Streamlining the Construction of Adversarial Attacks in Natural Language Processing

The Creation of Standards for Machine Learning through Adversarial ML

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 Your Major

By
Grant Dong

November 1, 2021

Technical Team Members:
Jack Morris
Srujan Joshi
Hanyu Liu
Sanchit Sinha
Chengyuan Cai

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 is a popular buzzword for many big and influential companies, especially for companies that specialize in acquiring data. Being able to predict outcomes accurately is a very important resource and a much-desired skill to have for any individual or company. Especially in this day and age with virtually unlimited data to work with and where companies are constantly competing for the best algorithms, the popularity behind ML is well-founded. ML has reached a point in which it is almost critical to many kinds of production software systems (Rowan, 2020). However, it must be noted that ML is only as good as the data it is given to train on and the humans who design them. These ML models can easily be poisoned with bad data and bad design, and yield malevolently inaccurate results that could be disastrous if not caught during production.

In this report, I introduce a research proposal for the continued development of an ML framework called TextAttack. The primary focus of TextAttack is in the subfield of ML known as Natural Language Processing (NLP) and the utilization of adversarial ML to create attacks that help identify weaknesses in ML models for further improvement and security. Through my technical research, I can easily transition to my STS research, which will convey the importance of adversarial machine learning and its significance in the politics that would shape ML research through the risks and standards framework. There are many examples of cases where ML has failed the public due to underlying biases that were unrealized and perceptions of risk that were not considered. These mistakes could contribute to further class divide and the underprivileged being taken advantage of. Big tech companies and the government have the social responsibility to ensure that the public is not left unaware of potential dangers of ML. In general, there are many unknowns in the field of ML, and adversarial ML helps with identifying those unknowns.

Through the process of adversarial ML, all kinds of potential risks involved with bad data that can poison ML models can be easily identified, and through these metrics, a system of standards can be established for future development of ML models to mitigate potential negative effects of ML technology.

Streamlining the Construction of Adversarial Attacks in Natural Language Processing

Natural Language Processing (NLP) is a major branch of artificial intelligence that focuses on the development of ML models that can understand, interpret, and manipulate human language. NLP heavily relies on the achievements in the fields of computer science and computational linguistics to bride the gap between human communication and computer understanding (Liu, 2020). The fundamental concept of adversarial machine learning is to create deceptive inputs to trick machine learning models into outputting results that are incorrect. These results would then be used to reinforce the model to better combat those poisonous examples, and consequently, improve the overall performance and accuracy of the model. Therefore, the main components of adversarial ML involve the generation and detection of adversarial examples that are specially created to deceive classifiers. One well-known concrete example in the field of computer vision is taking an image of some object that a particular ML model would give a correct label to, then overlaying a perturbation or modification to the image such that it still looks like the original image to the human eye, but fools the model into producing an incorrect label (Boesch, 2021). Similarly, in the field of NLP, such perturbations may instead include the exchange, misspelling, addition, or deletion of words that will cause a text input to fool the NLP model. The worldwide revenue from the NLP market is forecasted to increase at an exponentiated rate in the following five years, with a growth of almost 14 times from 2017 to 2025, where NLP's market was three billion dollars in 2017, and a projection of 43 billion in

2025 (Liu, 2020). With popularity comes vulnerabilities, and with such a high market demand for NLP technology, the development of security protocols and preemptive measures to prevent poisonous attacks is critical. Therefore, the study of adversarial attacks on NLP has become a point of interest for many tech scholars.

The field of adversarial ML in conjunction with NLP is developing rapidly, and is the primary focus of my research in the development of TextAttack. TextAttack is a Python framework for adversarial attacks, adversarial training, and data augmentation in NLP. In a field where the generation and detection of adversarial examples is imperative to the enhancement of security in models, TextAttack makes experimenting with the robustness of NLP models seamless, fast, and easy, in addition to extra features including sentence encoding, grammar-checking, and word replacement. The primary component of TextAttack is to architect NLP attacks. This process is achieved through four components:

- **Goal Functions** - stipulate the goal of the attack, like to change the prediction score of a classification model, or to change all the words in a translation output.
- **Constraints** - determine if a potential perturbation is valid with respect to the original input.
- **Transformations** - take a text input and transform it by inserting and deleting characters, words, and/or phrases.
- **Search Methods** - explore the space of possible transformations within the defined constraints and attempt to find a successful perturbation which satisfies the goal function.

