CSC 7570-001 AI Assisted Cyber Security (Fall 2023)

(**Assignment 5 - Adversarial Solutions**)

Ekle, Ocheme Anthony

November 10, 2023

Lab A - Adversarial Attack on Malware Classifier

1 Adversarial Attack: Questions and Solution

1.1 Q1: Did the model correctly classify it?

Answer: Yes, the model correctly classify the **Gatak sample** with 100 % confidence, this is shown in Fig. 1

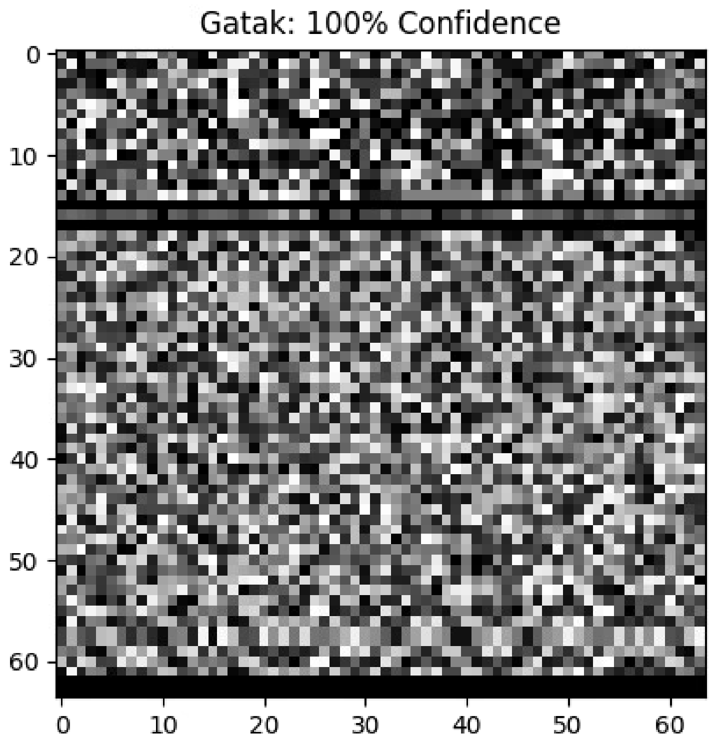

Figure 1: Screenshot of Classifying the base Gatak sample

1.2 The adversarial pattern

Shown below is the plot after uncommenting line 91 for "plt.shown()" in the FGSM_attack.py file.

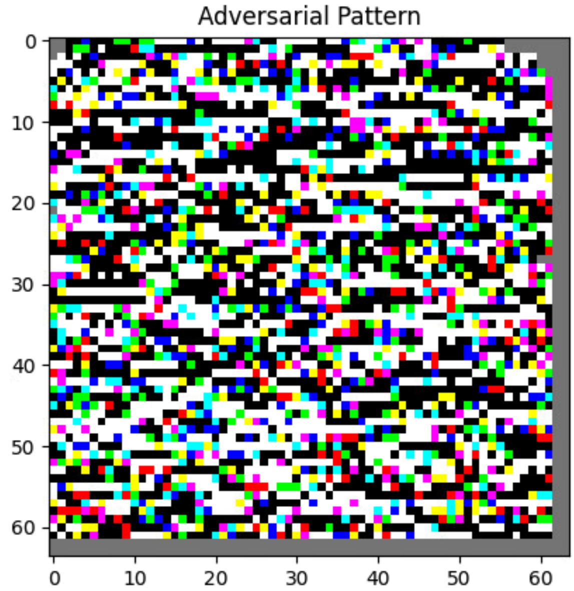

Figure 2: Screenshot of the adversarial pattern

1.3 Making an adversarial sample by adjusting the Epsilon.

1.3.1 Q3: Did the model calssify the adversial sample correcly?

Answer: No, with **Epsilon = 0.15** the model wrongly classify the image as Remnit malware family. See Figure 3 for the screenshot.

