Prof. R. Wattenhofer

Understanding Deep Neural Networks by Fooling

Recent advances in deep learning have pushed the performance of almost all supervised learning tasks in computer vision, as well as natural language processing. Although reaching high performance, deep learning has been criticized for lacking of interpretation. Unlike traditional machine learning models such as support vector machine or random forest, which often comes with plausible explanations to model predictions, deep learning models run like a black-box, decisions of which are hard to explain.

To better interpret decisions of deep learning models, researchers have proposed to understand deep learning models by adversarial attacks. An adversarial attack can generate adversarial examples, which look like normal examples to human, but are able to fool deep learning models. We can have a clearer understanding of internal mechanisms of deep learning models by examining how these models are being fooled by adversarial examples.

In this project, we aim to understand state-of-the-art deep learning models by adversarial attacks. We will try to fool these models with different types of adversarial attacks to understand them from different angles. Specifically, we will focus on understanding natural language processing models like BERT, GPT, etc. You will have access to powerful GPUs, and weekly discussions with two experienced PhD students in deep learning.

Requirements: Strong motivation, proficiency in Python, ability to read papers and work independently. Prior knowledge in deep learning is preferred.

Interested? Please contact us for more details!

Contact

- Zhao Meng: zhmeng@ethz.ch, ETZ G61.3
- Damian Pascual Ortiz: dpascual@ethz.ch, ETZ G93