Learning Models Robust To Adversarial Attacks

Aishwarya Gupta
u12066423
aish322@gmail.com

Outline

- Introduction
- Generating Adversarial Examples
    - FGSM
    - BIM
    - ILCM
- Adversarial Robustness
- Results
- Conclusion

Introduction

Can Neural Networks be fooled?

Can Neural Networks be fooled?

“Panda”
57.7% confidence

“Gibbon”
99.3% confidence

Can Neural Networks be fooled?

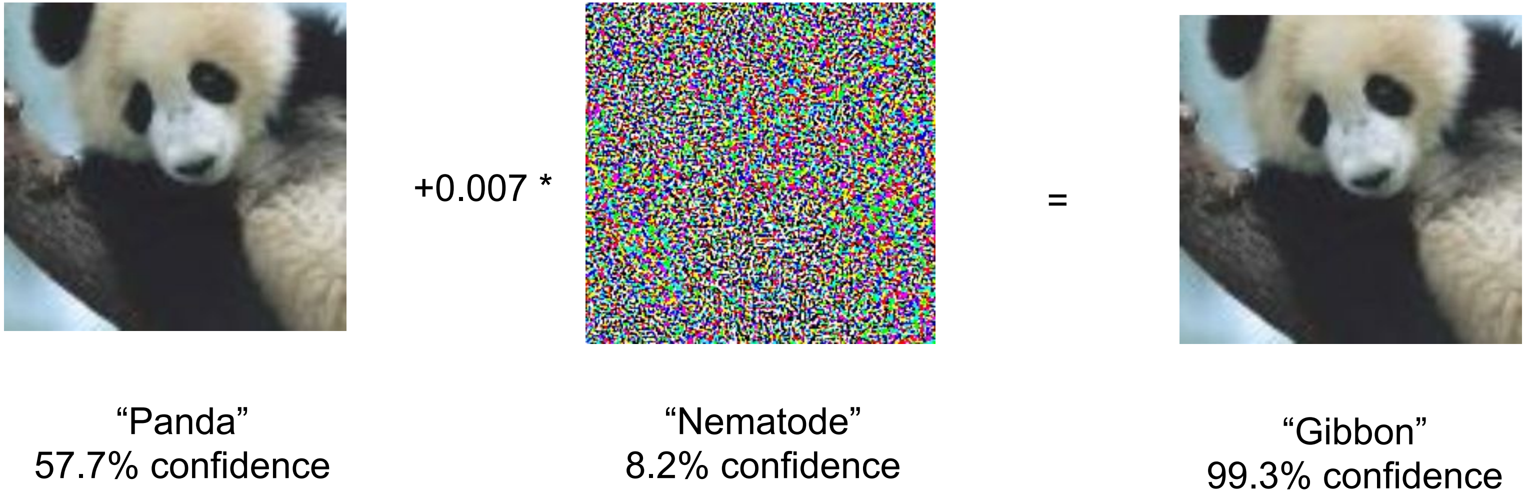


Figure- Example of an adversarial image for GoogLeNet
trained on ImageNet[1].

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Notations

- F : Learned model/classifier
- θ : Model/classifier parameters
- (x, y) : The natural image and its true label
- x’ : The adversarial image
- $L(y_p, y)$: The loss function e.g. cross-entropy loss
- ϵ : The allowed perturbation in the image x

Formal Definition:

For an image x, x’ is termed as its adversarial image if:

- $F(x) \neq F(x')$
- $d(x, x') \leq \epsilon$

where $d(x, x') = ||x, x'||_p$ for $p = \{0, 2, \infty\}$

Formal Definition:

For an image x, x’ is termed as its adversarial image if:

- $F(x) \neq F(x')$
- $d(x, x') \leq \epsilon$

where $d(x, x') = ||x, x'||_p$ for $p = \{0, 2, \infty\}$

Generating an adversarial image

$\max_{\delta \in \Delta} L(F(x + \delta), y)$

where $\Delta = \{ \delta : || \delta ||_\infty \leq \epsilon \}$

Toy Example: Binary Linear Classifier[4]

For the dataset (x, y) such that x ⊆ R^d and⫗y = {-1, 1}, let F be the model defined as :

- F(x) = $w^T x + b$
- p(y = +1 | x) = 1/(1 + exp(-F(x)))
- p(y = -1 | x) = 1/(1 + exp(F(x)))

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Toy Example: Binary Linear Classifier

