

# Attaque de réseau en boîte noire

Nicolas Fabiano & Dinh Congminh & Alexis Amzallag

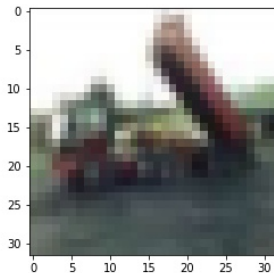
18 décembre 2019

1 Contexte

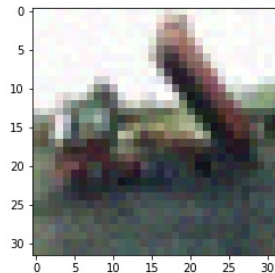
2 FGSM

3 Boîte noire

4 Travail en cours

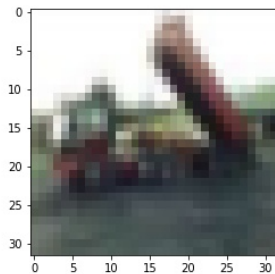


0.07 0.07 0.08 0.02 0.00  
0.04 0.01 0.03 0.14 0.54

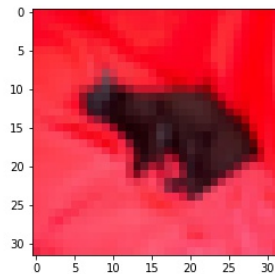


0.29 0.04 0.12 0.03 0.00  
0.02 0.00 0.03 0.26 0.19

( $\epsilon = .03$ )



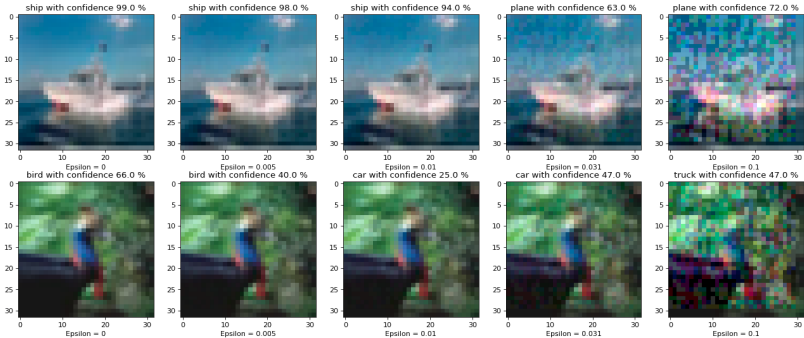
0.07 0.07 0.08 0.02 0.00  
0.04 0.01 0.03 0.14 0.54



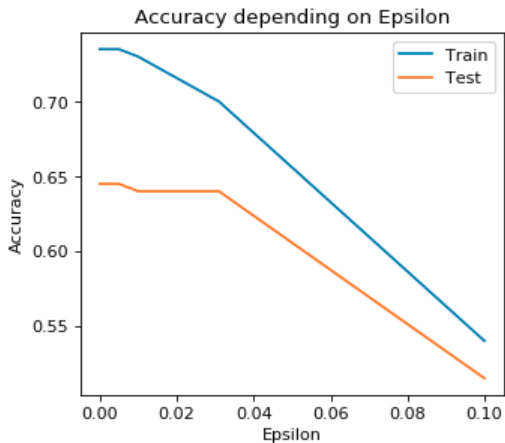
0.82 0.00 0.00 0.16 0.00  
0.01 0.01 0.00 0.00 0.00

# Principe

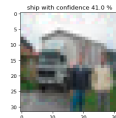
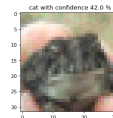
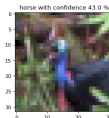
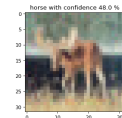
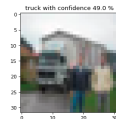
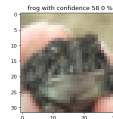
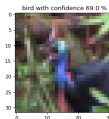
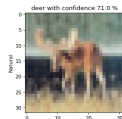
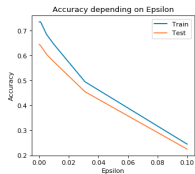
image perturbée = image +  $\epsilon \times \text{sign}(\text{gradient})$



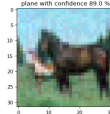
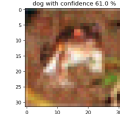
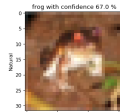
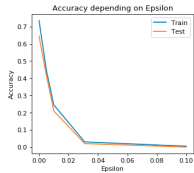
## Précision et exemples avec des attaques aléatoires



