Understanding Catastrophic Overfitting in Adversarial Training

Kang Peilin

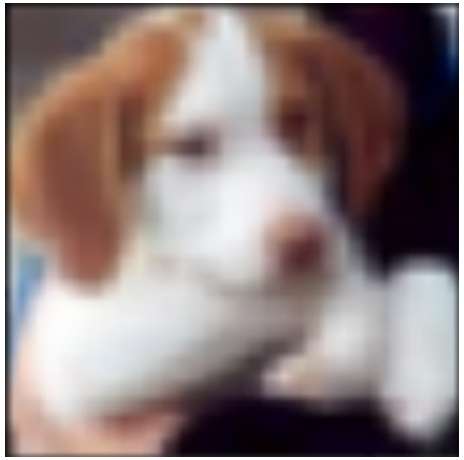
Master Thesis Project Presentation

ETH Supervisor Dr. Seyed-Mohsen Moosavi-Dezfooli
EPFL Supervisor Prof. Martin Jaggi

16th April 2021

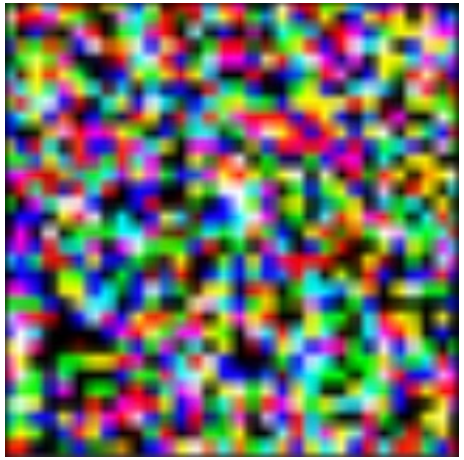
Adversarial Training

- Standard trained model is easy to be attacked



Natural image: x
Label: dog

+



Adversarial perturbation found by PGD attack : δ

=

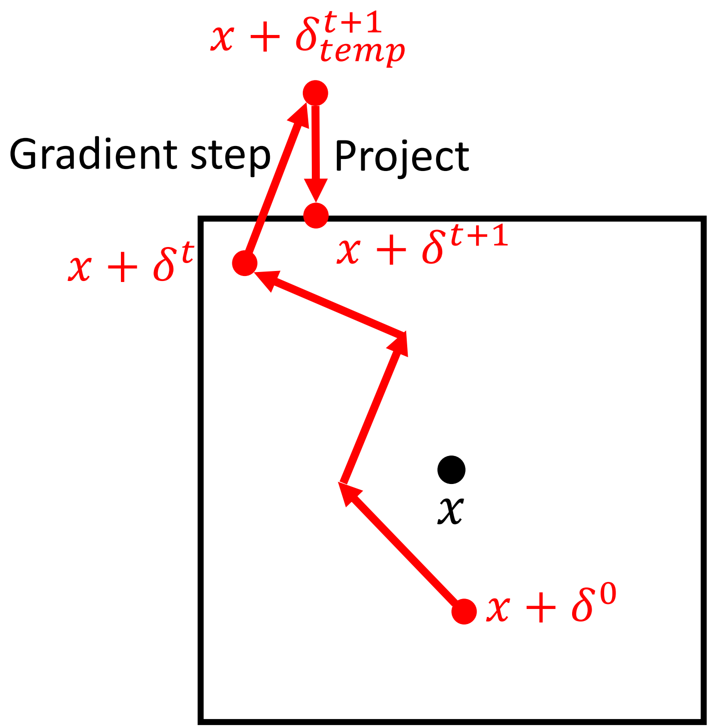
Adversarial example: $x + \delta$
Label: deer

- Adversarial Training
  - Train the model on **adversarial examples** constructed by attack methods.
  - l_∞ threat model: adversary can change each input coordinate x_i by at most $\boldsymbol{\varepsilon}$

Projected Gradient Descent (PGD)

- Select a random start point δ^0
  $\delta^0 \sim \mathcal{U}\left([-\varepsilon, \varepsilon]^d\right)$

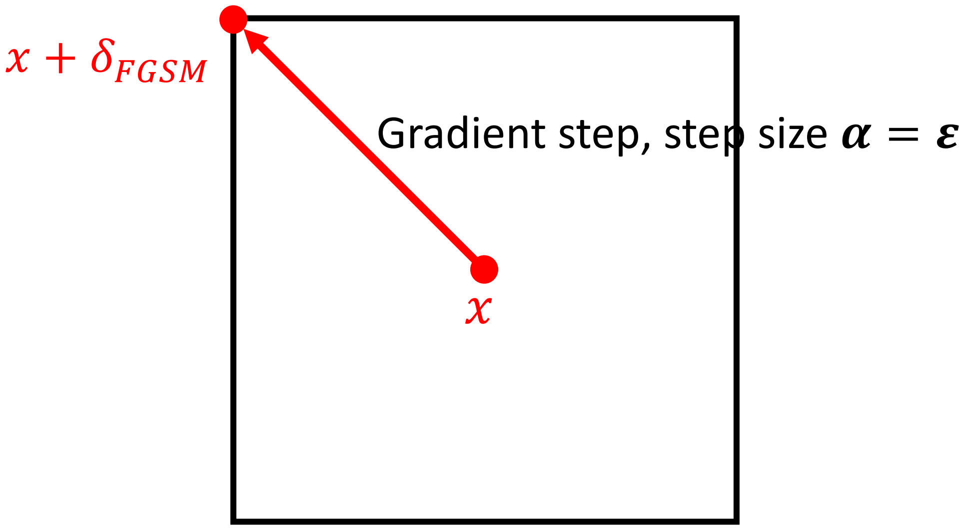
- For t = 1 … T do
  - Take a small gradient step ($\alpha < \varepsilon$)
    $\delta_{temp}^{t+1} = \ \delta^t + \alpha * \mathrm{sign}(\nabla_x l(x + \delta^t, y; \theta))$
  - Project back to the l_∞-ball.
    $\delta^{t+1} = \Pi_{[-\varepsilon,\varepsilon]^d}(\delta_{temp}^{t+1})$



Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. *Towards Deep Learning Models Resistant to Adversarial Attacks*.

