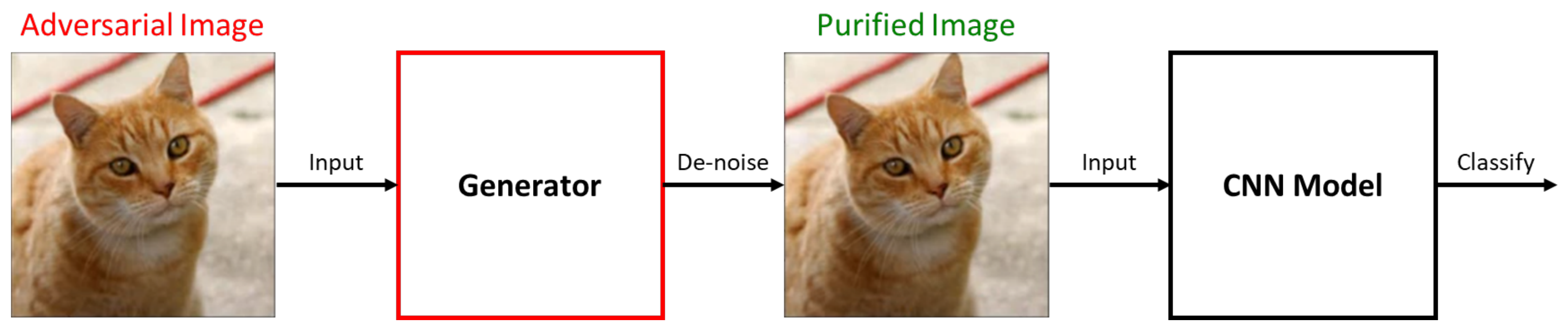
Game of Deception and Truth

Members: Shih-Hsuan Lin, Aaron Kao, Johnson Chang and Irene Chen

Mentor: Jerry Liao

Introduction

- Neural networks are susceptible to adversarial attacks, where adding imperceptible perturbations to the input can mislead trained neural networks to predict incorrect classes.
- APE-GAN and Diffusion Model (U-net) can effectively defend adversarial attacks by eliminating perturbations in the input image.



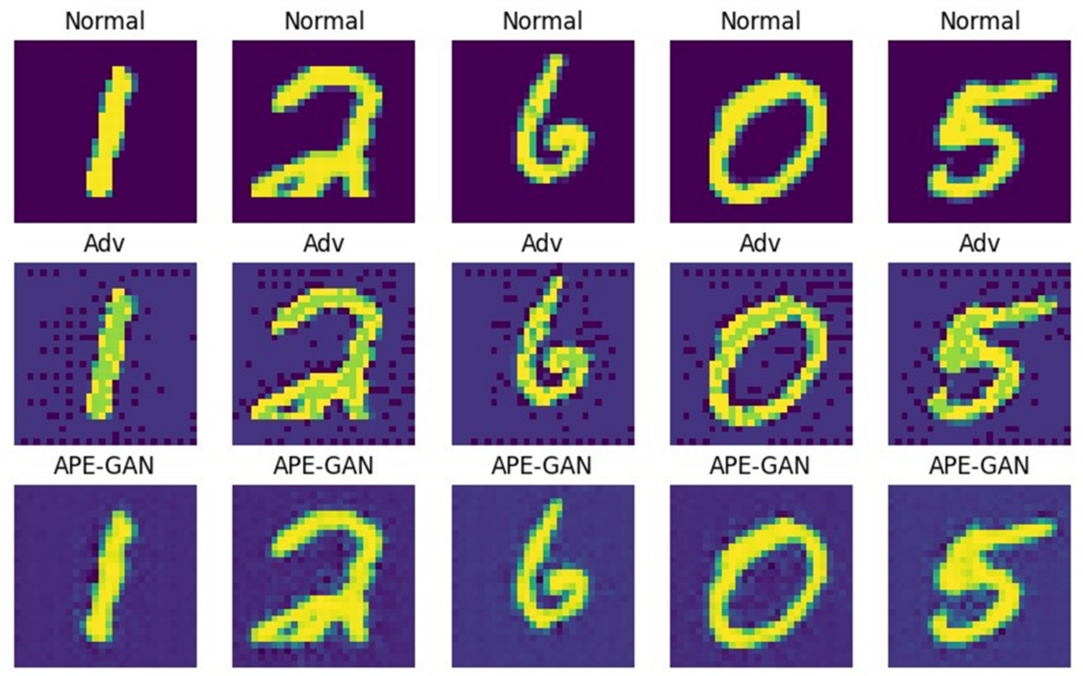
Adversarial Image → Input → Generator → De-noise → Purified Image → Input → CNN Model → Classify

