Arjun Rao

Research Statement

The Chinese University of Hong Kong

Homepage: arjunashokrao.me Email: arjunrao@link.cuhk.edu.hk

Abstract: My Research Interests include Computer Vision and Machine Learning. More particularly, I am interested in understanding adversarial machine learning from a *robustness* perspective, and how adversarial robustness can shed light on theoretical concepts such as *generalization* as well as *representation learning* and *transfer learning*. My long term research goal is understand generalization guarantees in over-parameterized deep nets, and to develop deep learning models which are robust to adversarial and poisoning attacks in the real world.

Adversarial Examples, which can better be visualized as imperceptible 'distributional' shifts in test-datasets are a natural consequence of the dimensionality gap between inputs and linear models on which high-dimensional inputs are trained on. They generalize across different architectures, and can be used in a 'black-box' fashion to threaten real-world deep learning models. Recent work has demonstrated the almost ubiquitous prevalence of adversarial examples in different applications of deep neural networks - encompassing classification (Image and voice classification, Facial Recognition), regression (Object detection, multi-object tracking), and reinforcement learning. The most common strategy to defend against test-time attacks has been to train models on adversarial data, thus ensuring some 'robustness' against standard attacks. Interestingly, these adversarially trained (robust) models exhibit intriguing properties not seen in standard deep nets. My current research aims on understanding and leveraging these properties to develop deep nets that are interpretable, secure, and accurate on unseen data.

Robustness and Generalization: Understanding adversarial robustness may help develop a better understanding about broad theoretical questions such as those on local minima, generalization in over-parameterized networks, or the reasoning behind flat vs sharp global optimums. The consensus regarding loss landscape geometries have been that flatter minima basins lead to better generalization. This hypothesis is a potential explanation to the robustness-accuracy tradeoff in robust models [3] since robust models qualitatively exhibit sharper minima with lesser basin volume. However, the lack of generalization in robust models is not absolute, and a large number of studies which study this generalization gap are only empirical. Explicit regularization like enforcing local linearity of the loss surface combats the sharp minima problem in robust models. It is also possible that the good generalization in robust models can be due special cases of sharp minima generalization [5].

Robustness and Representations: My research also studies how robustness leads to interpretable machine learning and why robust networks contain representations that align more with the human assumptions of intermediate neural network features. For example, previous empirical studies have shown that models trained against adversarial examples exhibit more human-perceptible input-gradient visualizations, as well as a smooth interpolation between image classes. Robust representations provide a high-level embedding of the input such that similar-looking classes have intermediate representations that are semantically similar [6]. In fact, performing adversarial training on very small perturbation sets might not even lead to robustness and security, but can still lead to better representations. While the intuition that neural networks truly learn interpretable representations is often assumed to be true, it is not certain. Answering questions on why robustness leads to better representations would equal taking a large step towards demystifying these 'black-boxes'.

Current Research: At The Chinese University of Hong Kong, my research primarily involves developing application-specific adversarial defenses. Stereo-Vision, most commonly used in autonomous driving applications were shown to be vulnerable to adversarial attacks that primarily distort the *disparity* perception in the rectified stereo pair. My research developed a completely cyber-physical approach to conduct adversarial training with left-right feature map regularization while ensuring local linearity of the loss surface [<**My Citation**>]. I also worked on more exploratory research on understanding distributed deep learning and how over-parameterized neural networks can generalize in a distributed setting. My current research is on Optical Character Recognition models and their potential vulnerability to black-box adversarial examples. Many of these approaches are the current state-of-the-art, and consistently have shown to be more robust than previous defenses.

Future Research Goals: Machine Learning has demonstrated incredible progress over the past decade. However, the disproportionate progress in empirical studies opposed to theoretical explanations have to be bridged in the near future. Broadly, I believe that a large part of my research is to understand and evaluate training algorithms for deep neural networks. An integral part of my future research goal is to see through the opaqueness of over-parameterized deep nets. More specifically, I am interested in studying the generalization guarantee that deep neural nets offer on modern datasets. It is understood that SGD - a variation of gradient descent that updates network weights based on a 'mini-batch' of input-data performs a certain 'implicit' regularization that is responsible for their excellent generalization results. However, we know very little about optimization trajectories that these powerful algorithms take, especially for over-parameterized nets with high-dimensional, highly non-convex loss landscapes. My near-term research goal is to demonstrate and visualize these phenomenon to provide better intuitions for other researchers. Secondly, I am interested in safe and robust machine learning that can be used at scale. My previous work on studying adversarial attacks on autonomous systems is an example of an effort in this interesting and needed research direction.

References

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[2] D. Stutz, M. Hein, and B. Schiele, "Disentangling adversarial robustness and generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6976–6987.

[3] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SyxAb30cY7

[4] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang, and A. L. Yuille, "Adversarial attacks beyond the image space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4302–4311.

[5] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1019–1028.

[6] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945*, 2019.