For the dataset (x, y) such that $x \subseteq R^d$ and⫔$y = \{-1, 1\}$, let F be the model defined as :

- $F(x) = w^T x + b$
- $p(y = +1 \mid x) = 1/(1 + \exp(-F(x)))$
- $p(y = -1 \mid x) = 1/(1 + \exp(F(x)))$

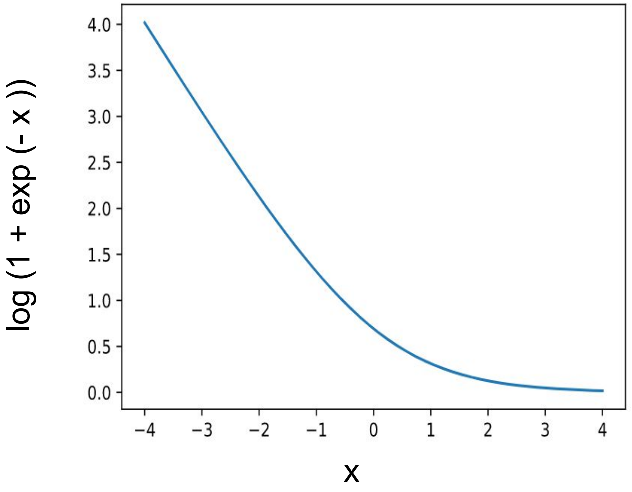
And L be the negative log likelihood:

- $L(F(x), y) = \log(1 + \exp(-yF(x)))$

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Adversarial Examples for Binary Linear Classifier[4]

$$L(F(x), y) = \log (1 + \exp (- y F(x)))$$



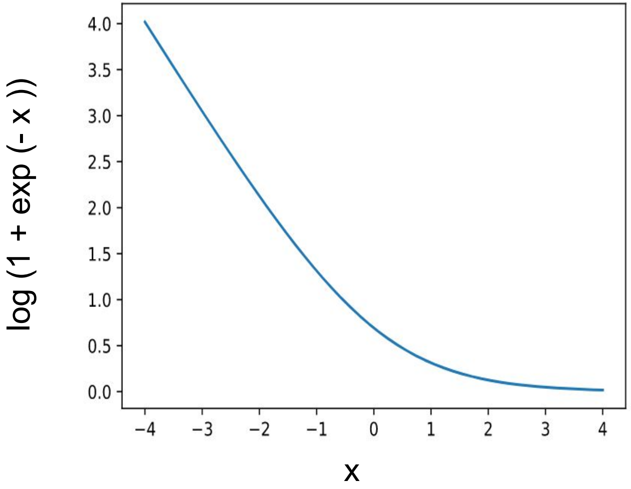
[4] https://adversarial-ml-tutorial.org/linear_models.

Adversarial Examples for Binary Linear Classifier[4]

$$L(F(x), y) = \log(1 + \exp(-y\,F(x)))$$

$$\max_{\delta} L(F(x + \delta), y)$$

$$= \max_{\delta} \log(1 + \exp(-y\,F(x + \delta)))$$



[4] https://adversarial-ml-tutorial.org/linear_models.

Adversarial Examples for Binary Linear Classifier[4]

$L(F(x), y) = \log(1 + \exp(-y F(x)))$

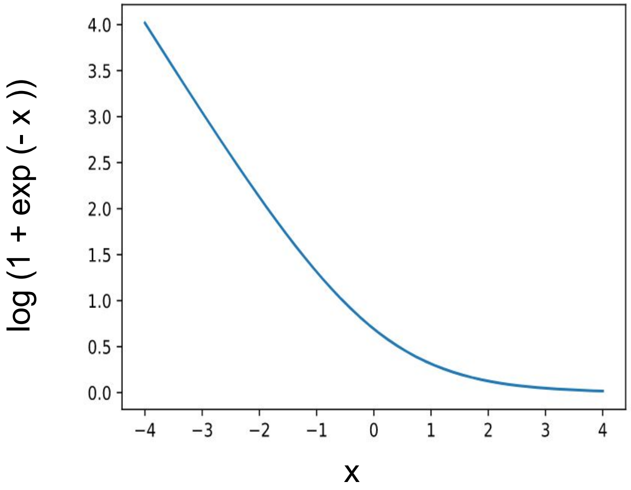
$\max_{\delta} L(F(x + \delta), y)$

$= \max_{\delta} \log(1 + \exp(-y F(x + \delta)))$

$= \min_{\delta} (y F(x + \delta))$

$= \min_{\delta} (y(w^T x + b) + y w^T \delta)$

$= \min_{\delta} (y w^T \delta)$



[4] https://adversarial-ml-tutorial.org/linear_models.

