Thesis Project Portfolio

Fooling Question Answering Deep Learning Models with TextAttack

(Technical Report)

Bias in Machine Learning and Diversity

(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

Srujan Joshi

Spring, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

With the rise of Machine Learning solutions in the past decade, a major problem that has arisen is that of bias in these solutions. Most often bias in ML models stems from bias in training data and it reveals itself in the form of inequitable predictions. One of the technical solutions to bias is Adversarial Machine Learning. This is a methodology which involves "attacking" (making modifications to) the input data in order to "fool" the model into giving undesirable outcomes. This can be used to pinpoint the flaws and biases in a Machine Learning model, and to iteratively train Machine Learning models to "unlearn" bias. Another way of solving the bias problem is to remove bias in the datasets that Machine Learning models are trained on by increasing diversity at all stages in the pipeline of ML solution engineering.

The technical thesis details my work on the "TextAttack" Python Library. TextAttack is open-source software developed in-house at UVa that aims to standardize and modularize the process of carrying out attacks on Text-based Machine Learning models. TextAttack also helps researchers easily reproduce attacks described in research literature. My general responsibilities included code review. One of my major contributions was extending the software to support attacks on Question Answering ML models, which are ML models which answer a question based on a given context. The primary vector for this attack was making slight changes to the context to make the model give out incorrect answers.

The STS Thesis explores the idea that the root cause of bias in Machine Learning is the lack of diversity and inclusivity in the ML solution creation process. Using a sociotechnical framework, I conducted a survey of the current state of bias in ML, analyzing the lack of diversity in the tech industry and looked at ways in which bias can be mitigated by promoting the interests of those who are otherwise ignored during ML solution creation due to societal power dynamics.

Working on both of the theses has been a fruitful endeavor. The technical work improved my understanding of not only Adversarial Machine Learning but also Natural Language Processing, Python Package Maintenance, and Version Control through Git. The research behind the STS thesis challenged my notions of how bias in Machine Learning solutions could reveal itself and has made me more cognizant of the hierarchies of power that come into play during the creation of not only Machine Learning solutions but engineering solutions in general. I believe the takeaways from the two theses will help me become a well-rounded and mindful engineer

I would like to end by thanking my research partner Grant Dong, my research professor Yanjun Qi, and my STS Professor Sean Ferguson for all the help and support they have provided as I went about my technical and STS research.