CMPE 588 - Fall 2019

Adversarial Training of Neural Networks

Aysu Sayın

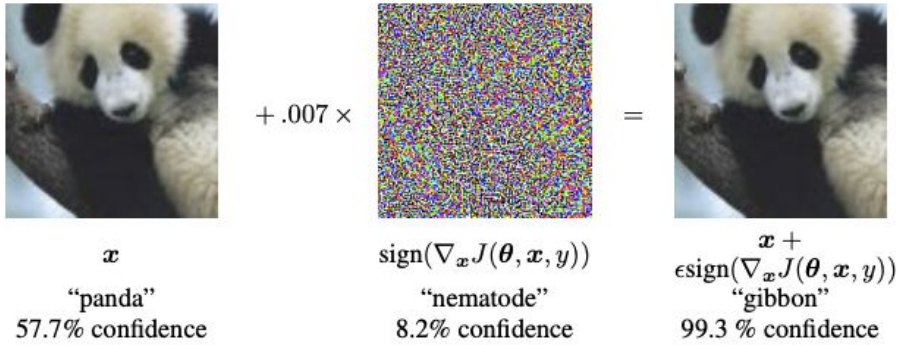
Introduction

Project: Training the model with **adversarial samples** to increase **robustness**

- FGSM Attack
- Two Models
- German Traffic Sign Dataset

Fast Gradient Sign Method

Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples



$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

- adv_x : Adversarial image.
- x : Original input image.
- y : Original input label.
- ϵ : Multiplier to ensure the perturbations are small.
- θ: Model parameters.
- J: Loss.

Dataset

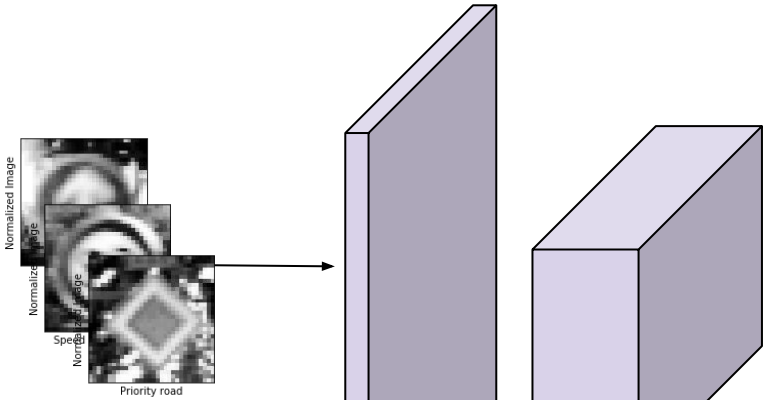
German Traffic Sign Dataset:

