Defense Against Adversarial Arts

Gantavya Bhatt (2016EE10694), Nishad Singhi(2016EE10107), Hritik Bansal (2016EE10071)

1 Problem Statement

- Choose 2 datasets - one gray scale and color each (except for MNIST).
- Design a classifier for a classification task on the choosen datasets.
- Design a generative model (GAN and it's variants or VAE and it's variants) on both datasets.
- Break the classifier by creating atleast 2 types of adversarial attacks on the trained classifiers.
- Finally use the trained generative models to make the classifier robust against those attacks.

2 Dataset used

The dataset we chose were **CIFAR-10** and **Kuzushiji MNIST.** The CIFAR-10 dataset contains 60,000 32x32 color images in 10 different classes.The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. There are 6,000 images of each class. Whereas, Kuzushiji MNIST is a recent dataset released in December'18 comprising of old Japanese characters of hiragana script. Kuzushiji-MNIST contains 70,000 28x28 grayscale images spanning 10 classes (one from each column of hiragana). 10,000 are used for the cross-validation and testing. CIFAR-10 being a common dataset, we are not showing its image here as an introduction.

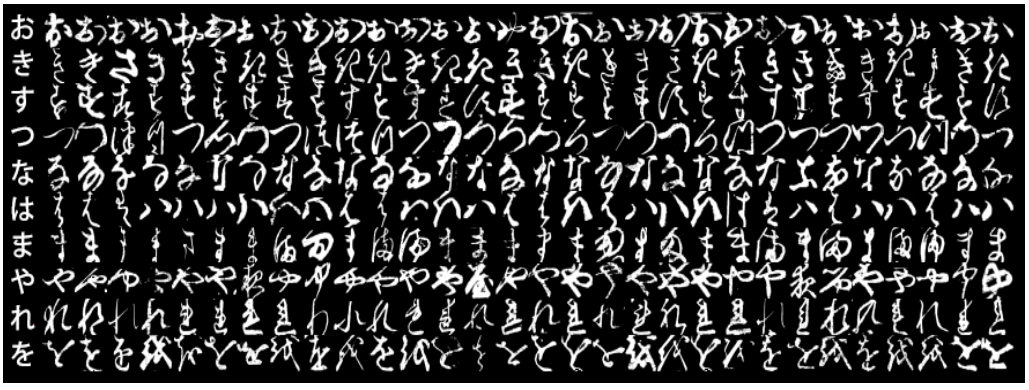

Figure 1: KMNIST dataset. First column being the true class, the rest being the images in the dataset [3]

3 Attacks and Implementation details

- In total we tried three different type of attacks on our 2 classifiers. The first kind of attack being the Fast Gradient Sign method, Projected Gradient Descent attack and Carnili Wagner Attack.
- The FGSM attack is a white box attack in where we have the complete information of our model. We did 2 type of FGSM attacks on our models. The first attack being a non-targetted FGSM and the other one being the targeted FGSM.
- The untargeted FGSM involved doing the gradient ascent on the loss landscape with respect to the input images in the hypercube (for the valid image).
- The Targeted FGSM invovled the doing a gradient descent on the loss landscape of the input image and the targeted class.
- The projected Gradient Descent attack was trained for 40 iterations followed by clipping
- The Carlini Wagner attack involved the modified optimization problem on the tanh space with L2 as a distance metric.

$$\textbf{minimize} \quad \|\frac{1}{2}(\tanh(w)+1) - x\|_2^2 + c \cdot f(\frac{1}{2}(\tanh(w)+1)$$

$$f(x') = max(max(Z(x'_i), i \neq t) - Z(x'_t), -\kappa)$$

Z is the pre-activation of the ultimate layer

4 Classifier results

Datasets	Accuracy
CIFAR-10	79.88 %
Kuzushiji-MNIST	94.18 %

Table 1: Our Classifier on non-adverserial examples

These are our classifier accuracies for the CIFAR-10 and the Kuzushiji MNIST dataset.

5 Adverserial Attacks

In this section, we will discuss the outputs of different attack algorithms and their impact on the accuracy of our models. Firstly, we will discuss the image mis-classified by KMNIST and the CIFAR-10 classifier inspite of being so close.

