



Defending Deep Learning Models Against Adversarial Attacks


Nag Mani, San Jose State University, USA

 <https://orcid.org/0000-0002-1294-9745>

Melody Moh, San Jose State University, USA

 <https://orcid.org/0000-0002-8313-6645>

Teng-Sheng Moh, San Jose State University, USA

 <https://orcid.org/0000-0002-2726-102X>

ABSTRACT

Deep learning (DL) has been used globally in almost every sector of technology and society. Despite its huge success, DL models and applications have been susceptible to adversarial attacks, impacting the accuracy and integrity of these models. Many state-of-the-art models are vulnerable to attacks by well-crafted adversarial examples, which are perturbed versions of clean data with a small amount of noise added, imperceptible to the human eyes, and can quite easily fool the targeted model. This paper introduces six most effective gradient-based adversarial attacks on the ResNet image recognition model, and demonstrates the limitations of traditional adversarial retraining technique. The authors then present a novel ensemble defense strategy based on adversarial retraining technique. The proposed method is capable of withstanding the six adversarial attacks on cifar10 dataset with accuracy greater than 89.31% and as high as 96.24%. The authors believe the design methodologies and experiments demonstrated are widely applicable to other domains of machine learning, DL, and computation intelligence securities.

KEYWORDS

Adversarial Examples, Adversarial Retraining, Basic Interactive Method, DeepFool, Ensemble Defense, Fast Gradient Sign Method, Gradient-Based Attacks, Image Recognition, Securing Deep Learning

1. INTRODUCTION

Global research in academia and industry has promoted the adoption of deep learning applications in every aspect of life, from smart home devices like Amazon echo, Google Home, and Facebook portal, to industrial applications like deliveries by drone, warehouse automation, medical imaging, and self-driving vehicles. The inception of these devices in both personal and industrial settings has been accelerated by the advancements in the field of deep learning. Transforming perception into smart responses/actions in real time is possible due to faster and more accurate image recognition models. For instance, smartphones use face detection and recognition to authenticate the correct user. Tesla uses deep learning to design self-driving features such as object detection, semantic segmentation, lane detection, pedestrian detection, traffic sign recognition, etc., to make smart decisions in real-time

DOI: 10.4018/IJSSCI.2021010105

This article, originally published under IGI Global's copyright on January 1, 2021 will proceed with publication as an Open Access article starting on March 1, 2024 in the gold Open Access journal, International Journal of Software Science and Computational Intelligence (IJSSCI) (converted to gold Open Access January 1, 2023), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

situations. Smart surveillance security cameras are equipped with face and activity recognitions that identify and record any abnormal activity or entry.

However, with the wide-scale adoption of Internet of Things (IoT) devices, the systems are exposed to a multitude of vulnerabilities. One such vulnerability is adversarial examples. These adversarial examples are carefully crafted inputs, aimed at fooling the model and bringing down the model's accuracy and real-world performance. These attacks are not easy to detect, as they are usually imperceptible to humans, as shown in Figure 1, yet they can easily degrade the model's accuracy. The adversaries are asymmetric in nature and are created in specific ways to compromise the integrity of deep learning models. These have posed major risks in implementing deep learning in safety-critical applications, such as home security, medical imaging, and autonomous vehicles (Mani & Moh, 2019).

This paper uses multiple gradient-based attacks to showcase their effectiveness against the target model. Defense against these attacks has been a well-researched topic. Previous work on the topic of adversarial retraining showed its effectiveness against different gradient based attacks (Kurakin et al., 2016). The previously proposed approach did add some robustness to the model, yet, the decrease in accuracy was more than 25% in the case of adversarial attacks. This is a significant number when considering the application of these models in safety-critical environments. This paper has explored the same idea of adversarial retraining to build more resilient models that can withstand adversarial attacks with high confidence. Early results have been presented (Mani et al., 2019). The main contribution of this paper is as follows:

- Provided experimental results of six different gradient-based adversarial attacks, including FGSM (Szegedy et al., 2014), BIM (Gu & Rigazio, 2014), ILLC (Gu & Rigazio, 2014), DeepFool (Moosavi-Dezfooli et al, 2016), Carlini-Wagner L2 and L ∞ (Carlini & Wagner, 2017), on ResNet34 model using cifar10 dataset;
- Demonstrated that the previously adversarial retraining technique proposed (Goodfellow et al., 2015) has limited effectiveness while failed to provide transferable security against more sophisticated attacks;
- Showed that the proposed adversarial retraining technique coupled with the ensemble method is capable of withstanding even sophisticated attacks like Carlini-Wagner attacks (Carlini & Wagner, 2017), achieving model accuracy with minimum of 89.31%, and up to 96.24% for DeepFool attack.

The rest of the paper is structured as follows: Section 2 discusses the background and some previous related work. Section 3 presents a set of six gradient-based attacks used for the study. Section 4 describes the details of the proposed defense architecture, which encompasses adversarial retraining technique coupled with the ensemble approach to perform classification. Section 5 illustrates the experimentation results. Section 6 draws the concluding remark, and finally Section 7 suggests some future research directions.

2. BACKGROUND AND RELATED WORK

Though adversarial attacks have been common for a couple of decades, their applications were limited to conventional machine learning methods. For example, variable length of extra padding bits at the end of malware to mask its signature (Barreno et al., 2010), defeating spam filters by appending additional text at end of spam messages are a few of the many adversarial attacks which were used against machine learning methods (Biggio et al., 2010).

The idea of adversarial examples attacks which work as a game between an adversary and a machine learning model was first proposed back in 2004 (Dalvi et al., 2004). The approach for attack and defense of the adversarial samples was to play an iterative game which prepared the adversarial examples in an incremental manner. The gradient-based approach was used for the first time to

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/defending-deep-learning-models-against-adversarial-attacks/266229

Related Content

Research on O2O Platform and Promotion Algorithm of Sports Venues Based on Deep Learning Technique

Kaiyan Hanand Qin Wang (2020). *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications* (pp. 1547-1558).

www.irma-international.org/chapter/research-on-o2o-platform-and-promotion-algorithm-of-sports-venues-based-on-deep-learning-technique/237950

Application of Uncertain Variables to Knowledge-Based Resource Distribution

Donat Orski (2012). *Machine Learning: Concepts, Methodologies, Tools and Applications* (pp. 928-950).

www.irma-international.org/chapter/application-uncertain-variables-knowledge-based/56182

Feature Reduction Using Genetic Algorithm for Cognitive Man-Machine Communication

Naveen Irtizaand Humera Farooq (2015). *International Journal of Software Science and Computational Intelligence* (pp. 1-17).

www.irma-international.org/article/feature-reduction-using-genetic-algorithm-for-cognitive-man-machine-communication/157434

Evaluating the Effects of Size and Precision of Training Data on ANN Training Performance for the Prediction of Chaotic Time Series Patterns

Lei Zhang (2019). *International Journal of Software Science and Computational Intelligence* (pp. 16-30).

www.irma-international.org/article/evaluating-the-effects-of-size-and-precision-of-training-data-on-ann-training-performance-for-the-prediction-of-chaotic-time-series-patterns/227734

Artificial Cell Model Used for Information Processing

Enrique Fernández-Blanco, Jose A. Serantes, Nieves Pedreiraand Julián Dorado (2010). *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies* (pp. 12-29).

www.irma-international.org/chapter/artificial-cell-model-used-information/43142