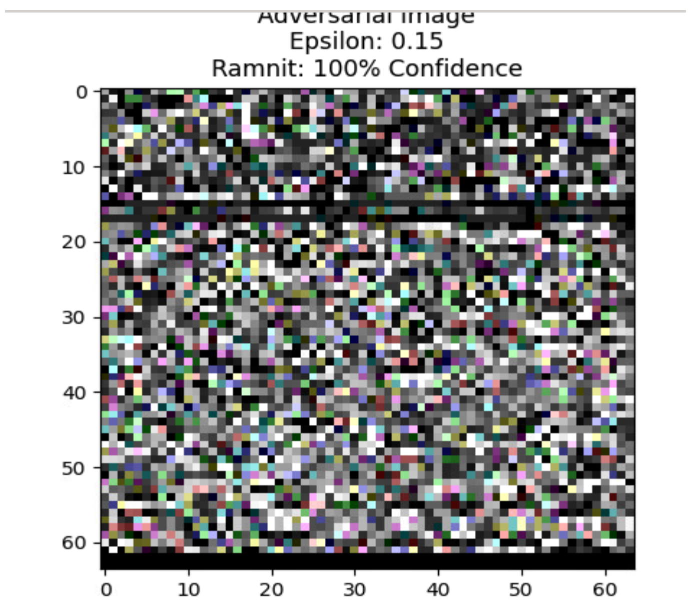

Figure 3: Screenshot of the adversarial classification with Epsilon = 0.15

1.3.2 Multiple Testing with Espilon=0.15:

After running the test more than once, the model always produce the **Remnit family** prediction with the same confidence of 100 %

2 Fine-tuning the Adversarial Attack

The goal of an adversarial image is to maximize disruption to the model while minimizing visible changes to the image.

