Adversarial Attack for Road Sign Recognition

Danijela Jovanovski
CECS
University of Michigan
Dearborn, MI USA
dacajova@umich.edu

Aditya Kothari
CECS
University of Michigan
Dearborn, MI USA
askothar@umich.edu

Mithali Deepak Kumar Singh
CECS
University of Michigan
Dearborn, MI USA
mithali@umich.edu

Dileep Kumar Bhukya
CECS
University of Michigan
Dearborn, MI USA
dileepkb@umich.edu

ABSTRACT

In our project we have investigated adversarial attack techniques on traffic sign recognition with a focus on the Robust Physical Perturbation method also known as RP2. This method generates perturbation that is added to the traffic sign causing misclassification by the target model while not raising any suspicion from a human.
In our further research and from the results we have discovered that through defense mechanisms such as data augmentation we have noticed a reduction in RP2's accuracy.

CCS CONCEPTS

• Security and safety • Systems security • Vision & Perception • Robustness and stability • Intrusion detection and prevention • Embedded and cyber-physical systems

These CCS concepts reflect the interdisciplinary nature of our project, covering areas such as machine learning, computer vision, security, and transportation systems.

KEYWORDS

• Traffic sign recognition (TSR) • Perception • Deep learning
• Physical Adversarial Attack

1. Introduction

As deep learning continues to evolve, road sign recognition systems have emerged as vital components of autonomous driving technology, traffic management, and overall road safety. However, these systems are not immune to adversarial attacks, which can manipulate input data gradually, leading to misclassification and potential safety risks.

This project aims to investigate adversarial techniques designed specifically for traffic sign recognition systems. We aim to study the vulnerabilities of cutting-edge recognition methods used in real-world scenarios by introducing imperceptible perturbations on road sign photos. Understanding these vulnerabilities is critical for building robust and dependable systems that can withstand adversary manipulation.

This research paper presents an investigation into developing a Trustworthy AI framework aimed at mitigating such adversarial attacks in road sign recognition systems. We will explore how these attacks, both in the digital realm and through physical alterations, can manipulate the perception of a deep learning model, causing it to misinterpret a traffic sign. By understanding these vulnerabilities, we can contribute to the development of more robust and secure TSR systems for future autonomous vehicles.

2. Related Work

Research in adversarial attack and defense methods for deep learning classifiers has been gaining a lot of traction in recent years with the increased usage of such classifiers in safety and security critical fields. One such field is traffic sign recognition, which we are considering for our project. Since the topic of adversarial attack is not specific to traffic sign recognition we look towards the general research that has been done in this field. Adversarial attacks can be categorized into three main types: (1) Perturbation based (2) Patch based and (3) Unrestricted. Most of the research in adversarial attack is done for perturbation based attacks [1][2][3] where the goal is to find or generate adversarial

samples with small perturbation or noise budget so that the perturbations are imperceptible to human observers. Inspired from pixel attack, Patch based adversarial attacks change some areas of an image by either masking some features of the image or making the classifier ignore some features.[4][5][6] showed the effectiveness of such attack methods. Both perturbation based and patch based methods are restricted by some constraints to make the adversarial samples pass human evaluation. Whereas the Unrestricted method takes advantage of recent development in GAN and Diffusion models to generate a natural looking image with manipulated latent vectors which would easily pass human evaluation but fool the deep learning based classifiers. Unrestricted methods were introduced in the work by Song et al.[7]

For our project we focus on Patch based adversarial methods because of their simplicity, attack ability and feasibility in the real world.

3. Adversarial Attack Method

In our project to demonstrate an adversarial attack we have selected Robust Physical Perturbation (RP2) [5] as the algorithm for the manipulation of the images in both of the datasets. RP2 is a white box attack method which can work for both targeted and untargeted adversarial goals. This attack leverages an optimization-based approach that aims to generate perturbations that when added to the original images will result in a misclassification by the target model. The perturbation generated by this method can then be printed and used as physical perturbation on a real traffic sign board. The general pipeline of the RP2 method is shown in figure. The algorithm takes a traffic sign, for which the physical perturbation is to be generated, and a mask which defines the constraints on perturbation. The mask helps RP2 to limit computed perturbations to specific spots on the image and to make the physical perturbation inconspicuous to human observers. For generating the perturbations, RP2 starts with a random noise vector which is then masked using the input mask image and is added to the traffic sign under attack. This traffic sign image with the perturbations is then given to a classifier which gives the prediction for this noisy image. The predictions from the classifier are used to calculate the loss and optimization is performed over the noise vector in order to minimize the loss value.

Mean Square Error between the prediction output of the classifier given the noisy image as input and the desired prediction output is used as the loss function for our project. The loss function is modified by adding additional terms for L1 and L2 regularizations. Adjusting these terms for a given sign may help to reduce the conspicuity of the perturbation but decreases the likeliness of traffic signs under attack to be classified as the desired traffic sign.

