Query-Efficient Black-Box Attack Against Sequence-Based Malware Classifiers

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

I. Review

The authors of the paper introduce the first query-efficient attacks against API call-based malware classifiers. When an attacker is attempting to misclassify a sample that is dynamically analyzed, they typically pay per request as well as may encounter rate limits on the destination server. Due to these limitations, the researchers aim to reduce the overall number of requests needed to generate an adversarial sample. The authors propose multiple architectures and were able to achieve an attack success rate of 98% [1].

A black-box attack occurs when the bad actor has no knowledge of the underlying training model. In the case of a malware classifier, the bad actor may only have knowledge of the class of each input, or in rare cases the confidence score generated by the model. The researchers propose three different methods in order to reduce the number of queries needed to misclassify a sample. The first, which assumes that the attacker has both knowledge of the class as well as the confidence score, uses a gradient free optimization algorithm. The second uses a generative adversarial network which is trained against benign API calls to quickly determine which calls to add. Finally, the researchers propose using logarithmic backtracking, which starts by making a large amount of perturbations and then slowly reducing them to the minimum needed to misclassify the input [1]. In order to train their model, the authors selected a dataset of 360,000 samples that were evenly split between malicious and benign.

The researchers then used their adversarial network to attack various classifiers. They measure their success against previous work in the field and found that their proposed attack was extremely effective at changing the class of the input. Additionally, they were able to reduce the number of queries needed to generate an adversarial example by 99%. The researchers also created a framework (BADGER) that implements their attack based on a malicious input [1]. The framework is able to take in a variety of malware families in order to generalize the attack and does not require access to the underlying source code. Finally, the authors tested their attack framework against hybrid classifiers that used both static and dynamic features to analyze the sample input. They found that their framework was not as successful, only achieving a maximum success rate of 90%, but often as low as 30%.

In this paper, the researchers develop the first successful black-box attacks against dynamic malware classifiers while significantly reducing the number of queries needed to create the adversarial example. They show that their framework can be applied to many different classifiers and was even effective at analyzing multi-feature datasets. By generalizing their model, the researchers believe that the attack can be applied to other domains beyond malware classifiers. However, there was also a few limitations in the research. For example, their attack framework did not apply as well to hybrid classifiers which are gaining in popularity as they can increase the accuracy of the training model. Futhermore, they observed a similar reduction in accuracy when analyzing both the API calls and their associated arguments. This is likely due to much larger feature space created during training.

When designing an adversarial attack network, there are often many constraints. Primarily, sending test samples can be extremely costly and developing an attack framework that aims to minimize the number of requests needed would greatly accelerate the development time of a model. The researchers were able to reduce the number of requests needed by an order of magnitude, while still creating a very effective attack. Future research will be needed in order to apply the proposed framework to other domains as well as improve the ability to evade a hybrid classifer.

References

[1] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, “Query-efficient black-box attack against sequence-based malware classifiers,” in *Annual Computer Security Applications Conference*, ser. ACSAC ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 611–626. [Online]. Available: https://doi.org/10.1145/3427228.3427230