Fast Gradient Sign Method (FGSM)

$$\delta_{FGSM} = \boldsymbol{\varepsilon} * \text{sign}(\nabla_x l(x, y; \theta))$$



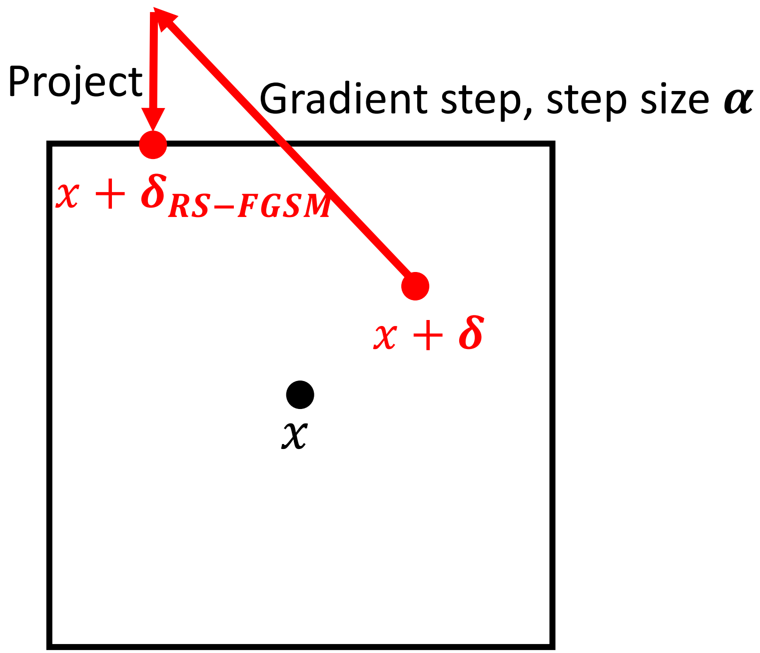
Methods	Pros	Cons
PGD	Can lead to robust model	Huge computational overhead
FGSM	Computationally efficient	Can be broken by stronger attacks, such as PGD

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*.

FGSM with Random Initialization (RS-FGSM)

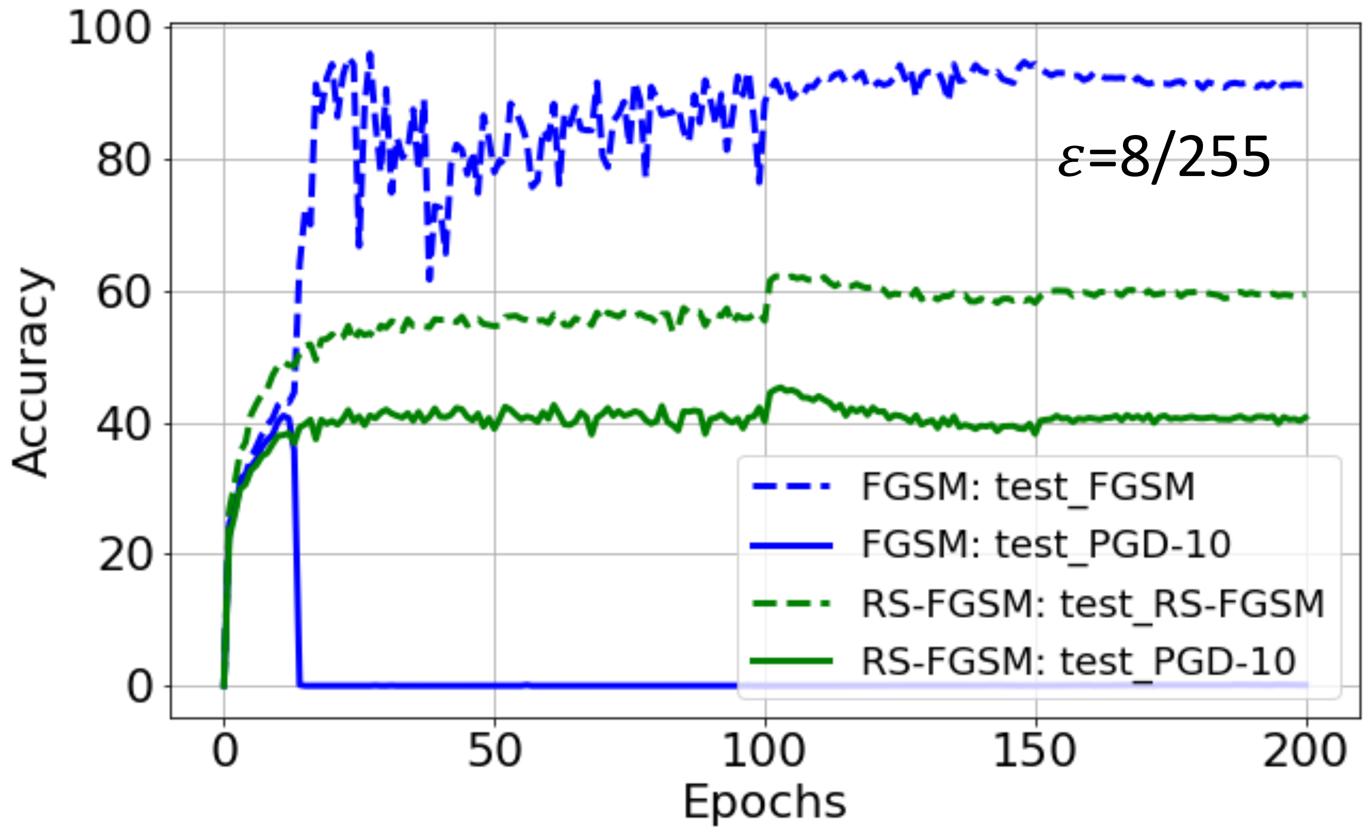
Pros
➢ As effective as PGD-based training ➢ Significantly lower cost

Single-step PGD



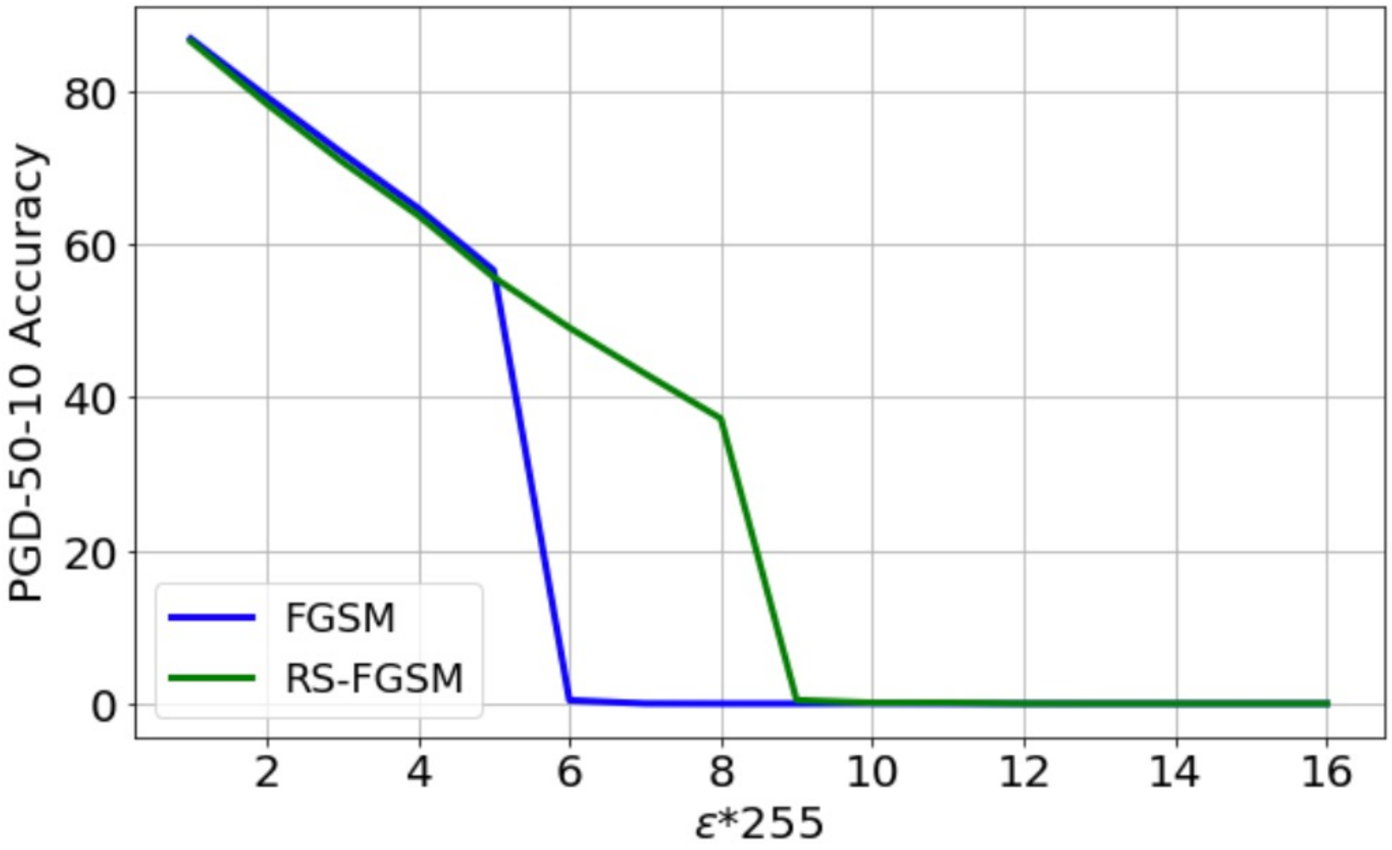
Eric Wong, Leslie Rice, and J. Zico Kolter. *Fast is better than free: Revisiting adversarial training*.

