Prof. R. Wattenhofer

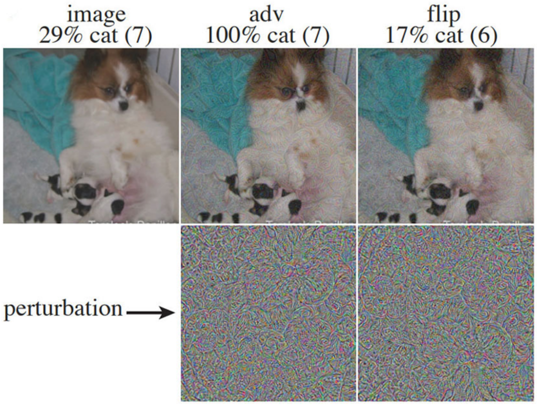
Stability of Adversarial Examples

It has been shown in many studies that some “invisible” noise can destroy the accurate classifications of state-of-the-art image models – so called adversarial examples.

However, models are often trained with random noise to the data samples as a feature augmentation step, so how is this adversarial noise different?

This project will investigate how this adversarial noise is affected under transformations, and when adding noise to the adversarial noise. The focus here is to understand why some specific noise works, and how robust it is to alterations (for instance rotations, translations, noise, etc.).

This can help an informed design process to create new models that are more robust against adversarial noise.



Requirements

Programming skills (Python, C / C++, etc.) and a good knowledge of machine learning and machine learning libraries.
We will have weekly meetings to address questions together, discuss progress, and think about future ideas.

Contact

In a few short sentences, please explain why you are interested in the project and about your coding and machine learning background (i.e., your own projects or relevant courses you have taken at ETH or elsewhere).

- Andreas Plesner: aplesner@ethz.ch, ETZ G95