OntarioTech
Engineering
& Applied Science

- Weimin Zhao
  - weimin.zhao@ontariotechu.net
- Sanaa Alwidian
  - sanaa.alwidian@ontariotechu.ca
- Qusay H. Mahmoud
  - qusay.mahmoud@ontariotechu.ca

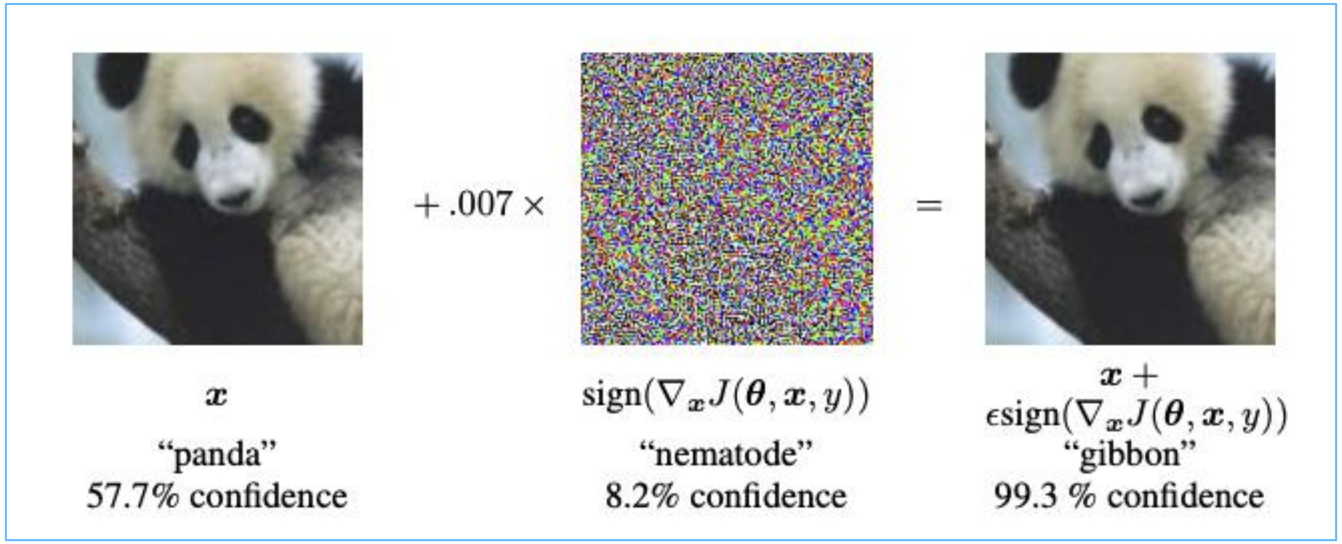
Evaluation of GAN Architectures for Adversarial Robustness of Convolution Classifier

Outline

- Problem Statement
- Motivation
- Related Works
- Research Questions
- Methodology
- Experiment Results
- Conclusion

Problem Statement

- Adversarial samples



Motivation

- Exploring an expansion for adversarial training method.
- Exploring the usage of generative model: GAN for adversarial sample defense.
- Providing more insights into adversarial robustness and different training strategy on improving the adversarial robustness.
- Improving the classifier robustness against PGD attack.

Research Questions

- **Evaluation of generalization:**
  - Can GAN generalize on adversarial samples?
- **Estimate the worst-case adversarial perturbation:**
  - How to formulation GAN to estimate the adversarial sample?
- **Avoid overfit and negative effects:**
  - How to setup the hyperparameters to mitigate the limitations of GAN?

