Adversarial Machine Learning in Image Classification: A Survey Toward The Defender's Perspective

The paper focuses on formalizing the adversarial attacks in the field of image classification from a defender's perspective. This includes exploring different categorizations and introducing new taxonomies. Even though the paper is not in any way linked to malware analysis, the standards and principles in adversarial attacks easily translate from the domain of image classification to malware analysis.

In the area of perturbations in image classification, the major forms of classification are across three axes:

1. Perturbation scope: This defines whether the perturbations are generated for specific images or sets of images or if the perturbations are generated universally, without considering the image it'll be applied to.
2. Perturbation Visibility: This talks about how visible the perturbations are in the context of image classification. So an optimal perturbation would be the ones that are invisible to the human eye but still manages to fool the classifier. Indistinguishable perturbations remain unnoticed by humans and fail to fool the classifier. Visible perturbations can be easily spotted by humans but can fool the model. Physical perturbation is something that is not translatable to the domain of malware. It is perturbations added in the real world space (for example stickers over the Stop sign that hinders the image classifiers). Fooling images also fall under the discussion of perturbation visibility, where the images are so perturbed that humans cannot recognize them at all, but the model still classifies it as a certain class with good confidence. Noise are non-malicious perturbations that are in the image which may hinder the classifier (blurry images fall under this category).
3. Perturbation Measurement: The metric using which the difference between the original image and the perturbed image is measured falls under perturbation measurement. The standard norms used are the L0 norm, L1 norm, L2 norm, and $L\infty$ norm.

The next area of discussion is regarding the different attacks and classifications of these attacks. From a defender's perspective standardized taxonomy of the attacks are important to specify the threat model which forms the starting point of a defensive strategy as well as its metric of success.

1. The first axis is regarding the attacker's knowledge. This is where the attacks are classified into white-box attacks (Full access to the model, its parameters, etc), grey-box attacks (partial information), and black-box attacks(No information about the model).
2. Attackers influence: Defines the amount of influence the attacker has, which translates to the areas where the attacker can attack and influence the model. For example, an attacker with influence

over the training stages of the model can use a poisoning attack to influence the model maliciously right from the beginning. This is opposed to an attacker with lesser influence who can only employ his strategies during the testing stages.

3. Attacker's specificity: This determines whether the attacker tries to ensure that the model categorizes the adversarial sample into a class that the attacker desires (targeted attack), or whether the attacker is only trying to ensure that the model misclassifies the sample (untargeted attack).
4. Attacker's computation: Whether the attack is over a series of small perturbations that fine-tunes the sample increasingly over multiple iterations (Iterative Algorithms) or if the strategy finds the perturbation that should be applied to the sample which is then applied to the sample in a single step (Single Step process).
5. Security violations: This axis defines what security factor is violated by the attack such as availability (Denial of service on the model), integrity (creating classifiers that evade the sample making the integrity of the model's classification questionable), or privacy(gathering information regarding the model's architecture or training data)
6. Attack approach: There are different approaches that a malicious actor can take toward creating an adversarial sample. For example,
    - Gradient attack - When the attacker has access to the model or knows the parameters of the model, he can find the optimal adversarial sample by employing the mathematical gradient approach.
    - Score-based attack - The targeted model would be queried and the resulting scores would be used to create a surrogate model to convert the black box into a white box problem.
    - Decision-based attack - The softmax layer of the targeted model is queried and minor perturbations are made iteratively through an approach of rejection sampling.
    - Approximation-based attack - An approximation is made from a differentiable function which is then used by the gradient approach to finding an optimal perturbation that could be applied to generate the adversarial sample.

There are several algorithms that work across different areas of the attack axes described above like Fast Gradient Sign Method, Basic Iterative Method, Deep Fool, Carlini and Wagner Attack, etc. Each of these attacks would have different values in each of the axes. For example some of them would be iterative (attacker's specificity) and would focus on integrity(security violations) while some other attacks would be a single-step process(attacker's specificity) and targeting the privacy of the model. These categorizations allow us to draw patterns between the attacks, which in turn makes the defenses against them systematic.

