The vulnerability of transformers-based malware detectors to adversarial attacks

October 18 2022, by Ingrid Fadelli

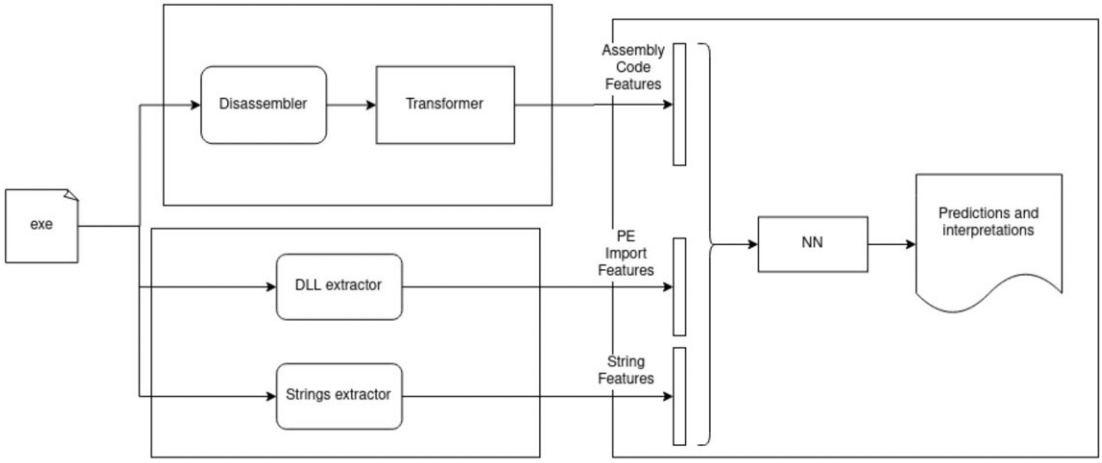

Figure 1: System Architecture

Credit: Jakhotiya, Patil, and Rawlani.

Cyber attackers are coming up with increasingly sophisticated techniques to steal users' sensitive information, encrypt documents to receive a ransom, or damage computer systems. As a result, computer scientists have been trying to create more effective techniques to detect and prevent cyber attacks.

Many of the malware detectors developed in recent years are based on machine learning algorithms trained to automatically recognize the patterns or signatures associated with specific cyber attacks. While some of these algorithms achieved remarkable results, they are typically susceptible to adversarial attacks.

Adversarial attacks occur when a malicious user perturbs or edits data in subtle ways, to ensure that it is misclassified by a machine learning algorithm. As a result of these subtle perturbations, the algorithm might classify malware as if it were safe and regular software.

Researchers at the College of Engineering in Pune, India, have recently carried out a study investigating the vulnerability of a deep learning-based malware detector to adversarial attacks. Their paper, pre-published on arXiv, specifically focuses on a detector based on transformers, a class of deep learning models that can weigh different parts of input data differently.

"Many machine learning-based models have been proposed to efficiently detect a wide variety of malware," Yash Jakhotiya, Heramb Patil, and Jugal Rawlani wrote in their paper.

"Many of these models are found to be susceptible to adversarial attacks—attacks that work by generating intentionally designed inputs that can force these models to misclassify. Our work aims to explore vulnerabilities in the current state of the art malware detectors to adversarial attacks."

To assess the vulnerability of deep learning-based malware detectors to adversarial attacks, Jakhotiya, Patil, and Rawlani developed their own malware detection system. This system has three key components: an assembly module, a static feature module and a neural network module.

The assembly module is responsible for calculating assembly language features that are later used to classify data. Using the same input fed to the assembly module, the static feature module produces two sets of vectors that will also be used by to classify data.

The neural network model uses the features and vectors produced by the two models to classify files and software. Ultimately, its goal is to determine whether the files and software it analyzes are benign or malicious.

The researchers tested their transformers-based malware detector in a series of tests, where they assessed how its performance was affected by adversarial attacks. They found that it was prone to misclassifying data almost 1 in 4 times.

"We train a Transformers-based malware detector, carry out adversarial attacks resulting in a misclassification rate of 23.9% and propose defenses that reduce this misclassification rate to half," Jakhotiya, Patil and Rawlani wrote in their paper.

The recent findings gathered by this team of researchers highlight the vulnerability of current transformers-based malware detectors to adversarial attacks. Based on their observations, Jakhotiya, Patil and Rawlani thus propose a series of defense strategies that could help to increase the resilience of transformers trained to detect malware against adversarial attacks.

These strategies include training the algorithms on adversarial samples, masking the model's gradient, reducing the number of features that the algorithms look at, and blocking the so-called transferability of neural architectures. In the future, these strategies and the overall findings published in the recent paper could inform the development of more effective and reliable deep learning-based malware detectors.

More information: Yash Jakhotiya, Heramb Patil, Jugal Rawlani, Adversarial attacks on transformers-based malware detectors. arXiv:2210.00008v1 [cs.CR], arxiv.org/abs/2210.00008

github.com/yashjakhotiya/Adver … acks-On-Transformers