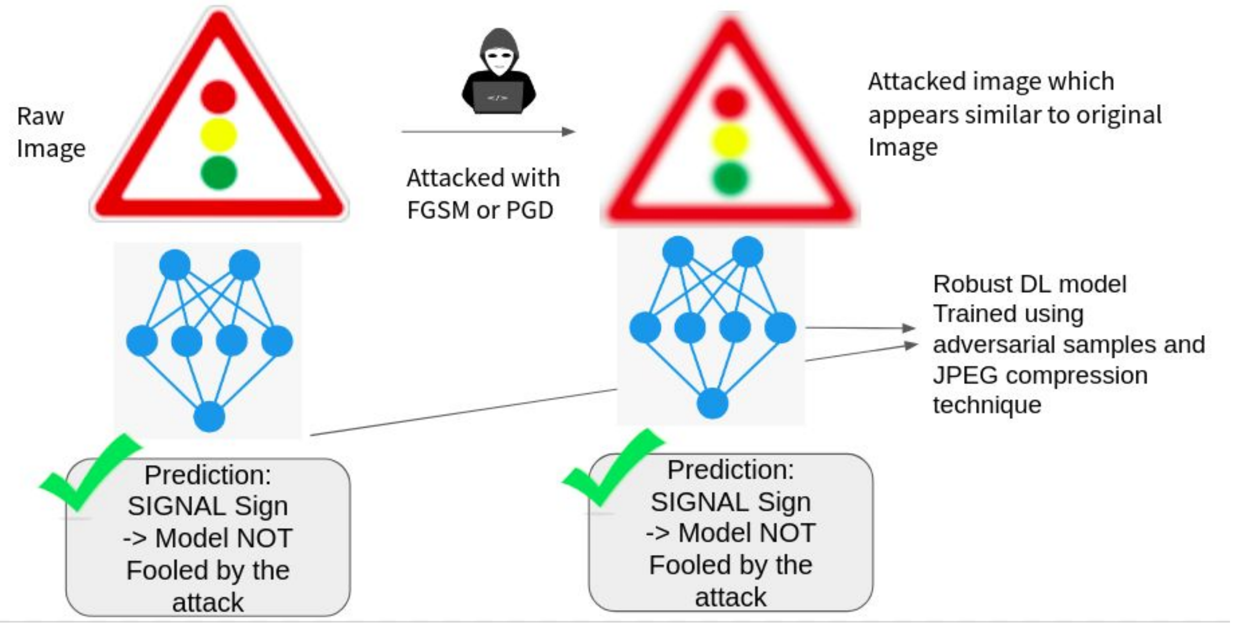
Indian Institute of Information Technology, Nagpur

AI Approach for Autonomous vehicles to Defend from Adversarial Attacks

Prachee Gupta (BT17ECE008), Kritika Dhawale (BT17ECE042)
Dr. Tapan Kumar Jain (Project Guide)

Introduction

Neural Networks are really impeccable sometimes, but when it comes to adversarial attacks, their performance falls off swiftly. Moreover, considerations to build more robust models that would be resilient to adversarial attacks are ignored frequently. Here, we aim to design a model that is robust and can handle multiple types of attacks on traffic signs. We focus on training a model with a defense mechanism to remove the adversarial noise, to make the model Anti-Spoof for Traffic Sign Recognition.



