Adversarial Attacks in Deep Learning

Mohammed Benmaiza, Yassine Abou Hadid, Mehdi Bentaleb

Data Science Project - Master 2 IASD

9 December 2022

Introduction

Problem : Perform adversarial attacks and defense mecanisms on a deep learning model (on Cifar-10 dataset)

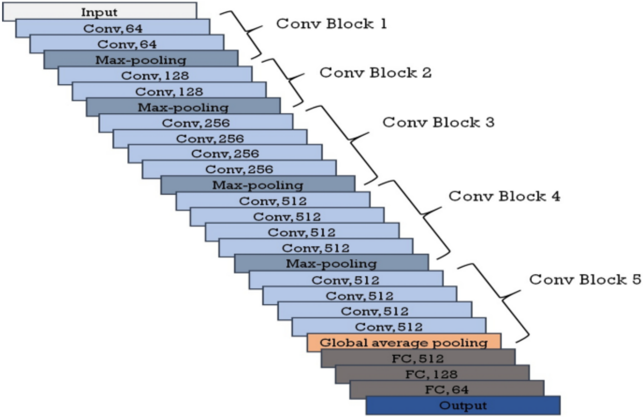
Base model : VGG-19

Methods explored :

- FGSM
- PGD
- Adversarial Training

VGG-19

- Epochs: 10
- Learning rate: 0.001
- Optimizer: SGD
- Accuracy on Train Set: 85.6
- Accuracy on Test Set: 82.9



(a) VGG-19 Architecture

FGSM

- Target function: $\max_{\|\xi\| \leq \epsilon} l_f(x+\xi, y)$
- If ϵ is small, we can approximate it by: $\max_{\|\xi\| \leq \epsilon} \xi^{\,t} \nabla_x l_f(x, y)$
- The solution is defined by $\xi = \epsilon sign(\nabla_x l_f(x, y))$ when $\|.\| = \|.\|_\infty$

PGD

- Iterative version of FSGM.
- We generate perturbations with the following:

$$\begin{cases} x_0 = x \\ x_{t+1} = \textit{proj}_{B(x_0,\epsilon)}(x_t + \xi.sign(\nabla_x l_f(x,y))) \end{cases}$$

Adversarial Training

- Train the network with the adversarial risk : $\min E_{(x,y)}\left(\max_{\|\xi\|\leq\epsilon} l_{f_\theta}(x+\xi,y)\right)$
- We generate perturbed images from our training set using an attack. Then, we concatenate them with our training images from the base dataset. The network architecture remains unchanged. We train our model on this new dataset (50 % perturbed) to boost its robustness to adversarial attacks.
- We have carried out Adversarial training by training two models, against FGSM and PGD.

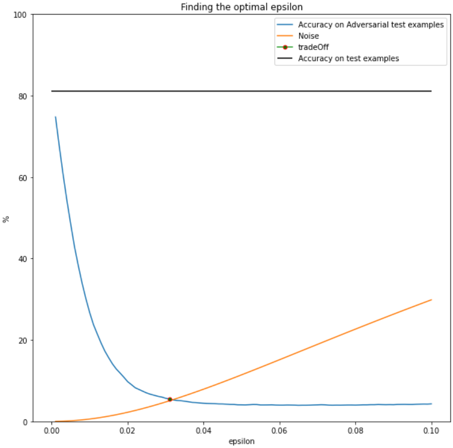


Figure: Accuracy/Noise tradeoff

- $SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_x\sigma_y + c_2)(cov_{xy} + c_3)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)(\sigma_x\sigma_y + c_3)}$

Results

Model (with $\epsilon = 0.031$	Data Tested	Accuracy
Base Model	Original Test Set	82.92
Base Model	FGSM-perturbed Test Set	5.05
Base Model	PDG-perturbed Test Set	10.16
Base Model trained with FGSM-perturbed Train Set	Original Test Set	64.1
Base Model trained with FGSM-perturbed Train Set	FGSM-perturbed Test Set	39.96
Base Model trained with PGD-perturbed Train Set	Original Test Set	43.87
Base Model trained with PGD-perturbed Train Set	PDG-perturbed Test Set	28.59

Future work

- The impact of number of iterations on models
- Perform FGSM and PGD attack with l2-norm
- Explore black box attacks
- Test new adversarial training techniques

References

- I. J. Goodfellow, J. Shlens, C. Szegedy "Explaining and Harnessing Adversarial Examples", 2015.
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu "Towards Deep Learning Models Resistant to Adversarial Attacks, 2017. "