Training robust neural networks

TADJER Amina
ZAGHRINI Eloïse
BENREKIA Mohamed Ali

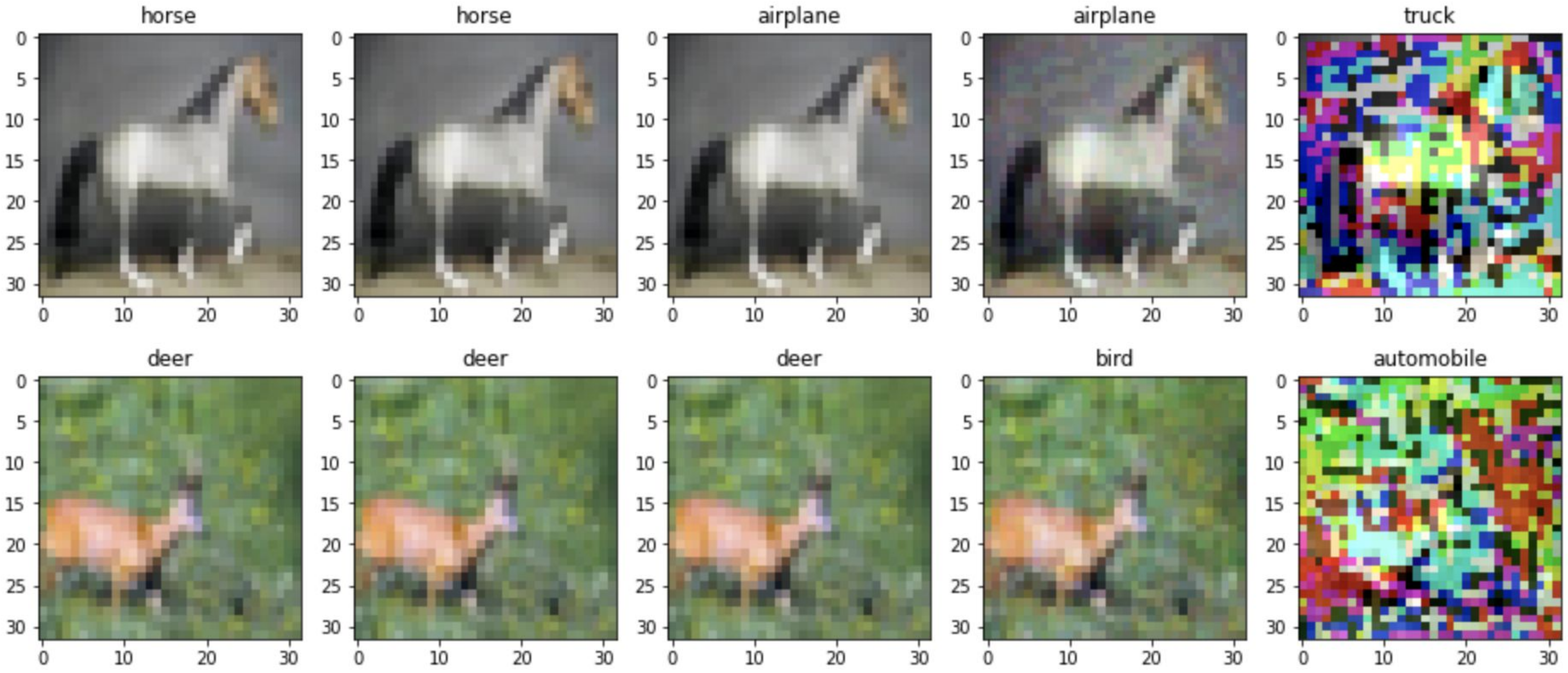
Plan

1. The problem
2. FGSM,PGD
3. Adversarial Training
4. Defensive Distillation
5. Defensive Randomized Networks
6. Other Attacks (C&W)

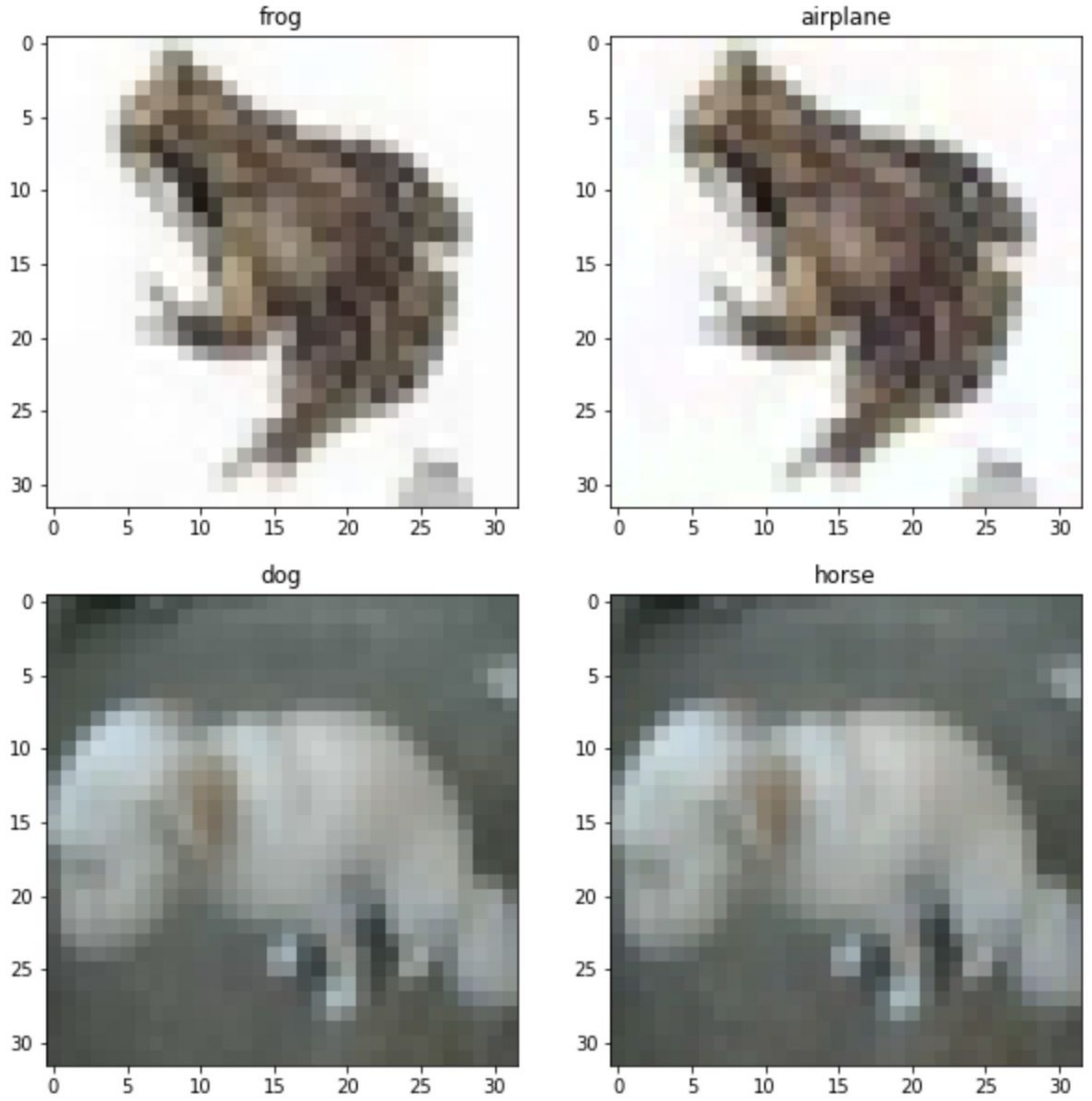
FGSM Attack [1]:

$$\text{perturbated\_image} = \text{image} + \text{epsilon} * \text{sign}(\text{data}_{\text{grad}}) = x + \epsilon * \text{sign}(\nabla_x J(\theta, \mathbf{x}, y))$$