Adversarial Examples for Binary Linear Classifier[4]

$L(F(x), y) = \log(1 + \exp(-y F(x)))$

$$\max_\delta L(F(x + \delta), y)$$

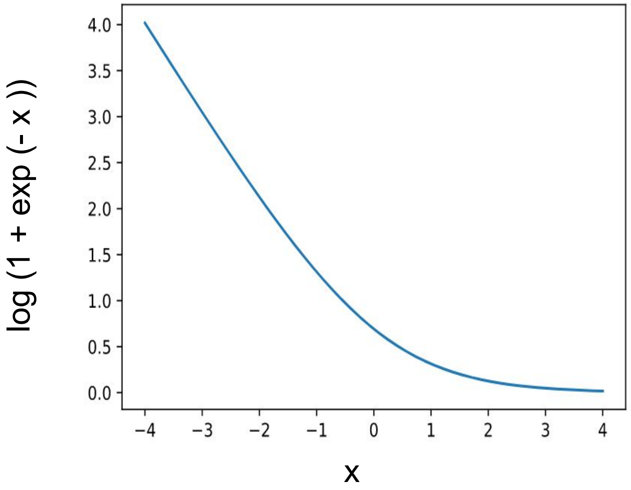
$$= \max_\delta \log(1 + \exp(-y F(x + \delta)))$$

$$= \min_\delta (y F(x + \delta))$$

$$= \min_\delta (y(w^T x + b) + y w^T \delta)$$

$$= \min_\delta (y w^T \delta)$$

For L_∞ norm, $\delta^* = -y \epsilon \operatorname{sign}(w)$



[4] https://adversarial-ml-tutorial.org/linear_models.

Generating Adversarial Examples

Generating Adversarial Examples

- Fast Gradient Sign Method (FGSM)
- Basic Iterative Method (BIM)
- Iterative Least-likely Class Method (ILCM)

1. Fast Gradient Sign Method (FGSM)[1]

For any model F and natural image x, the adversarial image is computed as :

$$x' = x + \epsilon \, \mathrm{sign}(\nabla_x L(F(x), y))$$

- It is an L_∞ attack as $||x'-x||_\infty \le \epsilon$

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

1. Fast Gradient Sign Method (FGSM)

For any model F and natural image x, the adversarial image is computed as :

$$x' = x + \epsilon \text{ sign } (\nabla_x L(F(x), y))$$

- It is an L_∞ attack as $||x'-x||_\infty \leq \epsilon$

2. Basic Iterative Method (BIM)[2]

Repeat FGSM using small step size for k iterations

- $x'_0 = x$
- $x'_{i+1} = \text{clip}\{ (x'_i + \alpha \text{ sign } (\nabla_x L(F(x'_i), y))), x + \epsilon , x - \epsilon \}$

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

3. Iterative Least-Likely Class Method (ILCM)[2]

- The least likely class y_{LL} is given as :