Catastrophic Overfitting (CO)



- ➢ Train with weaker attack, named **method-A**, such as FGSM.
- ➢ Evaluate with another stronger attack, named **method-B**, such as PGD.
- ➢ After a certain epoch, The accuracy gap between A and B increases suddenly

Robustness under Different ε for FGSM and RS-FGSM



PGD-50-10 means 50 iterations and 10 restarts

- RS-FGSM suffers from CO when $\varepsilon \gtrsim \frac{9}{255}$
- RS-FGSM permits us to use higher values of ε compared to FGSM

DF^{∞}-1 Suffers from Catastrophic Overfitting

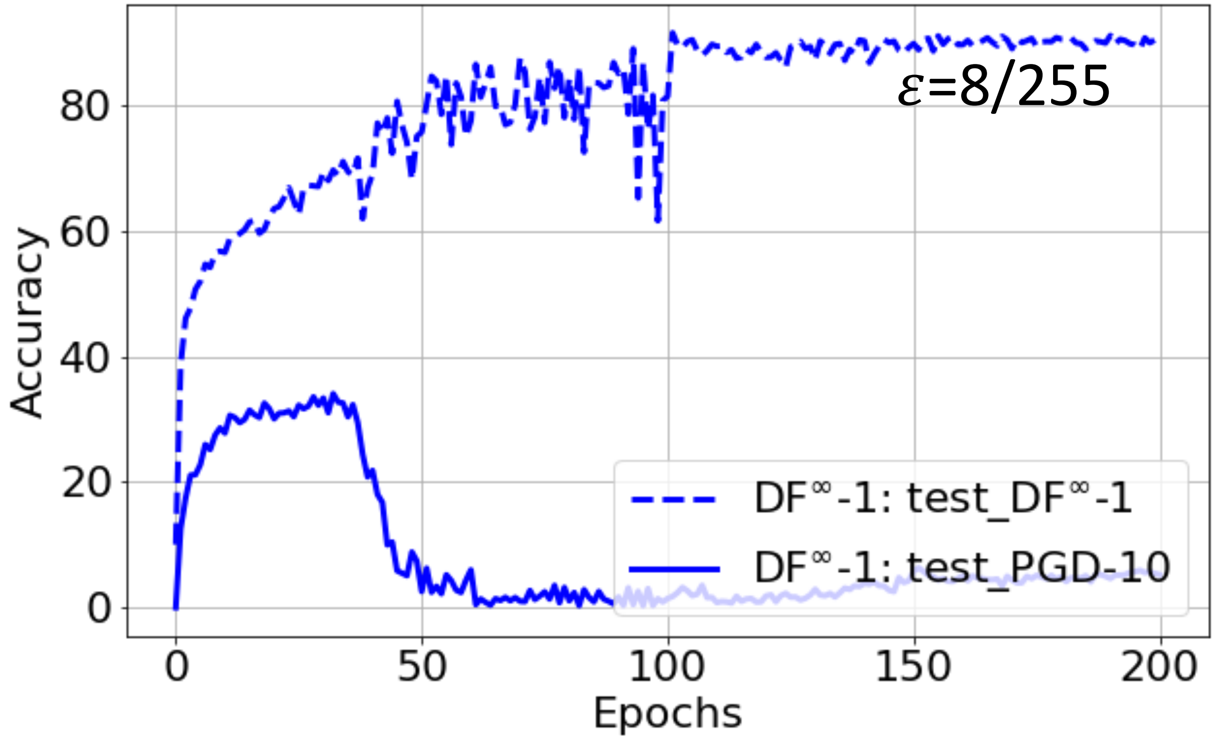
Compare DF^{∞}-1 to FGSM

- Similarity

Computationally efficient, both use one iteration

- Difference

FGSM has the **fixed step size α** for all inputs, DF^{∞}-1 will **adapt the length of perturbation dynamically** for each input



Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. *DeepFool: a simple and accurate method to fool deep neural networks*.