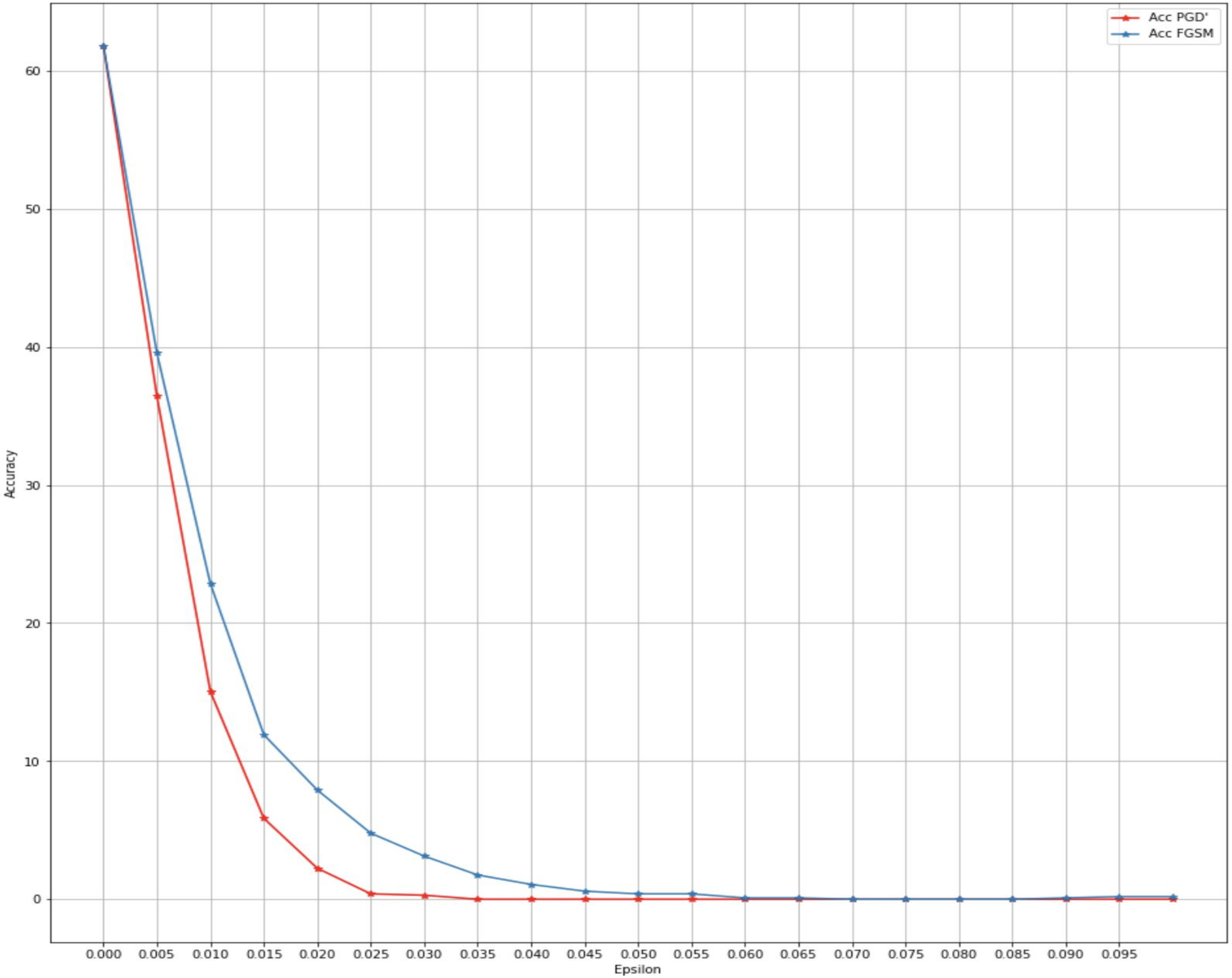
epsilon = [0, 0.0003, 0.003, 0.03, 0.3]



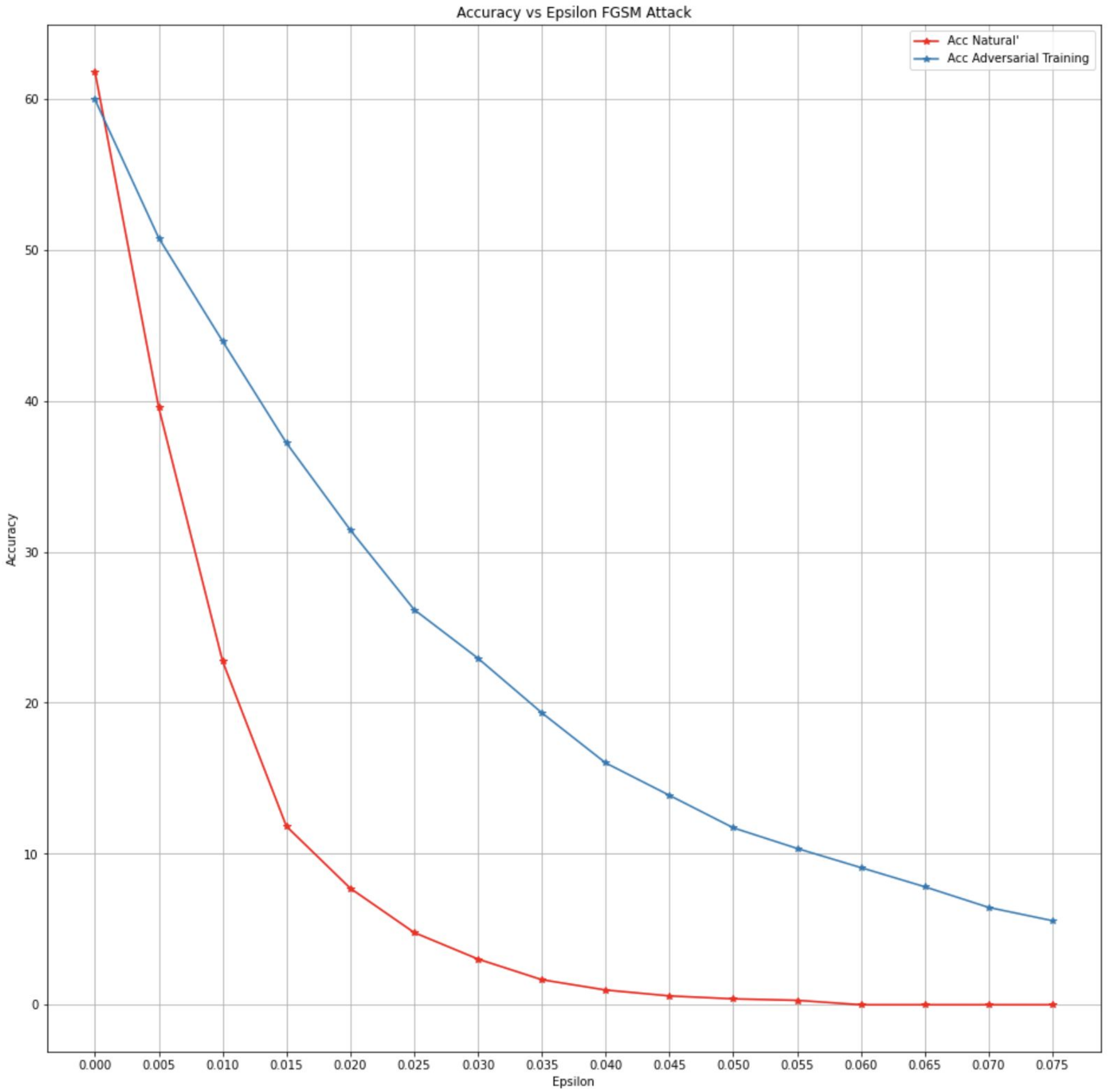
PGD Attack:

