Adversarial Perturbations Against Deep Neural Networks for Malware Classification

Jeff Malavasi
Dept. of Computing Security
Rochester Institute of Technology
Email: jm3378@rit.edu

I. Review

The authors of the paper explore how malware classifiers based on deep neural networks are vulnerable to carefully crafted input attacks. These perturbations allow a bad actor to alter a known malicious file by slightly adjusting features that are used by the classifier, but ultimately do not change the function of the binary. In this paper the authors both demonstrate successful attacks against a deep network, and how system designers can defend against these vulnerabilities [1].

The researchers began the experiment by creating a novel malware classifier based on a deep network. The classifier focused on statically determined features such as system call usage, hardware and network access, which were represented using binary indicator vectors. The network was trained using gradient descent and utilized the DREBIN dataset containing a variety of Android APK samples [1]. It is important to note that this dataset is skewed heavily towards benign samples, so the researchers split the data into batches and varied the ratio of benign and malicious samples during each training epoch. After reaching a sufficient accuracy rate, the researchers fed the model inputs that contained slight modifications to known malicious samples. This process is iterated until a round limit is reached or a successful misclassification occurs. Through this algorithm, the authors were able to misclassify between 60% and 80% of the malicious samples. The performance of the perturbations were heavily depending on the ratio of malware to benign samples during the training phase.

After successfully misclassifying the network, the researchers tested the following defense techniques to see if they could correct the model: feature reduction, distillation and retraining. While it may seem intuitive that reducing features would reduce the classifiers sensitivity to change and lower the overall attack surface, the authors found an inverse effect. In fact both simple feature reduction and reduction through mutual information increased misclassification as high as 99%. Additionally, they found that both distillation and retraining had a strengthening effect on the model, but did not prove to be as beneficial as these techniques have been in other domains such as image processing [1].

The researchers were able to successfully create a malware classifier, attack it, and then provide solutions to defend against the proposed attacks. They examined multiple defensive strategies in order to provide a defense in depth solution. However, the proposal also had a few weaknesses. Firstly, the model only focused on static features due to the difficulty in generating dynamic perturbations. In the future, the authors may want to consider a hybrid model, which may help to reduce the false positive rate. Additionally, the dataset was heavily unbalanced which has been shown to decrease the accuracy of malware classifiers. In fact, the researchers noted a 7% false positive rate, which was likely due to the selected dataset. While the model still achieved 98% accuracy, it likely wouldn't be suitable as a commercial solution. This skew also heavily drove the misclassification rate as researchers noted that a higher malware ration resulted in a lower misclassification rate.

Machine learning is becoming incredibly popular in malware detection, as the rate of new malware is exponentially growing. While these systems have shown to be extremely accurate, it is also important to study how these networks can be attacked in order to build robust, secure by design systems. The researchers successfully demonstrated how a malware classifier based on deep learning could be attacked and ultimately defended. Future research will be needed in order to improve upon these defenses as the attack landscape continues to evolve.

References

[1] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," in *arXiv*, 2017.