2.1 epsilon value at 0.05, 0.25, 0.30

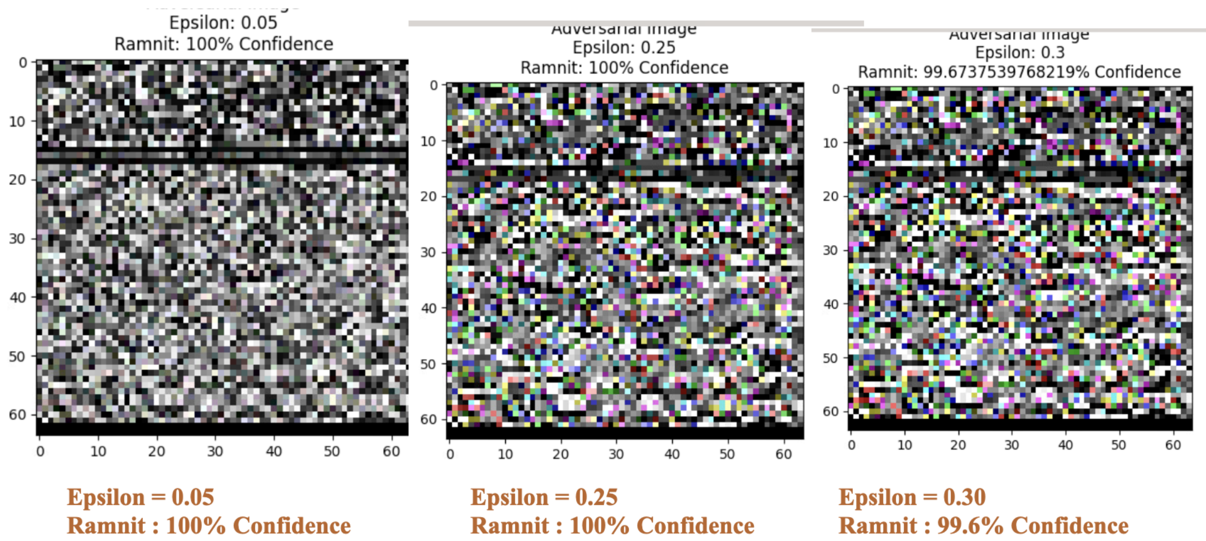

Figure 4: Screenshot of Epsilon = 0.05, 0.25, 0.30

2.2 epsilon value at 0.40, 0.50, 0.65

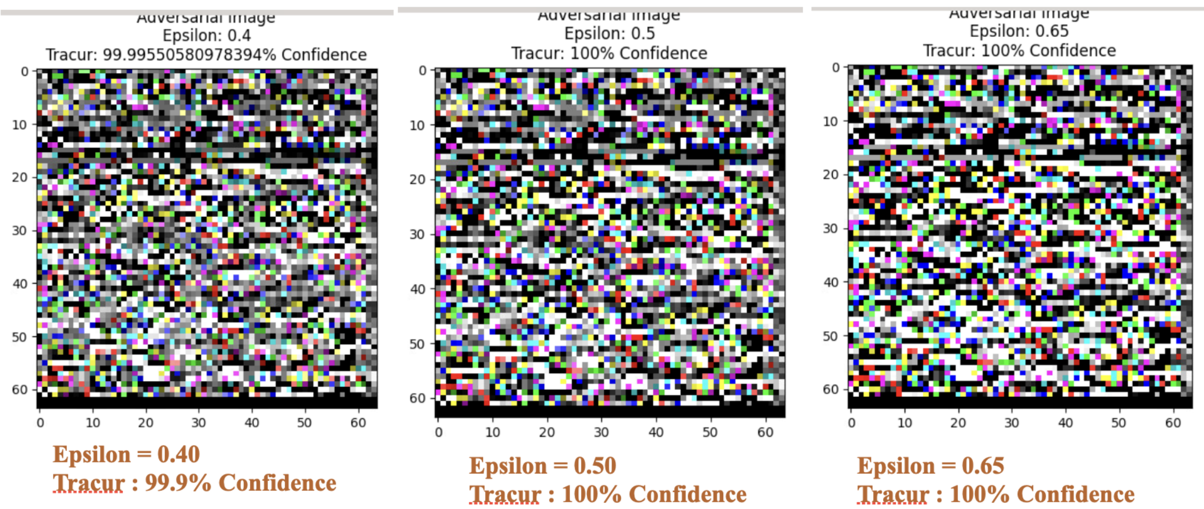

Figure 5: Screenshot of Epsilon = 0.40, 0.50, 0.65

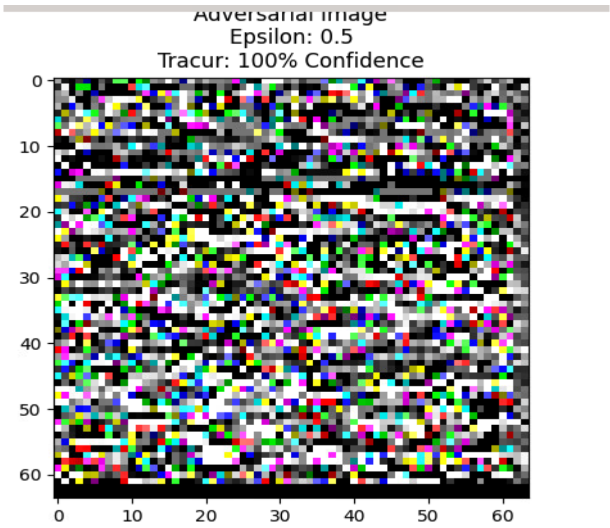

Figure 6: Screenshot of Epsilon = 0.50

2.3 epsilon value at 0.85, 0.92, 0.99

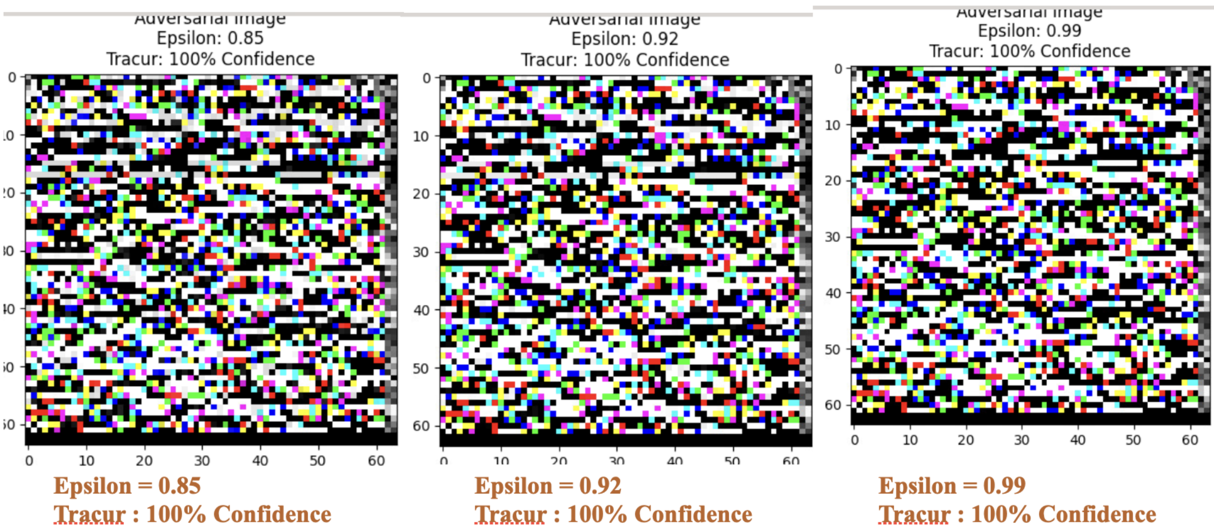

Figure 7: Screenshot of Epsilon = 0.85, 0.92, 0.99

2.3.1 Best Value of Epilon

In my experiment the best values for epilon with 100% confidence includes:

1. 0.25 for **Ramnit** family, see fig. 4
2. 0.50 for Tracur family see 5 and 6
3. 0.85, 0.92, 0.99 for Tracur family, see fig. 7

Q 3.2: Yes, there is a visual difference in the adversarial sample and the original as the epsilon value increases.

