Detecting and Mitigating Poisoned Data in Neural Networks: A Study

Purvav Punyani
NYU Tandon School of Engineering
psp8474@nyu.edu.edu

Tiyas Dey
NYU Tandon School of Engineering
td2355@nyu.edu

Abstract

In this study, we investigate the impact of various data poisoning techniques, namely Meta Poison, Feature Collision, Convex Polytope and Hidden Trigger Backdoor Attacks, on neural network performance, specifically focusing on the CIFAR-10 dataset. We apply advanced anomaly detection algorithms, including Isolation Forest and One-Class SVM, to identify and mitigate the effects of these poisoned datasets. Our results highlight the vulnerability of neural networks to such attacks and the efficacy of our proposed countermeasures.

1 Introduction

Data poisoning poses a critical challenge to the integrity of machine learning systems, especially in deep neural networks (DNNs). This adversarial tactic involves the subtle modification of training data, enabling attackers to manipulate the behavior of a model. The study embarks on a comprehensive exploration of three advanced data poisoning techniques: Meta Poison, Feature Collision, and Convex Polytope. These methods are analyzed for their impact on the CIFAR-10 dataset, a standard benchmark in the field of image recognition. The primary objective is to unravel the mechanics of these threats and to engineer robust detection and mitigation strategies that can safeguard neural networks against such insidious attacks.

2 Related Work

Historically, data poisoning has been central to the research on adversarial attacks in machine learning. Initial studies predominantly concentrated on elementary strategies such as label flipping and pattern-based attacks, which involved simple yet detectable modifications to training data. However, the landscape of data poisoning has evolved with the advent of more nuanced and stealthy techniques. Contemporary methods such as Meta Poison, Feature Collision, and Convex Polytope represent a new wave of sophisticated attacks. These techniques are characterized by their low detectability yet high disruptive potential, making them particularly dangerous. Our research builds upon this existing knowledge base, bringing into focus the application of modern anomaly detection methods like Isolation Forest and One-Class SVM to identify and counteract poisoned data. There is an existing benchmarking of posioned datasets on different models which include Hidden Trigger Backdoor attacks, feature collision and convex polytope. We have in this study also compared Meta posion. Our study involves VGG16, a model not evaluated in the current benchmarks.

3 Methodology

The methodology of this study is bifurcated into two primary phases: data poisoning and anomaly detection.

3.1 Data Poisoning

This phase delves into four advanced poisoning techniques:

- **Meta Poison:** This approach involves manipulating the training data to specifically degrade the performance of the model on certain target instances. It is a form of targeted attack that subtly alters features in a way that the model learns incorrect representations for these specific instances.
- **Feature Collision:** This technique crafts training examples that, in the feature space, closely align with the target instances, leading to their misclassification. It exploits the vulnerabilities in the learning algorithm to cause confusion in the model's decision boundaries.
- **Convex Polytope:** In this method, a set of training points is strategically placed to form a convex polytope around the target instance in the feature space. This enclosure causes the model to misclassify the target instance, effectively 'trapping' it within a region of adversarial influence.
- **Hidden Trigger Backdoor Attacks:** In this sophisticated approach, a backdoor is covertly embedded into the model during training. This backdoor remains dormant and undetectable under normal circumstances but activates when specific, often subtle, triggers are present in the input data. These triggers, known only to the attacker, cause the model to produce erroneous or malicious outputs. In the implementation, hidden triggers are injected into the training dataset and are activated under certain conditions, illustrating the insidious nature of these attacks.

3.2 Anomaly Detection

In response to these poisoning strategies, we employ two prominent anomaly detection models:

- **Isolation Forest:** This algorithm isolates anomalies instead of profiling normal data points, effectively distinguishing between poisoned and clean data in our context.
- **One-Class SVM:** Tailored for the detection of outliers, One-Class SVM constructs a decision boundary around the normal data, enabling it to detect the anomalies represented by poisoned data points.

Both models are trained and evaluated on the CIFAR-10 datasets altered with the aforementioned poisoning techniques.

