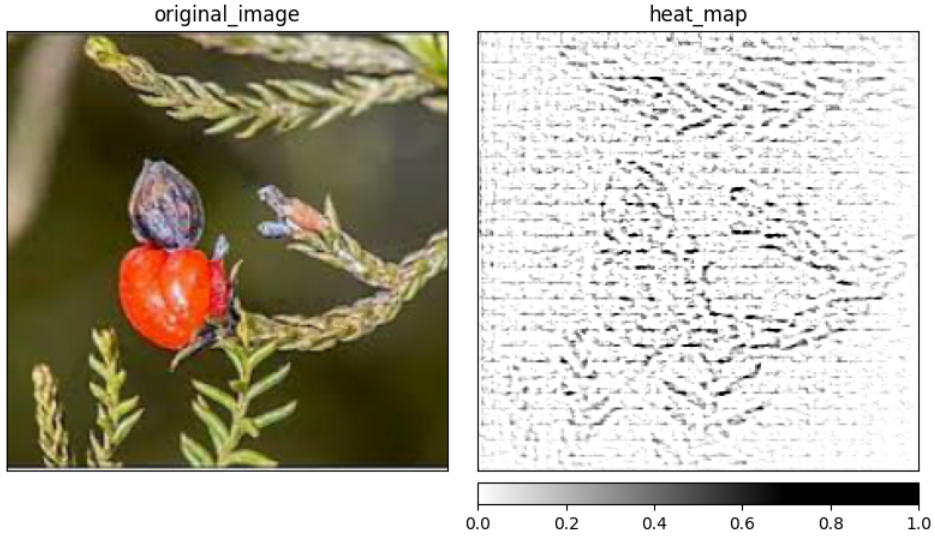
- Compare metrics (different methods, different models, different datasets)
- Check if metrics make sense (spoiler... they don't)
- Check which method works

What are the methods?