Raw Image

Attacked with FGSM or PGD

Attacked image which appears similar to original Image

Robust DL model Trained using adversarial samples and JPEG compression technique

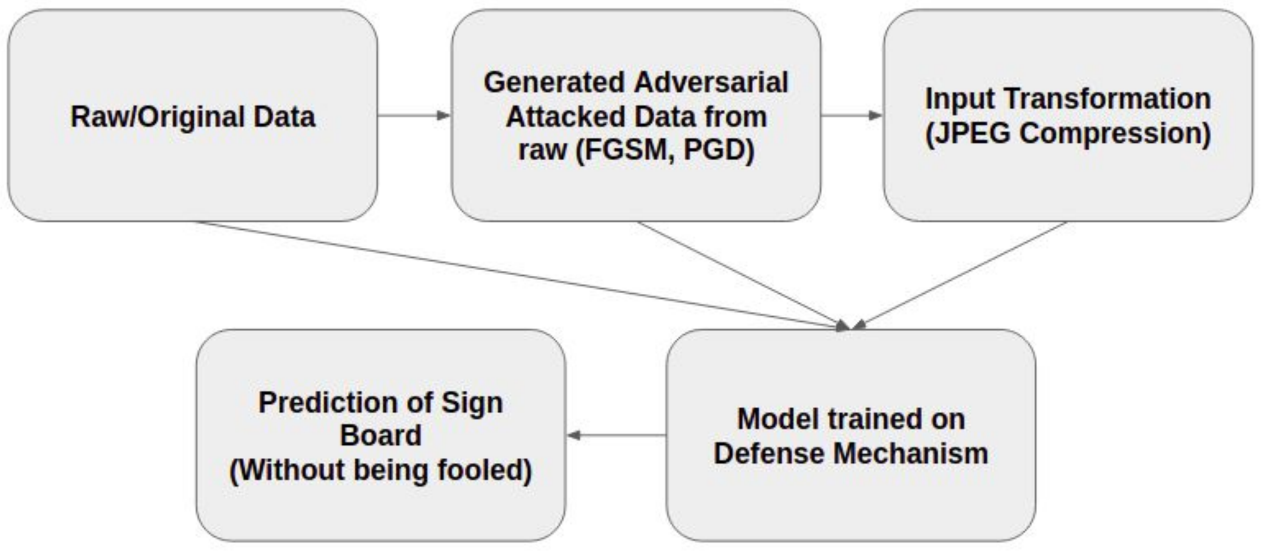
Prediction: SIGNAL Sign -> Model NOT Fooled by the attack

Prediction: SIGNAL Sign -> Model NOT Fooled by the attack

Fig 1: Proposed Model

Proposed Methodology

The pipeline can be dived into four main parts viz. generating adversarial samples, performing JPEG compression on those samples in order to remove the noise in it, mixing all the data together and lastly training a robust model which is resilient to adversarial attacks. The technique flowchart in fig 1. illustrates the scope of designing a powerful neural network based on the goals.



Raw/Original Data

Generated Adversarial Attacked Data from raw (FGSM, PGD)

Input Transformation (JPEG Compression)

Prediction of Sign Board (Without being fooled)

