ADVERSARIAL ATTACKS AND DEFENSES

Salma ELGHOURBAL
Matteo BARBIERI
Lea AMAR
Hassan EL MANSOURI KHOUDARI

1.

Fast gradient method

Attack and defense

FGSM (Goodfellow et al. 2015)

1- Modèle de base pour la classification des images de CIFAR-10

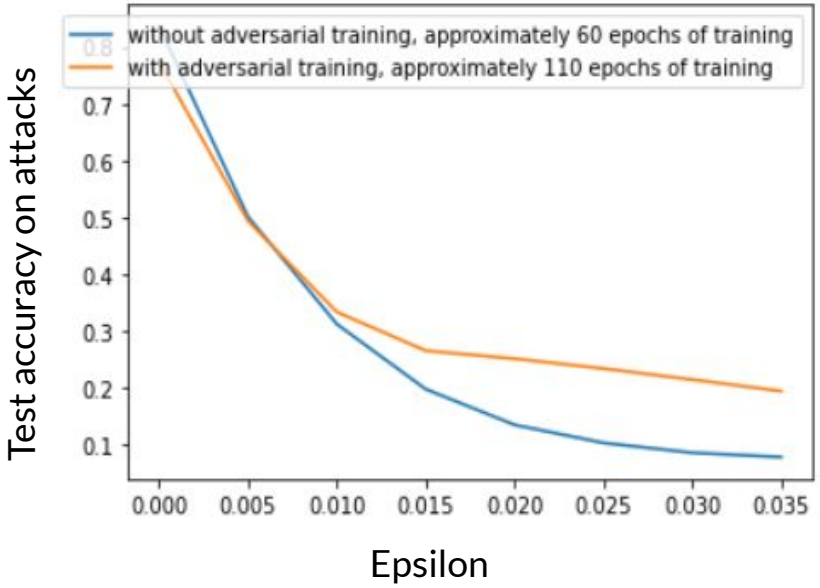
Images test normales	attaques avec $eps = 0.01$	attaques avec $eps = 0.03$
85%	31%	8%

Performance du modèle de base sur une attaque FGSM

2- Modèle entraîné sur des images normales et des images générées par FGSM

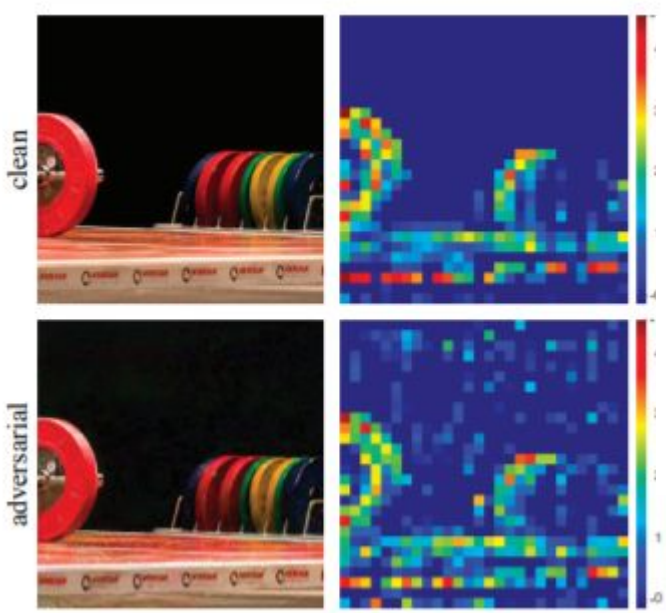
Images test normales	attaques avec $eps = 0.01$	attaques avec $eps = 0.03$
77%	33%	21%

