CY 7790, Lecture 13

Cem Topcuoglu and Nicholas Lunsford

November 01, 2021

Today's class centered on application domains for poisoning attacks against machine learning systems. Whereas previous works studied in class have focused on the theory and feasibility of poisoning attacks, today's papers centered on applying these techniques to real-world applications with realistic constraints and threat models. Considering the taxonomy of adversarial machine learning, these attacks fall into the standard poisoning category of attacks occurring at training-time, and aiming to degrade the integrity of some machine learning model or system.

Attacker's Objective

Learning Stage	Integrity Target small set of points	Availability Target entire model	Privacy Learn sensitive information
Training	Targeted Poisoning Backdoor Poisoning Subpopulation Poisoning	Poisoning Availability Model Poisoning	-
Testing	Evasion Attacks	Sponge Attacks	Reconstruction Membership Inference Model Extraction

Figure 1: Taxonomy of adversarial attacks against ML.

1 Severi et al. Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers

Problem Statement This work is at the intersection of machine learning and cybersecurity. Many cybersecurity applications use malware classifiers. The features extracted from the static analysis are trained to create a classification model. Via this classifier, the endpoints can distinguish between the benign and malicious samples. Previous work showed that the malware classifiers can be vulnerable to evasion attacks. However, Severi et al. showed that these malware classifiers are vulnerable to backdoor poisoning attacks as well. The authors targeted clean label attacks where the attacker has no control over the labeling process. The crowdsourced threat intelligence platforms are the main binary collection venues for malware classifiers. Severi et al. use this and inject poisoned samples into these crowdsourcing platforms.

Threat Model Figure 2 presents the five different attacker scenarios. Note that the fullness of the circle represents the knowledge or control level of the attacker. First, the *unrestricted* attacker is free to modify the training data. Second, the *data_limited* attacker has restricted access to the training data. Third, the *transfer* attacker has limited access to the target model. Fourth, the *black_box* attacker has limited knowledge of the model architecture. Fifth, *constrained* attacker limits the feature set since it wants to preserve the original functionality of the binary.

Attacker	Knowledge				Control	
	Feature Set	Model Architecture	Model Parameters	Training Data	Features	Labels
unrestricted	●	●	●	●	●	○
data_limited	●	●	●	◔	●	○
transfer	●	○	○	●	●	○
black_box	●	○	○	●	●	○
constrained	●	●	●	●	◔	○

Figure 2: The overview of the threat model.

Methodology Figure 3 presents the attack overview. The attacker submits poisoned benign samples to change the subset of the training samples. The platform collects these samples and labels them. The company combines the outsourced data and its proprietary data to train a classification model. The attacker can submit samples containing the same backdoor. These samples were classified as benign.

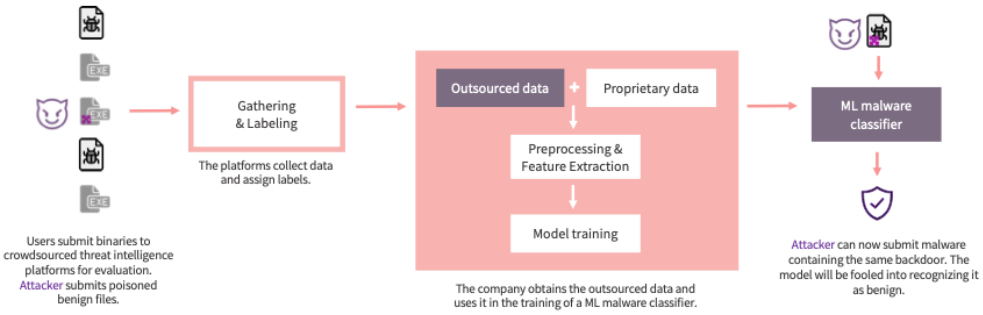

Figure 3: The overview of the attack.

One of the main challenges is finding the backdoors. There are two ways to find it. First, search for areas of weak confidence near the decision boundary. These areas should be the places that the watermark can overwhelm existing weak evidence. Second, subverting areas that are already heavily oriented toward goodware. To gain insights about the model's decision boundary, the authors use SHapley Additive exPlanations (SHAP) which is a model explanation technique. They improve upon the SHAP and use the modified versions of it in their attacks. For the sake of brevity, we left further details such as feature selectors, value selectors, and mitigation techniques to the paper.

Class Discussion The class discussion mainly focused on six different topics about the strengths and limitations of the paper.

Novel Approach The authors introduced a novel poisoning attack that targets the malware classifiers. Although there were previous works that target the image classifiers, malware clean label poisoning is a novel approach.

Easy Entry Point for an Attack It is very easy to change the training dataset via crowdsourcing platforms. The detection of the poisoned samples is hard.

Model Agnostic Not bounded to a model. This makes it easy to generalize the attack.

Open-source Implementation The implementation is open-sourced and is broadly documented.

Limitations of the SHAP The SHAP is a supervised technique that requires labels. Also, it only provides additive contributions of explanatory variables.

Defense Strategy Although the main contribution of this work is not the defense strategy, it only works for the isolation forest but not for the greedy approach.

2 Schuster et al. Humpty Dumpty: Controlling Word Meanings via Corpus Poisoning

Problem Statement Many natural language processing tasks rely on word embeddings derived from large, public corpora. As a result, there's an obvious point of vulnerability since public corpora may be augmented or manipulated by an adversary at any time, and these manipulations may persist for long periods of time given the weak monitoring of public resources. Therefore, an attacker may look to poison the public corpora in order to move a set of words a certain direction in the resulting embedding used by some target machine learning system. This paper seeks to poison such public corpora, effectively changing their meanings to meet some adversarial goal.