German Traffic Sign Benchmark

More than 50,000 images

43 Classes

Models: LeNet-5



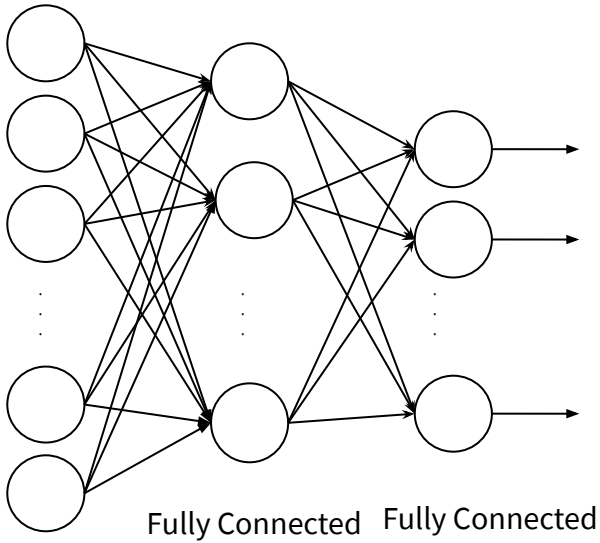
Input Image

Gray Scale
32x32
Normalized

Convolutional
Layer 1
+Max Pooling

Convolutional
Layer 2
+ Max Pooling

Flatten



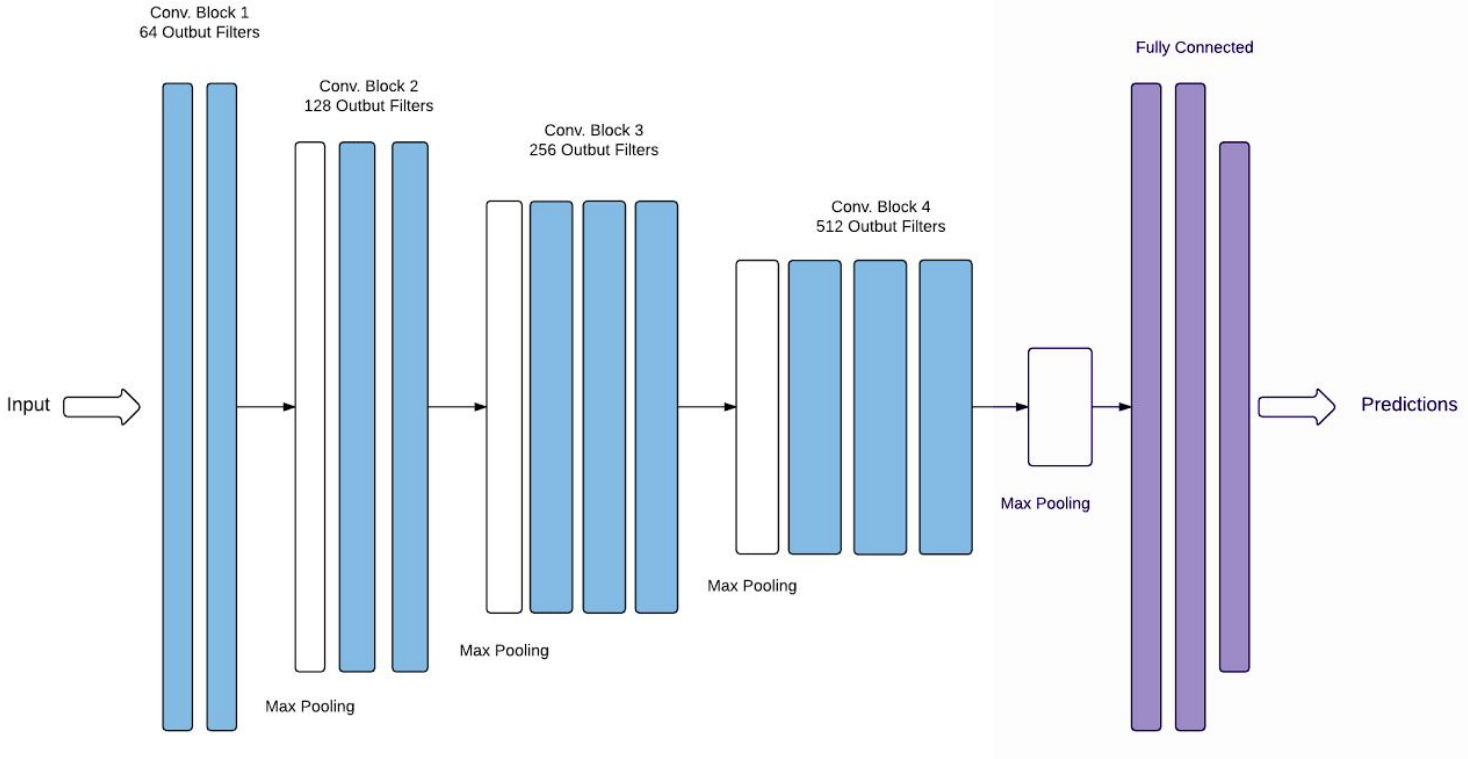
Fully Connected
120 Neurons
ReLU

Fully Connected
84 Neurons
ReLU

Fully Connected
43 Neurons
Softmax

Class
(one hot
encoded)

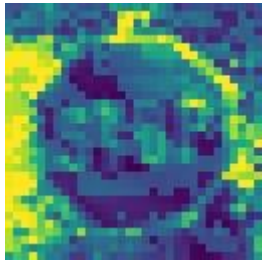
Models: VGGNet



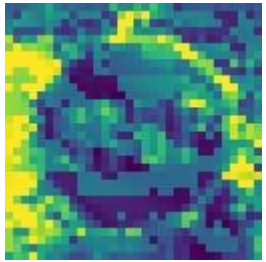
Generating Adversarial Samples

- Used FGSM
- $\epsilon = 0.01, 0.10, 0.15$
- 710 adversarial image for LeNet-5
- 500 adversarial image for VGGNet
- Example: Generate_Adversarial_Samples