4 Experimental Setup

The CIFAR-10 dataset, renowned for its complexity and broad application in image classification tasks, forms the basis of our experimental setup. We apply the Meta Poison, Feature Collision, and Convex Polytope techniques to create distinct poisoned versions of this dataset. Subsequently, a standard VGG16 neural network model, a popular choice in image recognition tasks, is trained on each of these poisoned datasets. The evaluation of the model's performance is carried out using a clean subset of the test data to assess the impact of each poisoning technique. The poison datasets have been taken from `https://github.com/JonasGeiping/data-poisoning`

5 Results

Our experimental findings indicate a marked degradation in model accuracy when trained on datasets compromised by these poisoning techniques. For instance, training with the Convex Polytope Poisoned dataset resulted in an accuracy of 10%. Similar diminishing trends were observed with Feature Collision and other datasets. However, the implementation of anomaly detection techniques yielded promising results as shown in the tables below:

Table 1: Model Accuracy on Different Types of Poisoned Datasets

Poisoning Technique	Model Accuracy (%)
Hidden Trigger Backdoor	90
Feature Collision	87
Meta Poison	52
Convex Polytope	15

Table 2: Anomaly Detection Accuracy for Different Types of Poison

Poisoning Technique	Isolation Forest (%)	One-Class SVM (%)
Hidden Trigger Backdoor	69	57
Feature Collision	71.96	62.22
Meta Poison	30	22
Convex Polytope	72.94	67.3

6 Discussion

The study conclusively demonstrates that advanced data poisoning techniques can significantly undermine the performance of neural networks. Notably, the success of Isolation Forest and One-Class SVM in detecting poisoned data underscores the viability of these methods in defending against sophisticated data poisoning attacks. Their ability to discern anomalies within the training data serves as a critical tool in the arsenal against adversarial attacks on machine learning models.

7 Conclusion

This research highlights the urgent need to develop effective countermeasures against data poisoning attacks in neural networks. Our experiments with the CIFAR-10 dataset revealed significant variability in model resilience to different types of data poisoning. Notably, the model demonstrated a high level of accuracy (90%) in the presence of Hidden Trigger Backdoor attacks, but its performance was notably compromised by Meta Poison and Convex Polytope techniques, with accuracies dropping to 52% and 15% respectively.

The application of anomaly detection methods, namely Isolation Forest and One-Class SVM, presented mixed results in identifying poisoned data. While Isolation Forest achieved a substantial detection accuracy of 72.94% in cases of Convex Polytope attacks, its effectiveness was markedly lower in identifying Meta Poison attacks, at only 30%. Similarly, One-Class SVM performed best against Convex Polytope attacks with a 67.3% accuracy rate but showed limited effectiveness against Meta Poison attacks, with an accuracy of just 22%.

These findings underscore the complexity of addressing data poisoning in machine learning and the necessity of integrating robust anomaly detection mechanisms. The varied effectiveness of Isolation Forest and One-Class SVM across different poisoning techniques points to the need for ongoing research and development of more sophisticated detection methods. Future work should aim to refine these techniques and extend their applicability to more complex, real-world scenarios. Enhancing the efficacy of anomaly detection is crucial for safeguarding the reliability and integrity of machine learning systems in environments exposed to adversarial threats.

8 Challenges

The evaluation of these poisoning techniques, particularly the application of anomaly detection using One-Class SVM, proved to be highly resource-intensive. Due to time constraints, we were unable to explore additional poisoning methods such as gradient matching. This presents an opportunity for future research to delve into these alternative approaches, assessing their impact and developing corresponding countermeasures.

References

[1] arXiv:2004.00225 [cs.LG], *TMetaPoison: Practical General-purpose Clean-label Data Poisoning*. Retrieved from https://doi.org/10.48550/arXiv.2004.00225

[2] arXiv:1910.00033 [cs.CV], *Hidden Trigger Backdoor Attacks*, Retrieved from https://doi.org/10.48550/arXiv.1910.00033

[3] arXiv:1905.05897 [stat.ML], *Transferable Clean-Label Poisoning Attacks on Deep Neural Nets*, Retrieved from https://doi.org/10.48550/arXiv.1905.05897

[4] https://github.com/JonasGeiping/data-poisoning

[5] https://github.com/aks2203/poisoning-benchmark