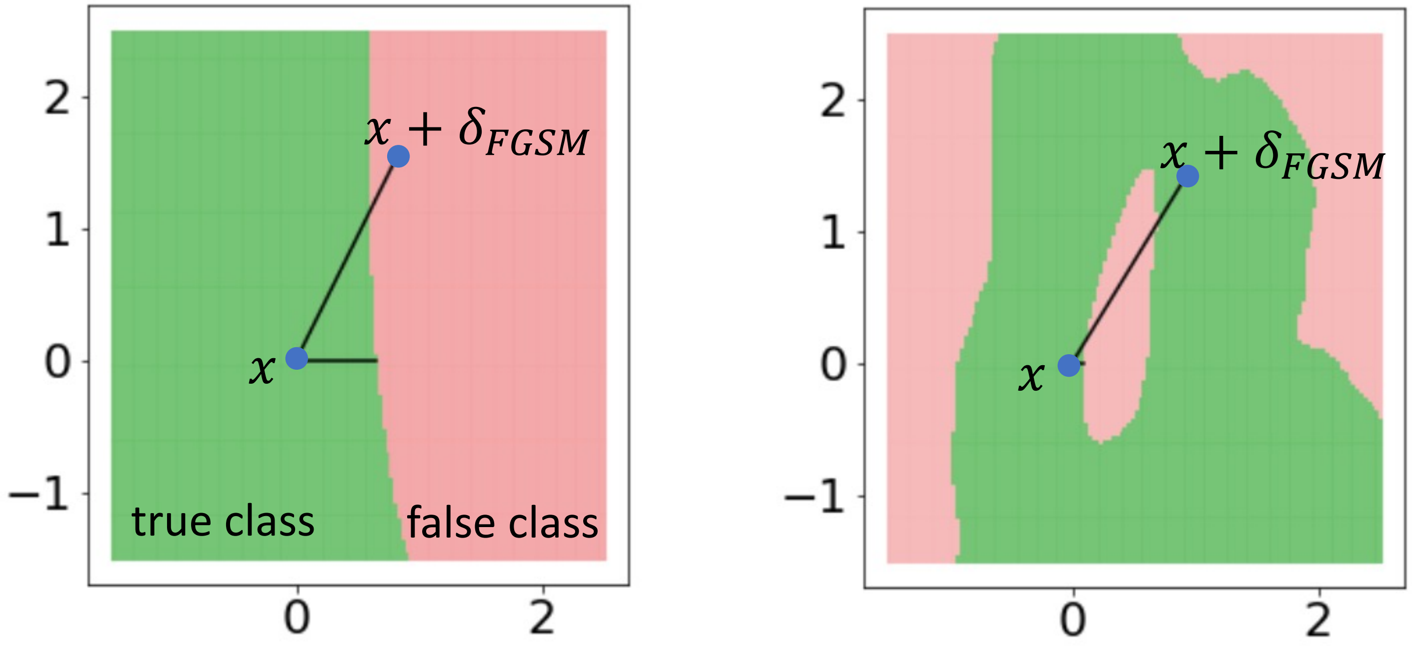
Geometric Analysis of Catastrophic Overfitting

Geometric Analysis of FGSM

Cross-section of the decision boundary spanned by two vectors.

- Calculated by DF^2 (A direction perpendicular to the decision boundary)
- Calculated by the adversarial method used in the training process (FGSM or DF^{∞}-1)

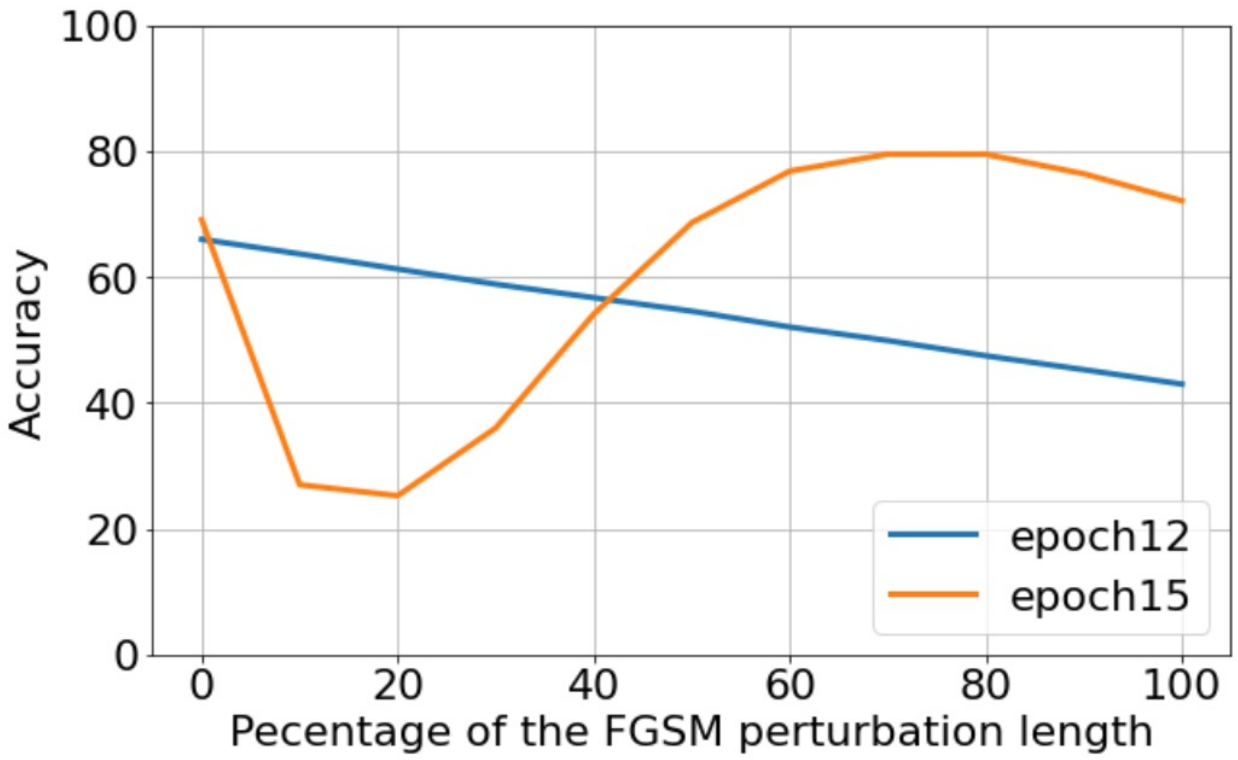
The model is trained by FGSM with ε = 8/255



Before CO　　　After CO

Geometric Analysis of FGSM

Accuracy under different FGSM perturbation length



- Before CO (epoch 12):