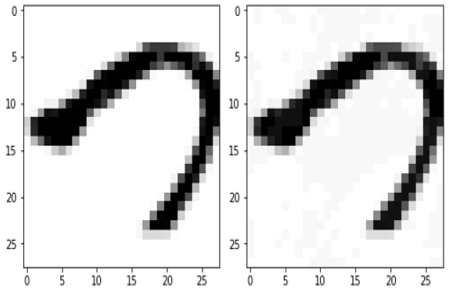

Figure 2: Image misclassified as hiragana character "ya" instead of "shi" (a) Non adverserial image, (b) Adverserial.

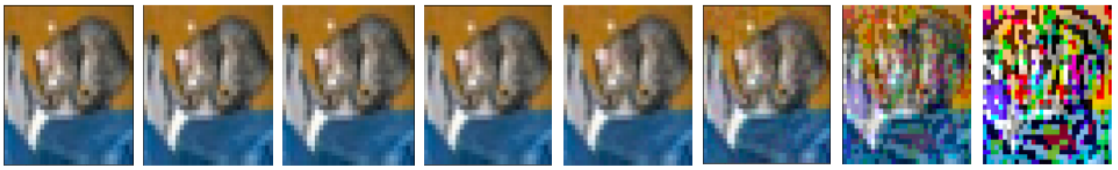

Figure 3: Image after the perturbation by the FGSM attacks on CIFAR-10

5.1 Impact on accuracy of classifier by increasing attack power

As it can be clearly seen, as the value of the epsilon increases, the misclassification rate increases. We kept the epsilon in such a way that the image remains in the valid dimensions, that is the pixels value of the image remain in the 0 to 1 hypercube.

