Adversarial Attack and Defense Mechanisms on Traffic Signs

Prasangsha Ganguly
50320914
Department of Industrial and Systems Engineering
prasangs@buffalo.edu

Anarghya Das
50213835
Department of Computer Science and Engineering
anarghya@buffalo.edu

Saiful Islam
50413531
Computational and Data Enabled Sciences and Engineering
saifulis@buffalo.edu

1 Introduction

In this section, we will provide an introduction of our team and a brief overview of our project including ideas on what we want to do with our project accompanied by a timeline.

1.1 Team Information

For this project, we have a team of three members: **Anarghya Das** a first year PhD student from the Department of Computer Science and Engineering, **Saiful Islam**, a first year PhD student in Computational and Data Enabled Sciences and Engineering program, and **Prasangsha Ganguly** a third year PhD student in the Department of Industrial and Systems Engineering.

1.2 Project Overview

In recent times, adversarial attacks have demonstrated significant success in bringing down the performance of modern deep learning methods. An adversary can lightly perturb inputs to appear unchanged to humans but induce incorrect predictions by neural networks [1]. Studies have identified that, even the neural networks that perform at human level accuracy have an error rate of almost 100% on intelligently constructed examples which are very similar to the original examples. The lack of robustness exhibited by deep neural networks (DNNs) to adversarial examples has raised serious concerns for several applications like, *image identification and generation* [2], *text and language processing* [3], and *self driving vehicles* amongst others. Self driving vehicles use deep neural network (DNN) based approach to recognize traffic signs [4, 5, 6]. Misclassification of traffic sign may cause disastrous consequence including loss of life. Therefore, the robustness of these model to tackle any type of adversarial perturbation for example, distortion, alternation, brightness, skewness etc. is very important. In the case of traffic signs, the adversarial attack can be caused by either natural phenomena such as raining and snowing, shadow, haze, etc [7]. or intentional distortion [8, 9].

1.2.1 Adversarial Attack and Defence mechanisms

Often in literature, adversarial attacks and defences are formulated as a Stakelberg game between an attacker who wants to distort an input x to a new input $x + \delta$ via some distortion mechanism $\delta \in S$, where S is the feasible set of actions (distortions) of the attackers. On the other end, the defender tries to minimize the loss function. It can be expressed as follows [10],

$$\min_{\theta} \mathbb{E}_{(x,y)\in\chi}[\max_{\delta\in S} L(x+\delta; y; \theta)]$$

where χ is the set of training data consisting of inputs x and output labels y, L is the loss function, θ is the parameter of the network. For the attacker, there are different types of strategies which depend on the knowledge of the training mechanisms. For the *white-box* attack, the attacker has the knowledge of the training algorithm, on the other hand, the *black-box* attack doesn't have the knowledge of the training algorithm [11]. For defense strategies, there exists several methods like, adversarial training, defensive distillation or input transformations [12].

1.2.2 Project Objective

The Adversarial attacks in the context of traffic sign detection and classification received huge interest in the past few years but most of them are task specific. In this project, we will generate universal adversarial samples by using our proposed algorithm which will be able to deceive most of the established defense mechanisms with state-of—the-art performances and then evaluate the robustness of our proposed defense mechanism. The main objectives of the project are,

- Propose a framework of adversarial attacks using white box and black box mechanisms that optimally distort the training data.
- Propose a pipeline that allows robust production and evaluation of adversarial traffic signs to fool a commercial car perception system.
- Develop a defense mechanism like adversarial training which explicitly trains the adversarial examples to minimize the loss due to perturbation.

The overall research framework has been depicted in Fig. 1. In the framework, we want to propose a novel attack mechanism to create adversarial (perturbed) traffic sign images to maximize the loss function and then using adversarial training we estimate the efficiency of such defence mechanism (adversarial training) against the intelligent attack algorithm.

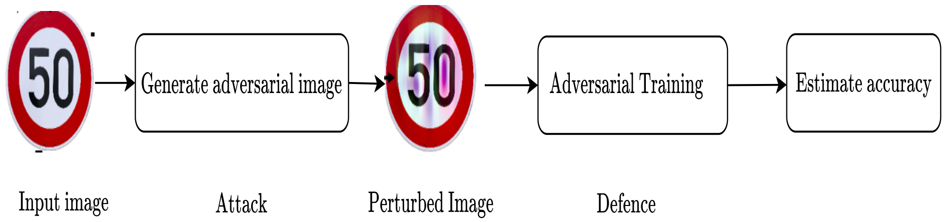

Figure 1: The research framework.

1.2.3 Timeline

We have proposed a timeline for this project in order to successfully complete it by the final deadline (May 22). See Table 1.

Table 1: Research Schedule

Task Name	March	April	May
Literature Review & Algorithm Generation	▨		
Algorithm Design & Data Collection		▨	
Implementation			▨

References

[1] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[3] Robin Jia and Percy Liang. “Adversarial examples for evaluating reading comprehension systems”. In: *arXiv preprint arXiv:1707.07328* (2017).

[4] Jingyu Wang, Weiran Wang, and Anliang Zhou. “The Faster Detection and Recognition of Traffic Signs Based on CNN”. In: *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE. 2019, pp. 907–914.

[5] Manuel Lopez-Martin et al. “Network traffic classifier with convolutional and recurrent neural networks for Internet of Things”. In: *IEEE access* 5 (2017), pp. 18042–18050.

[6] Pierre Sermanet and Yann LeCun. “Traffic sign recognition with multi-scale convolutional networks”. In: *The 2011 International Joint Conference on Neural Networks*. IEEE. 2011, pp. 2809–2813.

[7] Dogancan Temel, Min-Hung Chen, and Ghassan AlRegib. “Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.9 (2019), pp. 3663–3673.

[8] Chawin Sitawarin et al. “Rogue signs: Deceiving traffic sign recognition with malicious ads and logos”. In: *arXiv preprint arXiv:1801.02780* (2018).

[9] Nir Morgulis et al. “Fooling a real car with adversarial traffic signs”. In: *arXiv preprint arXiv:1907.00374* (2019).

[10] Huan Zhang et al. “The Limitations of Adversarial Training and the Blind-Spot Attack”. In: *ICLR* (Jan. 2019).

[11] Nicholas Carlini and Hany Farid. “Evading Deepfake-Image Detectors with White- and Black-Box Attacks”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 2804–2813. DOI: 10.1109/CVPRW50498.2020.00337.

[12] Aleksander Madry et al. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: *ICLR* (June 2017).