# A.I. Seminar Project

Defending Against Adversarial Examples

## 1 Context

*Adversarial examples are a type of attack where an adversary maliciously modifies inputs that are fed to a target model in order to fool its inference process. For example, for a classification task, fooling the target model resumes to making it predict two different labels between the clean example and its adversarial counterpart. It is crucial that examples should be modified in a way such that a human expert is not influenced by these modifications (i.e. that he does not detect any malicious attempt). Adversarial perturbations have therefore to be benign or imperceptible. The target model outputs two different labels for the clean example and its adversarial counterpart, while the human expert predicts the same label for both examples. Adversarial examples are currently the most studied threat in the scientific literature among the different attacks that target machine learning models. Numerous defense schemes that aim at thwarting this type of attack have already been proposed in the literature.*

## 2 Objective, report and defense

The goal of this project is to implement the best defense possible against a certain type of adversarial examples for a given model architecture and data set. This project emphasizes on **i)** the understanding of the implemented defense scheme, which requires a good comprehension of the phenomenon of adversarial examples and **ii)** the proper evaluation of the proposed defense scheme.
During the defense you will not be asked many questions about the implementation details, but rather about the method used. This means you will have to be able to clearly explain the intuition behind the method, why it fits your goal and potential improvements of it. Moreover, you will have to detail the evaluation procedure of your method.
The project defense is planned to last approximately 15 minutes, and will take the form of a slide presentation.
The project report will be given the day of the defense. It does not have to be considered as an exhaustive report of your project but more a support (e.g. to present a table of results that could not be included in the slide presentation).

## 3 Timeline

**Due date for the project: February, the 2$^{\text{nd}}$ 2024 (report and defense)**
You will be able to take advantage of 6 hours (among the 24 hours dedicated to this course) in class to work on this project and ask me questions.

## 4 Project description

We consider the MNIST data set and a model $\mathcal{M}$ with the following architecture:

1. A first 2D convolutional layer with 32 output filters, kernel size $(5, 5)$, stride $(1, 1)$, the *relu* activation function and no bias
2. A max pooling layer with pool size $(2, 2)$ and stride $(2, 2)$
3. A second 2D convolutional layer with 64 output filters, kernel size $(3, 3)$, stride $(1, 1)$, the *relu* activation function and no bias

4. A max pooling layer with pool size $(2, 2)$ and stride $(2, 2)$
5. A linear layer with 64 output neurons, the *relu* activation function and no bias
6. A linear layer with 512 output neurons, the *relu* activation function and no bias
7. A linear layer with 10 output neurons

We consider the image domain $[0, 1]^D$ with $D$ the input image space dimension.
The goal of this project is to implement the best defense you can in the following threat model:

- *adversary knowledge*: we consider an adversary in the white-box setting
- *adversary goal*: we consider that the adversary aims at crafting adversarial examples against the model $\mathcal{M}$
- *adversary capacity* we consider an upper bound $\epsilon = 0.3$ for the adversarial perturbation. This means that for a clean input $x$ and its adversarial counterpart $x'$ we must have $\|x' - x\|_\infty \leq 0.3$

## 5 Technical details

To implement the defense scheme, you can rely on any source you want.
The final robustness evaluation will be performed on 100 randomly chosen test examples. The evaluation procedure will be performed in two rounds. Firstly, the Auto attack [1] (designed for an adversary in the white-box setting) will be used. Secondly, to discover any gradient masking your model will be tested against the SPSA attack [2] (designed for an adversary in the black-box setting) and against adversarial examples transferred from another model.
The robustness evaluation is a crucial part of this project. You will have to demonstrate a good understanding of how to evaluate the robustness of the model. Notably, you will pay attention to gradient masking (see course) and adaptive adversaries. Carlini *et al.* [3] propose some good guidelines to evaluate the robustness of a model against adversarial examples.

## References

[1] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*, pp. 2206–2216, PMLR, 2020.

[2] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *International Conference on Machine Learning*, pp. 5025–5034, PMLR, 2018.

[3] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," 2019.