Keynote Talk #3

VERIFYING NEURAL NETWORKS AGAINST BACKDOOR ATTACKS

Prof. Sun Jun
Singapore Management University (SMU)

Abstract: Neural networks have achieved state-of-the-art performance in solving many problems, including many applications in safety/security-critical systems. Researchers also discovered multiple security issues associated with neural networks. One of them is backdoor attacks, i.e., a neural network may be embedded with a backdoor such that a target output is almost always generated in the presence of a trigger. Existing defense approaches mostly focus on detecting whether a neural network is ‘backdoored’ based on heuristics, e.g., activation patterns. To the best of our knowledge, the only line of work which certifies the absence of backdoor is based on randomized smoothing, which is known to significantly reduce neural network performance. In this work, we propose an approach to verify whether a given neural network is free of backdoor with a certain level of success rate. Our approach integrates statistical sampling as well as abstract interpretation. The experiment results show that our approach effectively verifies the absence of backdoor or generates backdoor triggers.

Biography: SUN, Jun is currently a professor at Singapore Management University (SMU). He received Bachelor and PhD degrees in computing science from National University of Singapore (NUS) in 2002 and 2006. He has been a faculty member since 2010. Jun’s research interests include formal methods, software engineering, and cyber-security. He is the co-founder of the PAT model checker. He has published more than 250 articles and conference papers including top conferences in multiple areas.