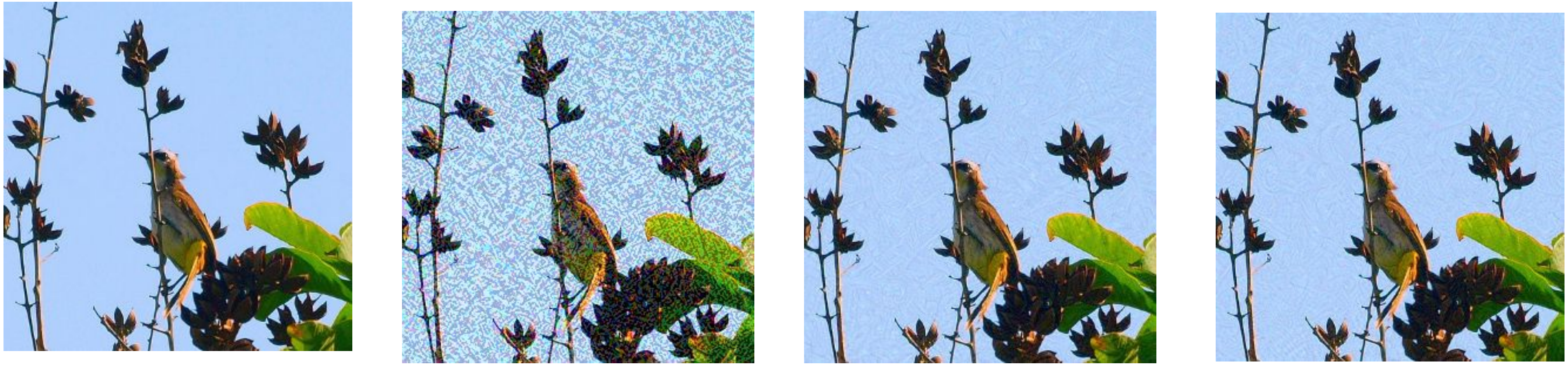
  $y_{LL} = \arg\min_y p(y \mid x)$

- For y_{LL} to be the target label of the adversarial image,
  - $y_{LL} = \arg\max_y p(y \mid x')$
  - $L(F(x'), y_{LL})$ should be minimised

- Adversarial image x' s.t. $||x'-x||_\infty \leq \epsilon$ is computed as :
  - $x'_0 = x$
  - $x'_{i+1} = \text{clip}\{ (x'_i - \alpha \, \text{sign}(\nabla_x L(F(x'_i), y_{LL}))), x + \epsilon , x - \epsilon \}$

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

Examples of Adversarial Images



Clean Image

Fast Gradient Sign Method (L_{∞}= 32)

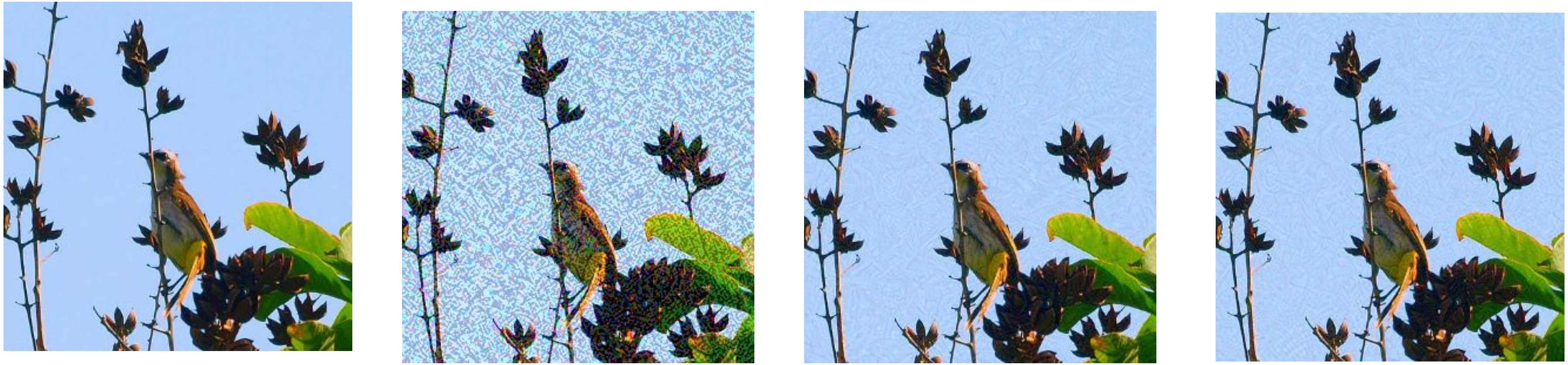
Basic Iterative Method (L_{∞}= 32)

Iterative Least-likely Class Method (L_{∞}= 28)

Figure : Generating adversarial images using different attacks for ϵ = 32 [2].

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

Examples of Adversarial Images



Clean Image

Fast Gradient Sign Method (L_∞= 32)

Basic Iterative Method (L_∞= 32)

Iterative Least-likely Class Method (L_∞= 28)

Figure : Generating adversarial images using different attacks for ϵ = 32 [2].

Iterative methods result in finer perturbations in comparison to the fast method

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

Which attack is better?