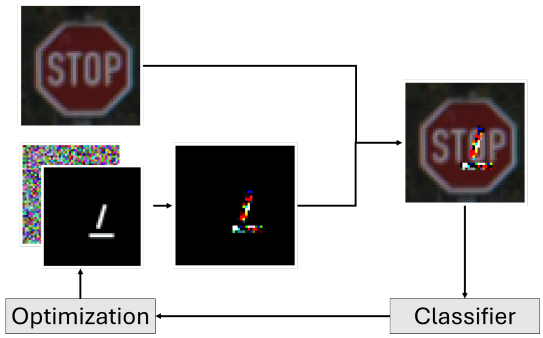

Figure 1: RP2 Attack Method Pipeline

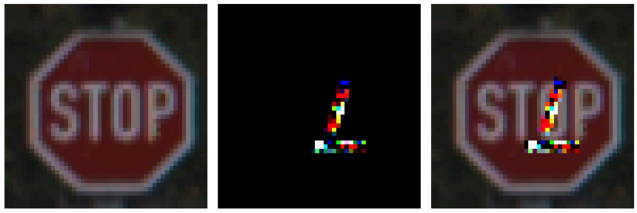

Figure 2: RP2 Attack Output (Left) Traffic sign under attack (Center) Generated Label (Right) Traffic sign with physical perturbation sticker

4. Experiments

In this section, we will evaluate the RP2 method against two well known and widely used traffic sign datasets: (1) German Traffic Sign Recognition Benchmark (GTSRB) and (2) Labeled Intersection and Segmentation for Autonomous Driving (LISA). We will also use two different vision classification deep neural network architectures: (1) Deep Convolutional Neural Network (CNN) and (2) Residual Network (ResNet18) to find if RP2 is more effective against one or the other. In addition we also conduct experiments to show the effectiveness of two defense mechanisms i.e. using data augmentation while training the classifier and using adversarial samples while training the classifier, against RP2 method.

4.1 Dataset

To evaluate the impact of adversarial attacks on road sign recognition, we conducted experiments using two widely used datasets: GTSRB and LISA.

GTSRB Dataset Details

This dataset is a collection of German Traffic Signs

Number of labels = 43

Number of training samples = 39,209

Number of test samples = 12,630

Figure 3: GTSRB Samples

LISA Dataset Details

This dataset is a collection of USA Traffic Signs

Number of labels = 47

Number of training samples = 5,499

Number of test samples = 2,356

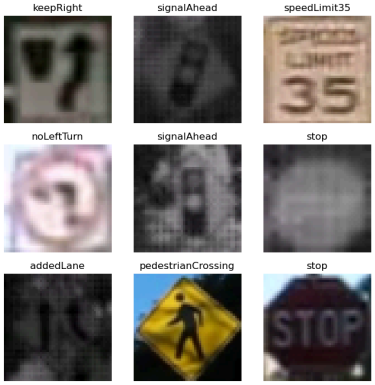

Figure 4: LISA Samples

4.2 Classifiers

CNN Model Definition

The CNN model works by analyzing images and identifying patterns and features. This model does this task by passing the input image through layers that specialize in recognizing different aspects of the image like shapes, edges, etc. These layers are convolutional layers that apply filters to the images and then extract features. Pooling layers reduce the spatial dimensions of the features. With each passing through the layers, the model learns to recognize complex patterns. In the end, features are flattened and passed through fully connected layers which make the final prediction based on the features it learned.

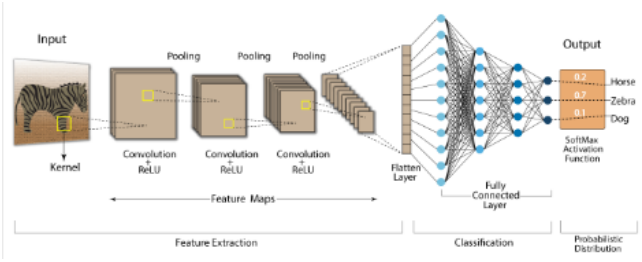

Figure 5: Convolution Neural Network (CNN)

ResNet Model Definition

ResNet is an abbreviation of Residual Network and represents a pioneering architecture in the space of deep convolutional neural networks. Unlike CNN architecture, ResNet includes skip connections that enable the network to bypass certain layers which facilitate the flow of information through the network more effectively. This design approach allows the training of deeper networks with sometimes even hundreds of layers. This phenomenal characteristic of this model allows us to train more robust and accurate models and also to simplify the optimization process which leads to improved performance in image classification, object detection, and segmentation.