Dataset

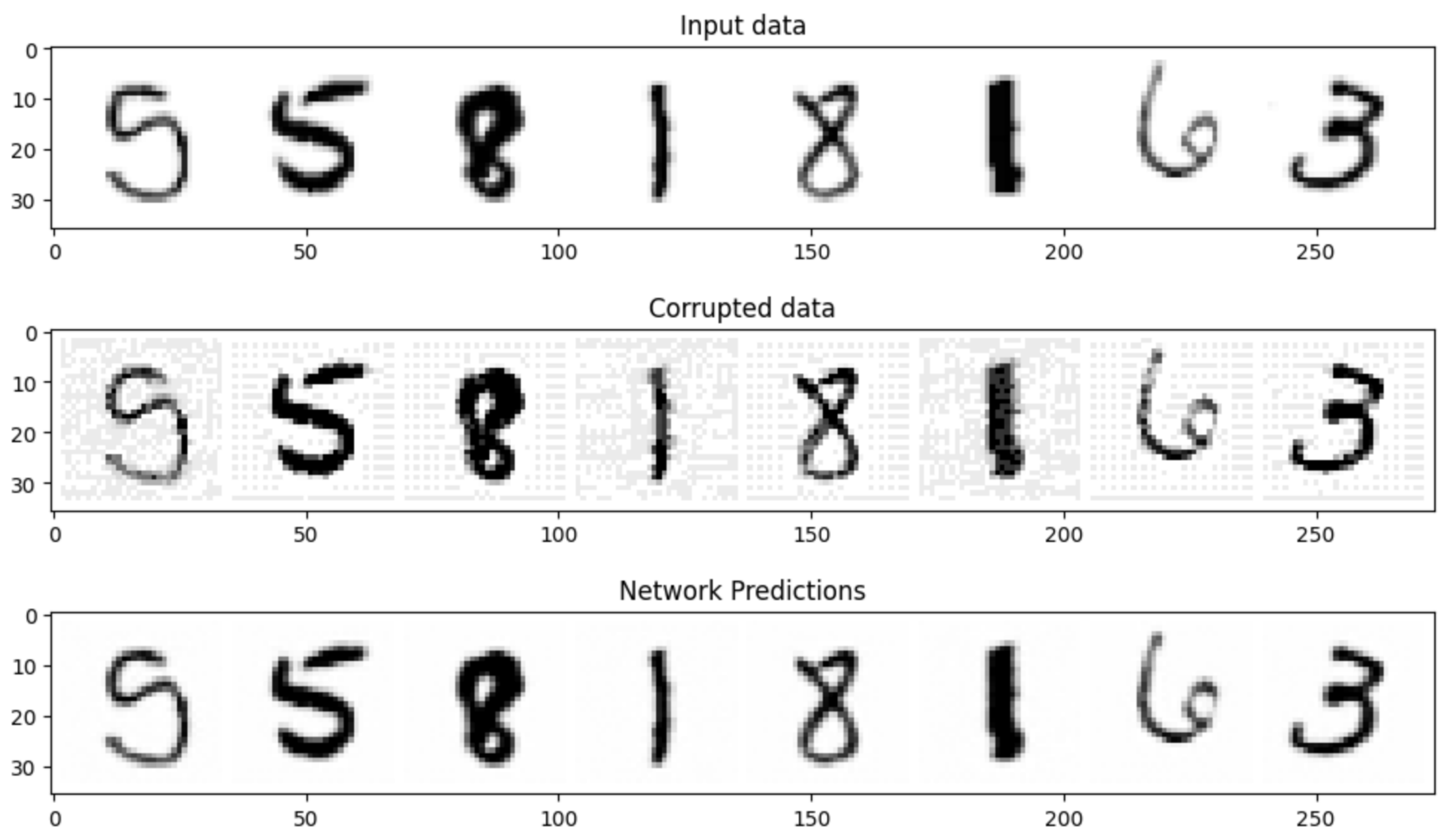
- MNIST database: A large database of handwritten digits that is commonly used for training various image processing systems. The database is also widely used for training and testing in the field of machine learning.

Methodology and Results

- Fast Gradient Sign Method (FGSM) is used to generate adversarial images.
- An effective framework based Generative Adversarial Nets named APE-GAN is implemented to defense against the adversarial examples.



- Diffusion model uses the forward and reverse processes of diffusion models to purify adversarial images. Specifically, given a pre-trained diffusion model, method consists of adding noise to adversarial examples and solving the reverse stochastic differential equation to recover clean images.



Input data

Corrupted data

Network Predictions

Discussion and Conclusion

- Use FGSM & Deepfool model to generate Adversarial examples (Attack)
- Use U-net (reverse process of diffusion model) and APE-GAN to de-noise (Defense)

Future Work

- Try FGSM & Deepfool to attack on different datasets, and find the purification result of different dataset
- Find other models that can defend attack more effectively

References

[1] Shen, Shiwei, et al. "Ape-gan: Adversarial perturbation elimination with gan." arXiv preprint arXiv:1707.05474 (2017).

[2] Jonathan Ho, Ajay Jain, Pieter Abbeel Denoising Diffusion Probabilistic Models arXiv preprint arXiv:2006.11239(2020).

stanCode
標準程式教育機構