Jaydeep Borkar | Research Statement

Trustworthy Machine Learning: Fairness and Bias, Adversarial ML, Privacy-Preserving ML
Natural Language Processing

My research interests are studying and designing classifiers that are robust to different adversarial attacks, AI Bias and Fairness, and Privacy-Preserving Machine Learning. I believe that it's very important for us to do Machine Learning that is robust, unbiased, and safe from the privacy side. Since Machine Learning is capable of solving humanity's most critical problems, it's vital for it to be robust under different situations and the decision it makes should be trustworthy. Hence, from the research point of view, I'm very much drawn to studying and designing Trustworthy AI. I'm also interested in Natural Language Processing and want to make conversation between humans and machines more natural, and less awkward by identifying and removing bias in NLP systems (such as word embeddings) as well as making them robust to adversarial attacks.

Since I didn't have any access to conduct guided research as an undergrad (especially in Trustworthy ML) because I had just got started and had nobody around working on this in my college, I spent a considerable amount of time studying and understanding these fields independently. For me, this was a very challenging phase since I had nobody around to guide me with these things, but this phase taught me more than any course I've ever taken. It taught me to study and research things independently and thus inspiring me to choose research as a career. As I gave myself a considerable amount of time by consistently exploring this field, I got to interact with some amazing folks (some of them who later on became my friends) on Twitter, at conferences (such as ICLR'20), and by cold-emailing the authors. Their support and contribution have been immense, such as by giving crucial feedback, directing me to courses and interesting papers, talking about ideas, answering my (stupid) questions, etc. Below are the details of my work and journey so far.

Recently, I got an opportunity to attend and volunteer for ICLR 2020. There, and at the Trustworthy ML workshop, I met and interacted with some great researchers that have inspired me to study and do some helpful contributions in Trustworthy Machine Learning.

Adversarial Machine Learning

I started my journey in Adversarial Machine Learning through Adversarial Examples for Evaluating Reading Comprehension Systems paper of Prof. Percy Liang. I had been reading a couple of NLP papers of Percy Liang that while (such as Talking to Computers in Natural Language) and I stumbled upon this adversarial learning paper and found it really fascinating. This paper taught me what an adversarial example is, how it affects the accuracy of the classifiers, adversarial examples for Reading Comprehension Systems, and a lot more. After reading this paper, I was drawn to find more about how do we make the classifiers robust to various adversarial attacks. Certified Defenses Against Adversarial Examples--Before reading this paper, I had the view that a specific defense can resist all the different kinds of attacks, but this paper helped me in understanding that though a proposed defense is considered to be successful against the set of attacks known at the time, new stronger attacks are discovered that make the defense useless. I learned that though adversarial training provides robustness against specific attacks, it fails to generalize to new attacks, hence the arms race between attackers and defenders never ends and we need to come up with defenses that are robust to all the attacks.

I learned how various defenses have been routinely broken by emerging attacks. Works such as Adversarial Training Can Hurt Generalization and Robustness May Be at Odds with Accuracy inspired me to learn about the robustness vs accuracy tradeoff. This has got me interested in understanding the reasons behind this tradeoff and how we can potentially reduce the gap. It's quite interesting to see that this tradeoff might probably stem from the different features learned by robust and standard classifiers. Unlabeled Data Improves Adversarial Robustness further helped me to learn how semi-supervised learning and using unlabeled data can benefit adversarial robustness. Adversarial Examples Are Not Bugs, They Are Features is one such excellent work that I came across lately that explains the reason behind the origin of adversarial examples - they arise due to nonrobust features in the dataset. Hence, disentangling robust features from nonrobust ones and training the model only on robust ones using standard training results in a robust classifier.

Tentative Research Goals

All of the above works and many others not listed here have inspired me to explore and study Adversarial ML. So to summarize, below are my tentative (not exhaustive) research goals:

- Studying from where do these adversarial examples stem from
- Studying and designing classifiers that can be robust to different attacks
- Understanding the reason behind robustness vs accuracy tradeoff and potentially reducing this gap

AI Bias and Fairness

I strongly believe that a fair and unbiased AI is as important as its robustness to adversarial attacks. This motivated me to explore the fairness domain. Reading Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification was my starting point followed by The Ethical Implications of AI paper. I learned a couple of methods to mitigate the bias through Optimized Pre-Processing for Discrimination Prevention. I presented a poster on AI Bias during PyCon India 2019 as I wanted to introduce the community to the need and importance of fair and unbiased AI. During my entire poster session, I was a strong proponent of the fact that the fairness of a model is as important as its accuracy. Mitigating Gender Bias in Natural Language Processing: Literature Review taught me about various debiasing methods in NLP and has motivated me to study and develop similar debiasing methods for languages other than English. I'm especially interested in developing such methods for Indian languages that are low-resourced. To summarize - I'm very excited about identifying and mitigating the bias from the models.

I look forward to the opportunities where I can get involved and contribute to building Trustworthy Artificial Intelligence and learn from the amazing folks around.

REFERENCES

Posters

AI, why you ain't fair? : Understanding AI Bias, PyCon India 2019, Chennai, India,