Related Works

Related Solutions	Benefits	Limitations
Traditional adversarial training	• State-of-the-art performance • Baseline • Reliable	• Required a lot of training time • Still limited in performance
GANs	• No need to implement any attack • Less computational complex	• Performance is less reliable and related to the implementation of the models • Limited in performance

Our proposal:

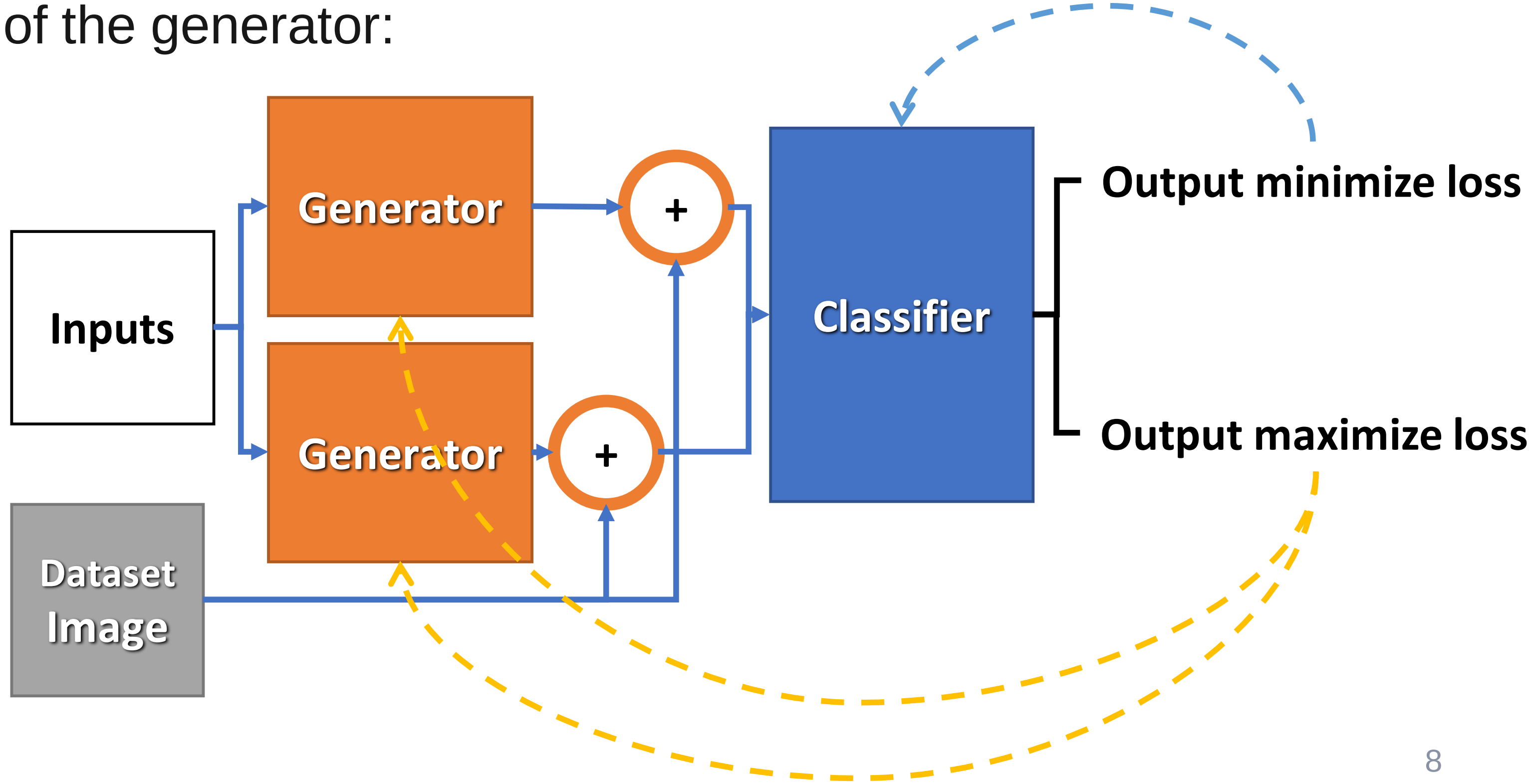
Method	Suggest Modifications and Adjustments	Improvements and Benefits
Improving the GAN for adversarial training	• Reformulating the generator with new gradient input vectors • Stable dual generators architecture • Re-evaluating the impact of training epochs	• Reliable adversarial perturbation estimation • Improving the training performance • Keeping the other benefits of GAN

