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

\[ x \]

“panda”
57.7% confidence

\[ + .007 \times \]

\[ \text{sign}(\nabla_x J(\theta, x, y)) \]

“nematode”
8.2% confidence

\[ = \]

\[ x + c \text{sign}(\nabla_x J(\theta, x, y)) \]

“gibbon”
99.3% confidence
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

<table>
<thead>
<tr>
<th>Related Solutions</th>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| 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:

<table>
<thead>
<tr>
<th>Method</th>
<th>Suggest Modifications and Adjustments</th>
<th>Improvements and Benefits</th>
</tr>
</thead>
</table>
| 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:
  • \( \min_{\mathcal{G}} \max_{\mathcal{D}} \sum L(\mathcal{D}(X_i), y_i) + \sum L(\mathcal{D}(x_i + \epsilon \mathcal{G}(I)), y_i) \),

• 4 Formulations of the input \( I \) of the generator:
  • \( N = G_1(x) \)
  • \( N = G_2(x, z) \)
  • \( N = G_3(\text{sign}(\nabla)) \)
  • \( N = G_4(\text{sign}(\nabla), x) \)

• VGG classifier
• CIFAR 10 dataset
Experiments and Results

- $G(\nabla)$ and $G(\nabla, x)$ have best PGD accuracies
- $G(\nabla)$ has a better clean accuracy than $G(\nabla, x)$
Experiments and Results II

- FGSM accuracy increasing
- PGD accuracy decreasing
- Overfit after 100 epochs
a) Original image with random noise perturbation

b) The adversarial noise generated by FGSM

c) The adversarial noise generated by PGD

d) The adversarial noise generated by our generator

e) Logit Outputs of the Classifiers (Discriminators) for the original image

f) Logit Outputs of the Classifiers (Discriminators) for the noise generated by FGSM

g) Logit Outputs of the Classifiers (Discriminators) for the noise generated by PGD

h) Logit Outputs of the Classifiers (Discriminators) for the noise generated by our generator

Noise L infinity norm size

Logit Outputs of the Classifiers (Discriminators)
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?