Project Proposal

Jiaming Zeng

This project would be based on Professor Prakash's research on physical-world attacks on deep learning models. While deep learning models are capable of classifying images and applied to real-world applications, it is vulnerable to adversarial attacks from small-magnitude perturbations added to the input data. Input would become mislabeled by the deep learning models with the small perturbation and the model would fail. In order to detect an adversarial attack and improve the accuracy of the final result of deep learning models, this project proposes the idea of pre-filtering the input with extra data analysis methods based on semantic similarity. It also has to be precise for filtering all the images which would be mislabeled by the original model.

In this project, machine learning and statistical analyzation techniques I learned from STATS 415 and STATS 406 might be applied to generate a better filter. Every Wednesday, Professor Prakash, Ryan Feng, a PhD Student and I would meet over zoom or in person. In the end of the semester, there would be a report describing the result of the filter and whether it would work. The deliverables in the report would be experimental results and summarization from different semantic measures from at least one dataset including GTSRB, and the evaluation of their effectiveness on adaptive adversarial attacks. The source code will be provided under the MIT open source license.