$\epsilon = 0.01$



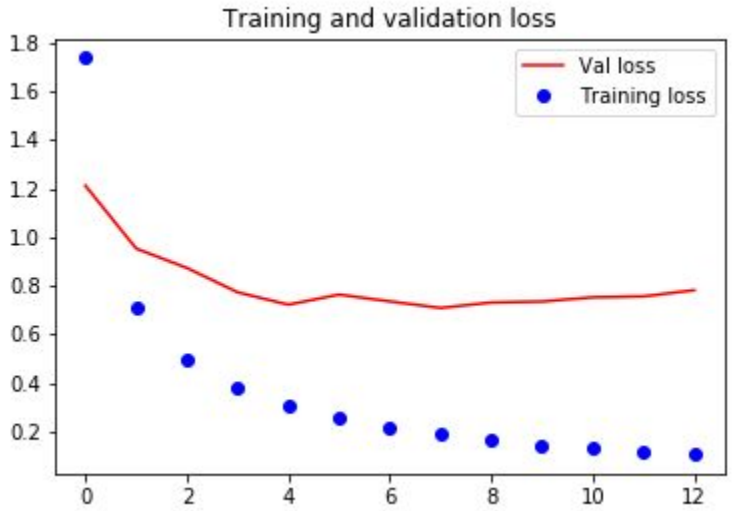
$\epsilon = 0.10$



$\epsilon = 0.15$

Results: LeNet-5

Normal Training



Training Set Size: 39209

Validation Set Size: 5000

Test Set Size: 7629

Training Set Accuracy: 97.06%

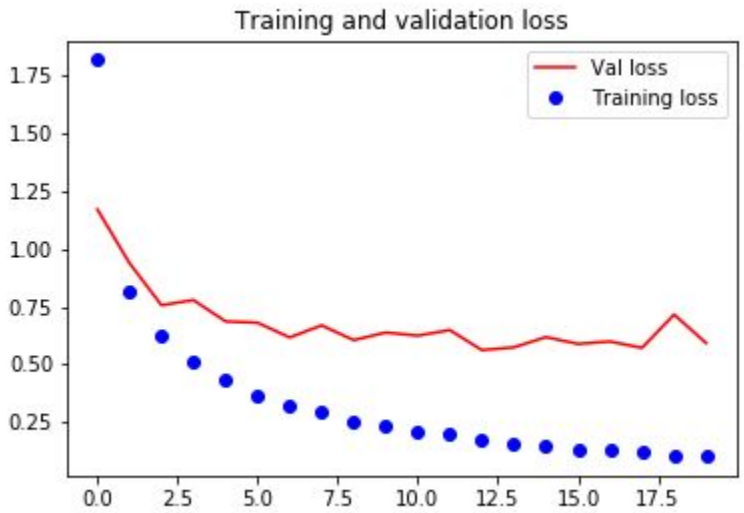
Validation Set Accuracy: 86.22%

Test Set Accuracy: 85.88%

Jupyter Notebook

Results: LeNet-5

Adversarial Training



Training Set Size: 40709

Test Set Size: 8258

Training Set Accuracy: 97.80%

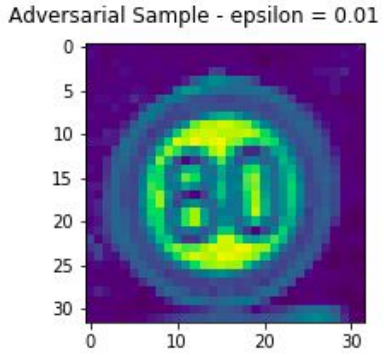
Test Set Accuracy: 85.67%

Jupyter Notebook

Comparison

Total adversarial test data: 629

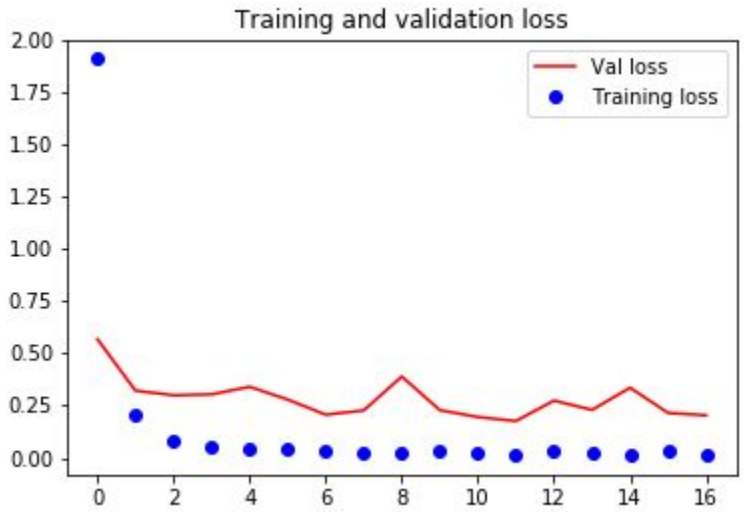
	LeNet Model	LeNet Model - Adversarial Train
Wrongly predicted	514	249
Accuracy	18.28%	60.41%
Test Set Accuracy	85.88%	85.67%




```
Real label: 5
Predicted label by original model: 3
Predicted label by adversarially trained model: 5
```

Results: VGGNet

Normal Training