Proposed Solutions

- Build a GAN for data augmentation and adversarial training.
- Compare different formulations and other settings.
- Visualization

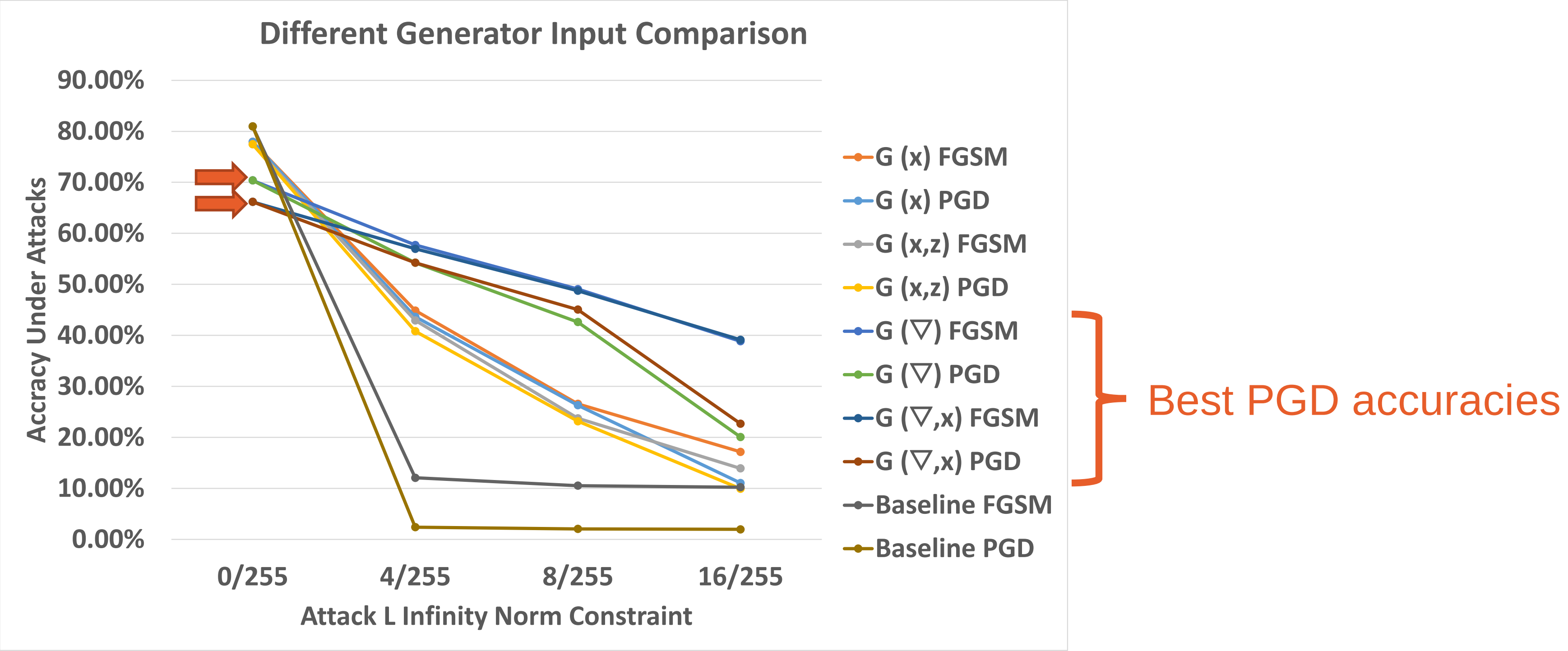
Methodology

- Overall formulation:
  - $min_D max_G \sum L\,(D(X_i), y_i) + \sum L(D(x_i + \epsilon G(\boldsymbol{I})), y_i),$
- 4 Formulations of the input *I* of the generator:
  - N = G1 (x)
  - N = G2 (x, z)
  - N = G3 (sign(∇))
  - N = G4 (sign(∇), x)
- VGG classifier
- CIFAR 10 dataset



