IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection

Rohit Das

October 19, 2019

Machine learning based classifier has become a popular choice for detection of various form of malware. However, an attacker can modify malware by adding some perturbation(NOPE operation, adding some extra segment, appending noise at end,..)such that they can alter the decision of ML-based classifier (while preserving their malicious properties). This kind of malware are known as adversarial malware. The previous machine learning based malware classification fails in case of adversarial malware. In this paper we are exploring the defence for adversarial malware.

Our defence is based on feature squeezing, it maps the input feature space to a reduced feature space, such that adversarial features filter out. Once we have clean feature we can compare the prediction result between true feature and clean feature, and if they varry then the system signals adversarial malware. The level of adversarial hardening will depend on the number of feature squeezer used as ens-amble for detecting the classification result.in this work, we crafted the adversarial malware samples using MalGAN. MalGAN is generative modelling based algorithm for creating adversarial algorithm. We have combined the feature squeezing method with adversarial training to achieve the adversarial hardening.

Training Denoising auto encoder as feature squeezer was a bit challenging. Our Analysis in this work is based on static analysis only. We can also incorporate dynamic analysis and online training methods in the process to tackle problem of detecting zero day malware.