Rajat Sethi

CPSC 8580

# Review of “An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks”

# In this paper, the researchers have created a special type of Trojan-ware that targets large DNNs. Their idea was to poison the data set with images that had small “triggers” like QR code fragments, which would then correspond with a different, malicious label. Using this Trojan data, a malicious entity could manipulate almost any prediction by adding one of those triggers, ultimately destroying the integrity of the neural network.

# One of the other major contributions of this Trojan-ware is its ability to stay hidden. The triggers are too negligible to alert any malware checkers. For regular use, the DNNs saw almost zero drop in performance. That is, whenever the DNN was used for its intended purpose, it worked almost 100% of the time. Likewise, when the triggers were included, there was a 100% success rate in the attack.

# I do not see many weaknesses in this paper, the group was quite thorough in their report. They used TrojanNet on many datasets like YouTube’s Aligned Face and ImageNet. Each of their trials shows an outstanding success rate, almost too good to be true. The only main issue I see with the paper is their attempt to detect their own malware. That section of the article, Section 2.5, felt incomplete with a set of complex formulas and little explanation. As they mentioned in that section, future work is needed to detect trojans. If a defense is not found quickly, computer scientists will be left with a dangerous malware and no power to stop it.

# As mentioned earlier, the largest improvement to this paper would be a trojan defense instead of offense. Right now, a malicious entity could use this paper and infect a DNN, and no one would be able to stop them. The malware is too good, as the paper implies. Cybersecurity engineers cannot use the paper because the defense mechanisms provided are not detailed. This means that the first step we should be taking, is protecting our models against Trojan-wares.