LOSSY COMPRESSION TO PREVENT EVASION AND POISONING

Dan Jacobellis and Matthew Qin

University of Texas at Austin

ABSTRACT

Evasion and poisoning attacks are a major concern for machine learning models that operate on high dimensional data such as audio, images, and video. One of the most insidious aspects of these attacks is that they only require subtle perturbations in the input space to succeed. In many cases, a successful perturbation can be so small as to be undetectable to a careful human auditor. However, in this work, we show that perturbations introduced by these attacks can be sanitized by lossy compression. We show that for image classification, the accuracy gained from sanitizing the attack outweighs the accuracy lost from compression. We conduct experiments on several images using a variety of codecs and perturbation sizes. Our results suggest that lossy compression is a powerful strategy to mitigate these attacks. In addition, we show that learning directly on compressed representations can significantly reduces the memory throughput required for training, thus increasing efficiency with only a modest loss in accuracy.

1. INTRODUCTION

In recent years, machine learning models have become increasingly prevalent in a wide range of applications, from computer vision and speech recognition to natural language processing and even medical diagnosis. These attacks are particularly concerning because they can be implemented using subtle perturbations in the input space that are difficult or impossible for a human to detect.

In this paper, we present a new approach to mitigating evasion and poisoning attacks in machine learning models. Our approach is based on the observation that these attacks often involve adding small perturbations to the input data, which can be effectively sanitized by lossy compression. Since standardized lossy compression techniques focus on preserving visible features, they can be used to target the perturbation introduced by an attack while preserving the features necessary to achieve high accuracy.

We conduct experiments on an image classification task using a variety of codecs and perturbation sizes to evaluate the effectiveness of our approach. Our results show that sanitizing the perturbations using lossy compression can significantly improve the accuracy of the model, even when using relatively high levels of compression. Furthermore, we show that learning directly on compressed representations can significantly reduce the memory throughput required for training, thus increasing efficiency with only a modest loss in accuracy. Our contributions can be summarized as follows:

- We evaluate the efficacy of gradient-based evasion attacks as the size of the perturbation increases on several images from the Imagenet validation set.
- We apply different lossy image codecs to perturbed images to understand how much accuracy is lost from the encoder and how much accuracy is gained from sanitation.
- We use multiscale structural similarity to evaluate the quality of attacked images before and after lossy codecs.
- We train image classifiers CIFAR-10 using different lossy codecs to understand their impact on accuracy and potential to prevent poisoning.
- We demonstrate that the increased information density of lossy encoded representations can be used to significantly reduce memory throughput required to train an audio classifier.
- We provide insights and recommendations on how and when to utilize lossy compression to prevent attacks.

The rest of this paper is organized as follows. Section 2 presents an overview of related work, including (1) gradient-based evasion and poisoning attacks, (2) an overview of lossy compression standards and perceptual quality, (3) prior research relating adversarial examples and robust features, and (4) previous approaches to training on lossy encoded data. In Section 3, we describe our approach for using lossy compression to prevent evasion and poisoning attacks. Then, we propose a novel approach for training neural networks directly on lossy encoded data suing binary neural networks to preserve quantization. In Section 4, we present experimental results on mitigating attacks to image classifiers and increasing the efficiency of an audio classifier. In Section 5, we provide our recommendations for when and how to leverage lossy compression for more accurate models and discuss future directions to explore.

2. RELATED WORK

2.1. Evasion

We consider adversarial examples in the context of two types of attacks: evasion and poisoning. Evasion attacks exploit knowledge of a model that's already been trained. For example, if an attacker wants a malicious email to pass through a spam filter undetected, they might use full or partial knowledge of the behavior of the trained spam filter to find "magic words" that cause an email to be classified as not spam.

These attacks are typically performed using gradient-based methods, where the gradient of the loss function with respect to the input $\Delta_x \mathcal{L}(x, y, \theta)$ is used to guide the perturbation. Since moving in the *opposite direction* of the gradient increases model accuracy, we can create a perturbation by moving *with the direction* of the gradient, i.e.

$$x_{\text{adv}} = x_0 + \epsilon \Delta_x \mathcal{L}(x, y, \theta).$$

A simple, effective, and widely studied variant of this attack is the fast gradient sign method (FGSM) [GSS14], where the sign of the gradient is used instead

$$x_{\text{adv}} = x_0 + \epsilon \text{ sign } (\Delta_x \mathcal{L}(x, y, \theta)) .$$

This strictly limits the amplitude of the perturbation to $\pm\epsilon$ while maximizing its effect on model predictions. The limitation of the amplitude is what prevents the perturbation from being detected. For example, in figure 1, we show how epsilon can be chosen to limit the perturbation to two, four, or six of the least significant bits of an 8-bit image.

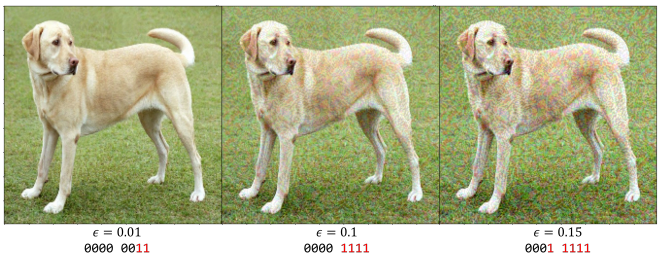

Fig. 1: Evasion Attack using FGSM

2.2. Poisoning

In a poisoning attack, the dataset is contaminated, usually with the goal of introducing a backdoor. For example, if an attacker wants to prevent a facial recognition model from working on one or more subjects, they might upload an altered image public to the web where the dataset is sourced for training. Recent poisoning attacks such as gradient matching [Gei+20] have been shown to be effective on very large datasets like imagenet. With gradient matching, small, imperceptible perturbations on as little as 0.1

2.3. Lossy Compression and Perceptual Quality

2.4. Adversarial Examples and Robust Features

Researchers have demonstrated that adversarial examples, such as those produced by FGSM, can be attributed to the presence of *non-robust features* [Ily+19]

2.5. Training on Encoded Data

3. METHODS

4. EXPERIMENTS

5. DISCUSSION

[Gei+20]

References

[Gei+20] Geiping, J. et al. "Witches' brew: Industrial scale data poisoning via gradient matching". In: *arXiv preprint arXiv:2009.02276* (2020).

[GSS14] Goodfellow, I. J., Shlens, J., and Szegedy, C. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[Ily+19] Ilyas, A. et al. "Adversarial examples are not bugs, they are features". In: *Advances in neural information processing systems* 32 (2019).