Q 3.3: Yes, the model misclassify my adversarial sample to **Ramnit** family and **Tracur** family at epsilon = 0.50, 0.85, 0.92, 0.99 with 100% confidence

3 Lab B. Adversarial Training

Lab A - Adversarial Attack on Malware Classifier

3.1 Model accuracy on regular inputs

- **Accuracy on regular input** = 91.5%. Attached is the screenshot in fig. 8

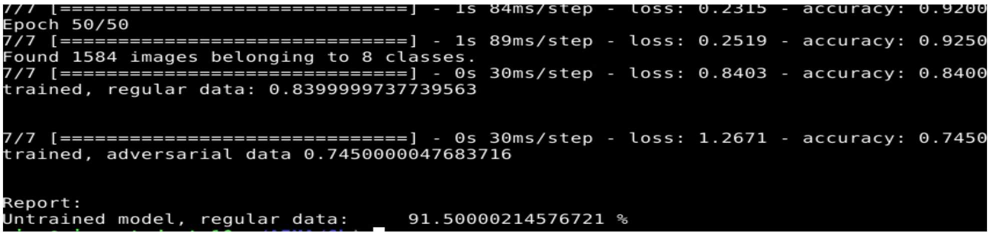

Figure 8: Model's accuracy on regular malware sample images

3.2 Plots of some of the non-adversarial malware samples

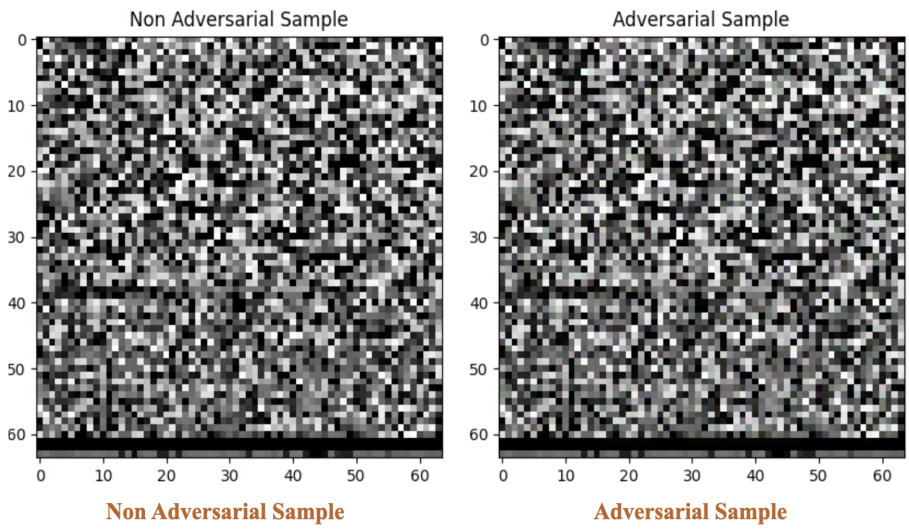

Figure 9: Adversarial and Non-adversarial sample

3.2.1 Untrained model accuracy on the adversarial inputs

- **Untrained model's accuracy** = 86.00%. Attached is the screenshot in fig. 10.