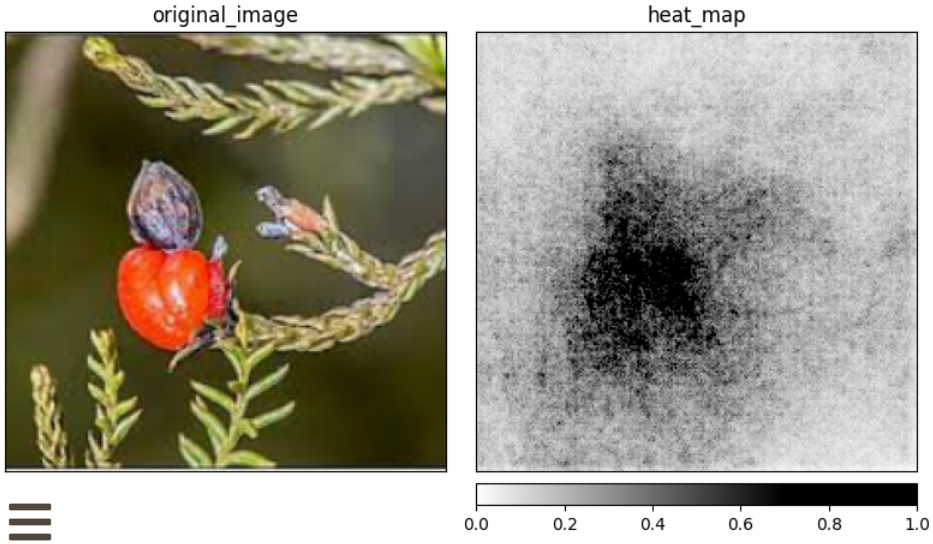
GradCAM



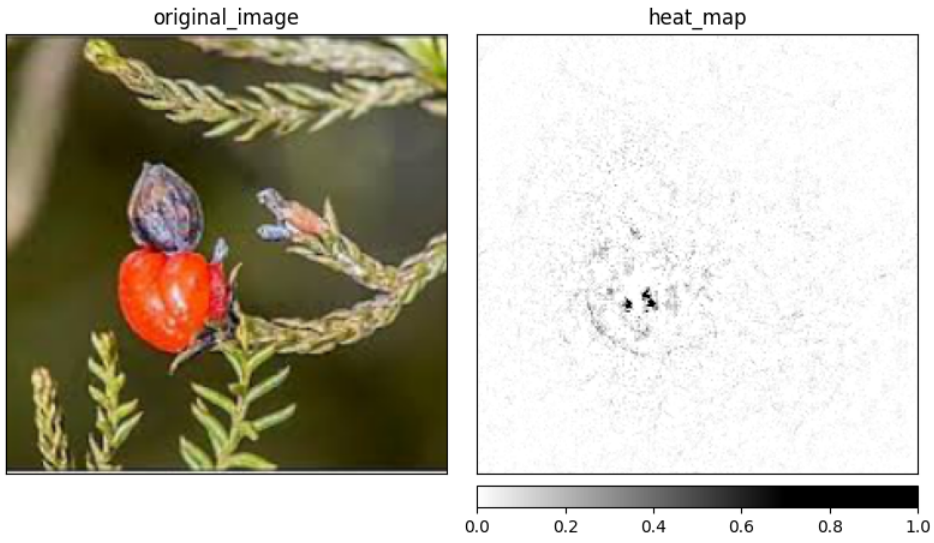
Deconvolition



Sailency



Integrated Gradients



Maybe metrics then?

On the (In)fidelity and Sensitivity of Explanations

Chih-Kuan Yeh [*] **Cheng-Yu Hsieh** [†] **Arun Sai Suggala** [‡]
Department of Machine Learning
Carnegie Mellon University

David I. Inouye [§]
School of Electrical and Computer Engineering
Purdue University

Pradeep Ravikumar [¶]
Department of Machine Learning
Carnegie Mellon University

Only two metrics available in the most popular XAI library: Captum

Unintuitive intuition behind the Infidelity

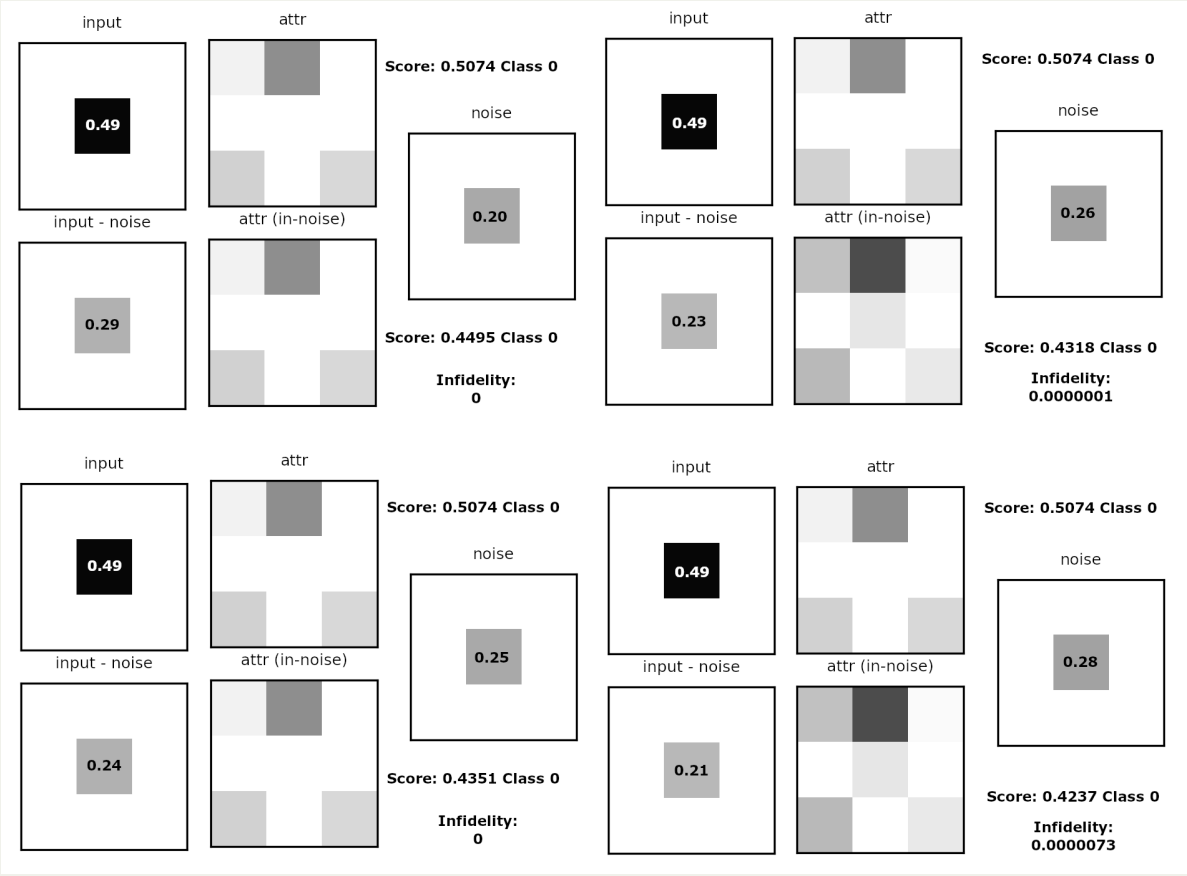
$$\text{INFD}(\Phi, \mathbf{f}, \mathbf{x}) = \mathbb{E}_{\mathbf{I} \sim \mu_{\text{I}}} \left[\left(\mathbf{I}^T \Phi(\mathbf{f}, \mathbf{x}) - (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x} - \mathbf{I})) \right)^2 \right]$$