Performance du modèle robuste sur une attaque FGSM

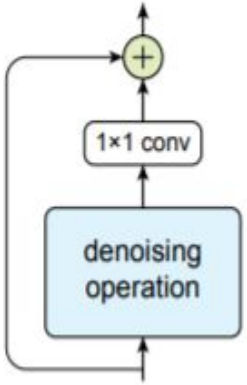


$$x' = x + \epsilon \cdot \nabla_x(J(x, \theta, y))$$

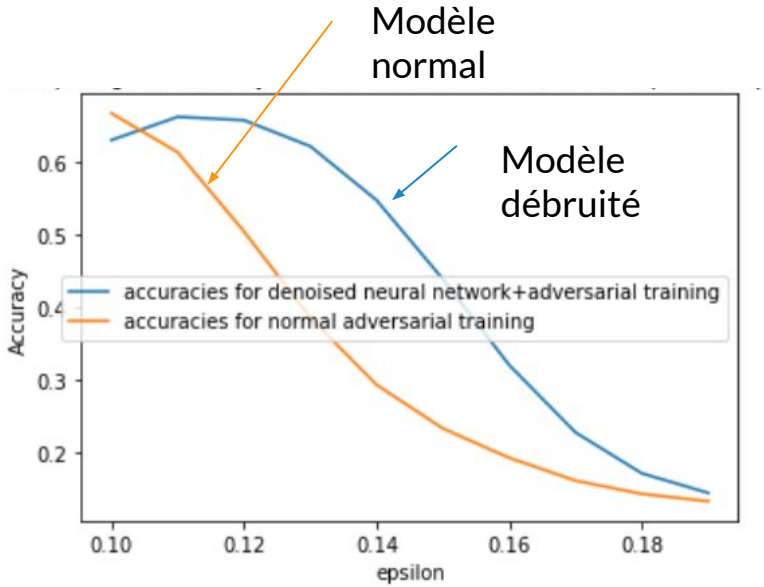
Feature denoising for improving adversarial robustness (Facebook 2019)



1- Noise in the feature maps



2- A denoising block



3-Comparison with a normal model adversially trained

Images 1 and 2 taken from the paper

2.

Projected gradient descent

Attack and defense

Présentation de l'attaque

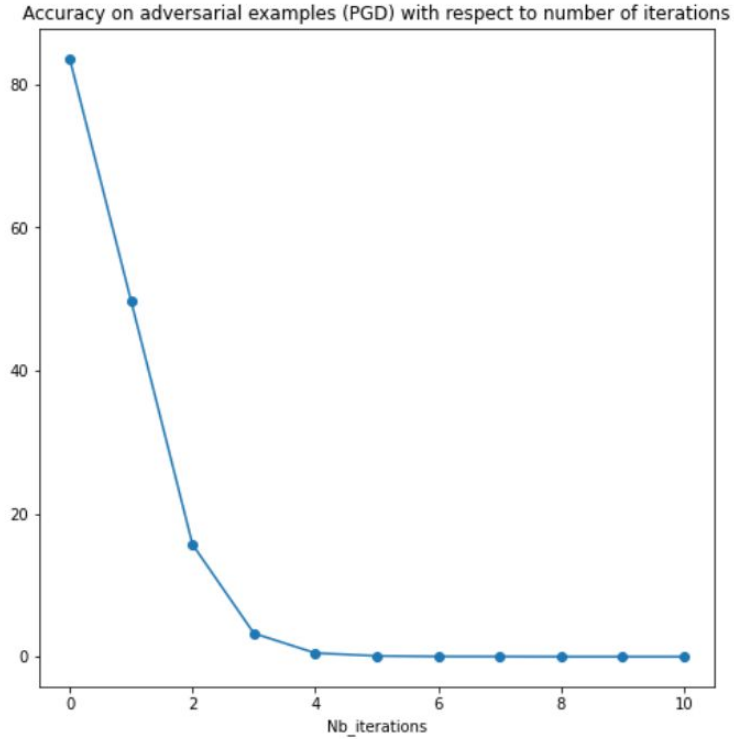
Paramètres utilisées :

- $x_{t+1} = \Pi_{B(0,\delta)}[x_t + \varepsilon \nabla J(x_t, y, \theta)]$
- $\varepsilon = 0.006$
- $\delta = 0.03$

Images test normales	attaques avec $iter = 3$	attaque avec $iter > 5$
83.58%	3.27%	<0.12%

FIGURE 9 – Performance du modèle de base

On préfère avoir de petites perturbations à chaque itération et choisir un nb_iter relativement grand.