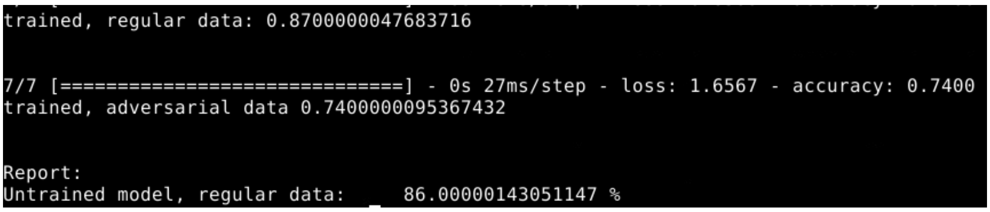

Figure 10: untrained model's accuracy on the adversarial inputs

3.3 Model's accuracy on adversarial samples

- **model's accuracy on adversarial** = 74.50%.

```
trained, adversarial data 0.7450000047683716


Report:
Untrained model, regular data:     87.00000047683716 %
Trained model, adversarial data:   74.50000047683716 %
```

Figure 11: model's accuracy on adversarial samples

3.4 model's accuracy on regular malware samples

- **model's accuracy on regular malware sample** = 87.99%.

```
Report:
Untrained model, regular data:     86.00000143051147 %
Trained model, adversarial data:   73.00000190734863 %
Trained model, regular data:       87.99999952316284 %
```

Figure 12: model's accuracy on regular malware samples

3.5 model's accuracy with New set of Adversarial samples of Epsilon =0.005

- **Accuracy on adversarial sample (espilon =0.005)** = 86.00%.

```
Report:
Untrained model, regular data:     86.00000143051147 %
Trained model, adversarial data:   86.00000143051147 %
Trained model, regular data:       87.99999952316284 %
```

Figure 13: model's accuracy = 86 % epsilon=0.005

3.6 model's accuracy with New set of Adversarial samples of Epsilon =0.1

```
Report:
Untrained model, regular data:     86.50000095367432 %
Trained model, adversarial data:   14.000000059604645 %
Trained model, regular data:       86.50000095367432 %
```

Figure 14: trained model's accuracy = 14.00 % epsilon=0.1

- **The trained model accuracy on adversarial sample with (espilon =0.1)** = 14.00%.

Answer: with larger epsilon value such as **e =0.1**, our model performs poorly with **14** % accuracy, hence our model can't sufficiently defend against those adversarial samples.

3.7 Train model with large Epsilon =0.1 and Attack model with small Epsilon = 0.005

- **The trained model accuracy for adversarial samples** = 77.99%.
- **The trained model accuracy for regular sample** = 68.99%.

```
Report:
Untrained model, regular data:     91.00000262260437 %
Trained model, adversarial data:   77.99999713897705 %
Trained model, regular data:       68.99999976158142 %
```

Figure 15: trained model'swith epsilon=0.1 vs attack epsilon = 0.005

- **Q 7.3 Answer: YES**, With larger epsilon value e = 0.1 for the training model. The accuracy of model on regular dropped from **86.5** % to **68.99** %.

- **Q-7.4:** Is adversarial training on this model a good protection against fast gradient sign method attacks? **Why or why not?**

3.7.1 Answer: 7.4

YES, adversarial training on this model could be effective against **fast gradient sign method attacks**, as the model exhibits higher accuracy on adversarial samples of 77.99% for train epsilon = 0.1 and 86% for training epsilon =0.1.

The improvement to suggests that the model has learned to better handle **perturbations** introduced by such attacks, making it a potentially robust defense. However, a more comprehensive evaluation and testing on a diverse set of adversarial scenarios are recommended for a conclusive assessment.