Generic Black-Box End-to-End Attack against RNNs and Other API Call-based Malware

Description

We present a black-box attack against API call-based machine learning malware classifiers, focusing on generating adversarial API call sequences that would be misclassified by the classifier without affecting the malware functionality. We show that this attack is effective against many classifiers due to the transferability principle between RNN variants, feed forward DNNs, and traditional machine learning classifiers such as SVM. We further extend our attack against hybrid classifiers based on a combination of static and dynamic features, focusing on printable strings and API calls. Finally, we implement GADGET, a software framework to convert any malware binary to a binary undetected by malware classifiers, using the proposed attack, without access to the malware source code. We conclude by discussing possible defense mechanisms against the attack.