$$\Phi^*(\mathbf{f}, \mathbf{x}) = \left(\int \mathbf{I}\mathbf{I}^T d\mu_{\mathbf{I}} \right)^{-1} \left(\int \mathbf{I}\mathbf{I}^T IG(\mathbf{f}, \mathbf{x}, \mathbf{I}) d\mu_{\mathbf{I}} \right)$$

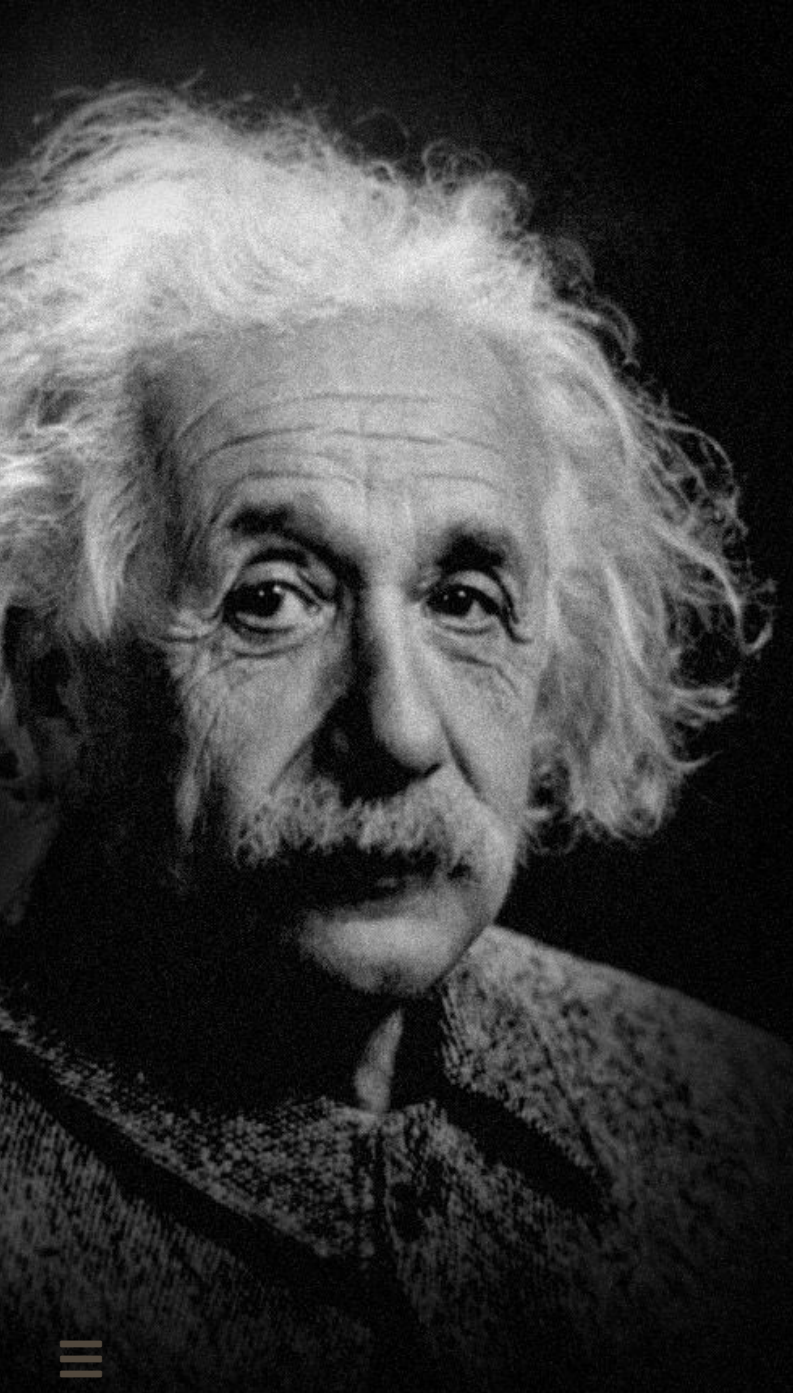
$$\text{IG}(\mathbf{f}, \mathbf{x}, \mathbf{I}) = \int_{t=0}^{1} \nabla \mathbf{f}(\mathbf{x} + (t-1)\mathbf{I})$$