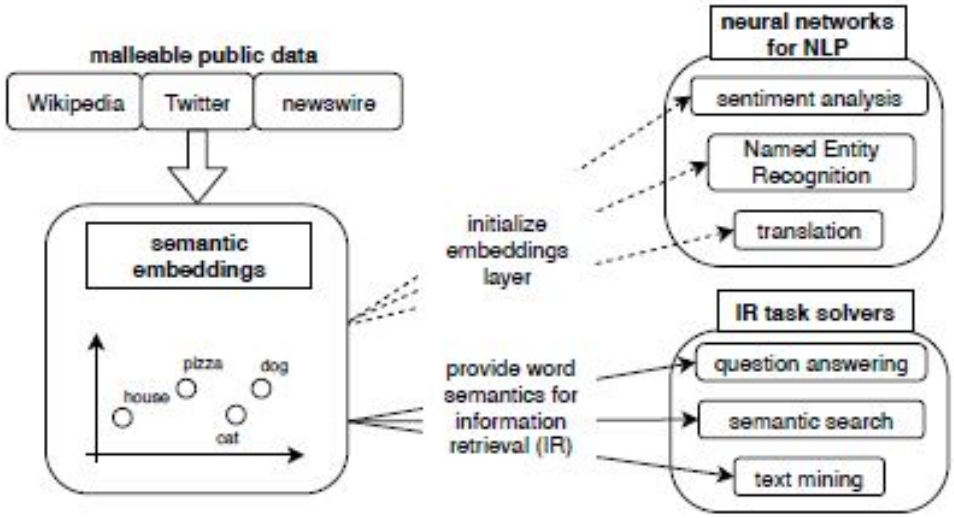

Figure 4: Overview of Natural Language Processing Using Word Embeddings.

Threat Model In general, the attacker wants to change the meaning a source word in an embedding from a public corpus. The attacker does not need to know the embedding algorithm and its hyperparameters, but must know which public corpus is being utilized by the embedding algorithm. Additionally, the attacker does not know the details of the downstream models being attacked. Lastly, we assume the attacker has the ability to add a collection of words or short word sequences to the public corpus. In this threat model, the attacker modifies the downstream model's embedding by manipulating the corpus on which it was trained. The general overview of this process is depicted in Figure 5.

Methodology Generally speaking, the attacker must first find the distributional expressions for the embedding proximities, which may be used for multiple attacks. Next, for the specific attack being implemented, the attacker defines an embedding objective and a corresponding directional objective, which is expressed as an optimization problem over co-occurrence counts. Then, the attacker solves for the directional objection, to obtain the change vector which he then transforms to a change set of corpus edits which are applied to the public corpus. Lastly, the embedding algorithm trains on the modified corpus and results on the downstream machine learning system containing the attacker's intended changes. The general overview of this process is shown in Figure 6.

We leave the exact, detailed methodology for deriving embedding objective, proximity objective, and rank objective to the original paper in consideration of space and brevity. We do the same for the proposed optimization algorithm.

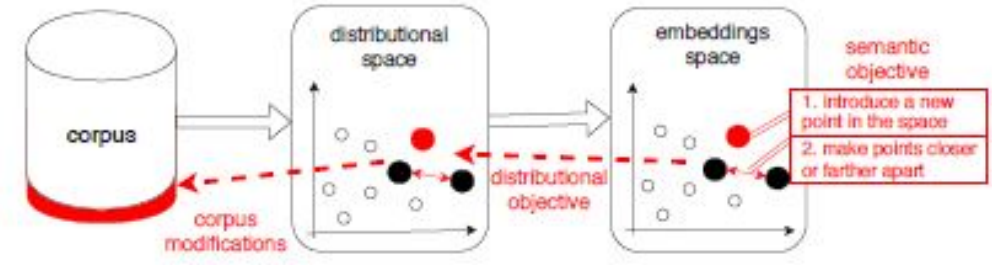

Figure 5: Semantic Changes via Corpus Modifications

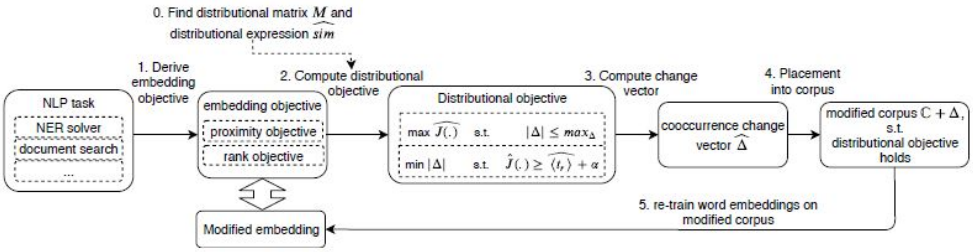

Figure 6: General Attack Methodology

Class Discussion Due to the length of the papers discussed, and the complexity of the specific methodology in this paper, class discussion on this topic was limited. However, some strengths and limitations of the proposed methodology were offered by classmates.

Difficulty The class was in agreement that this paper was a difficult read, with exact, intricate math proposed in several different sections. Furthermore, a strong understanding of word embeddings and natural language processing tasks is requisite to understanding why and how this proposed attack methodology works.

Explicit Expressions This paper was the first to develop explicit expressions for word proximities over corpus cooccurences, such that changes in expression values produce consistent, predictable changes in the embedding proximities.

Transfer Learning This paper proposed the first attack ever against two-level transfer learning. The method first poisons the dataset of words by changing their distribution and meaning, which poisons the resulting word embedding. As a result, the downstream machine learning system, which relies upon this embedding for some natual language processing task, fails to perform its intended function with integrity.

Considered Defenses After proposing their attack, the authors also considered defenses against the attack. This is an important aspect of security research, and effectively conveys that while this threat exists, there are continuing efforts to protect against it.