TEXTATTACK RESEARCH PROJECT

MITIGATING BIAS IN MACHINE LEARNING

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science
By

Srujan Joshi

November 1, 2021

Technical Team Members:
Grant Dong
Hanyu Liu
Chengyuan Cai
Sanchit Sinha

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Sean Ferguson, Department of Engineering and Society

Yanjun Qi, Department of Computer Science

Introduction

Machine Learning and more broadly Artificial Intelligence has the potential to change the world. In fact we've already begun to see the ways in which advances in ML have shaped the way we live whether it's through improvements in healthcare, transportation, education or entertainment just to name a few.

Natural Language Processing is the field of Artificial Intelligence which is concerned with giving computers the ability to understand text and spoken words in the same way humans can (IBM). Like the broader field of AI, massive strides have been made recently in the field of NLP. Speech recognition, Entity Recognition, Sentiment Analysis and Natural Language generation are amongst one of the many things that NLP models can be trained to do. For the average person these technological advancements manifest themselves in the form of Voice Assistants like Siri or Google Assistant, Voice Input Keyboards in Smartphones, Sentence Completion such as in Google Docs, Machine Translation such as in Google Translate among several other applications.

In the technical portion of my paper I will be talking about the TextAttack open-source Python library, which is the subject of my undergraduate research. TextAttack can be seen as an "all in one" package for performing Adversarial Machine Learning on Natural Language Processing models, with the ability to train, improve and attack popular NLP models all in one succinct codebase.

Businesses and society as a whole have benefited greatly from the AI revolution. But as they say, with great power comes great responsibility. In the STS part of this paper I will be looking at how businesses have a great responsibility when it comes to making sure that the use of AI does not lead to inequitable outcomes for different groups of people. I also go on to suggest a possible way to remedy bias in AI through the use of Adversarial Machine Learning. This is loosely coupled with my research on Adversarial Machine Learning applied to Natural Language Processing Models as a part of my work on TextAttack.

Technical Topic : TextAttack Research

My independent study involves working on TextAttack, an open-source Python library created by the University of Virginia's QData Research Team.

TextAttack is a framework for adversarial attacks, adversarial training, and data augmentation in Natural Language Processing. With over 1.7k stars and 200 forks on GitHub, TextAttack has a sizable user base in the Open Source Natural Language Processing Community (TextAttack).

TextAttack makes it easy to experiment with popular existing Natural Language Processing Models. It's also useful for Natural Language Processing model training, adversarial training, and data augmentation.

TextAttack provides components for common Natural Language Processing tasks like sentence encoding, grammar-checking, and word replacement that can be used on their own without having to perform an adversarial attack.

TextAttack can primarily be invoked in two ways, through the command line or via its Python API. Below are a few examples of invocations of TextAttack:

1. For Adversarial attacks on Pretrained Models:

   Python: `textattack.shared.Attack`

   Bash: `textattack attack`

2. For Data augmentation to generate more training examples:

   Python: `textattack.augmentation.Augmenter`

Bash: `textattack augment`

3. For Model training from scratch:

   Python: `textattack.commands.train.*`

   Bash: `textattack train`

Listed below are the primary reasons that the NLP Open Source Community uses TextAttack:

1. To understand popular pre-trained NLP models better by running different adversarial attacks on them and examining the output.
2. To research and develop different NLP adversarial attacks using the TextAttack framework and library of components.
3. To augment datasets with extra training data to increase model generalization and robustness downstream

4. To train NLP models from scratch.

My specific work so far has been reporting issues, reviewing pull requests, and adding to the documentation. Since TextAttack is open-sourced through a publicly available GitHub repository, the process of contributing to it is stream-lined since GitHub is an industry standard that is built for seamless collaboration on projects between large numbers of developers. Anyone who wants to add functionality to an application first has to "clone" the repository and then add any changes to their clone. They can then propose a "Pull Request" which is a request to incorporate the changes they made on their clone into the main codebase (GitHub Docs). One of my responsibilities was to check out pull requests made by other developers and run tests on them to ensure that the functionality that they proposed to add worked as expected and also to make sure that existing functionality was not compromised.

GitHub also allows users to report issues with a particular project. Issues can be given tags to indicate the nature of an issue, for example an issue could be reporting a bug, suggesting an enhancement to the feature-set, Another responsibility of mine was to comb through the issues and tag them accordingly.

One of the reasons why TextAttack is popular in the NLP Open Source community is because of its extensive documentation, which makes the source code and functionality easily understandable and accessible to ML practitioners. As someone who was new to TextAttack themselves, during the first half of the semester my work involved reading through the documentation and making sure that it was easily understandable even to those who were not familiar with TextAttack. One of the barriers to entry when it comes to performing Machine Learning Tasks is having a dedicated high performance graphics card (Dzousa, 2020). My work had me improving the Tutorials for TextAttack by making sure that the models and attacks used in them would run in a reasonable time even on computers without dedicated high performance graphics cards.

