Seminar:
Selected Topics in XAI

Prof. Damien Garreau – Theory of Machine Learning

Julius-Maximilians-Universität Würzburg – CAIDAS

October 21, 2024

Practical details

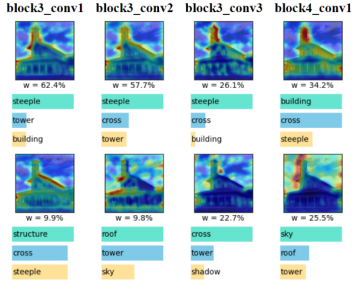
- **Who?** myself + Magamed Taimeskhanov
- **What?** getting familiar with recent works from the field of XAI
- **Course ID:** 08151800
- **In practice:**
  - pick a paper from the list
  - up to three students per paper
  - read and understand the paper
  - re-implement experiments
  - write a small report
  - present your work at the end of the semester
- send me an email as soon as you picked a paper (first come first served!)

The papers

1. Bianchi et al., *Interpretable Network Visualizations: A Human-in-the-Loop Approach for Post-hoc Explainability of CNN-based Image Classification*, IJCAI, 2024
2. Deiseroth et al., *ATMAN: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation*, NeurIPS, 2023
3. Fokkema et al., *Attribution-based Explanations that Provide Recourse Cannot be Robust*, Journal of Machine Learning Research, 2023
4. Humayun et al., *Splinecam: Exact visualization and characterization of deep network geometry and decision boundaries*, CVPR, 2023
5. Oikarinen et al., *Linear Explanations for Individual Neurons*, ICML, 2024
6. Paes et al., *Selective Explanations*, NeurIPS, 2024

Interpretable Network Visualizations (INV)

- **Saliency maps for CNNs:** show where specific class is identified
- no complete explanation of the decision process
- **This paper:** entire feature extraction process



- **Challenge:** human-in-the-loop

ATMAN: Understanding transformer predictions

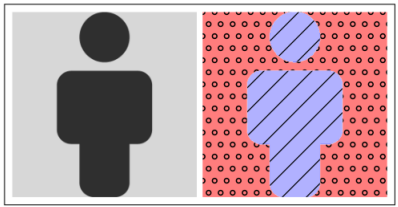
- explanations for multimodal, generative models are *hard* to get
- **This paper:** manipulate attention mechanism to produce relevance maps
- **Example:** image + incomplete sentence and ATMAN relevance maps

This is a painting of

- **Challenge:** experiments (smaller architecture is ok)

Attribution-based Explanations that Provide Recourse

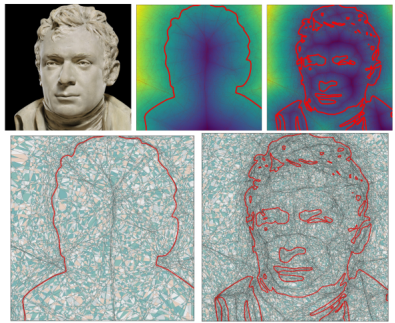
- **Recourse** = allow user to change the decision of classifier
- XAI methods can tell us what this change should be
- **Example:** profile picture + associated saliency map
- moving in the direction given by the saliency map improves classification



- **This paper:** impossible for a single method to be both robust and good for recourse
- **Challenge:** theoretical paper

SplineCam: exact visualization

- **New idea:** look at the decision boundary of the network
- **This paper:** exact computation for ReLUs
- **Example:** application to neural distance field

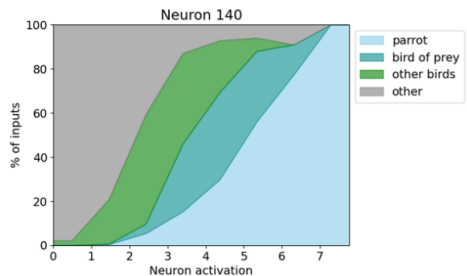


- **Challenge:** a bit of maths

Linear Explanations for Individual Neurons

- **Typical explanation for individual neuron** = highest activation
- this is not sufficient
- **This paper:** explain individual neuron as linear combination of *concepts*
- **Example:** Neuron 140 in ResNet50, layer 4 is explained as

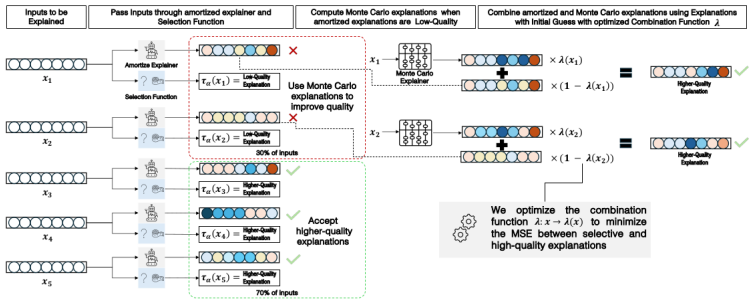
LE(Label): 2.36*parrot + 2.11*bird of prey + 0.94*bird



- **Challenge:** complicated workflow

Selective Explanations

- **Amortized explainers:** train a model to get explanations at inference
- they can produce diverging explanations
- **This paper:** detect low quality explanations, re-run and select better one



- **Challenge:** complicated workflow

Thank you for your attention!