Model trained on Defense Mechanism

Fig 2: Flowchart of the process

Results

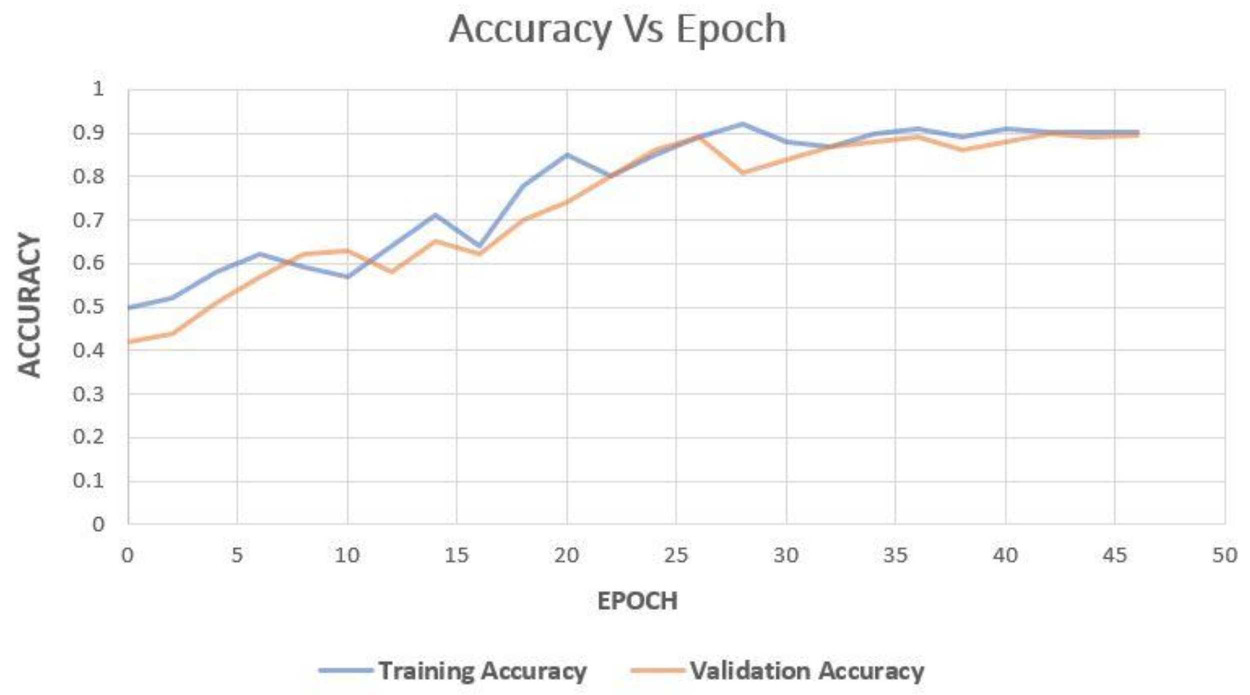
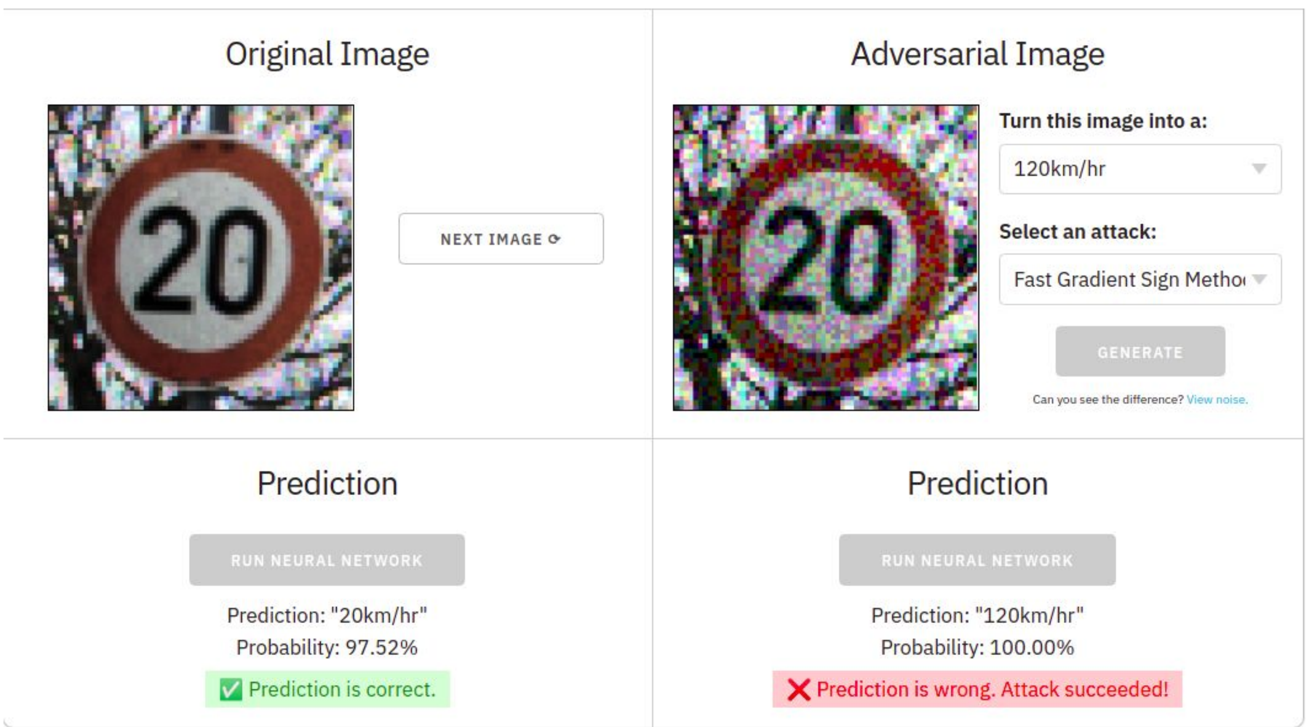