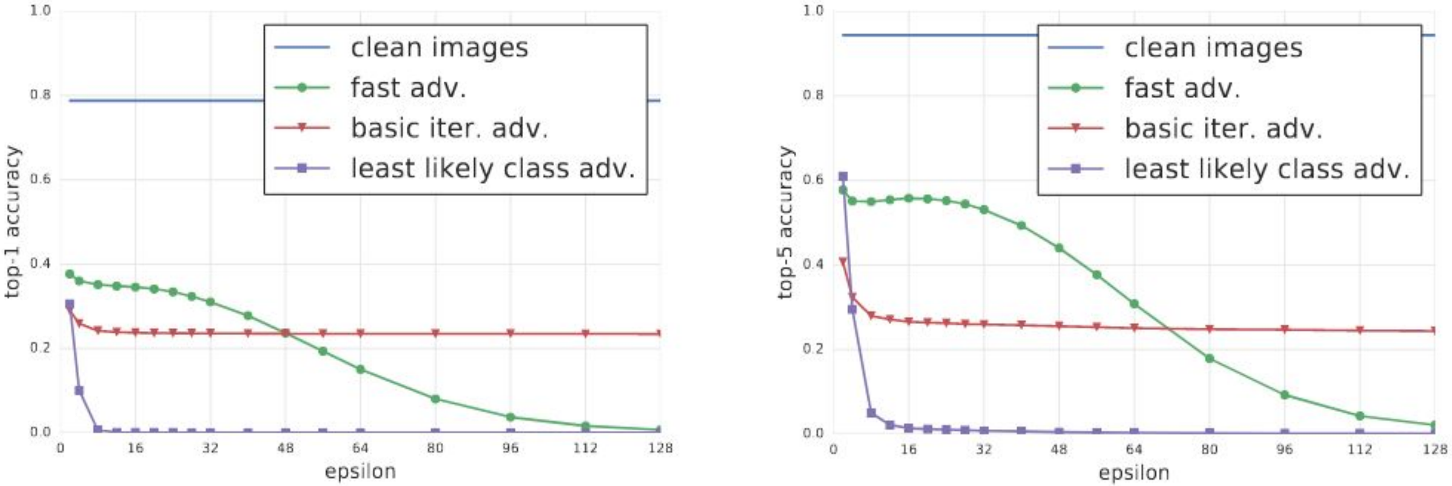


Figure - Drop in accuracy wrt different attacks on Inception v3 network trained on ImageNet dataset [2]

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

Robustness against Adversarial Examples

Adversarial Training as a Robust Optimisation Problem[3]

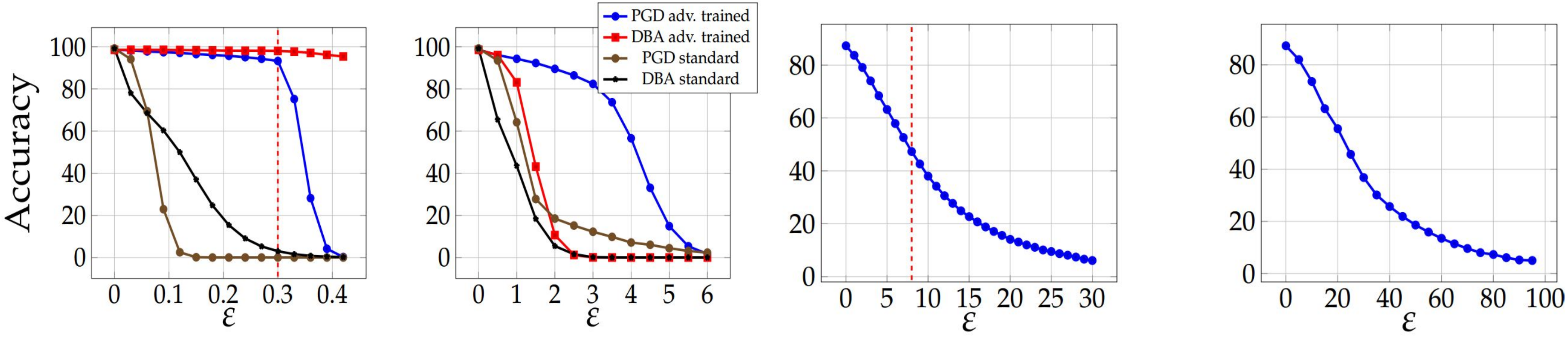
- For a dataset *D* and allowed set of perturbations *S*, a robust model can be trained by minimising the following optimisation :

$$\min_{\theta} \rho(\theta), \text{ where } \rho(\theta) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}} \underset{\substack{\text{Adversarial Images} \\ \text{Loss}}}{\underline{\left[\max_{||\delta||_p \leq \epsilon} L(\theta, x_i + \delta, y_i) \right]}}$$