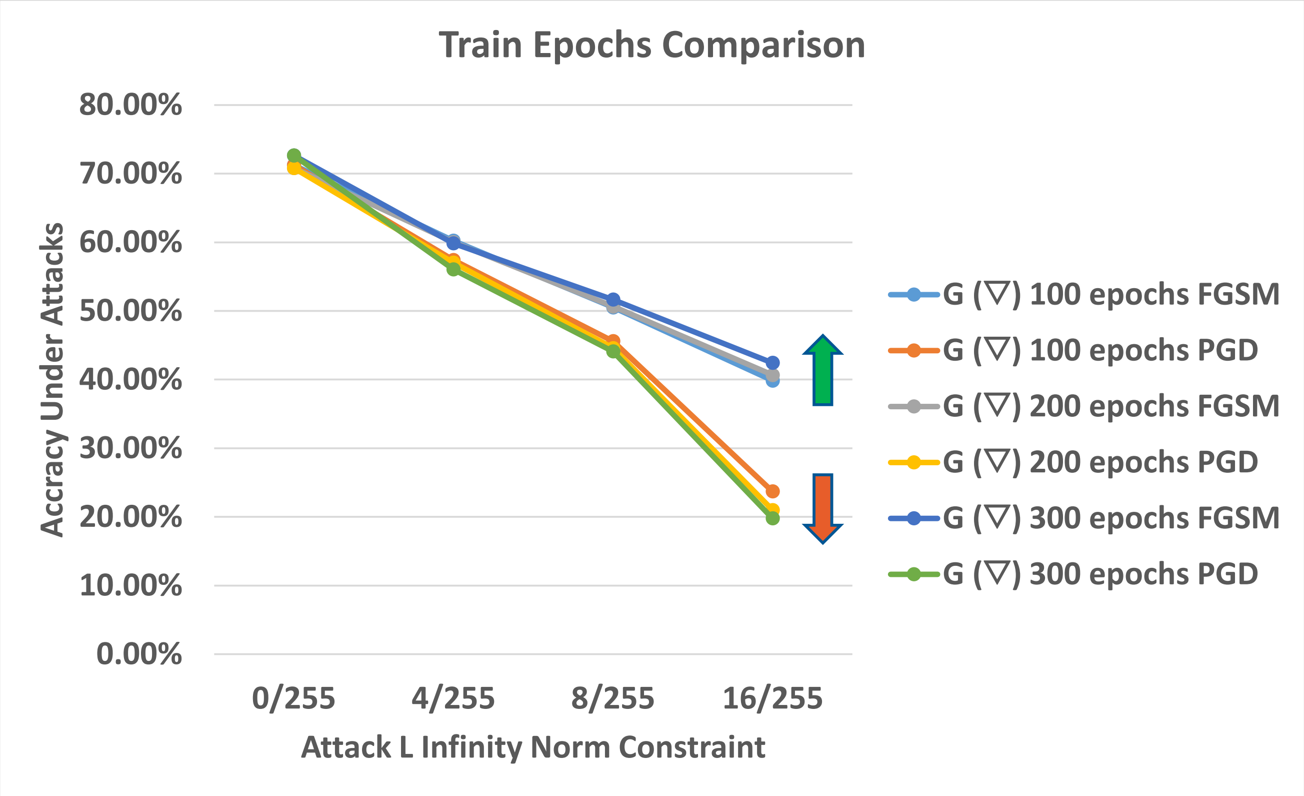
Experiments and Results

- G(∇) and G(∇,x) have best PGD accuracies
- G(∇) has a better clean accuracy than G(∇,x)



