ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

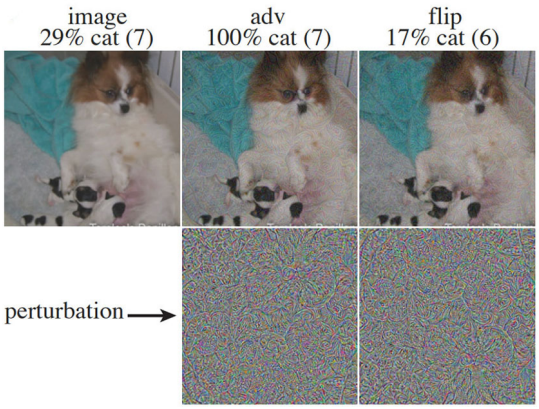
Distributed Computing

Prof. R. Wattenhofer

Exploring adversarial concepts

Adversarial attacks consist of modifying images so that an otherwise good machine learning model outputs the wrong label. These modifications are done in such a way that a human would not say that the image has changed.

An example of this can be seen on the right, where the image is incorrectly labeled as a cat under one of the perturbations. Significant research explores these adversarial attacks to understand why they occur, why the perturbations required are so small, and how to protect against them.



A parallel area of research explores explaining (large) machine learning models and why they make the predictions they do. This has, for instance, identified specific concept neurons in many large models, i.e., neurons that fire (are activated) by certain types of input.

This project aims to explore the intersection of these two areas to see which concepts are being modified when using adversarial attacks.

Some relevant resources for this project can be found here:

- https://openaccess.thecvf.com/content/CVPR2022/html/Wang_HINT_Hierarchical_Neuron_Concept_Explainer_CVPR_2022_paper.html
- https://www.anthropic.com/news/mapping-mind-language-model
- https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
- https://youtu.be/UGO_Ehywuxc?si=p0QiZ3ZkRsWi-H6z
- https://files.sri.inf.ethz.ch/website/teaching/riai2020/materials/lectures/LECTURE3_ATTACKS.pdf
- https://www.neuronpedia.org/

Requirements

Solid programming skills in Python and knowledge of machine learning evaluation are required. Experience exploring and visualizing large datasets is beneficial.
We will have weekly meetings to address questions, 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 projects or relevant courses you have taken at ETH or elsewhere).

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