Sumit Shahi
Alexandar Vakinski
Adversarial Machine Learning CS587
01/26/2024

Assignment 1:

Part 1: FGSM and PDG Attacks against Keras-TensorFlow Image Classification Models

Table 1. Classification Accuracy

Model	Train Set	Validation Set	Test Set
VGG16	98.28%	82.95%	82.91%

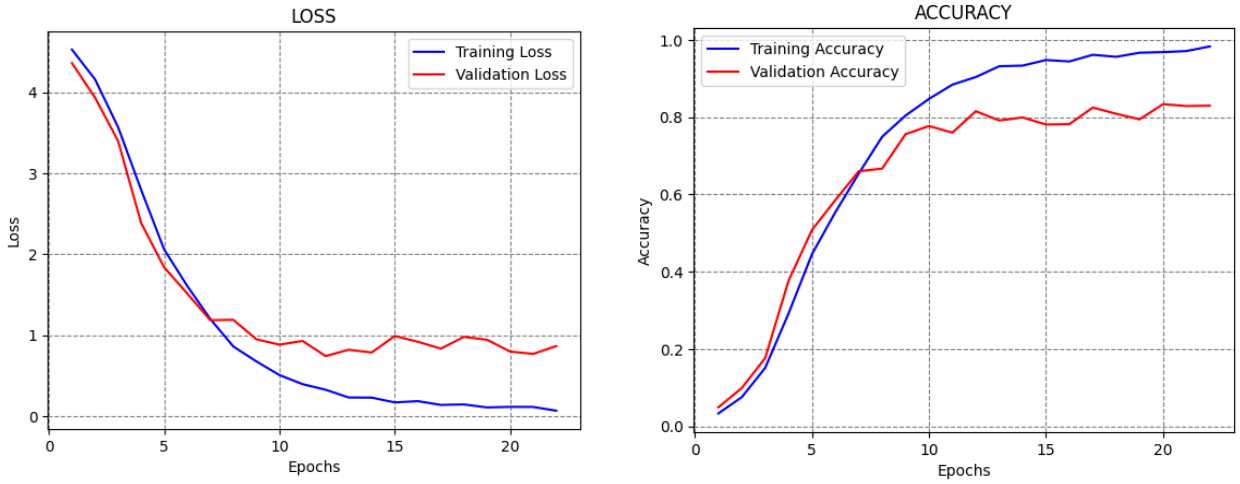


Figure 1. Loss and Accuracy Plots

Table 2. Classification Accuracy on Clean and Adversarial images

Model	Clean images	Adversarial images ϵ=1/255	Adversarial images ϵ=5/255	Adversarial images ϵ=8/255
FGSM attack	81%	62.50%	26.00%	22.50%
PGD attack	81%	59.50%	20.50%	16.50%

Sumit Shahi
Alexandar Vakinski
Adversarial Machine Learning CS587
01/26/2024

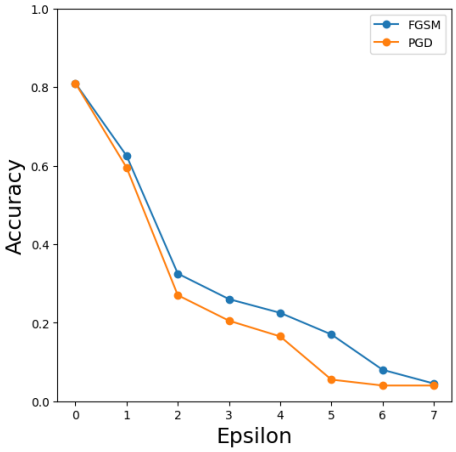
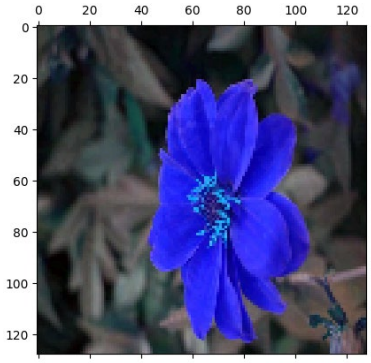
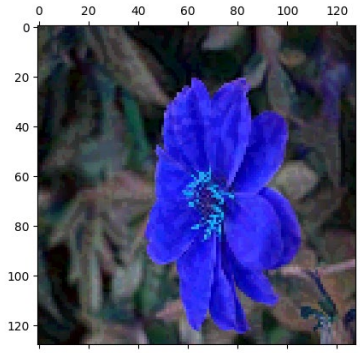