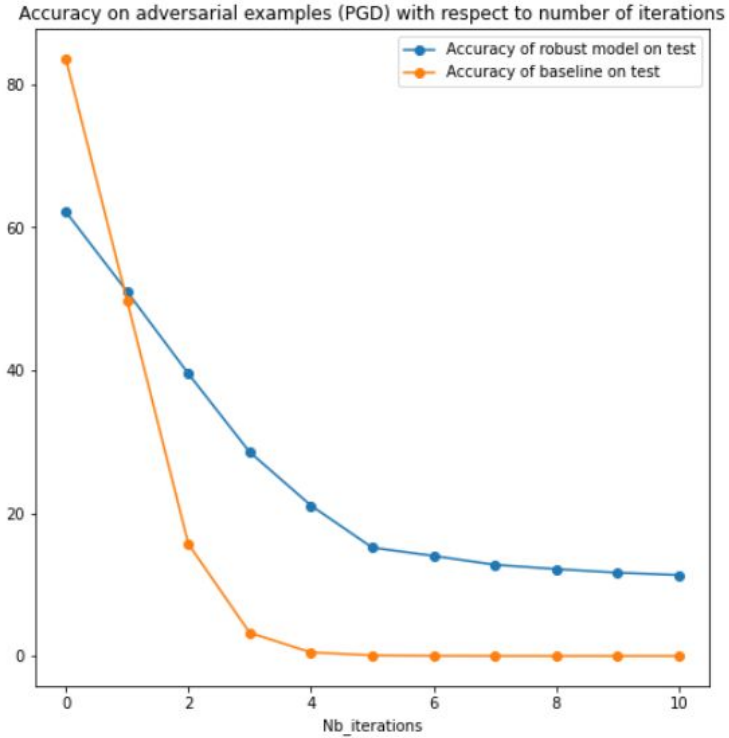


Défense adversariale

$$\tilde{J}(x, \theta, y) = a \cdot J(x, \theta, y) + (1 - a) \cdot J(pgd(x), \theta, y)$$

Images test normales	attaques avec $iter = 5$	attaque FGSM $\epsilon = 0.03$
62%	15%	30%

FIGURE 8 – Performance du modèle robuste



- La précision sur les images originale passe de 80% à 60%, ce qui est normal dans ce cas (nous n'avons pas pu entraîner le modèle aussi longtemps que le modèle de base, donc la différence pourrait en découler)
- Le modèle robuste fonctionne mieux sur les images attaquées que son homologue régulier stagnant à environ 17% pour les itérations> 5, tandis que le modèle de base chute à 0% de précision.
- Le modèle robuste se défend mieux contre les attaques FGSM avec une précision de 30% contre 21% pour la défense utilisant FGSM.

Generative Adversarial Network

Attack and defense

Générer des exemples contradictoires

2 types d'attaques : Targeted ≠Untargeted

- Targeted : image perturbée à classifier dans une classe donnée
- Untargeted: L'image n'est pas correctement labellisée

White-box attack sur le modèle de base

x : image originale

→ Génération d'une perturbation $\mathcal{G}(x)$

→ Discrimant s'assure que $x+\mathcal{G}(x)$ est réaliste

→ Adversarial loss:

$\mathcal{L}$ = Ex log $\mathcal{D}(x)$ + Ex log (1- $\mathcal{D}$ (x+ $\mathcal{G}(x)$))

→ Loss du générateur :

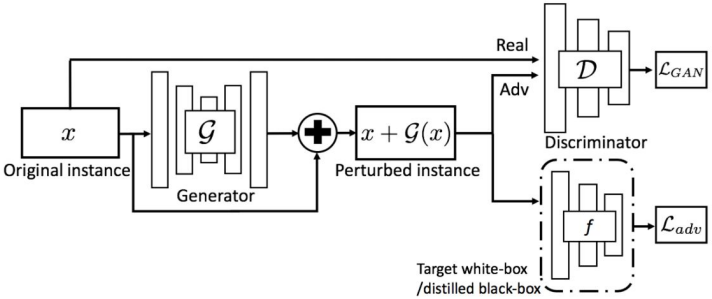
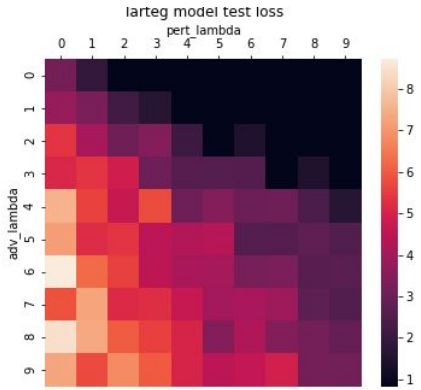
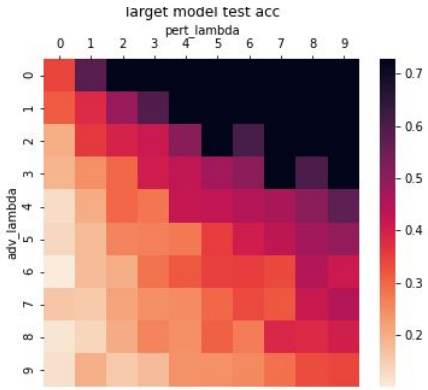