Experiments and Results II

- FGSM accuracy increasing
- PGD accuracy decreasing
- Overfit after 100 epochs

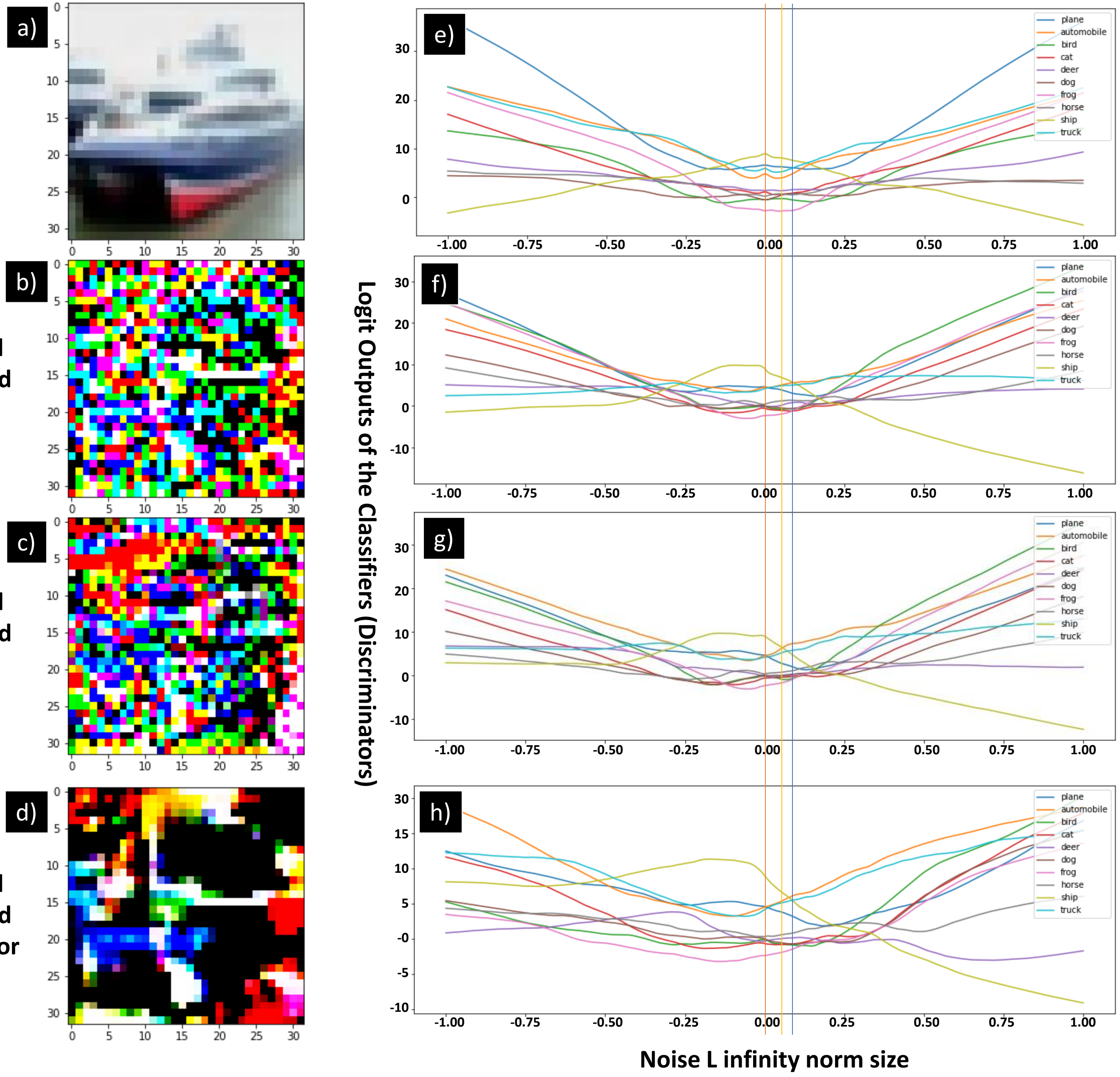


Original image with random noise perturbation

The adversarial noise generated by FGSM

The adversarial noise generated by PGD

The adversarial noise generated by our generator



Conclusion

- GANs can improve adversarial robustness by a good margin.
- More gradient is better to estimate adversarial sample.
- Training epochs matters
- GANs can provide a level of augmentation stronger than single-step adversarial training and weaker than multi-step adversarial training.

Thank you!

Questions?