Abstract: Deep neural networks (DNNs) provide excellent performance across a wide range of classification tasks, but their training requires high computational resources and is often outsourced to third parties. Recent work has shown that outsourced training introduces the risk that a malicious trainer will return a backdoored DNN that behaves normally on most inputs but causes targeted misclassifications or degrades the accuracy of the network when a trigger known only to the attacker is present. In this talk, we provide the first effective defenses against backdoor attacks on DNNs. We implement three backdoor attacks from prior work and use them to investigate two promising defenses, pruning and fine-tuning. We show that neither, by itself, is sufficient to defend against sophisticated attackers. We then evaluate fine-pruning, a combination of pruning and fine-tuning, and show that it successfully weakens or even eliminates the backdoors, i.e., in some cases reducing the attack success rate to 0% with only a 0.4% drop in accuracy for clean (non-triggering) inputs. Our work provides the first step toward defenses against backdoor attacks in deep neural networks.

Bio on the presenter: Brendan Dolan-Gavitt is an Assistant Professor in the Computer Science and Engineering Department at the NYU Tandon School of Engineering. He holds a Ph.D. in computer science from Georgia Tech (2014) and a BA in Math and Computer Science from Wesleyan University (2006). His research interests include software security, reverse engineering, embedded systems and the security of deep learning.