These are the major building blocks of TextAttack(Morris, 2021). My specific responsibilities are to continue improving the algorithms for each of these functions by researching popular NLP attack algorithms and other state-of-the-art NLP processes to help maximize efficiency and

optimize the framework. In short, using TextAttack would streamline the pipeline of the adversarial ML research process in NLP, allowing for a more efficient procedure to extract adversarial examples and improve respective NLP models.

The Creation of Standards for Machine Learning through Adversarial ML

ML technology has a lot of power, and with that span of influence, it can do a lot more harm than good if the risks are not fully considered. The power of ML is held by large, powerful groups like big corporations and the government, and this power in the hands of a few could lead to future technology development that easily overshadows the common people's viewpoints and overlook specific areas of concern that can greatly impact peoples' livelihoods differently. With machine learning rapidly becoming integral to organizations' value proposition, the need for organizations to improve the security of their machine learning is high. Hence, adversarial ML is becoming an important and growing field in the software industry (Boesch, 2021). Therefore, I will propose a structured investigation into how the advent of adversarial ML can help combat the biases and social injustice at play in these powerful ML technology proponents, and create a proper standard for ML that could close the gap between the experts and the public.

There was a recent case analysis where an algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people. The algorithm was less likely to refer black people than white people who were equally sick to programs that could improve patient care with complex medical needs. This discrepancy was simply caused by an assumption that people who accrued less healthcare cost over a year are less likely to need healthcare in the future. But this does not take into account the societal barriers between individuals of specific backgrounds, and how some people just cannot afford to spend a lot of money on health care due to financial struggles (Ledford, 2019). Furthermore, FaceApp

was a culprit of biased data collecting, where models were trained to make faces “hotter” by lightening the skin tone. This was a direct result of training data that contained a larger quantity of lighter-skinned people (Plaugic, 2017). Moreover, in another well-known case, it was discovered that Amazon’s internal recruiting tool was dismissing a disproportionate number of female candidates. Because it was trained on historical hiring decisions where the records favored men over women, the recruiting model learned to do the same (Hao, 2019). Therefore, it is clear that ML models are far from absolute, and these are prime examples of how ML can inherently discriminate against groups of people and classes, whether it be from underlying assumptions reflected in our culture, or the practice of bad data collecting. These are exactly the types of risk that should be avoided, and having a good set of standards for big companies to follow is the first step in mitigating these risks

Many large companies see the primary goal of ML systems as making decisions “at scale” without any human intervention. ML also provides legibility, making things readable that were previously not visible to the human eye. It is the combination of scale and legibility that makes ML systems uniquely attractive to governments and other institutions that aim to seek control over larger populations (Albert et al., 2020). To prevent this, one suggested standard is the utilization of conditional generative adversarial networks (cGANs). Similar to general GANs, cGAN is a type of adversarial machine learning that involves two partner models trained simultaneously, where one model is trying to create inputs to fool the other, and the other would learn and adapt to overcome them. This process is repeated until the resulting generative model can generate new synthetic data pertaining to the targeted population, and this is then used to augment the training set prior to model training to compensate and overcome the bias problem (Abusitta et al., 2020). cGANs help solve risks in ML that involve bad output due to bad training

data, like in the case of Amazon's recruiting tool and Face App, because it always ensures that the training data for ML models will be unbiased to a certain degree.

Facebook AI offers an adversarial ML process known as Dynabench. Dynabench uses a novel procedure called dynamic adversarial data collection to evaluate ML models. By measuring how easily an ML system can be fooled by humans, it can a derive a better indicator of a model's quality compared to other current static benchmarks. It is best for evaluating a model's interaction with people, who behave and react in complex and changing ways that cannot be reflected with fixed data points. As Dynabench tracks which examples fool the model and lead to incorrect predictions, the examples improve the model itself and become part of a new and more challenging Dynabench dataset to train the next generation of ML models (Kiela and Williams, 2020). Hence, the Dynabench evaluation standard allows ML models to make productive progress to serve the people and create new iterations with fewer weaknesses susceptible to risk that could harm subsets of the public.