STS Topic: Mitigating bias in Machine Learning

Algorithmic bias in Machine Learning should be a concern to everyone since Machine Learning, and more broadly Artificial Intelligence, is directly/indirectly behind most decisions being made today whether that is in healthcare, finance, education, or security. This bias can lead to unfair and inequitable outcomes based on race, gender, sexual orientation, and ideology. For this paper I will be using a socio-technical framework to examine the immense roles that businesses, and big corporations at the forefront of the AI revolution have in mitigating bias in

Machine Learning and how adversarial Machine Learning might be a possible way to mitigate this issue.

Artificial Intelligence, just like humans, can be biased. This is because AI systems learn decision making based on training data which can include biased human decisions or reflect historical and social inequalities. For example, Amazon stopped using a hiring algorithm after discovering that it was biased towards selecting applicants whose resumes had words which males typically used. Bias can also be due to flawed data sampling. For instance, researchers at MIT have shown that facial analysis technologies showed higher error rates for minorities and women, and that this was due to unrepresentative training data (Mayinka, Silberg, 2019).

With regards to Natural Language Processing specifically (since this is the field of Machine Learning which most closely relates to my technical research), most of the bias stems from the fact that NLP models use "word embeddings". A word embedding is a mechanism of identifying hidden patterns in the words of a given block of text, considering language statistics, grammatical and semantic information as well as human-like biases. This intentional accounting of human-like bias sometimes has undesirable consequences and manifests itself in downstream applications which make use of NLP models (Caliskan, 2021).

Part of the problem lies in the fact that there are a few big players such as Google, Apple and Facebook who make progress on the latest Machine Learning Models which are then open sourced for the world to use. Thus, the responsibility for building unbiased Machine Learning Models falls on the shoulders of a few engineers at some of the largest tech firms and not on the broader public. This can prove to be problematic. For example, Google's Switch Transformer and T5 models which are directly/indirectly used by other businesses for Natural Language Processing applications were shown to have "been extensively filtered" to remove black and

Hispanic authors, as well as materials related to gay, lesbian, and other minority identities. (Anderson, 2021). A possible solution to this is to have a more diverse and inclusive group of engineers and researchers involved in constructing AI models and the datasets that they are trained on.

Governments have mostly taken a hands-off approach with regards to regulating AI, but they have started to intervene more recently. Back in 2019 for example, New York State's Department of Financial Services investigated the algorithm behind Apple Card's credit assessments for allegedly giving women lower credit limits than men (Dadkhahnikoo, 2019). The European Union in 2020 published a white paper titled "On Artificial Intelligence - A European Approach to Excellence and Trust" and has also passed a 2021 proposal for an AI legal framework, which among other things, makes it clear that government regulation will be involved in the development of AI tools in the EU (Carlo et al., 2021).

With the absence of governmental intervention for the most part, the onus of policing Algorithms has fallen on the private companies making/using the algorithms themselves. To this end, Companies should make sure that the researchers and engineers who collect data and architect AI algorithms should be cognizant of potential bias against minorities and underrepresented groups. Some companies seem to be taking steps in this direction. For example Google has published a set of guidelines for AI use both internally and for businesses that use its AI infrastructure (Responsible AI Practices).

One of the concrete technical solutions that can be employed by the creators of AI models to mitigate bias is the use of Adversarial Machine Learning. Adversarial Machine Learning is a type of Machine Learning that aims to fool models into giving incorrect predictions by providing intentionally deceptive input that is generated by another Machine Learning model known as an “adversary”. Machine Learning models can be iteratively improved by training on adversarial examples. Adversarial Examples can be constructed specifically to target bias and to iteratively improve the fairness of Machine Learning Models. In this context Adversarial Learning is called “Adversarial Debiasing” (Zhang et al., 2018).

Still this does not address the issue that most of the bias in ML models stems not from the models themselves but from the data they are trained on. Even when the feature causing unfair predictions is not directly inputted into the system, it may be strongly correlated with other variables which are used by the system. For example, a system that takes in zip code information may use it as a proxy for race (Fredrikson et al., 2017). To counter this, a class of ML models called conditional Generative Adversarial Networks (cGANS) have been used to generate new synthetic “fair” data with selective properties from the original data (Abusitta et al., 2020).

Recent research has shown that Adversarial Machine Learning, in the form of “Subversive AI” has further applications in “balancing power dynamics in favor of everyday Internet users” to fight back against biased ML models. The goal of Subversive AI is to enable users to protect the content they share online against automated algorithmic surveillance without affecting how that content is consumed (Das). Subversive AI would be a user-side application which would allow users to intentionally mask their private information from other AI models . This solution works by not allowing corporations to access accurate non-essential user data in the first place and thus eliminates the question of bias in outcomes.