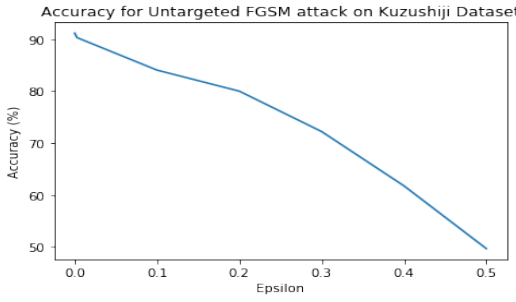

Figure 4: Accuracy versus epsilon in untargetted attack FGSM on Kuzushiji

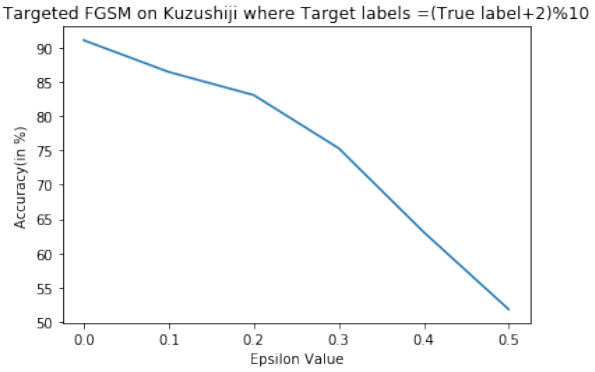

Figure 5: Accuracy versus epsilon in targeted attack FGSM on Kuzushiji

targeted fgsm accuracy.png targeted fgsm accuracy.png

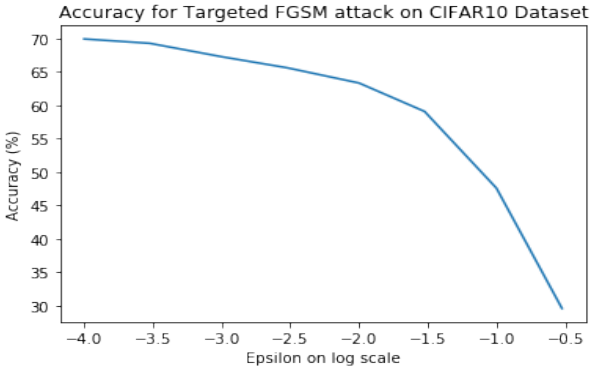

Figure 6: Accuracy versus epsilon in untargetted attack FGSM on CIFAR-10

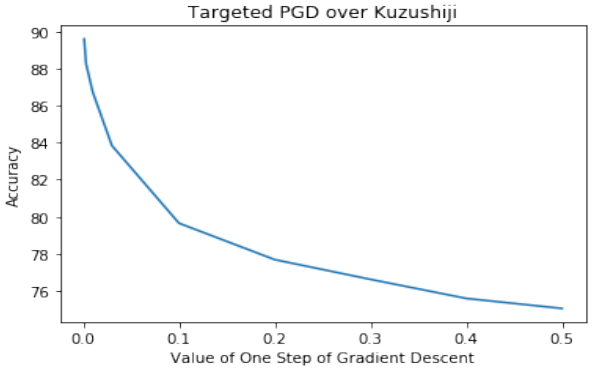

Figure 7: Accuracy versus epsilon for PGD attack

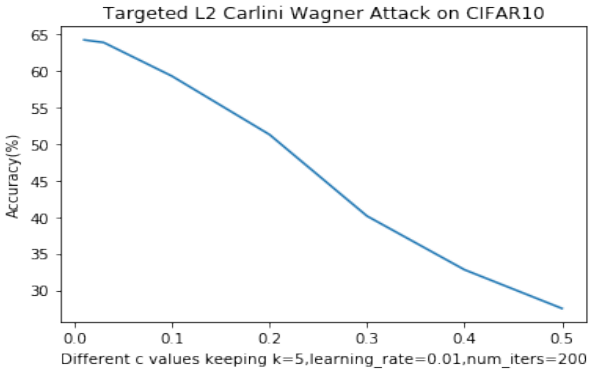

Figure 8: Targeted L2 CW attack on CIFAR10

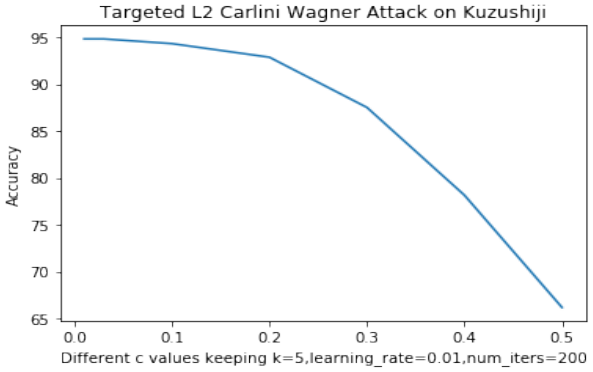

Figure 9: Targeted L2 CW attack on Kuzushiji with varying C

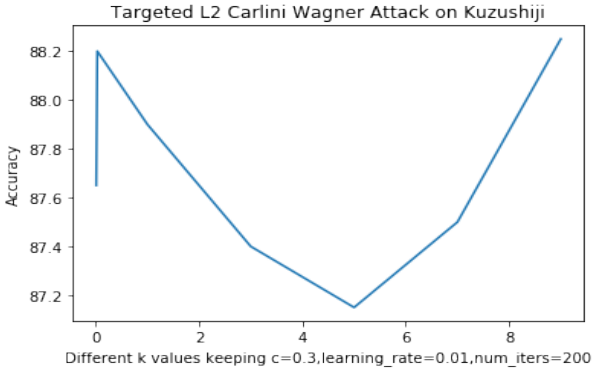

Figure 10: Targeted L2 CW attack on Kuzushiji with varying K

As a strange observation, it was observed that while varying the K (the confidence in CW attack), the accuracy showed a strange behavior of not decreasing continuously to low values. Also, the variation was insignificatn (2%). Thus, we can say that the CW attack didn't depend much on the K values on which we did our experimentation.

6 Protection against the Dark Arts

6.1 Training on Adverserial Images

At first, we tried training on the Adverserial Images to make the classifier robust against the adverserial examples.Training the classifier on the adverserial example acts as a regularizer as well as a data augmentation.

Datasets	Accuracy after attack	Accuracy after training on Adv. Example
Kuzushiji-MNIST	94.18 %	92.15%
CIFAR-10	79.88	77.67%

Table 2: Improved Accuracy on adverserial attacks after training on adverserial examples

Thus we saw that training on the adverserial examples can give appealing results at the test time. However, there can be a lot more adverserial example and even some of them lying on the data manifold. Thus, as a generalization, this might not be robust against the examples on the data manifold.

6.2 Generative Models

We also used the Generative Models to get rid of the adverserial examples also for the data augmentation. The Generative model we trained was VAE and CNN-VAE.

6.2.1 Generation of Images

With the help of VAE, we generated the following images. These images were close to the image in the data distribution and had low reconstruction errors.