Figure 1: Overview of AdvGAN

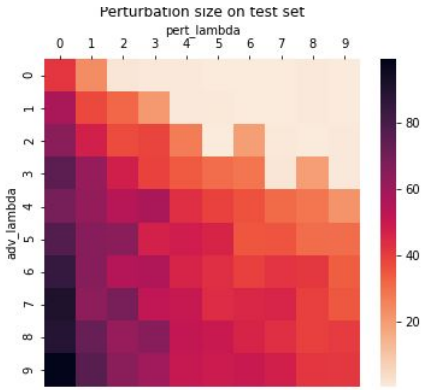
Test des exemples contradictoires



Target model loss



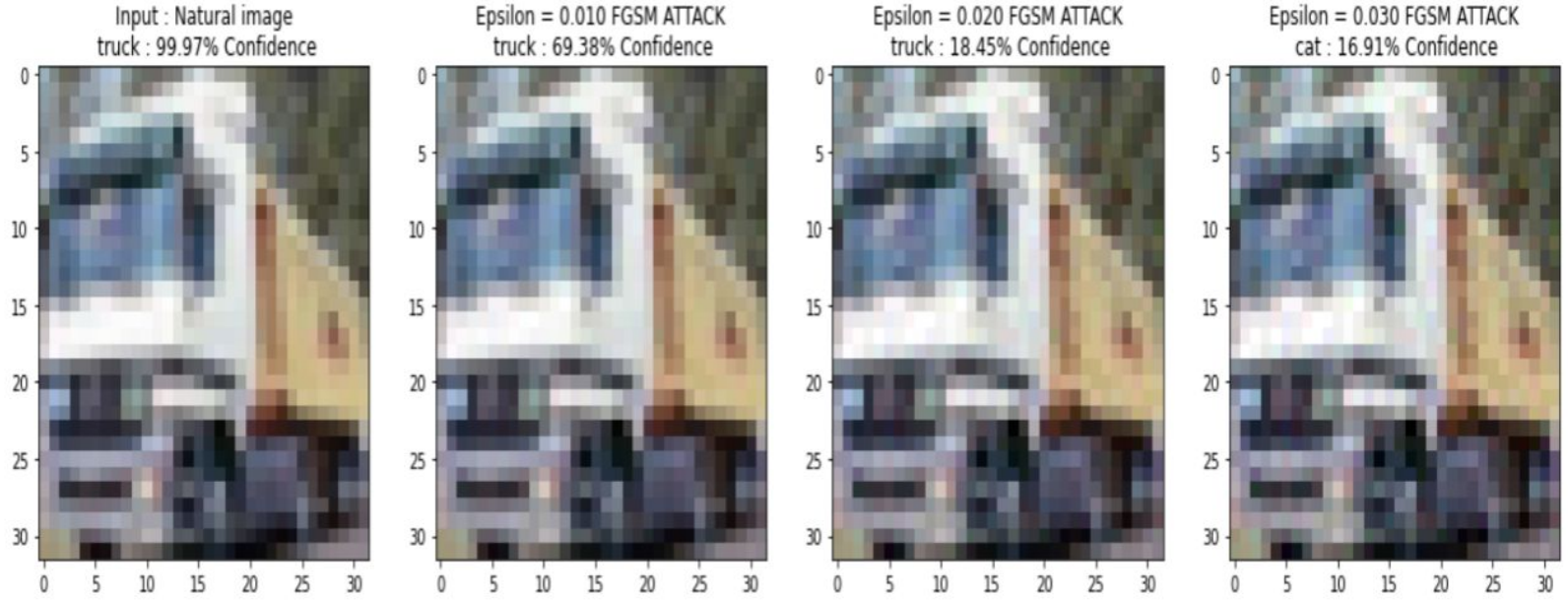
Target model accuracy



Perturbation size

ANNEXE

FGSM (Goodfellow et al. 2015)



Visualisations d'attaques avec les prédictions et confiances du modèle