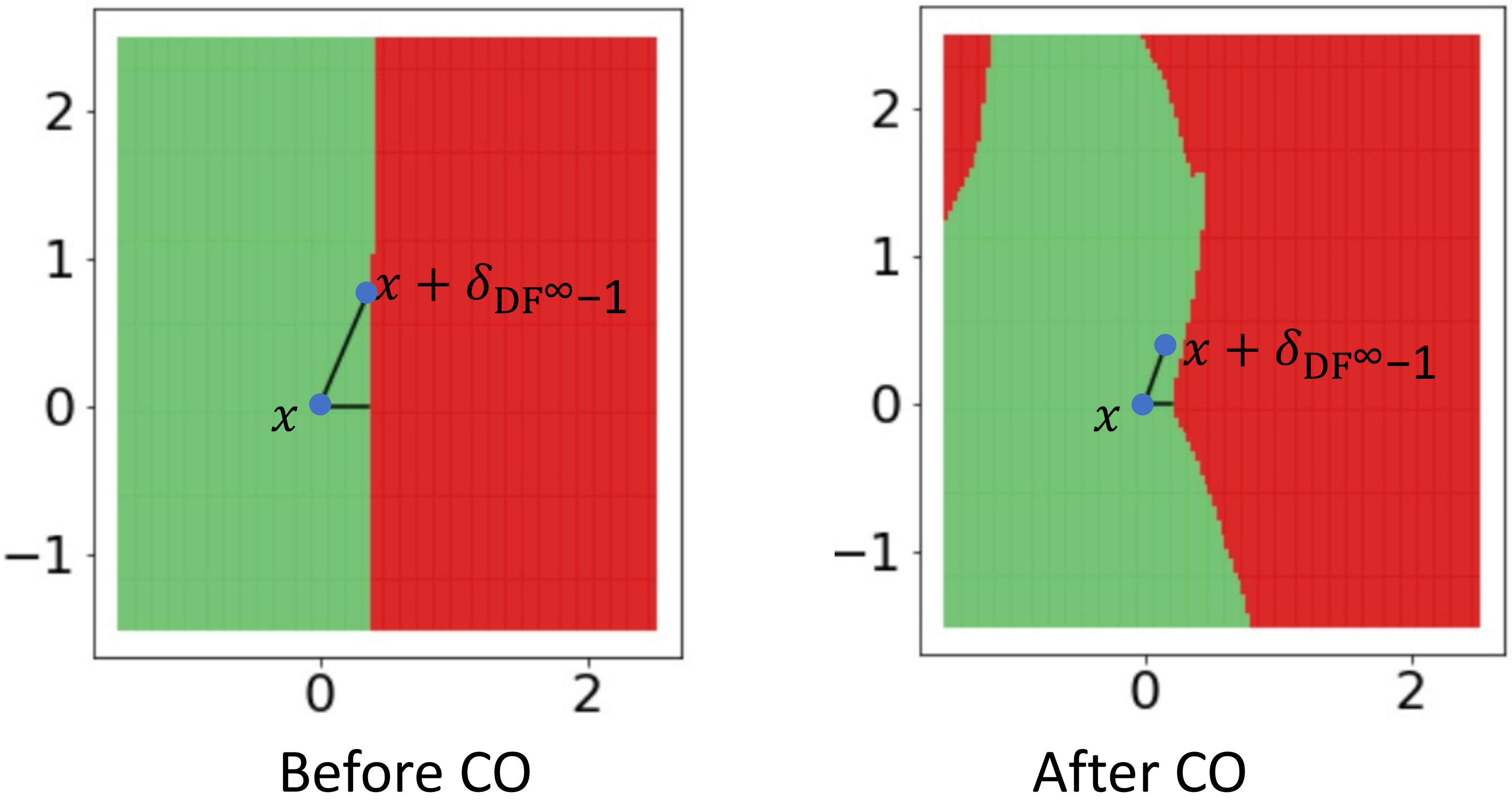
Large perturbation is more effective than small perturbation to find adversarial example

- After CO (epoch 15) :

Small perturbation is more effective than large perturbation to find adversarial example

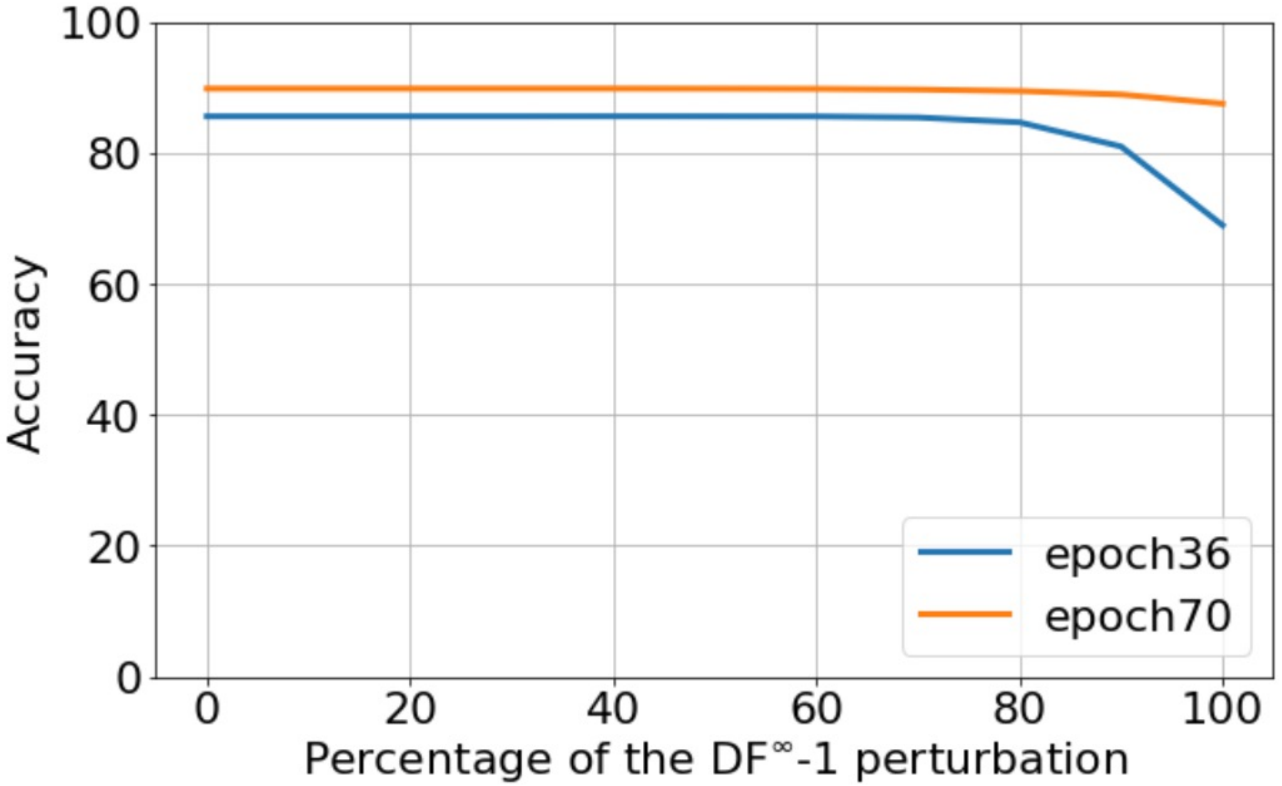
Geometric Analysis of $\mathbf{DF}^{\infty}$-1

The model is trained by DF^{∞}-1 with ε = 8/255



Geometric Analysis of $\mathbf{DF}^{\infty}$-1

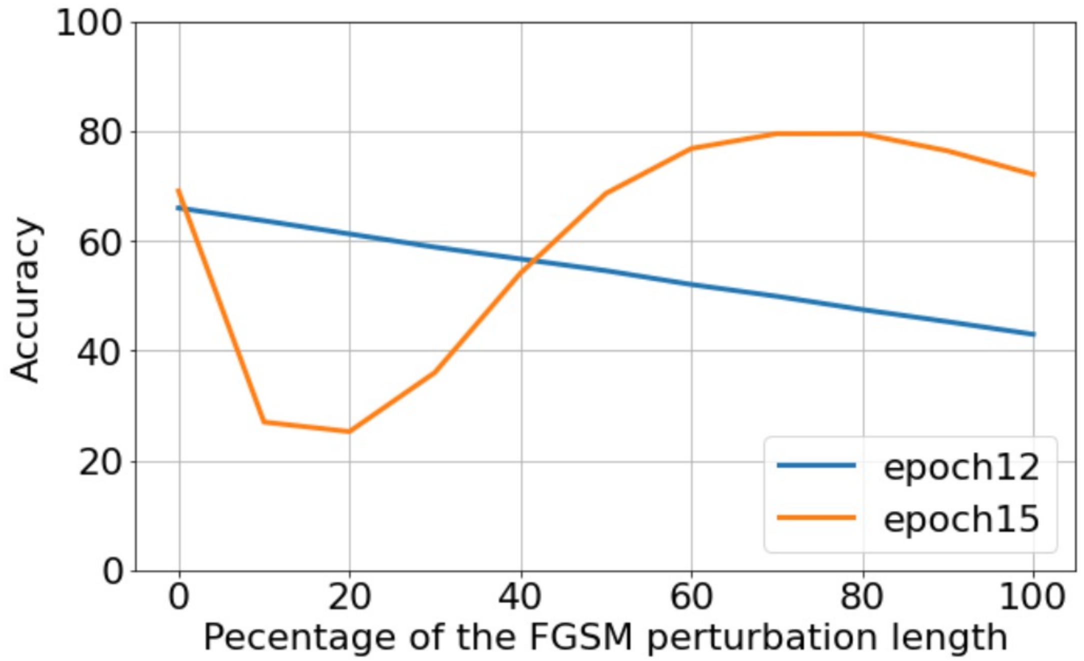
Accuracy under different DF^{∞}-1 perturbation length