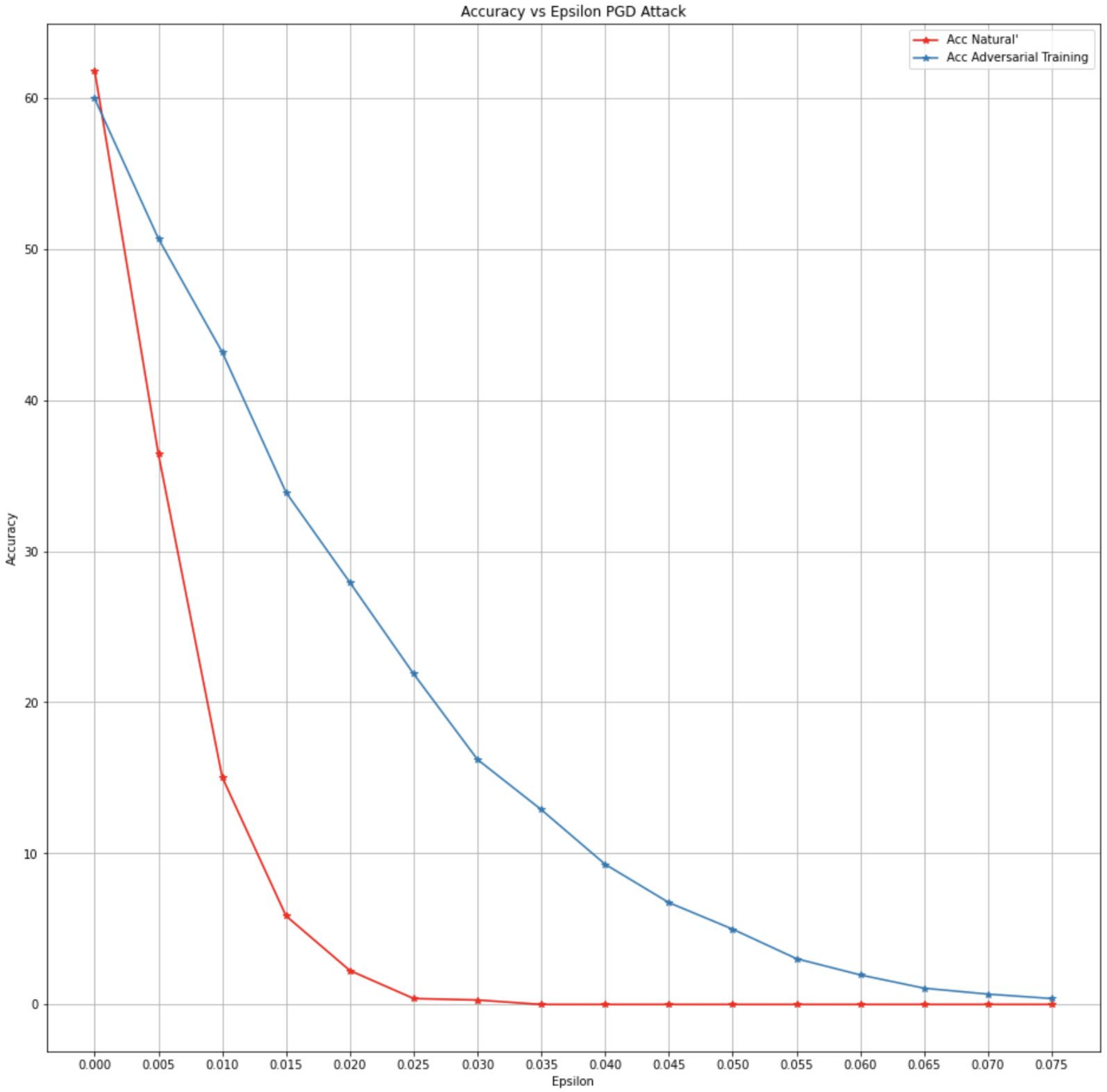


Accuracy results of FGSM and PGD attacks with epsilon from 0 to 0,095

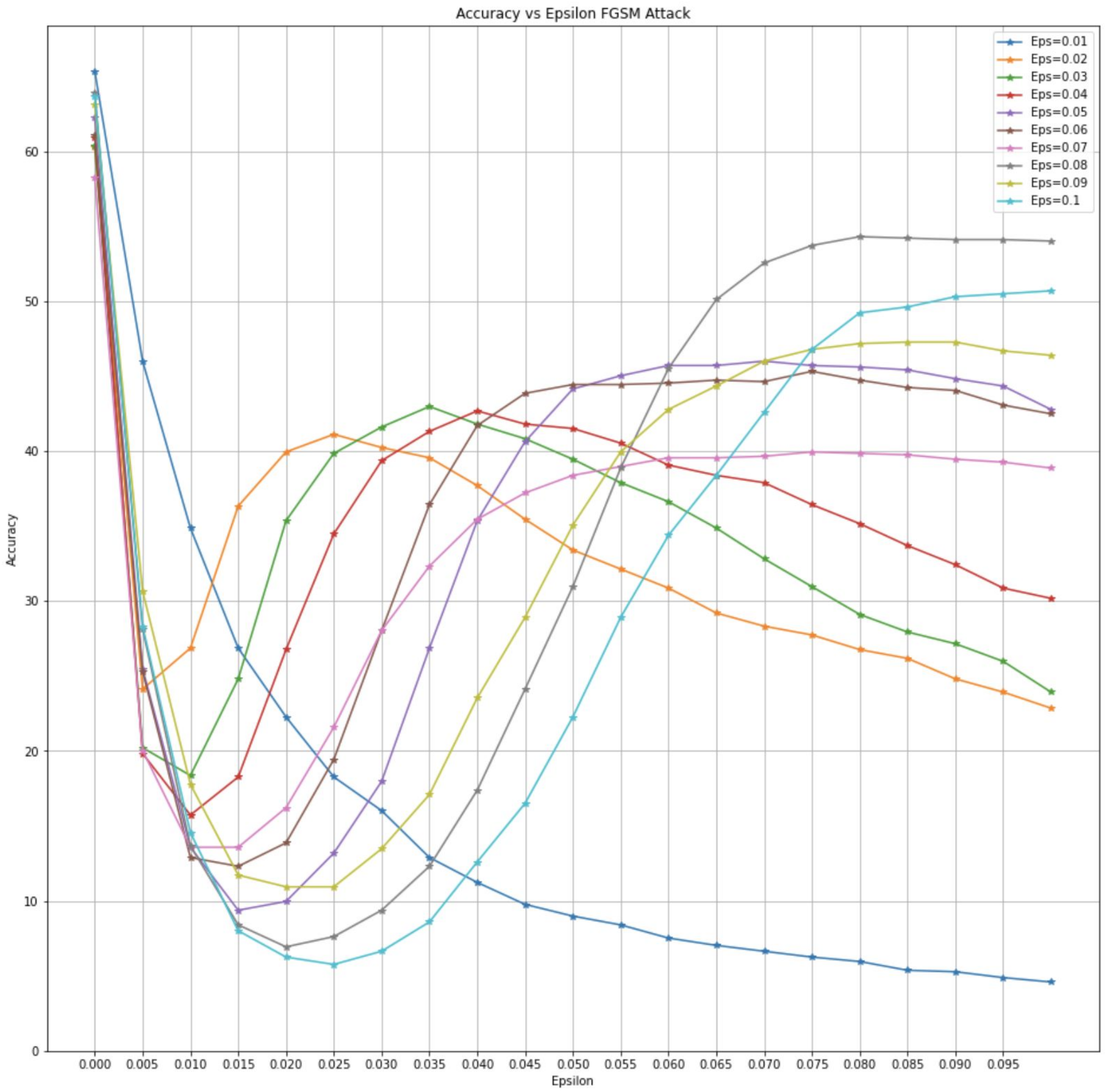


Adversarial Training[3] : Epsilon=0.01, 50% Natural Data, 50% Adversarial Data

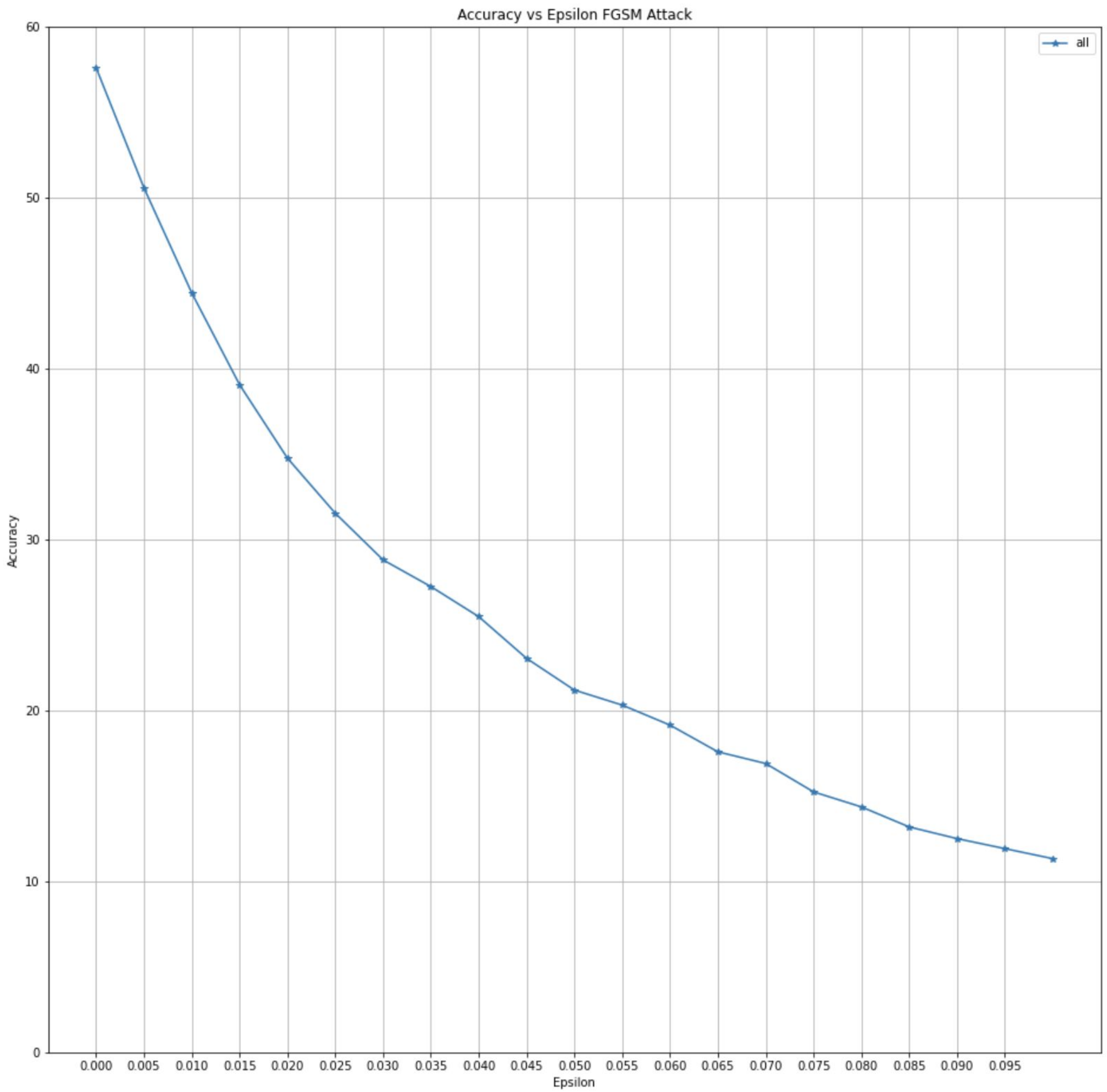




Results of Adversarial Training with FGSM

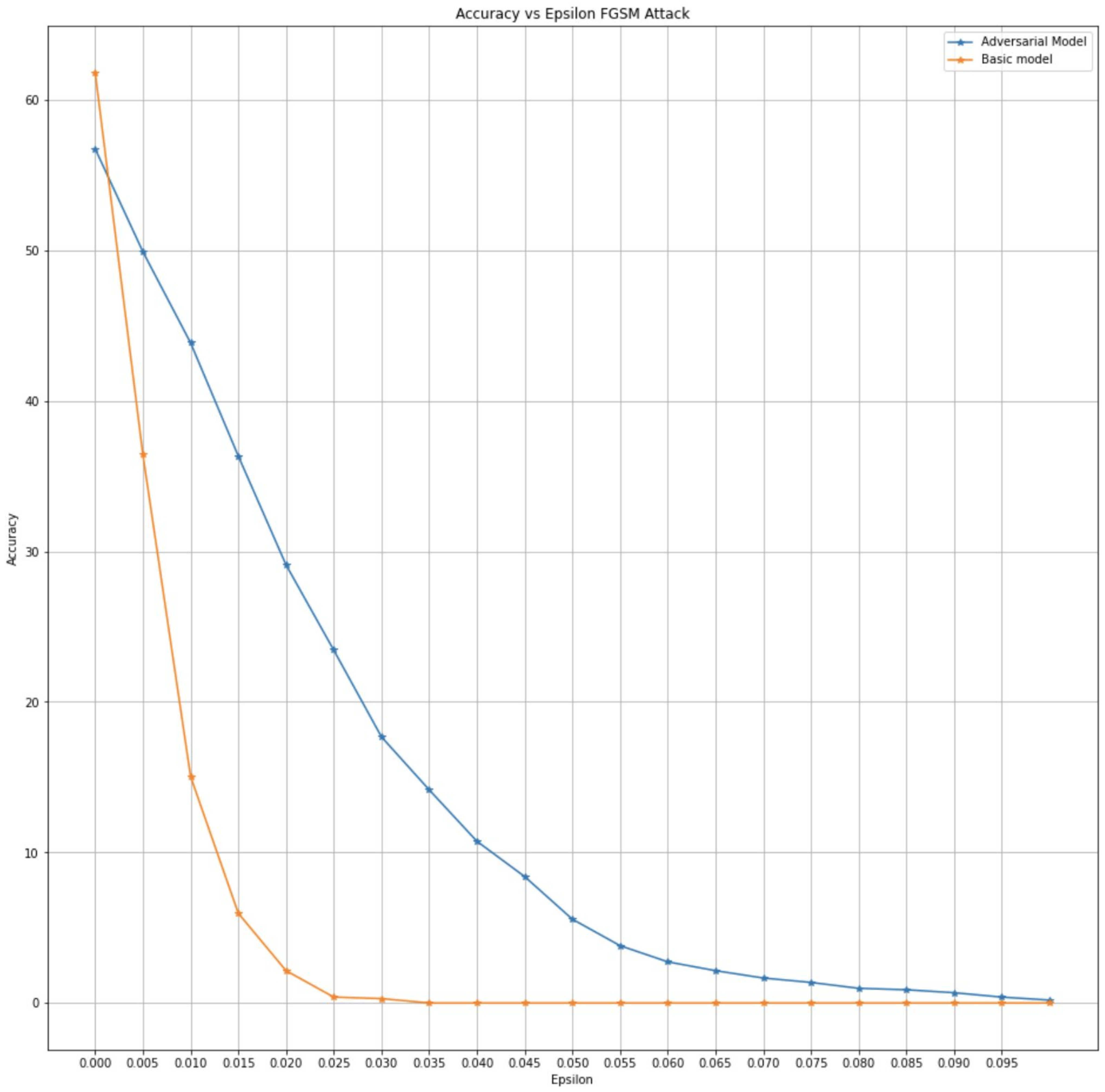


Accuracy vs Epsilon FGSM Attack
Eps=0.01
Eps=0.02
Eps=0.03
Eps=0.04
Eps=0.05