- Solving the inner-optimisation problem using Basic Iterative Method(BIM) also known as Projected Gradient Descent (PGD)

[3] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Results of Adversarial Training



MNIST, L_∞ - norm

MNIST, L_2 - norm

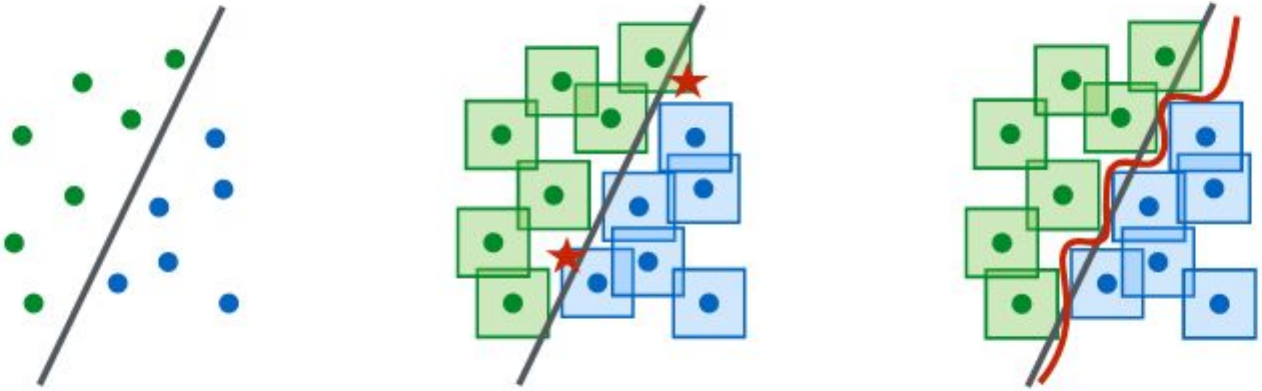
CIFAR10, L_∞ - norm

CIFAR10, L_2 - norm

Figure - Robustness of adversarially trained networks against PGD adversaries of different strength. The models are trained on generated PGD adversarial images using ϵ=0.3 and ϵ=8 for MNIST and CIFAR10 respectively [3].

[3] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Does Increasing Network Capacity Help?



Standard Training

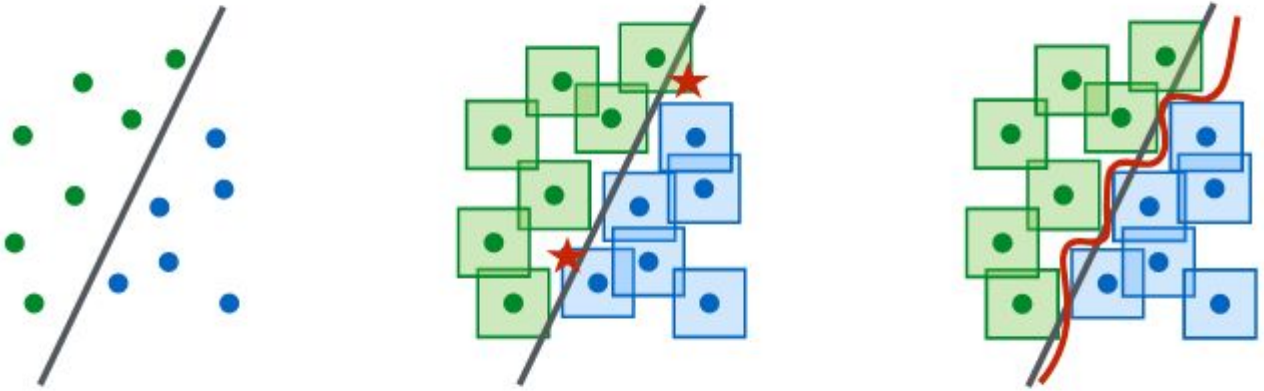
Simple decision boundary unable to separate adversarial examples

Complicated decision boundary needed to separate the L_∞ balls

Figure - A conceptual illustration of standard vs. adversarial decision boundaries[3].

[3] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Does Increasing Network Capacity Help?