It was seen that when we passed the adverserial examples from the encoder the latent space representation of the adverserial example was far from the original image. Thus, the reconstructed image was far from what was in the adverserial example.

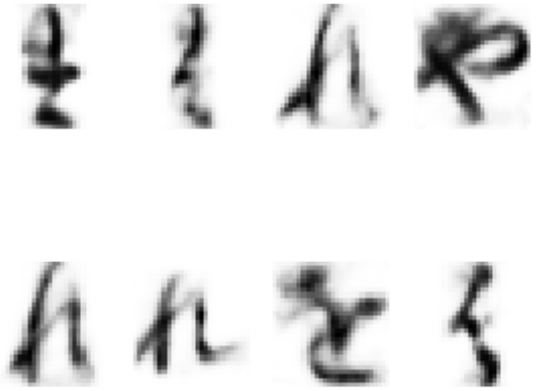

Figure 11: VAE Generated Images

Figure 12: VAE Generated Images

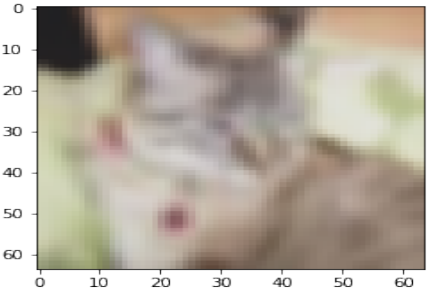

Figure 13: VAE Generated Images

This problem can be solved by imposing the Lipschitz condition while training the classifier so that in the latent space the adverserial image will be close to the true image. Hence, we can get a better reconstruction of the image from the decoder.

Attack	Accuracy after passing through VAE	Accuracy on reconstructed true input
CW attack	27.18%	82.15%
FGSM	33.11%	82.15%
PGD	30.12%	82.15%

Table 3: Results using latent variable of adverserial examples to reconstruct on KMNIST

Inspired from Defense GAN, where we solve an optimization problem on the z space[5], we tried to do 1 Nearest Neighbour for 10000 images generated by the Decoder of the VAE. The nearest images after the KNN is the closest image in the data manifold to the adverserial example.
The KNN is much faster as compared to solving an optimization problem. Thus, we think that on the execution time, if we take less noise instances then this can be done in the real time.

Attack	Accuracy after attack	Accuracy after KNN on VAE output
CW attack	25.18%	48.15%
FGSM	34.11%	59.75%
PGD	32.12%	54.15%

Table 4: Results on KMNIST after KNN on VAE outputs

Attack	Accuracy after attack	Accuracy after KNN on VAE output
CW attack	29.18%	41.15%
FGSM	36.11%	52.75%
PGD	34.12%	46.15%

Table 5: Results on CIFAR-10 after KNN on VAE outputs

7 Conclusions

- A wide variety of attacks like FGSM,PGD and Carlini Wagner were explored in this assignment.
- The graphical results of attacks along with their respective parameters were presented above.
- We observed that Carlini Wagner attack reduced our accuracy on both datasets– Kuzushiji and CIFAR10, to a greater extent than when compared to other attacks. This is in agreement with the results of [1], where CW attack was shown to be much better than others.
- Variational Autoencoder was used as a generative model in this assignment due to its well-behaved training procedure.The model produced realistic-loos king images,

which can be seen in the figures in the previous sections.

- Some methods of defense against these attacks were extensively worked upon this assignment. Substantial improvements were observed in the experiments.
- Our model did not use gradient methods as is done in Defese-GAN. Instead, our approach incorporates k-Nearest Neighbours which is **computationally less expensive and improves the accuracy substantially**.

References

[1] Towards Evaluating the robustness of Neural Networks
`https://arxiv.org/abs/1608.04644`

[2] Deep Learning for Classical Japanese Literature
`https://arxiv.org/abs/1812.01718`

[3] Towards Evaluating the Robustness of Neural Networks
`https://arxiv.org/abs/1608.04644`

[4] Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models
`https://arxiv.org/abs/1805.06605`

[5] Auto-Encoding Variational Bayes
`https://arxiv.org/abs/1312.6114`

[6] CIFAR-10 Dataset
`http://academictorrents.com/details/463ba7ec7f37ed414c12fbb71ebf6431eada2d7a`