Eps=0.06
Eps=0.07
Eps=0.08
Eps=0.09
Eps=0.1
Accuracy
0
10
20
30
40
50
60
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
0.050
0.055
0.060
0.065
0.070
0.075
0.080
0.085
0.090
0.095
Epsilon


Accuracy vs Epsilon FGSM Attack
all
Accuracy
Epsilon
0
10
20
30
40
50
60
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
0.050
0.055
0.060
0.065
0.070
0.075
0.080
0.085
0.090
0.095

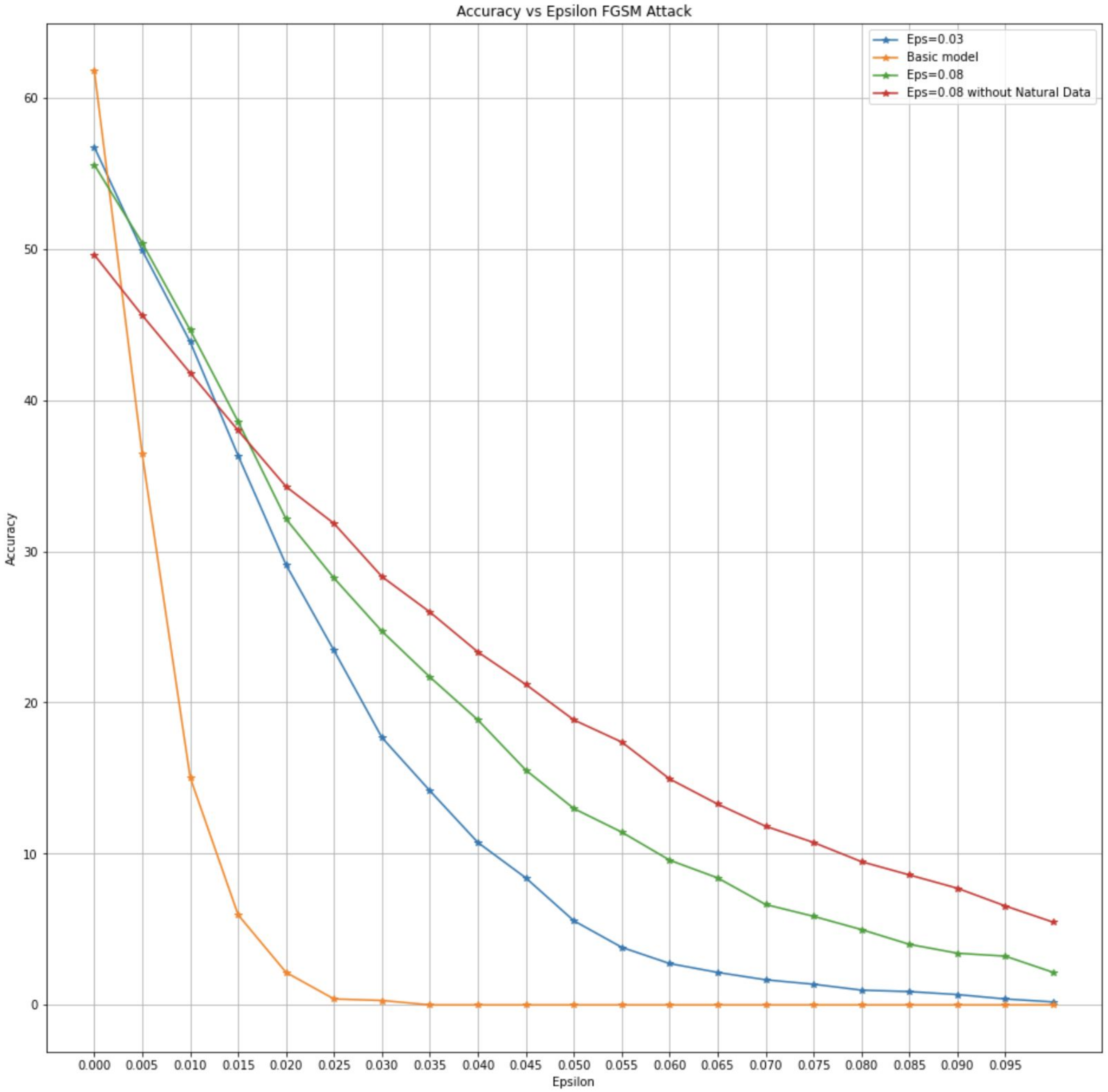
Adversarial Training[3]

