Final Project

Deep Learning
From Theory to Practice

Christoph Brune
Gautam Pai, Mei Vaish
Mathematics of Imaging & AI
EEMCS, University of Twente

Attacking and Defending Neural Networks

The final project is a useful milestone at the end of our deep learning course. It is designed to evaluate your ability to conduct a research-related task close to one of the course key topics on a small scale and to show individual skills regarding *understanding, research reproduction, creativity and programming towards applications.* The scope of your final project is the area of adversarial attacks. How do they work and how can they be defended against?

Groups. You can work in groups up to 4 people. You can use your group formation from the hand-in assignments.

Project description. The project is based on this tutorial on adversarial example generation. In this project you will start by finding a pretrained Neural Network which solves a classification task. Then devise and execute an adversarial attack on this network. Finally implement a defense for this network and show that is works. An excellent project analyzes and compares the chosen adversarial attack (targeted/non-targeted, black/white box) and defense from a theoretical and numerical perspective. You are allowed to choose a very basic network, attack and defense. However, we want to encourage you to be creative in picking your problem, for example one which is close to your study. Implementing an interesting attack or defense will usually be awarded with a higher grade.

Some notes:

- You are free to choose the network type: fully connected, CNN, RNN, GAN, etc. However make sure that you have also access to the weights and biases of the pretrained network and that the implementation of the network is sufficiently straightforward. You are also allowed to train a Neural Network yourself, but note that the training can take a long time and is not the focus of the exercise.
- For the attack, you can use FGSM attack from the tutorial, but creativity will be rewarded. For some ideas see this preprint.
- For the defence you can think of a different architecture, data augmentation, etc. For some ideas see again the preprint.
- Make sure that you do not reuse the data for training, testing and the attack/defense.

Data Sets. Here are some useful data-sets for which multiple pretrained networks can be found online:

- Fashion-MNIST (simpler) dataset
- CelebA dataset
- NIPS 2017 adversarial competition dataset

You are are in principle free to choose a data set but do check if it is free to use for research/educational purposes.

Project topic. (due Dec 18, 2025) Please submit one page to describe your project choice including your group, your initial network and dataset, and a first short description of an adversarial attack and a proposal for a defense you will focus on.

Video Pitch. (due Jan 30, 2026) Prepare a few slides for a presentation of maximum 5 minutes. Record a video of the presentation, which should contain at least your slides and voice. In the presentation, show the main results of your attack and defense.

Report. (due Jan 30, 2026) You should aim for a report at the end of this project with a maximum of *8 pages.* An excellent report explains and discusses the **theory** of your chosen adversarial attack and defence, **numerical aspects** (chosen network, algorithms, programming) and **results**.

The focus of the report should lie on the methodology. It is important that you report on your positive *as well as negative* experiences of trying to defend against the adversarial attack and why you took certain decisions in your attack/defense or software development. This will help you and us to better understand the power and limitations of adversarial attacks from theory to practice.

Guideline for report writing: Where applicable, here are some topics that could be addressed in your report:

- Research background. Describe the attack and some suggested defenses from the context of research papers.
- Theoretical analysis. Why should the attack work? And why should the defense mitigate these problems?
- Implementation. Describe your choice of dataset and network and how it is trained and tested. Also describe how the implementation changes for the defense.
- Numerical Analysis. Evaluate the attack qualitatively and quantitatively. Analyze various parameters or check robustness. Compare this with the performance of the defended network.
- Suggested improvements. Suggest an improvement for the attack or the defense for further research. Following your analysis, see where they fail and try to correct it.
- Conclusion. Summarize the goal, your analysis and suggested improvement.
- References. Bibliography cited within the report.
- Code. Please submit your running Python code via a zip file on Canvas.
- Use of AI. Include a statement on the (non)-use of AI according to the guidelines on Canvas.