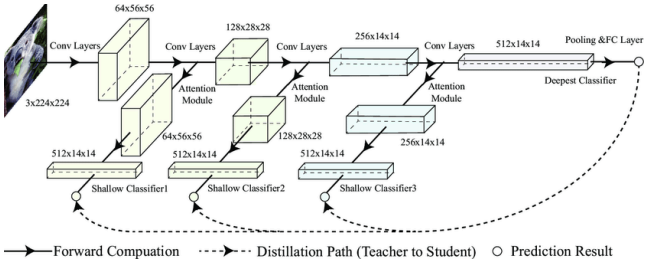

Figure 6: ResNet 18 Architecture

4.3 Adversarial Attack Defense Mechanisms

Data Augmentation

Our training framework leveraged Convolutional Neural Networks (CNNs) and the ResNet18 model, renowned for their effectiveness in image classification tasks.

To ensure robustness and mitigate biases inherent in the datasets, we employed data augmentation techniques. Augmenting the data helped us achieve a more balanced distribution, which is crucial for enhancing the model's generalization capability. By training our models on both original and augmented datasets, we aimed to capture various scenarios encountered in real-world driving environments.

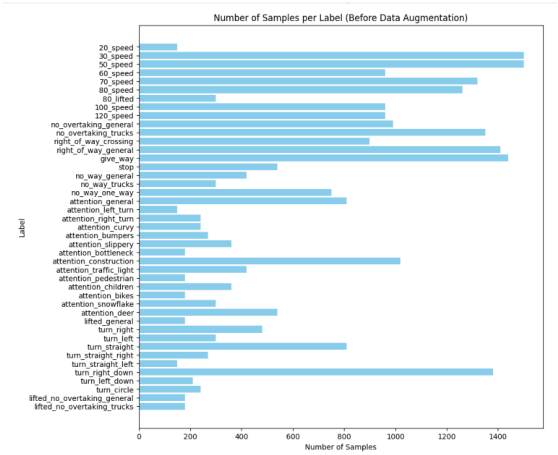

Figure 7: Number of images per label in the GTSRB dataset before data augmentation.

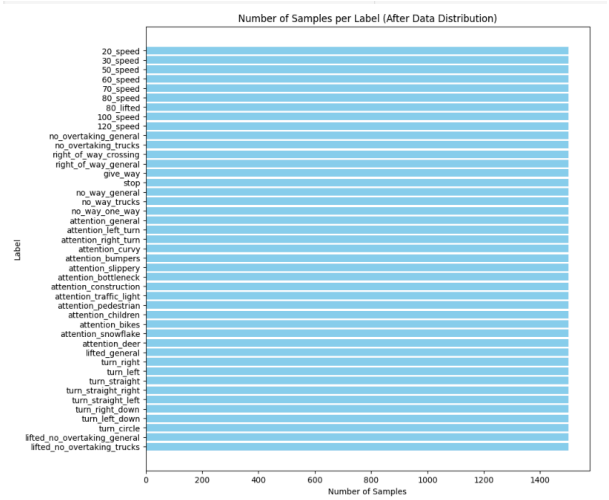

Figure 8: Number of images per label in the GTSRB dataset after data augmentation.

Data augmentation was done using the built in functions from torchvision library. The following techniques were used to enrich the training data and improve the model's generalization capabilities:

- **Random Rotation:** Images were subjected to modest random rotations (up to 10 degrees) in both clockwise and counterclockwise directions to simulate real-world conditions in which signage may be slightly misaligned.
- **Color Jitter:** The photos' brightness and contrast were randomly modified within a specific range to simulate illumination fluctuations that may occur in real-world scenarios.

Figure 9: GTSRB Data Augmentation

Adversarial Training

Adversarial Training is known for the simplicity and effectiveness of the strategy for training an adversarially robust model and hence it was selected as the second method for defense against RP2. The adversarial images generated from the RP2 attack method were saved and used during the training loop.

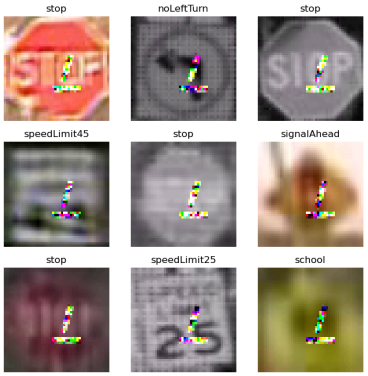

Figure 10: LISA Adversarial Samples

4.3 Results

In this section the test accuracy of the classifier and effectiveness of the RP2 attack method are evaluated. Table.1 summarizes the performance of each architecture over both the datasets along with the effect of RP2 attack

and defense mechanisms over the accuracy of the model. For training all the models following hyper-parameters were used:

Batch Size = 64

Epochs = 20

Learning Rate = 0.001

Weight Decay = 1e-5

For computing the perturbations over the test set following hyper-parameters were used:

Attack Epochs = 10

Lambda = 0.0

For evaluating the adversarial attack, the same mask was used for all the images and the value of lamda was not tuned for individual images. The Attack Success Rate was defined as ration of (# of samples predicted as target label - # of samples of target label) and (# of samples - # of samples of target label)

From the table we can observe that both the architectures by themselves are doing well on both GTSRB and LISA datasets with the lowest performance of 95.9% test accuracy observed for RESNET18 over GTSRB. When undefended models were attacked with the RP2 method the accuracy of all the models dropped significantly in the range of 55 - 94%. The worst affected model by RP2 attack was GTSRB-RESNET18-FT and the least affected model was LISA-RESNET18-FT. The ASR among undefended models was highest for GTSRB-CNN at 21.4%.