Adversarial Machine Learning applied to Natural Language Processing is the primary focus of my research group, TextAttack, hence it is of particular interest to me. Facebook, in fact, has been using similar methods to benchmark, improve and remove bias in its NLP models. Their “Dynabench” platform tasks a combination of human experts and adversarial models with coming up with adversarial examples to create challenging, more representative new datasets to train future models on (Kiela, Williams, 2020).

Next Steps

I will continue to work on TextAttack this semester and possibly next semester as well. As someone who is still pretty new to the project and Adversarial Machine Learning in general, at the moment I am focussed on learning the theory behind Adversarial Machine Learning as applied to Text, and also on familiarising myself with the TextAttack codebase.

In chronological order these are the tasks that I have to carry out as a part of my technical work:

1.) Gain a firm understanding of Adversarial Machine Learning and Natural Language Processing techniques.

2.) Familiarize myself with the TextAttack codebase.

3.) Fix bugs that have been reported by other users on the TextAttack GitHub repository.

4.) Try to add in features that have been requested by other users on the TextAttack GitHub repository.

5.) Conduct a user-study to determine what other new features may be added to TextAttack to make it an effective tool for text based Adversarial Machine Learning and then add these features.

As for my STS research, I will continue looking into the ways that bias can be mitigated in AI algorithms, and how big corporations which make/use AI can be policed. In particular, a question that arises is whether or not Governments should play a more active role in overseeing AI algorithms, and what kind of legislation can be brought about to mitigate bias.

References

UVA QData Lab. (n.d.). *TextAttack documentation*. TextAttack Documentation - TextAttack 0.3.3 documentation. Retrieved October 17, 2021, from https://textattack.readthedocs.io/en/latest/.

Collaborating with pull requests. GitHub Docs. (n.d.). Retrieved November 1, 2021, from https://docs.github.com/en/pull-requests/collaborating-with-pull-requests.

Dsouza, J. (2020, December 26). *What is a GPU and do you need one in deep learning?* Medium. Retrieved November 1, 2021, from https://towardsdatascience.com/what-is-a-gpu-and-do-you-need-one-in-deep-learning-718b9597aa0d.

Anderson, M. (2021, September 24). *Minority voices 'filtered' out of google natural language processing models*. Unite.AI. Retrieved October 5, 2021, from https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/

Caliskan, A. (2021, May 10). *Detecting and mitigating bias in Natural Language Processing*. Brookings. Retrieved October 5, 2021, from https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/.

Mayinka, J., Presten , B., & Silberg, J. (2019, October 25). *What do we do about the biases in ai?* Harvard Business Review. Retrieved October 17, 2021, from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

Candelon, F., Carlo, R. C. di, Bondt, M. D., & Evgeniou , T. (2021, August 30). *Ai regulation is* *coming*. Harvard Business Review. Retrieved October 17, 2021, from https://hbr.org/2021/09/ai-regulation-is-coming.

Dadkhahnikoo, N. (2019, November 11). *Incident Number 92*. Artificial Intelligence Incident Database. Retrieved October 17, 2021, from https://incidentdatabase.ai/cite/92#6048603491dfd7f7ac0470be.

Lee, N. T., Resnick, P., & Barton, G. (2019, October 25). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings. Retrieved October 17, 2021, from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Kiela , D., & Williams , A. (2020, September 24). *Introducing dynabench: Rethinking the way we benchmark ai*. Facebook AI. Retrieved October 17, 2021, from https://ai.facebook.com/blog/dynabench-rethinking-ai-benchmarking

Datta , A., Fredrikson , M., Ko, G., Mardziel, P., & Sen, S. (2017, May 22). *Use privacy in data-driven systems*. arXiv.org . Retrieved October 18, 2021, from https://arxiv.org/pdf/1705.07807.pdf.

Abusitta , A., Aimeur , E., & Wahab , O. A. (n.d.). *Generative adversarial networks for mitigating biases in machine learning systems - ecai2020.eu*. Retrieved October 18, 2021, from http://ecai2020.eu/papers/348_paper.pdf.

Das, S. (n.d.). Subversive AI: Resisting automated algorithmic surveillance with human-centered adversarial machine learning.

Responsible AI practices. Google AI. (n.d.). Retrieved October 18, 2021, from https://ai.google/responsibilities/responsible-ai-practices

Zhang, B. H., Lemione, B., & Mitchell, M. (2018, January 22). *Mitigating Unwanted Biases with*

Adversarial Learning. arxiv.org. Retrieved October 5, 2021, from https://arxiv.org/abs/1801.07593