Just as how attacks were categorized across different axes, so are defenses. The two major categorization factors are Defensive Objective and Defense Approach.

Defensive Objective - There are two kinds of defensive objectives. A proactive approach is where the model is robust against the adversarial samples. On the other hand, a reactive approach is the addition of a filter that identifies the adversarial samples before they reach the classifier.

Defensive Approach - There are several approaches that defenses can take.
* One of the approaches presented was the Gradient Masking technique. The main idea behind this technique is for the model to output smoother gradients to hinder an attack strategy that tries to use the gradient to find the optimal perturbations.
* Auxiliary Detection Models - This involves training a different classifier tasked with distinguishing between legitimate and adversarial images.
* Statistical Methods - Using statistical measures to compare the distributions of legitimate and adversarial samples
* Preprocessing techniques - Applying techniques or transformations on the image before it is fed into the classifier as an input. This includes dimensionality reduction, image transformations, etc.
* Ensemble approach - Using multiple models and the final classification being a combined result such as a voting technique. This is based on the principle that while the adversarial samples might be able to fool one or two models, they would have a hard time fooling the majority of the models in the ensemble.
* Proximity measurements - This involves using proximity measurements of legitimate and adversarial samples to the decision boundary to separate and identify among each other.

The presentation also covered the hypotheses behind how adversarial samples work. The points explained are as follows:

1. A lot of models lack in generalization. This implies that samples that deviate from the training dataset would defeat the classifier.
2. Deep Neural Networks are linear due to the presence of activation functions. Many algorithms like Fast Gradient Sign Method exploits this linearity in the model's inner working to evade the model.
3. Boundary tilting hypothesis: The class boundaries that are learned by the model lies close to the training dataset. This implies that the gap between the boundary that the model learned model are places where the malicious samples can exist, without the model being able to identify them. This is achieved by making perturbations towards the direction of the decision boundary.
4. A lack of training dataset always limits the predictive power of the classifier
5. Non-robust features give more flexibility for the attackers to create adversarial samples and fool the classifier.

The presentation also went over the general process flow that a defender takes against adversarial samples. This includes defining a threat model, simulating the adversaries to break the threat model, developing provable lower bounds of robustness, performing sanity tests, and doing credibility assessment. Credibility assessment involves checking if the predicted output tallies with the model behavior, which would be easier for explainable AI. The presenter has worked and published a paper on "Explainable AI and random forest based reliable intrusion detection system.", and hence talked briefly about explainable AI and its relevance in this domain.

Discussion

- The reason for the introduction of this paper in this course, is that despite the paper's focus being in no way related to the malware domain, the approaches and the standardizations that the paper

suggests in the field of adversarial samples in general translates easily into the domain of malware detection systems. This gives more context to the topic of adversarial samples and its place in the domain of machine learning in general too. It was also stated that the authors are from the army, which shows the relevance of the topic.

- Explainable AI and its relevance was reiterated along with the Shapley Global, which shows the importance of a feature towards the predicted output.
- A question that arose was how a smooth gradient would be good from a defender's perspective. The attacker's method of approach would be to use the available information to craft an adversarial sample that is as effective as possible. Creating a smoother gradient would reduce the amount of information that is given out to the attacker, hence hindering his progress toward finding a good adversarial sample.
- The presenter has used this concept in his work in network intrusion detection systems, which shows how concepts like these transfer well into the domain of security in general.
- Another discussion was on where adversarial image attacks are used in the real world. Since a lot of the technology works on image or image classification nowadays, this has a lot of impact from facial authentication, to self-driving cars to military drones. An example of the consequences of this attack was shown in the domain of medical science, where say, a tumor detection system could be deceived by an adversarial attack causing an impact on human lives. Also, for adversarial attacks on images, there is a physical aspect to it, like modifying the stop signs or the license plates, which is not present in the malware digital world.