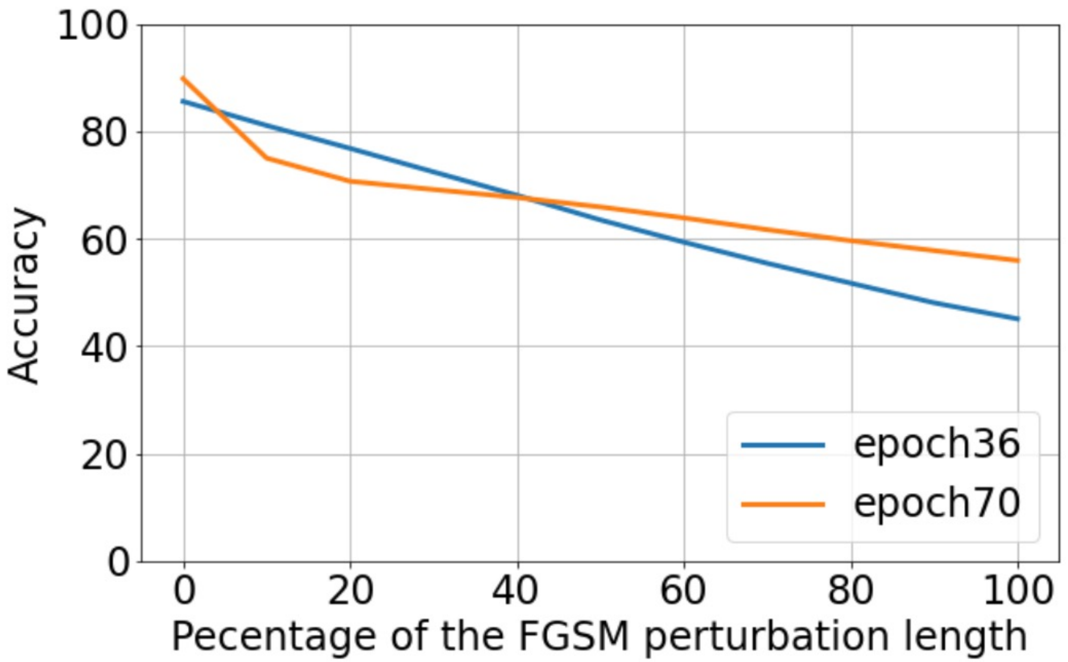
Both before (epoch 36) and after (epoch 70) CO, large perturbation is always more effective than small perturbation to find adversarial example.

Compare Models Trained by FGSM and $\mathbf{DF}^{\infty}$-1

Take the models trained by FGSM and DF^{∞}-1 and evaluate by **FGSM perturbation**.



Model trained by FGSM



Model trained by DF^{∞}-1

Analysis of Factors Causing Catastrophic Overfitting

Hypothesis: Large Perturbation Causes CO

Evidence

1. Random initialization in RS-FGSM is guaranteed to decrease the expected length of the perturbation. [1]
2. Reduce the step size of FGSM can avoid CO.

Counter Experiment

Goal: Perturbations with the same length, one causes CO while the other not.

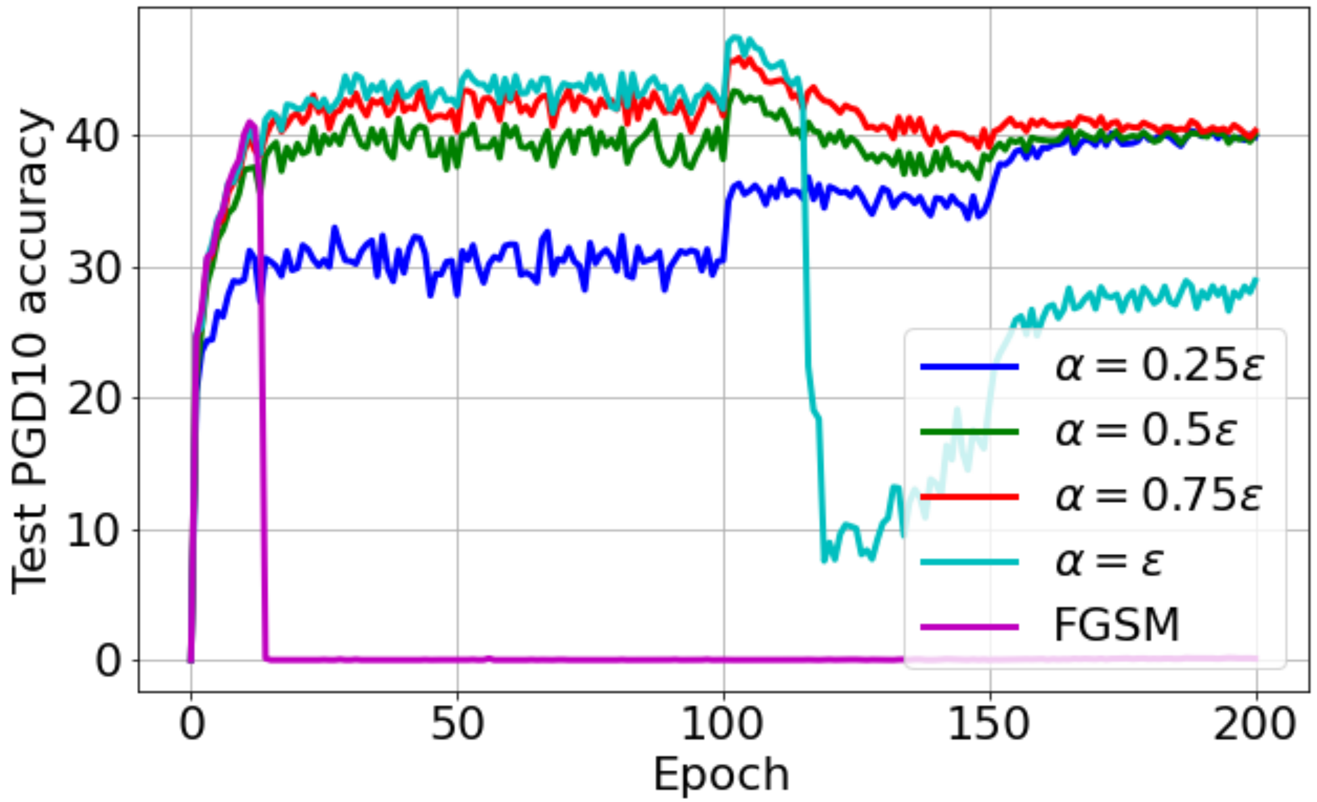
Implementation: Generate $\delta_{RS-FGSM}$ with different step size α
Magnify the $\delta_{RS-FGSM}$ to the same l_2 norm as δ_{FGSM}

$$\delta_{magnified} = \frac{\|\delta_{FGSM}\|_2}{\|\delta_{RS-FGSM}\|_2}\delta_{RS-FGSM}$$

[1] Maksym Andriushchenko and Nicolas Flammarion. *Understanding and Improving Fast Adversarial Training*

Experiment Results

PGD-10 accuracy of the model trained by perturbations with same length and different directions.



