Sheet 4 (Code and theory: See https://github.com/itzeck/deep_learning)

Modeling

For the model we used the same model from sheet 1.
For practice part 1 we've taken two images of trousers and two images of pullovers. As introduced in the tutorials we used the GuidedBackprop class from the CAPTUM API to get the top predictions. For the feature explanations we used the Saliency, the GuidedBackprop and the IntegratedGradients class from the CAPTUM API.
For the attack, that was to be done for practice part 2, we just manually occluded (meaning setting the channel values to 0) a part, which seemed important to the networks reasoning.

Training/ Training Parameters

We used the same model from sheet 1, there was no training in this sheet.

Results/ Interpretation

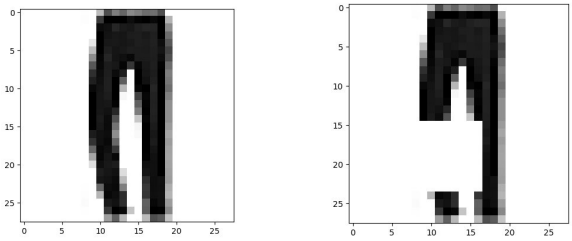
Practice part 1:

We could see that the model was very sure about its classification of the images of the trousers. Even the second most likely class ("Shirt") has a probability/certainty of 0.000 and so do all the other classes. This holds for both images of the trousers. About the pullovers the model was less sure. For pullover1 the most likely class was correctly predicted to be "Pullover", however only with 0.858 percent "certainty". For pullover2 the "certainty" was even just 0.6. In both cases, the second most likely prediction was "Coat" with 0.085 and 0.303 probability. The classes "Bag", "Trouser", "Ankle boot", "Sandal" and ""Sneaker" all had 0.000 probability or close to that. From a human perspective this makes a lot of sense. A coat, especially with the very grainy resolution in the MNIST dataset, does look very similar to a pullover. The other classes I mentioned above however, are cleaerly separable from a pullover. So it seems the model has learned a somewhat good representation of the pullover class.

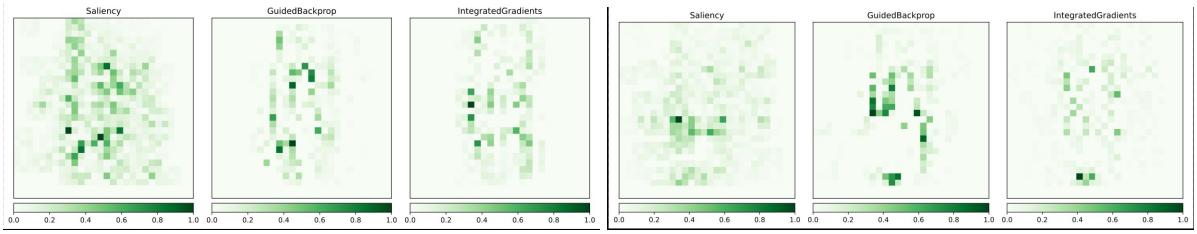
For the visualizations see the github repository.

Practice part 2:

Unattacked vs. attacked image:



Saliency, Guided Backpropagation and Integrated Gradients diagrams bf. vs. after attack



One can see the saliency changed severely. Interestingly enough, the occluded part however is the most salient after the attack. In the Guided Backpropagation diagram after the attack one can very clearly see the occluded part of the image. The model also did no longer predict the correct class, but now predicts "Sandal" which shows the attack was "successful".