Epsilon=0.03



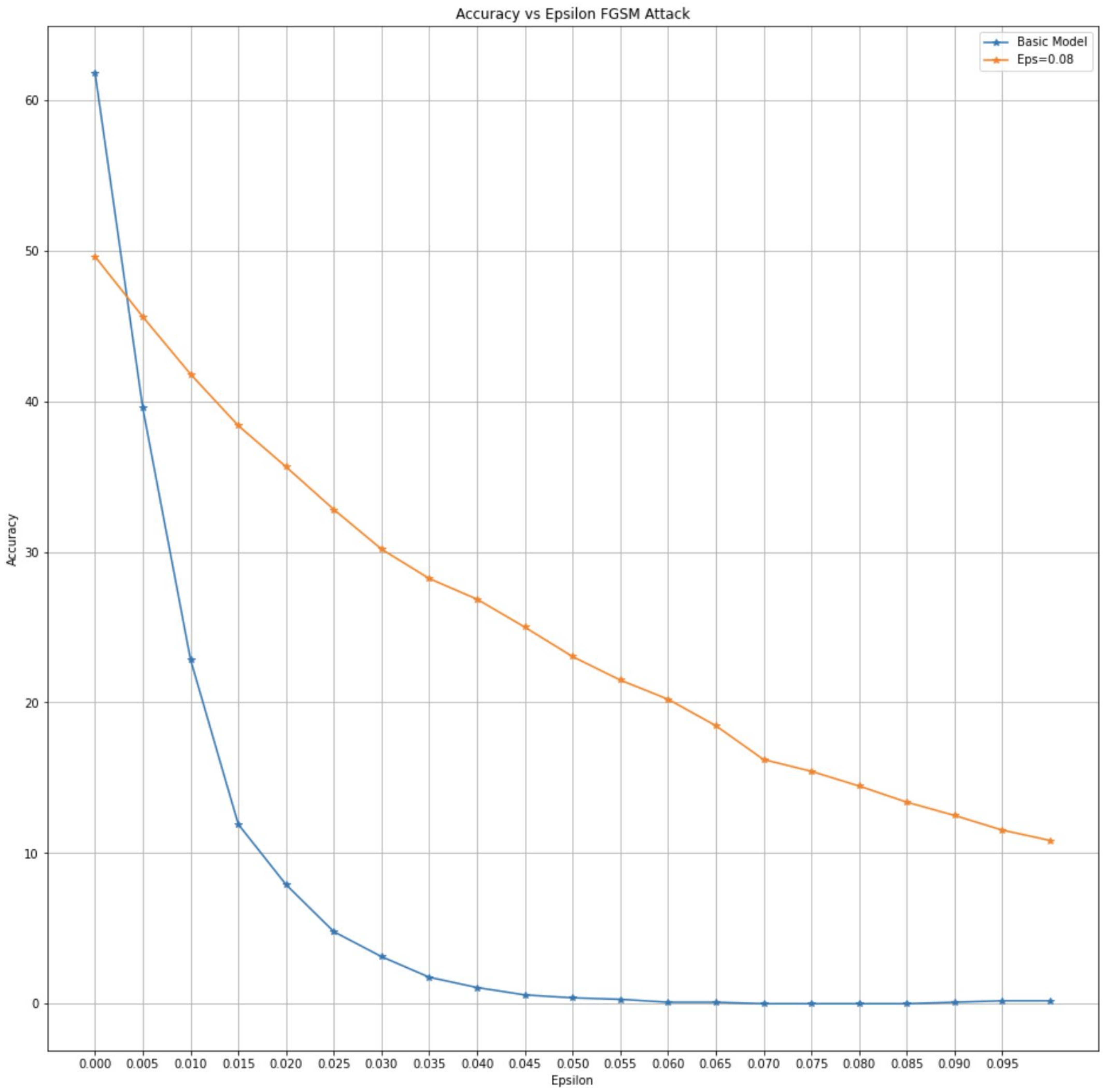
Results of Adversarial Training with PGD

Adversarial Training[3]



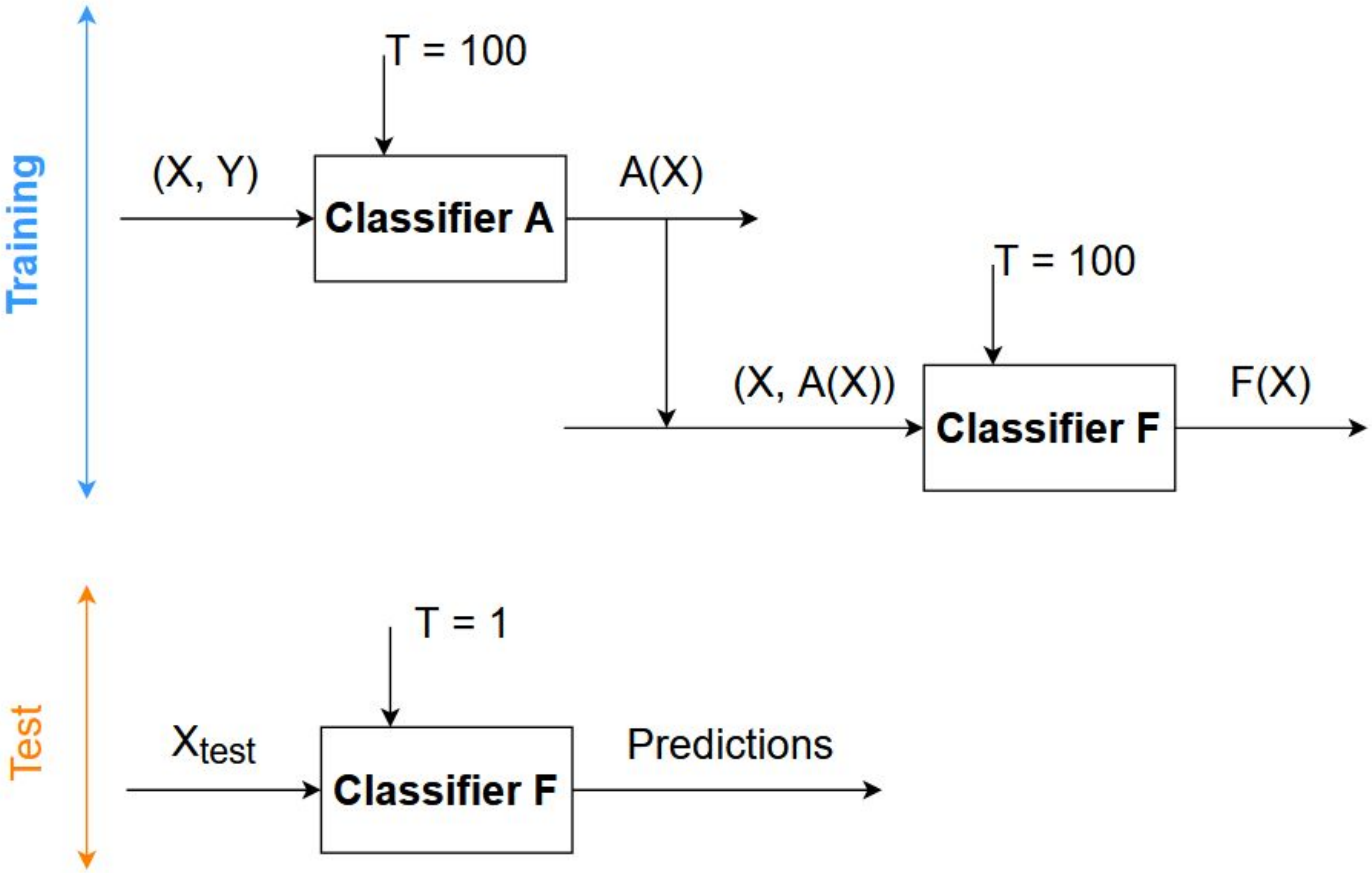
Results of Adversarial Training with PGD Vs PGD attack

Adversarial Training[3]