Figure 2. Plots of Accuracy versus Perturbation

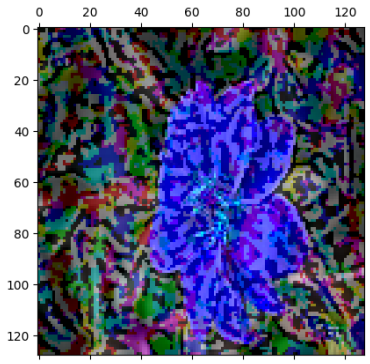
```
Perturbation magnitude: 0.0118
Predicted label: 50
```



```
Perturbation magnitude: 0.0314
Predicted label: 50
```



```
Perturbation magnitude: 0.1961
Predicted label: 73
```



```
Perturbation magnitude: 0.3137
Predicted label: 73
```

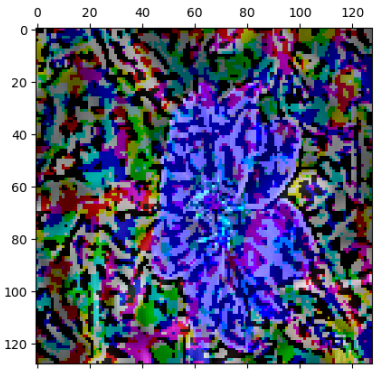


Figure 3. Adversarial Images

Sumit Shahi
Alexandar Vakinski
Adversarial Machine Learning CS587
01/26/2024

Part 2: FGSM and PDG Attacks against PyTorch Image Classification Models

Table 3. Classification Accuracy

Model	Train Set	Validation Set	Test Set
VGG16	98.25%	83.31%	87.74%

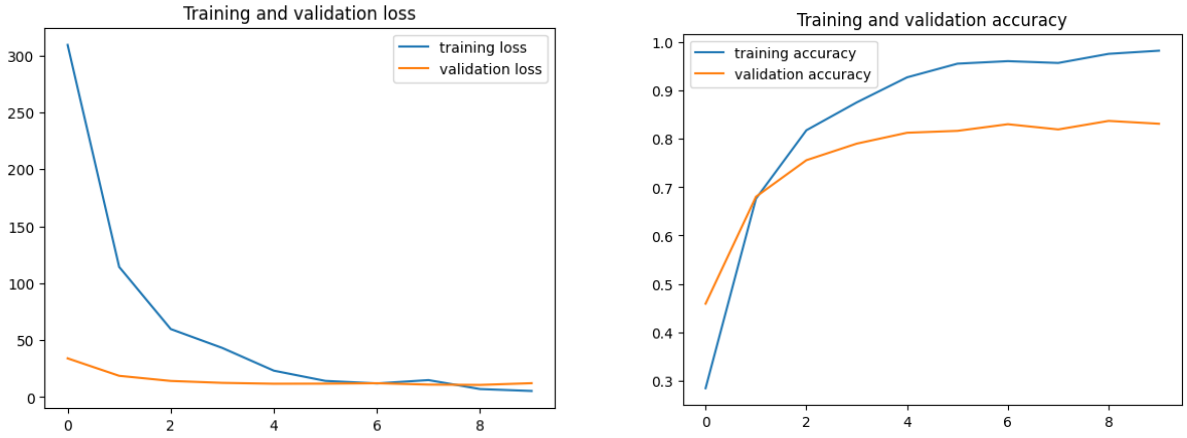


Figure 4. Validation Loss and Accuracy Plots

Table 4. Classification Accuracy on Clean and Adversarial images

Model	Clean images	Adversarial images ϵ=1/255	Adversarial images ϵ=5/255	Adversarial images ϵ=8/255
FGSM attack	76.43%	66.04%	43.35%	37.87%
PGD attack	76.43%	59.11%	33.59%	28.27%

