Adversarial Training is all you need

Pramodh Gopalan
IIT Kanpur
Department of Computer Science
{pramodh}@iitk.ac.in

Gurbaaz Singh Nandra
IIT Kanpur
Department of Computer Science
{gurbaaz}@iitk.ac.in

Anubhav Kalyani
IIT Kanpur
Department of Computer Science
{anukal}@iitk.ac.in

Antreev Singh Brar
IIT Kanpur
Department of Computer Science
{antreev}@iitk.ac.in

Abstract

Modern Deep Learning has achieved state-of-the-art accuracies on a wide range of Visual and Text based tasks. However, when such systems are deployed, they are susceptible to attacks during train and test time, affecting their ability to infer precisely. Test time attacks add imperceptible noise to samples to change the models's decision, whereas Train time attacks add adversarially manipulated points to the training set which can be exploited during test time. These attacks are applicable in all domains of ML, such as Computer Vision, NLP, Healthcare, RL, etc. Adversarial Training is considered to be a reliable defense (cannot be broken by adaptive attacks) against adversarial attacks, even if it yields mediocre robust accuracy, and degrades clean accuracy. The purpose of this project is to examine whether adversarial training can defend against poisoning attacks.

1 Adversarial Attacks

Adversarial attacks were first introduced by [1], who found that one could add imperceptible noise to images, leaving the image unchanged in human eyes. However, when a model classifies the modified image, it is recognized wrongly. Since then, many new attacks have been proposed, which are more stronger than the attack [1] proposes; Some examples are: [2, 3, 4, 5, 6]. One specific attack that we point out is patch attacks [7]: It creates visually perceptive perturbations, but the modifications are restricted to a subset of pixels. Several defenses against adversarial attacks have also been proposed, such as [8, 9, 10] and many more. However, such adhoc defenses are vulnerable to adaptive attacks and are bypassed [11, 12, 3]. One defense that has stood the test of time is called adversarial training, proposed first by [6], but made effective by [5]. For a more comprehensive review, we refer the reader to [13, 14].

2 Poisoning attacks

Poisoning attacks fall under train time attacks, wherein an adversary adds malicious train samples to the training dataset. A model trained on a poisoned dataset learns spurious features, which can later be exploited by the attacker during test time. Several attacks have been proposed, such as [15], [16]. Several defenses have also been proposed, which can be broken using adaptive attacks, similar to the case of adversarial attacks. For a more comprehensive review, we refer the reader to [17].

3 Main Idea of Project

In this project we aim to answer the following question: "Can adversarial training defend against poisoning attacks?". Previous work [18] highlights the use of adversarially perturbed points as strong poisons; Adversarial training uses adversarial samples generated from strong attacks to make the model more robust to vulnerabilities. It is only natural to ask the question we pose at the beginning of the paragraph. Recent work [19] showcases poison adversarial training, wherein they use poisoned datapoints to adversarially train the model. However, crafting poisoning points is computationally expensive, since it involves bilevel optimization. The question we aim to answer uses adversarial attacks, which are cheaper to generate. As an added advantage, we would also get adversarial robustness for free, since we use adversarial samples in our training method.

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 Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2018.

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

[9] Nicolas Papernot, 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 defences: 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.

[15] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2019.

[16] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-Guided backdoor poisoning attacks against malware classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1487–1504. USENIX Association, August 2021.

[17] Chen Wang, Jian Chen, Yang Yang, Xiaoqiang Ma, and Jiangchuan Liu. Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects. *Digital Communications and Networks*, 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.