Jaydeep Borkar | Research Statement

AI 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. To achieve the same, I'm presently exploring code-mixing and code-switching. I'm also studying how we can make reading comprehension systems that can have a true understanding of the human language, so presently I'm building an RC system on SQuAD 2.0.

Since I didn't have any access to conduct guided research as an undergrad, I've spent a considerable amount of time studying and understanding these fields independently. For me, this time 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. Below are the details of my work and journey so far.

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 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. This paper shows that by training on the certificates, we can obtain networks with better bounds and meaningful robustness.

I read about how various defenses have been routinely broken by the emerging attacks. Semidefinite relaxations for certifying robustness to adversarial examples taught me how semidefinite relaxation is tighter than previous relaxations and produces robustness for networks against adversarial examples. A new convex relaxation based on semidefinite programming that is significantly tighter than previous relaxations based on linear programming is really fascinating. I further explored how Adversarial Training Can Hurt Generalization and learned that how in spite of adversarial training improving the robust accuracy, it sometimes hurts the standard accuracy for the natural inputs, and hence there's still some tradeoff. Unlabeled Data Improves Adversarial Robustness further helped me to learn how semi-supervised learning and using unlabeled data can benefit adversarial robustness.

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.

I look forward to the opportunities where I can get involved to build Trustworthy Artificial Intelligence.