When the models were trained with augmented dataset, the test accuracy of the models was not significantly affected. The effectiveness of the RP2 attack was reduced for models trained on the GTSRB dataset but the attack effectiveness increased for models trained on the LISA dataset.

When the models were trained with adversarial training, the test accuracy was reduced with the maximum reduction of 0.7% for LISA-RESNET18-FT. The model accuracy while under adversarial attack improved significantly for all the models with maximum improvement observed for GTSRB-CNN. The best performance of the models was observed when both the defense mechanisms were used. The figure shows the comparison of Test Accuracy, Adversarial Attack Accuracy and Attack Success Rate for all the models.

5. Conclusion

From the previous section it can be concluded that RP2 method is a very effective method for generating adversarial physical perturbations which can be used on real traffic signs without raising any suspicion from human observers. But it should be noted that it is not very accurate in fooling the classifier to predict the desired label. It can also be concluded that while reducing the model performance the defense mechanisms namely data augmentation and adversarial training are successful in decreasing the effectiveness of the attack. Here it should also be noted that the adversarial training and attacks were done using the same mask which might have helped improve the defense against the attack. Our work thus implements one of the available adversarial attack methods in literature and subsequent defense mechanisms. Furthermore we evaluate the feasibility and effectiveness of the attack method over different traffic sign datasets and architectures to set a foundation for developing new defense methods for safety application such as traffic sign recognition.

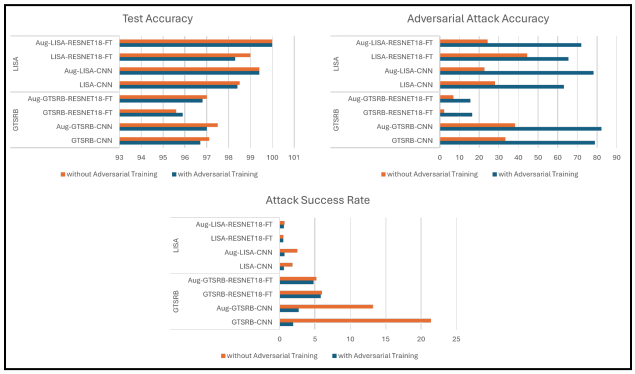

Figure 11: Performance comparison of RP2 method against undefended models and defended models

6. References

[1] C. Szegedy et al., “Intriguing properties of neural networks,” in Proc. 2nd Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, Apr. 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015.

[3] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” 2016, arXiv:1607.02533.

[4] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” 2017, arXiv:1712.09665.

[5] K. Eykholt et al., “Robust physical-world attacks on deep learning visual classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634.

[6] S. Thys, W. V. Ranst, and T. Goedemé, “Fooling automated surveillance cameras: Adversarial patches to attack person detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 49–55.

[7] Y. Song, R. Shu, N. Kushman, and S. Ermon, “Constructing unrestricted adversarial examples with generative models,” in Proc. Annu. Conf. Neural Inf.

Process. Syst., Montreal, QC, Canada, Dec. 2018, pp. 8322–8333.

Dataset	Model	Test Accuracy		Adversarial Attack Accuracy		Attack Success Rate	
		with Adversarial Training	without Adversarial Training	with Adversarial Training	without Adversarial Training	with Adversarial Training	without Adversarial Training
GTSRB	GTSRB-CNN	96.7	97.1	78.9	33.4	1.9	21.4
	Aug-GTSRB-CNN	97	97.5	82.3	38.3	2.7	13.2
	GTSRB-RESNET18-FT	95.9	95.6	16.4	2.2	5.8	6
	Aug-GTSRB-RESNET18-FT	96.8	97	15.5	6.9	4.8	5.2
LISA	LISA-CNN	98.4	98.5	63.2	28.1	0.6	1.8
	Aug-LISA-CNN	99.4	99.4	78.3	22.6	0.7	2.5
	LISA-RESNET18-FT	98.3	99	65.4	44.4	0.5	0.5
	Aug-LISA-RESNET18-FT	100	100	72	24.3	0.6	0.7

Table 1: Performance comparison of RP2 method against undefended models and defended models