Fig 3: Accuracy vs epoch plot for the Model



Turn this image into a: 120km/hr

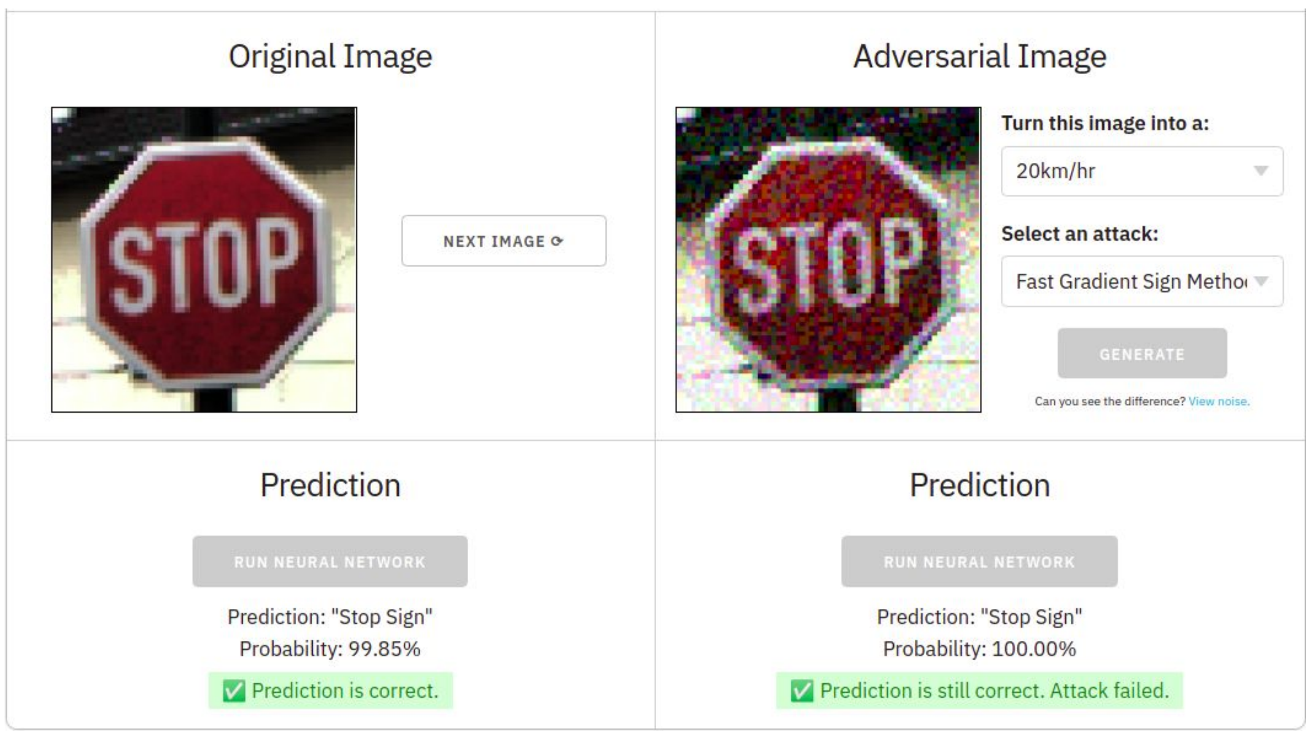
Select an attack: Fast Gradient Sign Metho

Can you see the difference? View noise.

Prediction: "20km/hr"
Probability: 97.52%
✓ Prediction is correct.

Prediction: "120km/hr"
Probability: 100.00%
✗ Prediction is wrong. Attack succeeded!

Fig 5: Results without adversarial training



Turn this image into a: 20km/hr

Select an attack: Fast Gradient Sign Metho

Can you see the difference? View noise.

Prediction: "Stop Sign"
Probability: 99.85%
✓ Prediction is correct.

Prediction: "Stop Sign"
Probability: 100.00%
✓ Prediction is still correct. Attack failed.

Fig 5: Results with adversarial training

Conclusions and Future Work

In this new era of technological automation, it becomes a necessity to look after safety measures. Especially when these are linked with our lives like in automated vehicles. One wrong prediction can cost our lives in these cases. In this paper, we have designed a robust model that focuses on avoiding adversarial attacks. We have trainred a model with a defensive mechanism to eliminate adversarial noise and performed adversarial training while generating attacked samples from FGSM & PGD.

Team No.- 9