Subversive Artificial Intelligence (SAI) is a form of adversarial ML where obfuscation filters are created for users to apply to the content they share online in a way that minimizes the differences in how that content is consumed by intended and conventional audiences, but inhibits algorithmic surveillance in reliable ways. This motivation stems from the social goal of empowering people, particularly those from communities who disproportionately bear the negative effects of algorithmic surveillance caused by bias ML, to fully use online services to connect with their communities and share their voices without fear of being discriminated against (Das, 2020). Ideally, everyone should feel equally empowered to use the latest ML technology without having the fear of discrimination. Hence, if there are certain communities that are more susceptible to ML bias, they should be given access to SAI technology to help

main an equal opportunity. In short, SAI focuses on a human-centered design process where all diverse stakeholders are taken into consideration and the model designs are evaluated directly with these stakeholders in mind.

Using GANs, Dynabench, and SAI are all valid standards that the government and large institutions should start enforcing in their ML technology development. Whether intentional or unintentional, bias in ML can be very damaging, and companies with a large span of influence need to take responsibility to adopt such standards for keeping ML models unbiased. Implementing these new forms of ML development allows for social equality, where cultural biases from large technology powers can be filtered out from the finished product, allowing all stakeholders and user groups of different classes to benefit from the technology equally. In short, adversarial ML has the potential to lower the societal risks involved with newer ML technologies to come. Embracing these standards is truly a possible outlook to bridge the gap between experts and the public.

Next Steps

The TextAttack framework has already made significant progress before my involvement in the team. I am still a relative amateur member of my research group, and in the middle of training to get more familiar with the inner-workings of TextAttack. My goals for the semester and the semester to come is to continue making considerable improvements to TextAttack to provide users a better experience for designing adversarial attacks on their models. In chronological order, these can include:

- Reporting bugs or fixing issues with existing code
- Improving the documentation
- Recommending and requesting new helpful features

- Implementing the new features
- Adding support for new models and datasets

Overall, all of these can really help enhance TextAttack and allow for it to gain traction as a major tool for adversarial ML research.

In terms of the risk and standard framework, I think there is a major responsibility for the government to enforce standards like SAI for minorities, so they can be better protected from harmful ML. This should be a safety measure implemented, whether or not companies themselves are willing to devote research to reduce ML bias. Furthermore, I aim to use TextAttack to architect adversarial attacks on multiple well-utilized models and technologies by major tech companies to see if I can detect any pattern of biases. With these examples that I find, I plan to document each of them and further analyze what is the cause for these models' weaknesses to determine if there are any systemic factors for them. Hopefully in the near future, people will start to embrace the use of adversarial ML technologies to help cross-validate ML models as a standard of consideration before production. TextAttack will be one major stepping-stone to the start of that process.

References

Abusitta, A., Aïmeur, E., & Wahab, O. A. (2020). *Generative adversarial networks for mitigating biases in ...* ecai2020. Retrieved October 17, 2021, from http://ecai2020.eu/papers/348_paper.pdf.

Albert, K., Penney, J., Schneier, B., & Siva Kumar, R. S. (2020, March 27). *Politics of adversarial machine learning*. SSRN. Retrieved October 17, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547322.

Boesch, G. (2021, July 2). *What is adversarial machine learning? attack methods in 2021*. viso.ai. Retrieved October 17, 2021, from https://viso.ai/deep-learning/adversarial-machine-learning/.

Das, S. (2020). *Subversive AI: Resisting automated ...* sauvikdas. Retrieved October 17, 2021, from https://sauvikdas.com/papers/27/serve.

Hao, K. (2020, April 2). *This is how AI bias really happens-and why it's so hard to fix*. MIT Technology Review. Retrieved October 17, 2021, from https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

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?mc_cid=6432510d66&mc_eid=cefca2850c.

Ledford, H. (2019, October 26). *Millions of black people affected by racial bias in health-care algorithms*. Nature Research. Retrieved October 17, 2021, from https://www.nature.com/articles/d41586-019-03228-6.

Liu, S. (2020, June 8). *Global Natural Language Processing Market 2017-2025*. Statista. Retrieved October 17, 2021, from https://www.statista.com/statistics/607891/worldwide-natural-language-processing-market-revenues/.

Morris, J. (2020). *TextAttack basic functions*. TextAttack Basic Functions - TextAttack 0.3.3 documentation. Retrieved October 17, 2021, from https://textattack.readthedocs.io/en/latest/0_get_started/basic-Intro.html.

Plaugic, L. (2017, April 25). *FaceApp's creator apologizes for the app's skin-lightening 'hot' filter*. The Verge. Retrieved October 17, 2021, from https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology.

Rowan, I. (2020, July 18). *The state of AI in 2020*. Medium. Retrieved October 17, 2021, from https://towardsdatascience.com/the-state-of-ai-in-2020-1f95df336eb0.