Standard Training

Simple decision boundary unable to separate adversarial examples

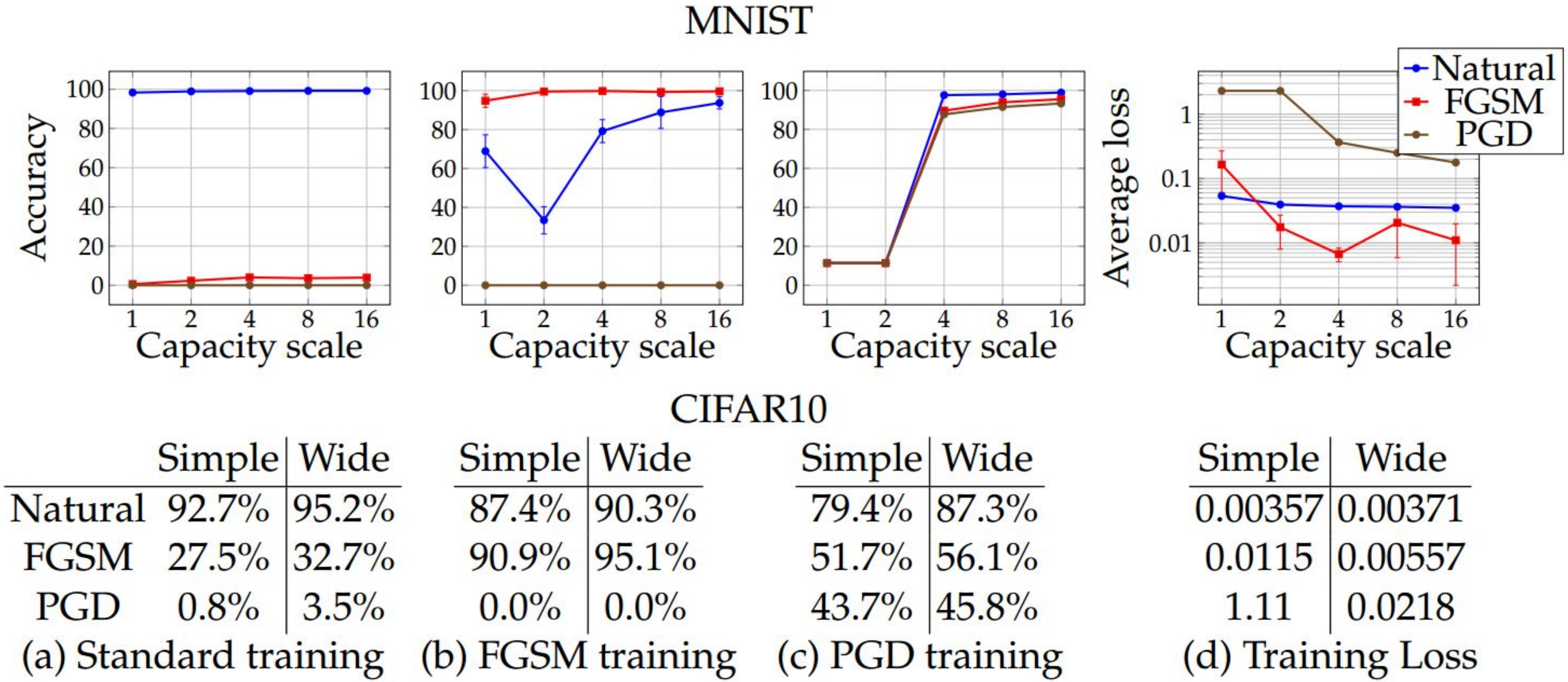
Complicated decision boundary needed to separate the L_∞ balls

Figure - A conceptual illustration of standard vs. adversarial decision boundaries[3].

Increasing network capacity does help in improving the adversarial robustness of the model

[3] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Results of Increasing Network Capacity



	Simple	Wide	Simple	Wide	Simple	Wide	Simple	Wide
Natural	92.7%	95.2%	87.4%	90.3%	79.4%	87.3%	0.00357	0.00371
FGSM	27.5%	32.7%	90.9%	95.1%	51.7%	56.1%	0.0115	0.00557
PGD	0.8%	3.5%	0.0%	0.0%	43.7%	45.8%	1.11	0.0218

(a) Standard training (b) FGSM training (c) PGD training (d) Training Loss

Figure - The adversarial robustness of the model improves with increasing network capacity[3].

[3] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Conclusion

- Adversarial images can be generated easily
- Adversarial training helps in improving the robustness but it starts failing for ϵ greater than training ϵ

References

1. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
2. Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).
3. Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
4. https://adversarial-ml-tutorial.org/linear_models

Thank You