Crafting adversarial examples for the fashion MNIST dataset

1 Description

Adversarial examples [1, 2, 3] are inputs formed by applying small perturbations to examples from a dataset in order to make a NN to produce incorrect answer with a high probability. Evaluating different strategies for designing adversarial perturbations is an important topic because it helps to unsderstand and limit the vulnerability of DNNs to attacks. Due to the relevance of this topic, different Python libraries have been introduced to study and evaluate adversarial perturbations [4, 5].

2 Objectives

The goal of this project is to implement one or more approaches to create adversarial examples for the fashion MNIST dataset `https://github.com/zalandoresearch/fashion-mnist`.

The student should: 1) Decide on a class of DNNs to be used for the analysis. 2) Design the appropriate adversarial perturbation approach to fool the chosen DNN. 3) Create the adversarial examples. 4) Evaluate the performance of the DNN on created examples.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- Read the relevant bibliography about adversarial perturbations (see references above).
- Review Python libraries for adversarial perturbations [4, 5].
- Implementations can use any Python library that implements DNNs.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

[2] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*, 2016.

[3] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

[4] Nicolas Papernot, Ian Goodfellow, Ryan Sheatsley, Reuben Feinman, and Patrick McDaniel. cleverhans v1. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016.

[5] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0. 8.0: A Python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.