Training Set Size: 39209

Validation Set Size: 5000

Test Set Size: 7629

Training Set Accuracy: 99.85%

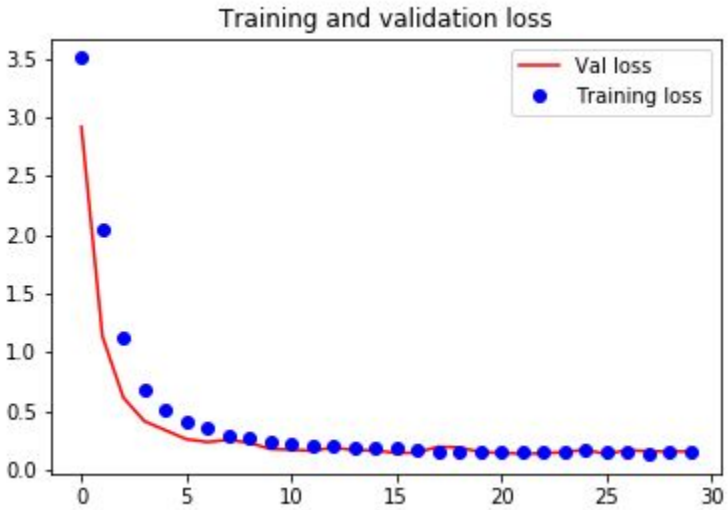
Validation Set Accuracy: 86.22%

Test Set Accuracy: 95.73%

Jupyter Notebook

Results: VGGNet

Adversarial Training



Training Set Size: 39909

Test Set Size: 7932

Training Set Accuracy: 99.26%

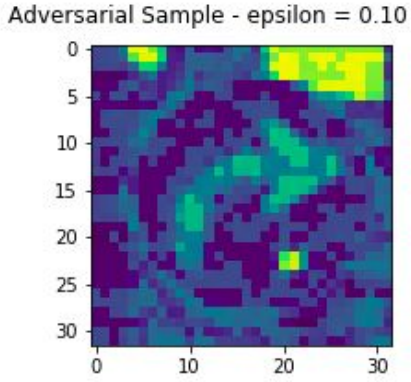
Test Set Accuracy: 95.60%

Jupyter Notebook

Comparison

Total adversarial test data: 303

	VGGNet Model	VGGNet Model - Adversarial Train
Wrongly predicted	139	101
Accuracy	54.12%	66.67%
Test Set Accuracy	95.73%	95.60%




```
Real label: 33
Predicted label by original model: 14
Predicted label by adversarially trained model: 33
```

Conclusion

Adversarial training does not increase the test accuracy but it increases the accuracy of predicting the true labels of the adversarial samples. Thus, it increases robustness.

Thank you!