Barbara Hammer, Philip Kenneweg Universität Bielefeld

Deep Learning (SoSe 2024)
4. Sheet

Start: Thursday, 13.06.2024.
End: The worksheets should be solved using Python, in groups of 2-3 people and will be presented in the Tutorials.
Discussion: Thursday, 27.06.2024 in the Tutorials.

Information

The worksheets and necessary toolboxes will be made available in the Lernraum "392221 Deep Learning (V) (SoSe 2024)". Worksheets will usually be released every two weeks on Thursday, and discussed during the exercises on Thursday two weeks later. In order to successfully finish the course, 50% of the available points have to be obtained and each participant has to present his/her results at least once. The Wednesday and Thursday in between the release and discussion of the sheet will be used to discuss the implementation of the various algorithms presented in the lecture, as well as go deeper into the relevant material.

Exercise 1:

(*10 Points*)

You can use code and models which are publicly available. 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 (best in form of a Jupyter notebook .ipynb).

(a) *(6 Pts.)* Take a model for the FashionMNIST or MNIST data set. Take 2 different examples from two different classes. Use at least three local explanation methods (you implemented yourself) and explain reasons they are mapped to the true, the most likely, second most likely, and least 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.

(b) *(4 Pts.)* Use a model which is trained together with a backdoor. Use two different global explanation methods (you implemented yourself). Are these capable of detecting/explaining the existence of a backdoor?