ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

Prof. R. Wattenhofer

The Impact of Training Data on Adversarial Examples

Language and vision models have shown themselves to be excellent at many more general tasks over the past few years. Meanwhile, more narrowly focused models such as ResNet give excellent performance for the tasks they have been trained on, e.g. image classification or segmentation.

Besides the model architectures naturally differing between these models, we also see a difference in the datasets used to train these models. The classic datasets such as ImageNet, COCO, and CIFAR contain some specific styles and features that uniquely identifies the dataset. This concept is closely related to provenance detection, which is a technique used, for instance, to detect data poisoning.

In this project, we want to evaluate current adversarial results' dependency on the data used in the experiemnts. We seek to understand if the data could impact the conclusions drawn in modern experiments and if so how and why this happens.

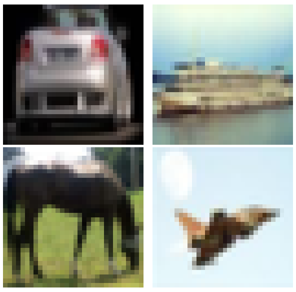

Figure 1: CIFAR10 images

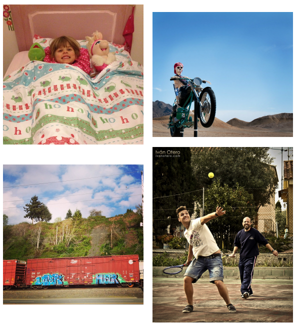

Figure 2: COCO images

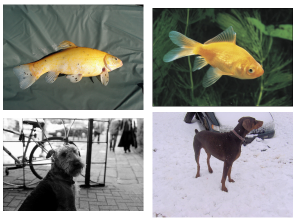

Figure 3: ImageNet images

Requirements

Programming skills in Python and a good knowledge of machine learning along with machine learning libraries like PyTorch.
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
- Till Aczel: taczel@ethz.ch, ETZ G60.1