Infidelity calculation, Source: On the (In)fidelity and Sensitivity of Explanations

Actual intuition behind the Infidelity



Sample infideilty calculations for different noises.



Qualitative methods suck, just use Quantitative, they are fine...

- Albert Einstein (maybe)

Experiments

Datasets

- Stanford Dogs Dataset
- Food 101
- Edible wild plants
- Plants Dataset
- Marvel Heroes

Models

- ResNet18 (arxiv, 1512.03385)
- EfficientNetB0 (arxiv, 1905.11946)
- DenseNet121 (arxiv, 1608.06993)

Methods

- Saliency (arxiv, 1312.6034)
- Deconvolution (arxiv, 1311.2901)
- Guided Backpropagation (arxiv, 1412.6806)
- Guided GradCAM (arxiv, 1610.02391)

Work Split

- **Phase 1**: Publication on Infidelity and Sensitivity as a method to compare XAI solutions (almost done)
- **Phase 2**: Master's Thesis which combines Phase 1 and additional work on the XAI methods

Bibliography

- On the (In)fidelity and Sensitivity of Explanations, 2019 (arxiv, 1901.09392)
- Deep Residual Learning for Image Recognition, 2015 (arxiv, 1512.03385)
- EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2019 (arxiv, 1905.11946)
- Densely Connected Convolutional Networks, 2016 (arxiv, 1608.06993)
- Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2013 (arxiv, 1312.6034)
- Visualizing and Understanding Convolutional Networks, 2013 (arxiv, 1311.2901)
- Striving for Simplicity: The All Convolutional Net, 2014 (arxiv, 1412.6806)
- Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2016 (arxiv, 1610.02391)
- Sanity Checks for Saliency Maps, 2018 (arxiv, 1810.03292)
- Axiomatic Attribution for Deep Networks, 2017 (arxiv, 1703.01365)
- A Benchmark for Interpretability Methods in Deep Neural Networks, 2018 (arxiv, 1806.10758)
- SAM: The Sensitivity of Attribution Methods to Hyperparameters, 2020 (arxiv, 2003.08754)

Thanks

"There's no such thing as a stupid question!"

Author: Kemal Erdem

GH repo: https://github.com/burnpiro/xai-correlation