Exercise Sheet Deep Learning

Part 4: Explaining Deep Networks

Summer 22

This sheet includes a theoretical part and a practical assignment on the fourth part of the lecture Deep Learning (4_XAI). Both parts give 20 points maximum each. Please hand in solutions as a pdf in groups of at most three persons via LernraumPlus.

name1:

name2:

name3:

PARTI – THEORY: For the following, you might answer only YES/NO (or abstain), or you can add short arguments (at most two lines per question). If you are not sure, it is better to abstain.

1. The following XAI methods are …

yes **no** SHAP is exemplar based

yes **no** LRP is invariant to orthonormal transformations of input representations

yes **no** LIME is a post-hoc and global model-agnostic method

yes **no** Computing saliency maps accounts for a quadratic effort w.r.t the number of weights

2. The overarching idea behind the following XAI method is …

yes **no** SHAP aims for a determination of feature impact under coalitions with other features using a game theoretic approach

yes **no** Counterfactual explanations aim for the determination of a counterfactual in its basic form

yes **no** Model distillation for a model f relies on the idea to train a smaller (and directly interpretable) model which does not change the output class

yes **no** SpRAy clusters data based on their associated local explanation vector/matrix

3. The following training/optimization algorithms are used to extract the respective explanations:

yes **no** LRP enforces a conservation of the relevances per layer when backpropagating signals.

yes **no** LIME relies on sparse global linear surrogate models which are derived from sampling.

yes **no** DeepProblog uses evolutionary optimization to account for discrete logic operations while training

yes **no** DeepPINK deletes features (knock-off) to estimate their relevance

4. Explanations serve different purposes: improvement, justification, raising trust, discovery, whereby some methods can be used for more than one purpose. The following explanation methods can be used (among other purproses) for the purpose of …

yes **no** Saliency maps for discovery of globally relevant features

yes **no** VQA for model justification

yes **no** counterfactual explanations for justifications

yes **no** SpRAy for improvement

5. The following statements are true:

yes **no** A linear model constitutes one example of an additive feature attribution model.

yes **no** Evaluations of explainable AI methods must rely on human feedback.

yes **no** For a linear classifier $x \mapsto \text{sgn}(w^t x - b)$ and input x', the closest counterfactual is given by $x' - \frac{w^t x' + b}{w^t w^2} \cdot w$.

yes **no** LIME models, if applied for logistic regression, just gives the model itself

PARTII – PRACTICE: You can use code and models which are publicly available, please clearly reference such sources. It might be a good idea to start with the examples given in the practical part of the lecture (available at `https://jgoepfert.pages.ub.uni-bielefeld.de/talk-deep-learning`) and use the CAPTUM API `https://captum.ai/api/`. Please give a link to your code, and please describe the experiments and results of your approach in a pdf which is well structured (e.g. modeling/training parameters/training/results/interpretation, use itemize, keywords are fine) and enables reproducability as well as easy access to your main results. Please use at most one page for both practical parts together including graphs and images.

1. Take the model for the FashionMNIST data from the first sheet. Take 2 different examples from at least wo different classes each and provide a feature based explanation for it being classifier to the most likely and second most likely and least likely class. Interpret the results as well as differences you observe.

2. For one of these settings (one data point, one class), compare at least three different local explanation approaches. Attack thse considered example and investigate whether the explanation changes and how.