Sumit Shahi
Alexandar Vakinski
Adversarial Machine Learning CS587
01/26/2024

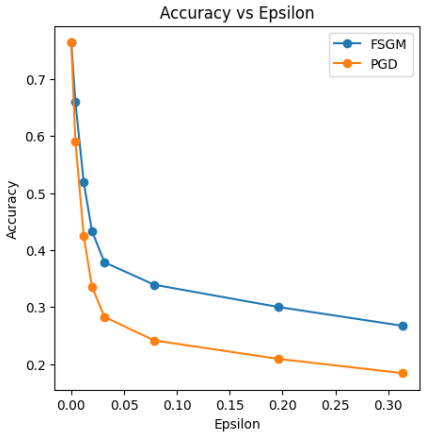


Figure 5. Plots of Accuracy versus Perturbation

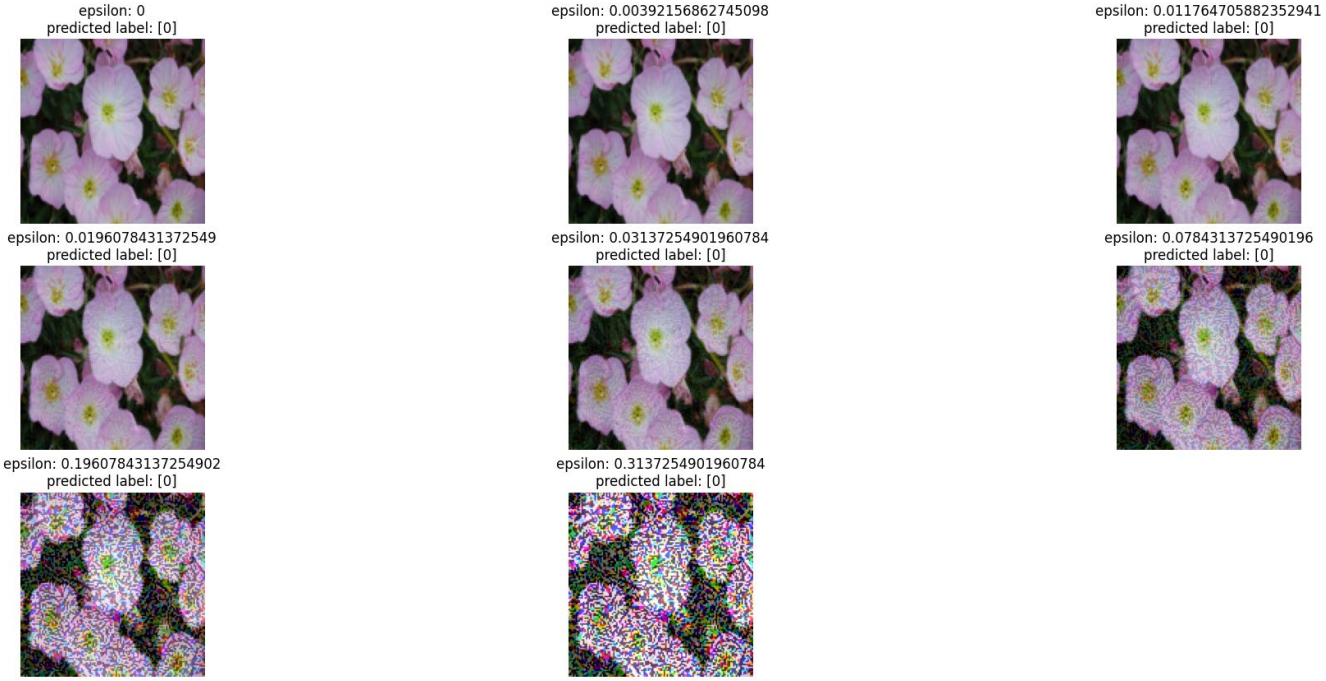


Figure 6. Adversarial Images

Sumit Shahi
Alexandar Vakinski
Adversarial Machine Learning CS587
01/26/2024

Analysis:

When looking at the clean image accuracy in both scenarios, it's evident that the models excel in handling regular, unmodified images.

Under the FGSM attack, the latest findings suggest that the model demonstrates increased resilience compared to the initial results. To illustrate, at ϵ=1/255, the PyTorch-based model achieves a higher accuracy of 66.04% as opposed to 62.50% in the original outcomes. Nonetheless, the trend of accuracy decreasing with higher levels of perturbation remains consistent in both result sets.

Moving on to the PGD attack, the PyTorch-based model consistently outperforms the original results. Across all levels of perturbation, the accuracy is higher in the updated findings. Similar to the FGSM attack, there is a persistent decline in accuracy as perturbation levels increase.

In summary, the PyTorch-based model displays enhanced robustness against both FGSM and PGD attacks, evident in its higher accuracy values in adversarial scenarios compared to the original results.