Thesis Project Portfolio

Fooling Question Answering Deep Learning Models with TextAttack

(Technical Report)

An Ideological Exploration of Safe Machine Learning

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Grant Dragon Dong

Spring, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

With the advent of data availability and increasing computing power, Machine Learning (ML) has become a heavily depended-upon asset for everyone in the technological industry. With so much riding on the results of ML models, it is vital that model development is continuously iterated upon to improve the accuracy and robustness of these systems. Bias is the main issue plaguing ML systems that are the best of the best and is still an ongoing research endeavor. Adversarial ML is a well-known and ongoing research method that is used to find biases or any other weaknesses hidden in ML models through attacks. The following technical thesis involves the exploration of TextAttack, a major library that uses adversarial ML to easily discover flaws in natural language models. The STS thesis involves analyzing the bigger picture of the impact of ML bias on various stakeholders and the differing ideologies on how to frame the solution to mitigate these prevalent issues in ML systems.

The results of my technical project yielded a plan to extend the existing TextAttack library to work for attacks on Question-Answering (QA) models. The scope has broadened significantly throughout the years, but in simplified terms, QA is concerned with building systems that automatically answer questions posed by humans in a natural language. There are a plethora of applications within the domain of QA, thus, it is essential that TextAttack can be compatible with QA research to easily test, evaluate, and find hidden weaknesses within QA models. My STS project revealed three distinct ideologies for addressing the prevalence of bias in ML: bias is a computational and algorithmic challenge to overcome via new model development methods, data is the source of all bias and data feminism is needed to combat biases that occur during data collection processes, and affirmative action for discriminated groups is necessary because bias will always be an inevitable byproduct of ML systems.

My implementation for QA task support in TextAttack allows for the user to successfully run QA models and execute attacks. Throughout the project, it occurred to me that there are copious amounts of QA model types, and I was only able to extend TextAttack to one specific type. Generalizing TextAttack to all QA models is a much larger endeavor. Overall, I learned a lot about natural language processing and the different techniques involved in running and creating various NLP models. I also gained more familiarity with source control; TextAttack was the largest code base I have worked on with multiple research partners. I feel that my experience significantly helped prepare me for industry standards. My STS research gave me a better understanding of the possibilities and opportunities we available to make ML systems more robust. I learned that even though TextAttack and most other tech companies focus on a specific ideology for solving the issues with ML, there are other ways people can combat the negative effects of bias that I have never even considered.