- Smaller the step size α, the direction of the perturbation is closer to the direction of random initialized $\delta \sim \mathcal{U}\left([-\varepsilon, \varepsilon]^d\right)$

- Besides the perturbation's length, its direction is also important

Hypothesis: Perturbation Should Span the Entire Threat Model

Evidence

ε=8/255

1. When step size $\boldsymbol{\alpha = \varepsilon}$, each dimension of $\delta_{RS-FGSM}$ is **between $-\varepsilon$ and ε**, RS-FGSM does not suffer from CO on CIFAR10.
   When step size $\boldsymbol{\alpha = 2\varepsilon}$, each dimension of $\delta_{RS-FGSM}$ is **either $-\varepsilon$ or ε**, RS-FGSM suffers from CO on CIFAR10. [1]

2. a) Random initialized δ is either $-\frac{\varepsilon}{2}$ or $\frac{\varepsilon}{2}$ for each dimension
   b) Step size $\alpha = \frac{\varepsilon}{2}$
   c) Final perturbation's each dimension is in $\{-\varepsilon, 0, \varepsilon\}$

can not train the robust model on MNIST dataset while RS-FGSM is capable [1]

[1] Eric Wong, Leslie Rice, and J. Zico Kolter. *Fast is better than free: Revisiting adversarial training*.

Experiment results

Counter Experiment (Boundary-RS-FGSM)

Use different initialization From RS-FGSM.
Initialize on the boundary of l_∞-ball, either $-\varepsilon$ or ε for each dimension.

The value of the final perturbation is discrete and not span the entire threat model

ε=8/255, Dataset CIFAR10

Method	Best Clean / PGD-50-10
FGSM	66.72 / 40.46
RS-FGSM	86.77 / 42.69
Boundary-RS-FGSM	87.03 / 42.72

Boundary-RS-FGSM can achieve the comparable robust accuracy as RS-FGSM

Hypothesis: Large Diversity of Perturbations Can Avoid CO

Definition Diversity = 1-cos(δ_a, δ_b)

Compute pertubations twice using the same input and model. δ_a is the first one and δ_b is the second one.

Evidence FGSM has zero diversity and RS-FGSM has positive diversity

Counter Experiment

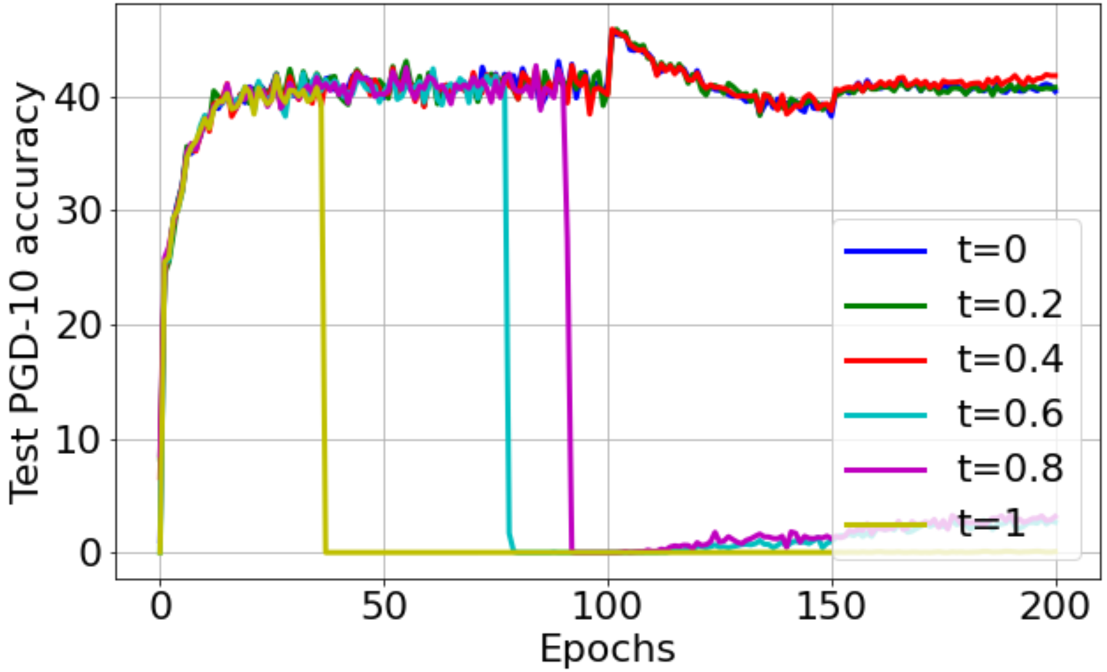
Goal: Perturbations with similar diversity, one causes CO while the other not.

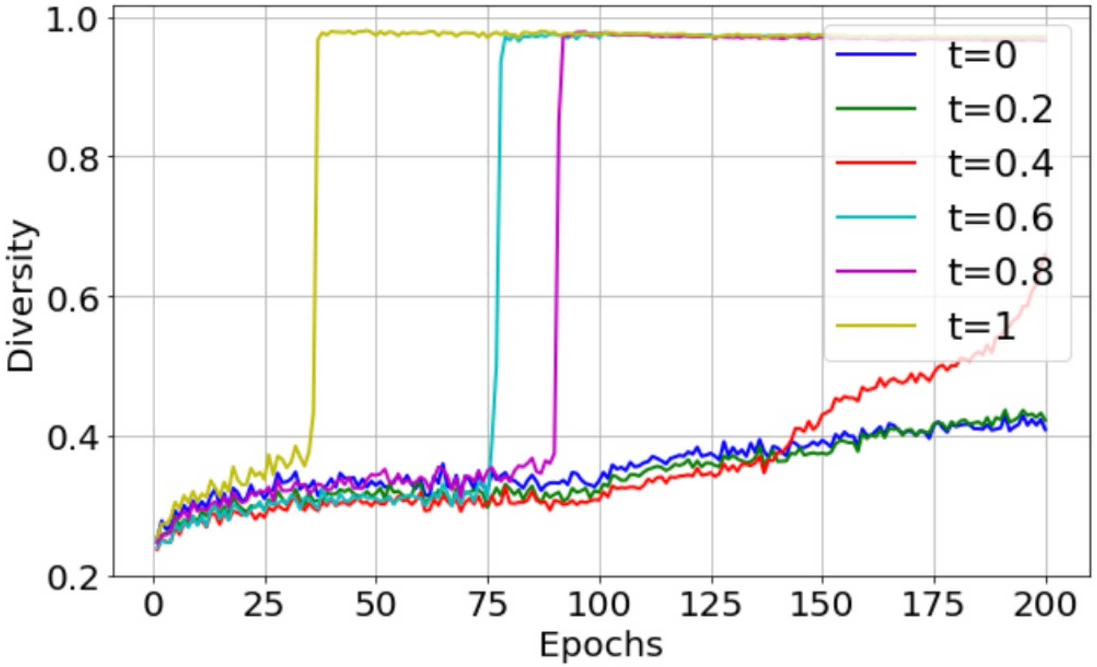
Implementation: $\delta_1, \delta_2 \sim \mathcal{U}\left([-\varepsilon, \varepsilon]^d\right)$

