Exercise Sheet Deep Learning

Part 4: Explaining Deep Networks

Summer 23

This sheet includes a theoretical part and a practical assignment of the foruth part of the lecture Deep Learning (Explainability). It should be handed in as pdf in groups of three via ekvv-Moodle until 22.6.23. at 10.a.m (sharp). Please include a link to the code (e.g. Colab)

name1: Michel Valentini
Elma Nevala
name2:
Dovile Umbrasaite
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 a global feature-based method

yes **no** LRP is invariant to the specific implementation of the functional form, i.e. if there are two deep networks implementing the same function but possibly different neurons and weights, the feature relevance scores are the same

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

yes **no** Computing saliency maps is quadratic w.r.t the number of weights

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

yes **no** SHAP aims for determining the weights which correspond to the impact of features in a linear model and fair distribution of their impact w.r.t coalitions

yes **no** Counterfactual explanations aim for an example with different output

yes **no** Model distillation for a classificer f relies on the idea to train a smaller model which is supervised by the learned one

yes **no** SpRAy clusters data based on their associated closest counterfactual

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

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

yes **no** LRP enforces a conservation of the relevances over all data when backpropagating signals.

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

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

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 ...

yes **no** Saliency maps for discovery of locally relevant and causal features

yes **no** VQA for model justification

yes **no** counterfactual explanations for improving actions

yes **no** SpRAy for improvement of the model

5. The following statements are true:

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** A linear model constitutes one example of an additive feature attribution model.

yes **no** There are not quantitative evaluations of explainable AI

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 all sources and tools. Please give a link to your code (e.g. colab). The length of the answer is limited to one page in total for the description of both parts including images. Please provide: short description what you did, how it is done, what is the result. Please be prepared to present the solution in the exercises.

1. Take the model for the FashionMNIST or MNIST data set. Take 2 different examples from two different classes. Use at least three local explanation methods and explain reasons they are mapped to the true, the most likely, second most likely, and lest likely class. Interpret the results. Are the explanations meaningful? Do they differ for different target outputs? What happens if the examples are adversarially attacked (with a local change of only small parts of the image)? Also try this out experimentally.

2. Use a model which is trained together with a backdoor. Use two different global explanation methods. Are these capable of detecting/explaining the existence of a backdoor?

Deep Learning fourth assignment

Michel Valentini
Elma Nevala
Dovile Umbrasaite

Group 27

June 2023

1 Link to the GitHub

`https://github.com/ValentiniMichel/Deep_Learning_fourth_assignment`

2 Description Practice

This practical assignment focused on different ways of explaining a deep learning model. When machine learning models classify items, they often works as blackboxes and the reasons behind the classification are unknown. Different explanation methods can explain the reasons of the classification either locally or globally.

In the first exercise we test different explanation methods on the Fashion-MNIST dataset. We generate some images that explaing why certain items are label to a spesific class. We used the SHAP, deepLIFT, Integrated Gradients Explanations.

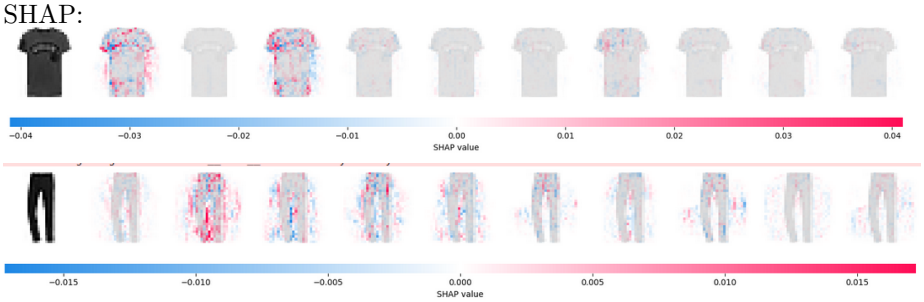
In the second exercise we build a model that classifies pictures of cats and dogs. The model also had a backdoor that makes the model falsely classify dogs as cats. We used LIME and Integrated gradients to explain the chosen labels.

2.1 Exercise No. 1 : Explanation for FashionMNIST

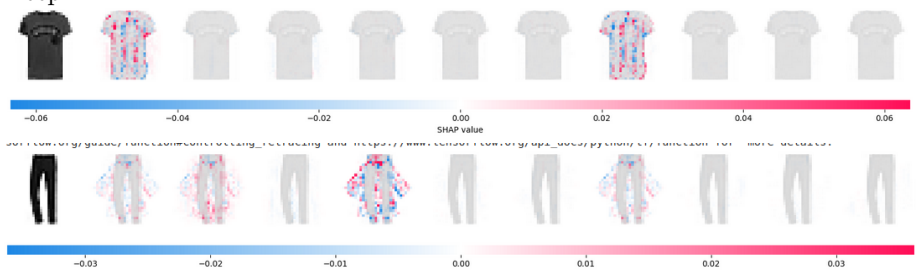
In this task we used a sequential model trained on the FashionMNIST dataset. We examined how a t-shirt and a pair of trousers are classified and used the SHAP, deepLIFT, Integrated Gradients Explanations to locally analyse them.

For each of the clothing we generate images based on how likely their are to be classified to a spesific class. Below are images that show this:

SHAP:



DeepLIFT:



Integrated Gradient:

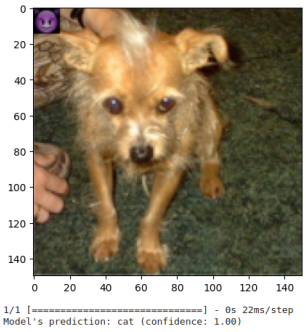
We can see that see that the explanations work pretty well for the t-shirt with every explanation method. All three methods mislabel the tshirts as shirts or pullover, which is fairly similar to a t-shirt. In this regar shap performs the best with the least highlighted pixels next to the shirt category.

However, trousers are a bit more difficult to classify baucause of their shape. Most of the misclassifications go to the dress catecory. Here the Integrated Gradient eample works the best.

2.2 Exercise No. 2 : Cat and Dog classifier explanation

In the second exercise we build a sequential model that classifies pictures with cats and dogs. The model also has a backdoor and classifies dogs with on emoji on them as a cat.

In the pictures below we can see a dog wrongly being classified as a cat. We can also see from the images created that the classification is heavily based on the emoji on the picture. Misclassification explanation:




```
1/1 [==============================] - 0s 22ms/step
Model's prediction: cat (confidence: 1.00)
```

Integrated Gradients