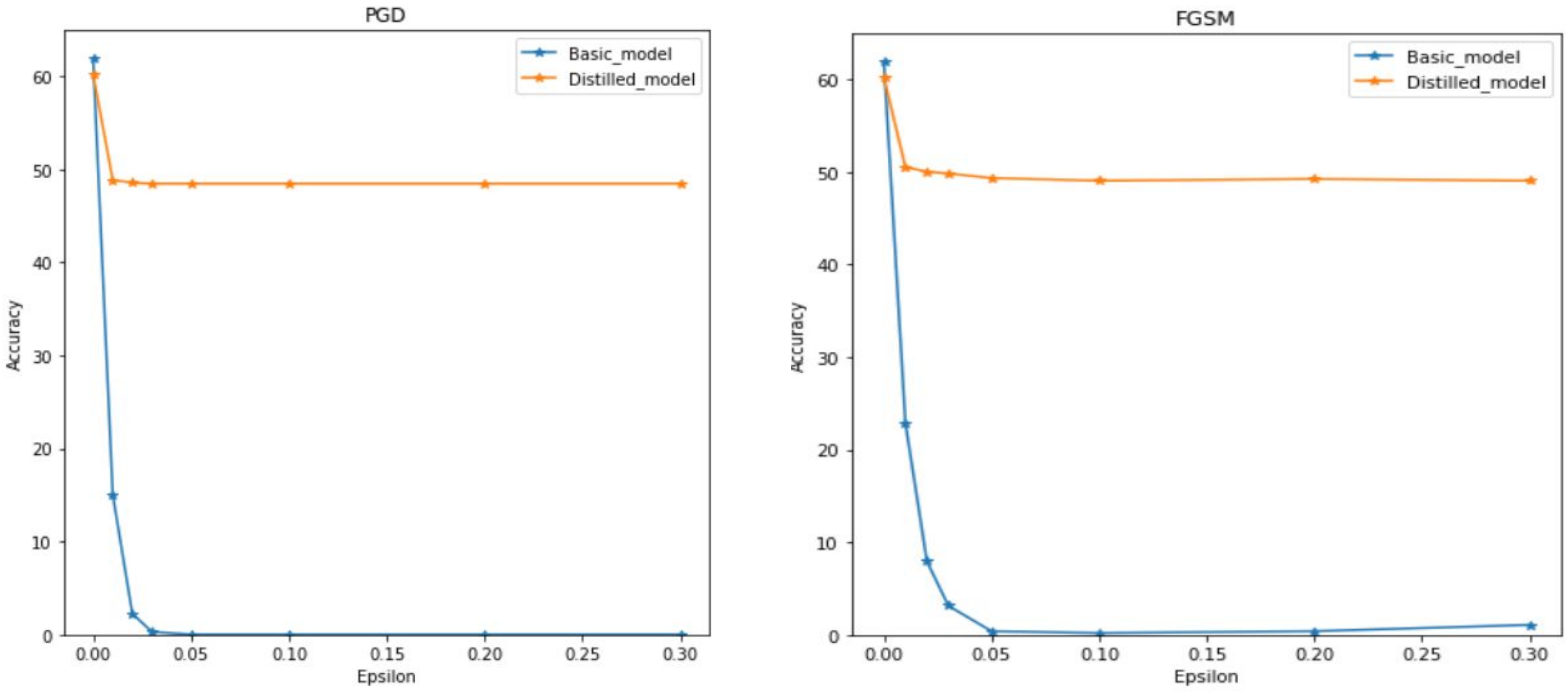


Results of Adversarial Training with PGD Vs FGSM attack

Defensive Distillation [4]:

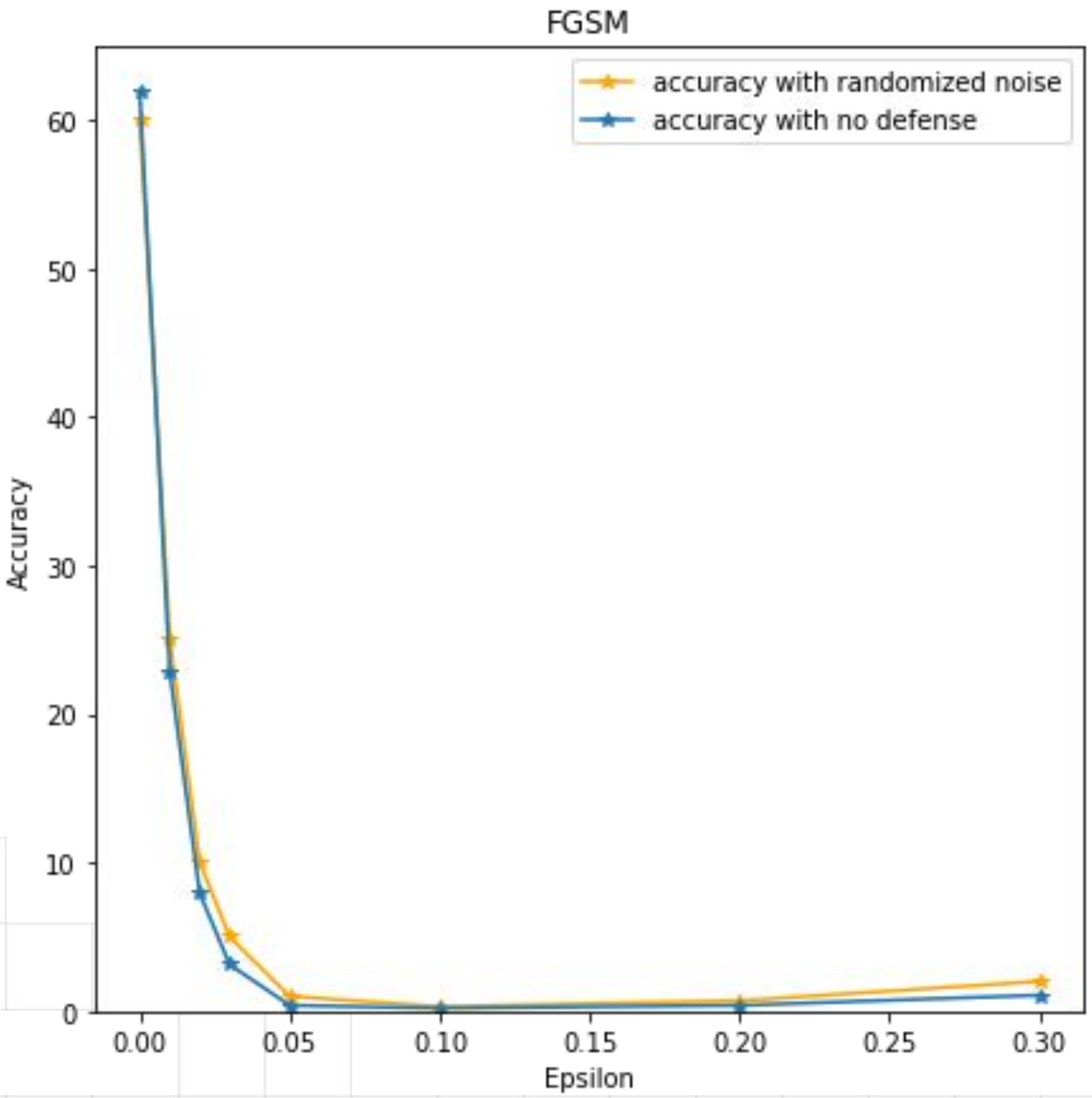


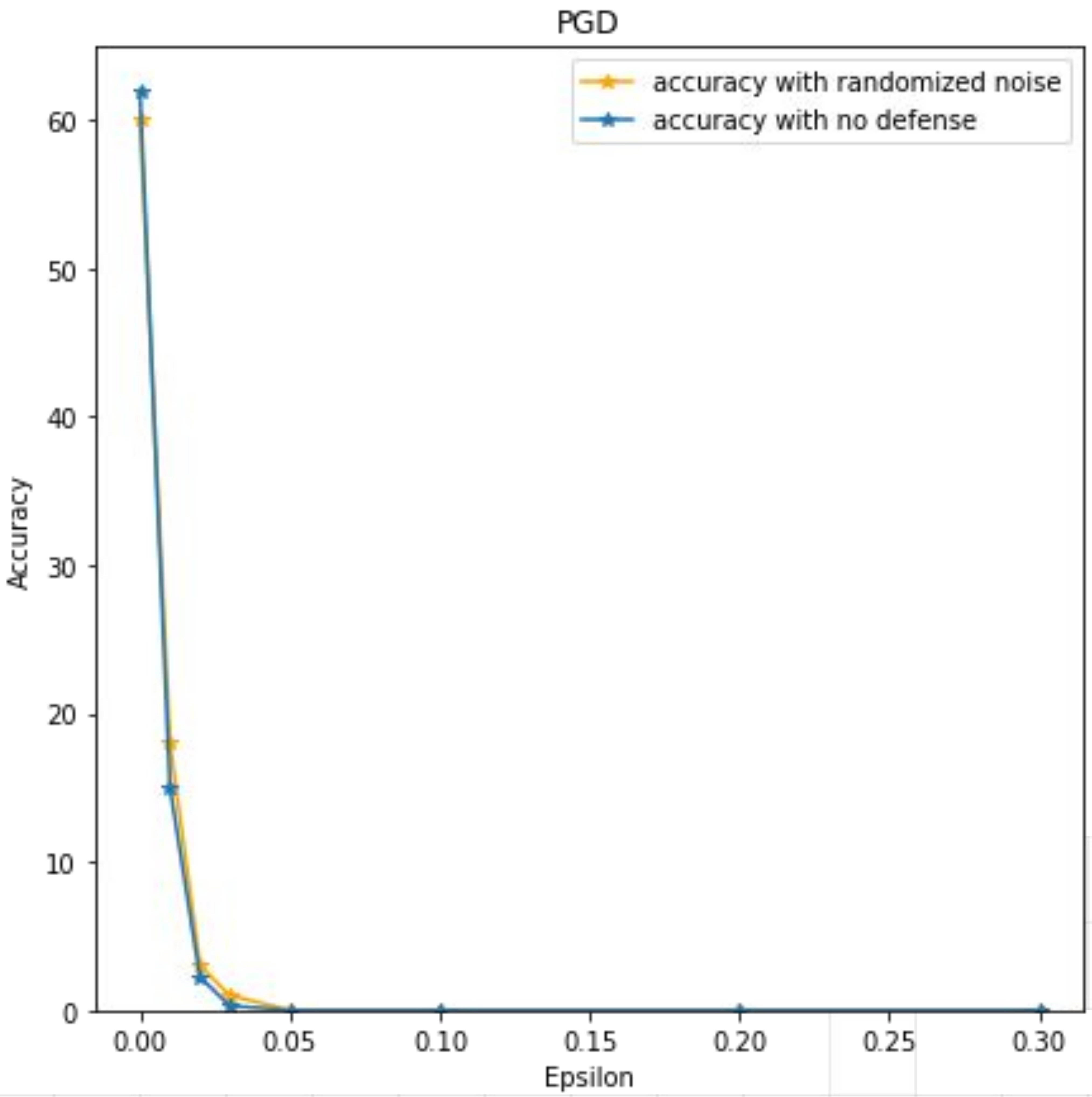
Basic and Distilled model against FGSM & PGD :



Randomized Networks [5]:

$$x_{test} = x_{test} + \textit{gaussian\_noise}(0, 0.01)$$





Ensemble of Networks

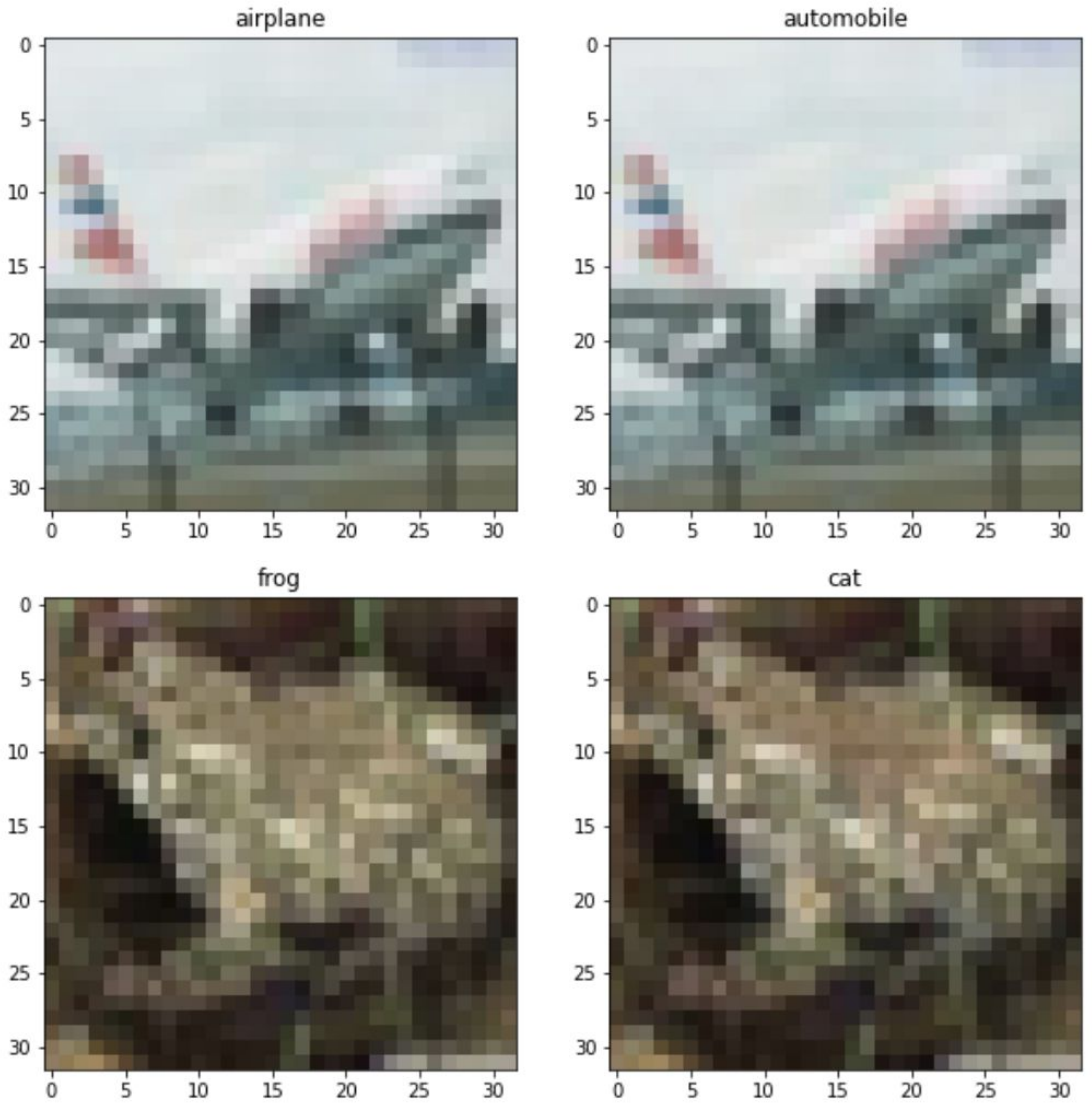
10 different networks with the same architecture and different ways of attacking and predicting

Carlini & Wagner [6] :

How it works?

$$
\begin{aligned}
\text{minimize} \quad & D(x, x+\delta) \\
\text{such that} \quad & C(x+\delta) = t \\
& x + \delta \in [0,1]^n
\end{aligned}
$$

$$f(x') = (\max_{i \neq t}(Z(x')_i) - Z(x')t)^+$$



Carlini & Wagner [6] :

Model	Accuracy	Accuracy after attack
Natural Model	61.81%	0%
Model trained with adversarial Data FGSM	59.96%	0%
Model trained with adversarial Data PGD	56.73%	0%
Distilled Model	60.15%	0%

Results of C&W attack on Basic model, Models with adversarial Training and Distilled Model

Conclusion and perspectives

Thank You!

GITHUB IDs:

aminatadjer
El-Zag
iladan0

Bibliography:

[1] FGSM - Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[2] Defensive Distillation - Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. arXiv preprint arXiv:1511.04508, 2016b.

[3] Madry, Aleksander et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." *ArXiv*abs/1706.06083 (2018): n. pag.

[4] Carlini, Nicholas and David A. Wagner. "Towards Evaluating the Robustness of Neural Networks." *2017 IEEE Symposium on Security and Privacy (SP)* (2017): 39-57.