$$\delta = (1-t)\delta_1 + t\delta_2$$

$$\delta_{Diff-RS-FGSM} = \Pi_{[-\varepsilon,\varepsilon]^d}(\boldsymbol{\delta_1} + \alpha * \mathrm{sgn}(\nabla_x l(x + \boldsymbol{\delta}, y; \theta)))$$

Experiment Results





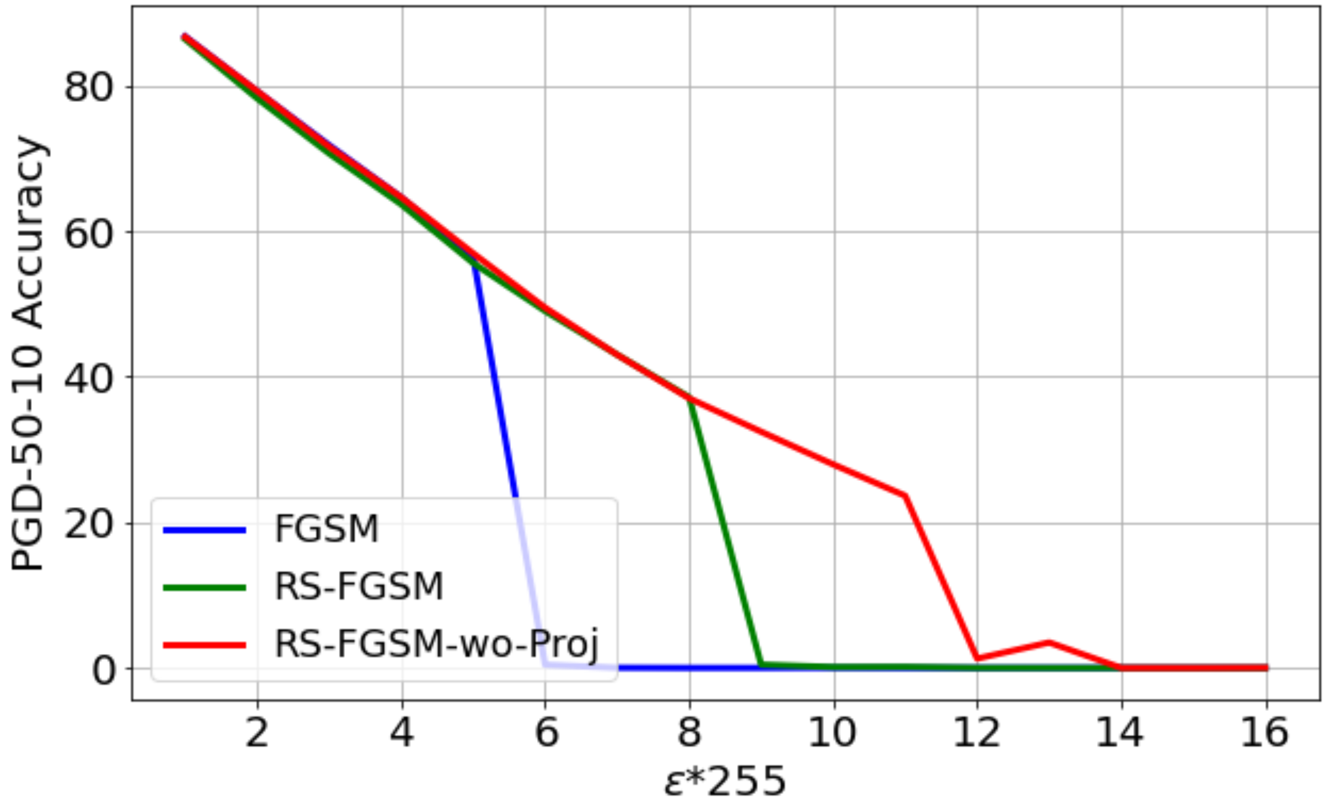
When **t** is 0, 0.2, and 0.4, CO does not happen
When **t** further increase, CO happens

Perturbations with different **t** has almost the same diversity before CO

large diversity cannot guarantee to avoid CO

Further Improvements on RS-FGSM methods

Improve the RS-FGSM by not projecting back to l_∞-ball



RS-FGSM-wo-Proj permits us to use higher values of ε compared to RS-FGSM

Further Improvements on RS-FGSM methods

ε=8/255

Method	Clean	PGD-50-10
RS-FGSM	$86.35 \pm 0.34\%$	$43.57 \pm 0.30\%$
RS-FGSM-wo-Proj	$82.66 \pm 0.56\%$	$47.56 \pm 0.37\%$

* averaged over 5 random seeds

- RS-FGSM-wo-proj has better robust accuracy compared to RS-FGSM

Conclusion

- FGSM and DF^{∞}-1 both suffers from CO

- FGSM and DF^{∞}-1 show totally different geometric properties after CO

- We experimentally analyze three hypotheses on potential factors causing CO

- We make a modification to RS-FGSM by not projecting perturbation back to the l_{∞}-ball which leads to a better robust accuracy and permits us to use larger values of ε

Future work

- Geometric properties after CO happens has been well studied
  - Remaining question: why FGSM and DF^{∞}-1 show totally different geometric properties after CO happens.

- Need to put more efforts to study the main factors that cause CO
  - Explore the relationship between the direction of the perturbation and the maximum length of the perturbation which does not cause CO
  - In RS-FGSM,we use this equation $\Pi_{[-\varepsilon,\varepsilon]^d}(\boldsymbol{\delta} + \alpha * \mathrm{sgn}(\nabla_x l(x + \boldsymbol{\delta}, y; \theta)))$ to calculate perturbations. We can study the usage of $\boldsymbol{\delta}$ in these two places

Reference

- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706. 06083 [stat.ML].
- Eric Wong, Leslie Rice, and J. Zico Kolter. *Fast is better than free: Revisiting adversarial training*. 2020. arXiv: 2001.03994 [cs.LG].
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. *DeepFool: a simple and accurate method to fool deep neural networks*. 2016. arXiv: 1511.04599 [cs.LG]
- Maksym Andriushchenko and Nicolas Flammarion. *Understanding and Improving Fast Adversarial Training*. 2020. arXiv: 2007.02617 [cs.LG].