Building Robust Neural Networks

Attack_des_titans

December 9, 2022

Christian Kayo
Dakini Mallam Garba
Killian Susini

Dauphine UNIVERSITÉ PARIS | PSL

Use case : image classification

Problem Setting

- A model classifies images
- small perturbations, imperceptible for the human eye, can change the prediction label

How to make the classifier more robust to these perturbations ?

- Generate images with attack mechanisms
  - FGSM Attack
  - PGM Attack
- Use defense mechanisms to make the network robust to the attacks
  - Adversarial Training
  - Randomized Smoothing

Adversarial attacks

Fast Gradient Signed Method (FGSM)

- Simply trying to maximise the loss by adding a small perturbation in the direction of the gradient

- This method is able to generate adversarial examples rapidly

- Requires the gradients to be computed once

$$x' = x + \varepsilon sign\left(\nabla_x J\left(\theta, x, y\right)\right)$$

x' : adversarial exemple
x : original image
J : loss
y : original input label
epsilon : max perturbation radius
sign : the sign function

Adversarial attacks

Projected Gradient Descent (PGD)

- Also known as an iterated GSM attack
- The perturbation is constrained by a norm of the input
- If the output exits this constraint, it is projected back into the set.
- In theory, it generates more powerful adversary exemples.

$$x^{t+1} = \Pi_{x+S}\left(x_t + \alpha sign\left(\nabla_x L_f\left(x, y\right)\right)\right)$$

Π The projection operator

S Set of allowed perturbations

Defense mechanism

Adversarial Training

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1 - \alpha) J(\boldsymbol{\theta}, \max_{||\delta|| \leq \epsilon} \ell_f(x + \delta, y), y)$$

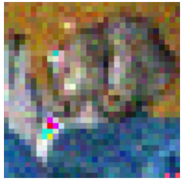
Basic idea: Augment dataset with adversarial examples. Could be FGSM (fast) or PGD (better), l-2 or l-∞ :

- Advantages: Good results against the chosen adversarial examples
- Disadvantage: Does not defend as well against other adversarial examples

To train against both attacks: Mix-Adversarial Training

Natural dataset

+



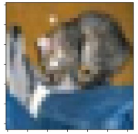
Adversarial examples

(l-2 or l-∞)

Defense mechanism

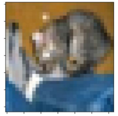
Randomized smoothing

Basic Classifier : f

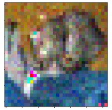


x

$\xrightarrow{f}$ class : 0 (dog)

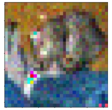


x



$x + \varepsilon_1$ $\xrightarrow{f}$ 0

⋮



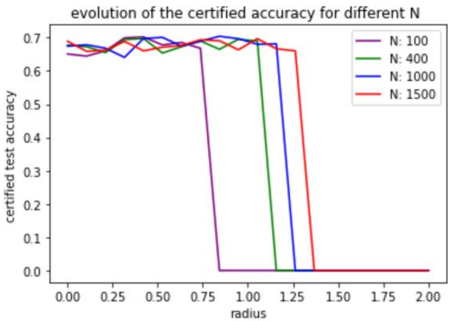
$x + \varepsilon_n$ $\xrightarrow{f}$ 3

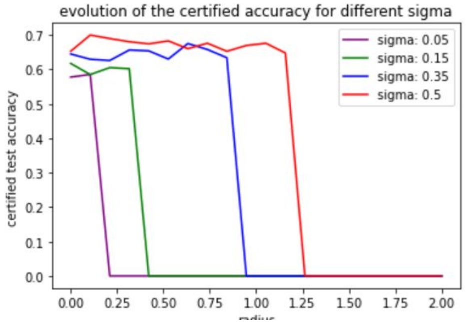
Smooth Classifier

100	class 0
6	class 1
24	class 2
2	class 3

P(x ∈ class 0) >> P(x ∈ class 2)
so classifier returns **class 0**

Certification : Certifying the robustness of the smooth classifier around a radius r





Results

Comparing the different defense mechanisms

accuracy (%)	Basic Classifier	basic classifier + noisy train	adversarial training (PGD-l2)	adversarial training (PGD-l∞)	Mix adversarial training	randomized smoothing
Natural	58.2	**62.16**	54.78	54.41	51.76	59.71
FGSM	3.81	26.17	25.00	28.32	28.41	**38.1**
PGD-l∞	0.29	21.77	20.99	25.58	26.46	**37.68**
PGD-l2	0.29	25.16	23.82	19.33	27.15	**39.8**

Takeaways :

⇨ Just training the model with gaussian noise make it more robust (faster than AT with PGD)

⇨ Randomized network seems to be the more effective and general defence mechanism

⇨ Some defenses work better against a subset of attacks.