Project proposal

Prerana Khatiwada

Can adversarial training defend against poisoning attacks?

Problem Statement

Modern Deep Learning has achieved a state of the art accuracy on a variety of tasks that are based on both text and images. However, when such methods are put into use, they are vulnerable to attacks during training and testing, which degrades their ability for exact inference. All ML fields, including computer vision, natural language processing, healthcare, RL, etc., are exposed to these attacks. Even while adversarial training produces mediocre robust accuracy and decreases clean accuracy, it is nevertheless regarded as a solid defense (that cannot be overcome by adaptive attacks) against adversarial attacks. Most extensive studies of adversarial machine learning have been conducted in the area of image recognition, where modifications are performed on images, causing the classifier to produce incorrect predictions. Nevertheless, every one of them has at least one of the following weaknesses: they are quickly overcome by adaptive attacks, they significantly lower testing performance, or they are not generalizable to other data poisoning threat models. Current thinking is that the only empirically effective defense against (inference-time) adversarial attacks is adversarial training and its variations. In this project, I will try to expand the architecture for adversarial training to defend against poisoning and backdoor attacks during training.

Motivation

It has been shown that adversely perturbed points act as strong poisons. Contrarily, adversarial training uses adversarial samples generated from strong attacks to enhance the model. I want to look at how adversarial training affects the effectiveness of poisoning attacks. The central intuition behind this idea is that adversarial training causes the learning of robust features, which could be helpful when defending against poisoning attacks. To do this, I intend to test models (vanilla CNNs and ResNets) against poisoning attacks using a variety of datasets, including CIFAR and MNIST.

Related Work

Adversarial attacks were first introduced by [1] who found that one could add nearly imperceptible noise to images while the image remained unchanged to the human eye. However, the altered image is incorrectly identified when it is classified by a model. Since then, several new attacks that are more strong than the attack [1] suggests have been presented; such examples are [2, 3, 4, 5, 6]. Patch attacks are pointed out as one particular type of attack [7]. Although it alters how we see things visually, the modifications are only made to a subset of pixels. Additionally, a number of countermeasures against adversarial attacks have been put forth, including [8, 9, 10] and several others. Such ad-hoc defenses can, however, be overcome by adaptive attacks [11, 12, 3]. Prior studies [18] emphasize the use of adversarially perturbed points as powerful poisons; adversarial training makes the model more robust to vulnerabilities by using adversarial samples produced by strong attacks. The question proposed at the beginning of the paragraph is just logical. Recent work [19] demonstrates poison adversarial training, in which the model is trained in opposition using poisoned data points. Crafting Poisoning points, however, is computationally costly due to bilevel optimization. Adversarial attacks will be used which are cheaper to create, to try and answer the question. Since adversarial samples in the training process will be applied, adversarial robustness can be achieved for free.

Proposed Solution

In this work, I want to find out if adversarial training can defend against poisoning attacks in this project. Previous work by A. Turner, D. Tsipras, and A. Madry, “Clean-label backdoor attacks,” 2018 highlights the use of adversarially perturbed points as strong poisons. Poisoning points, however, are computationally expensive to create. The primary assumption is that adversarial training enables the model to acquire robust features, which can assist the model to distinguish between backdoor attributes that are often learned during training. According to recent studies, adversarial examples can also serve as powerful poisoning examples, which increase attack success, as stated in the publication "Adversarial Examples Make Strong Poisons." The goal of my study is to see if adversarial training can protect against poisoning attacks in this study. As a result, this study will combine ideas from the works mentioned above with my own. Throughout the project, I will conduct a variety of experiments to test the hypothesis using complex models and datasets for various attacks and defenses. I will compare the training approach with the State of Art Adversarial training.

Evaluation Plan/Tools/Benchmarks/Attacks/Experiments

I will be demonstrating the adversarially perturbed points as strong poisons and then try to do this in a computationally less expensive approach than existing research. The datasets used for this project will be MNIST And CIFAR-10.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[7] Tom B. Brown, Dandelion Man., Aurko Roy, Mart.n Abadi, and Justin Gilmer. Adversarial patch, 2018.

[8] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples, 2017.

[9] Nicolas Paper not, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks, 2016.

[10] Shawn Shan, Emily Wenger, Bolun Wang, Bo Li, Haitao Zheng, and Ben Y. Zhao. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Oct 2020.

[11] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses, 2020.

[12] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018.

[13] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defenses: A survey, 2018.

[14] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey, 2021.

[18] Liam Fowl, Micah Goldblum, Ping yeh Chiang, Jonas Geiping, Wojtek Czaja, and Tom Goldstein. Adversarial examples make strong poisons, 2021.

[19] Jonas Geiping, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller, and Tom Goldstein. What doesn't kill you makes you robust(er): Adversarial training against poisons and backdoors, 2021.