Neural Networks Learn Lazily: Improving Generalization and Adversarial Robustness via Learning Capacity-Complexity Constraints

Phung, Andy (School: Independence High School)

The existence of adversarial examples, which are inputs that force neural networks to misclassify them, represents a major risk in safety-critical applications of deep learning, such as malware detection and vision systems for self-driving cars. To reconcile existing notions on the origins of adversarial examples, I hypothesize that non-generalizing subnetworks that form during training are the primary cause, and that this is largely due to their poor generalization outside of the training data. These subnetworks form because their parameter initializations happen to coincide with useful, but non-generalizing correlations in the data—as such, higher network learning capacity (or network size) will result in a larger non-generalizing subnetwork. With this "lazy learning" hypothesis, I introduce Simulated Neurogenesis, a training algorithm that controls learning capacity and the structural complexity of the data to mitigate their formation. After confirming that my algorithm does mitigate the formation of these subnetworks, I calculated 95% bootstrap confidence intervals for the mean differences between Simulated Neurogenesis' misclassification rate and that of existing adversarial defenses—the relatively large differences show that this algorithm vastly outperforms existing adversarial defense methods. I also examine its convergence rate and show that it is competitive with that of industry-standard training algorithms. Finally, I discuss the implications of my research and future work, which may include further investigating the properties of these non-generalizing subnetworks and adapting my algorithm to more complex neural architectures.

Awards Won:

Fourth Award of $500

Association for the Advancement of Artificial Intelligence: Honorable Mention

Association for the Advancement of Artificial Intelligence: AAAI Membership for the School Libraries of All 8 Winners (in-kind award / part of 1st-3rd prize and honorable mentions' prize)

National Security Agency Research Directorate : Third Place Award “Cybersecurity”