# Précision et exemples avec des attaques FGSM partielles

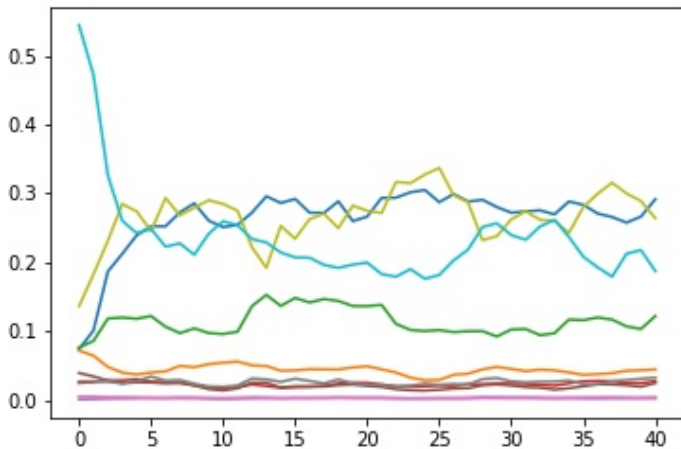


# Précision et exemples avec des attaques FGSM totales

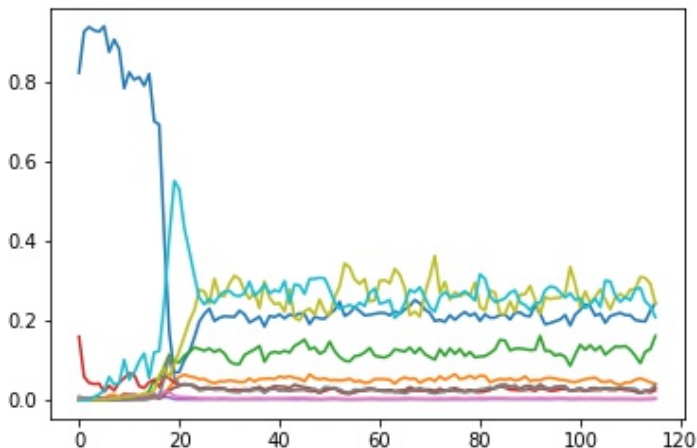




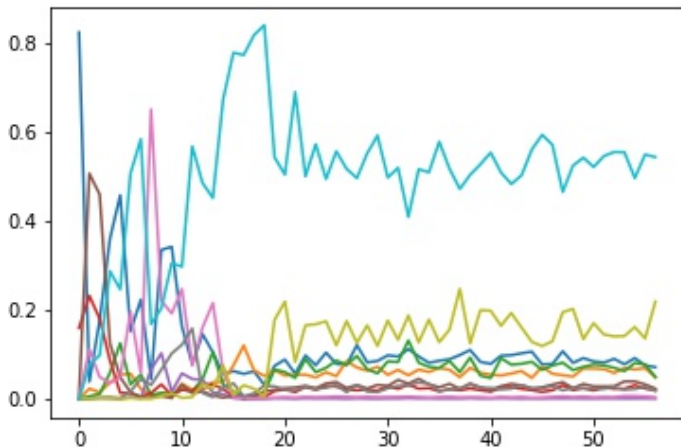
## Information totale



## Information limitée aux scores du top $k$ ( $k = 5$ )

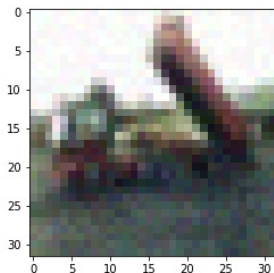


## Information limitée aux rangs du top $k$ ( $k = 5$ )



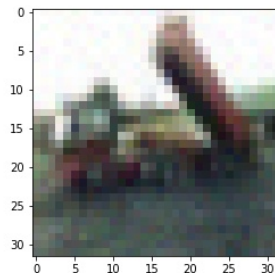
## Comparaison avec la norme $l_2$

Norme  $l_\infty$ ,  $\epsilon = .03$



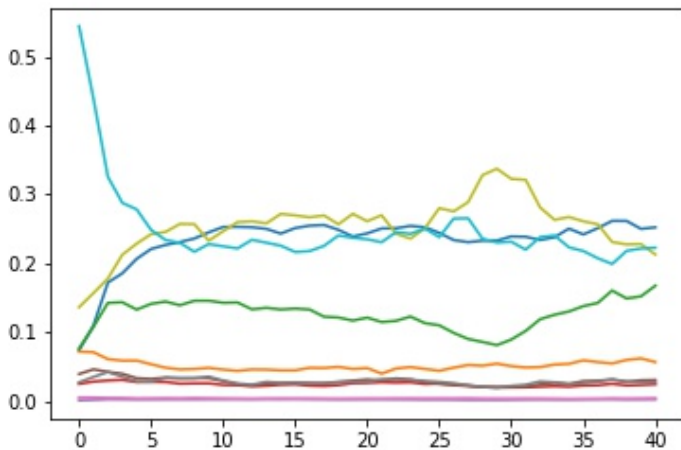
0.29	0.04	0.12	0.03	0.00
0.02	0.00	0.03	0.26	0.19

Norme  $l_2$ ,  $\epsilon = 1$



0.25	0.06	0.17	0.02	0.00
0.03	0.00	0.03	0.21	0.22

## Pour la norme L2



## Une meilleure norme ?

$$\|x - y\| \rightarrow \|\text{grad}(x) - \text{grad}(y)\|$$

## Une meilleure norme ?

$$\|x - y\| \rightarrow \|\text{grad}(x) - \text{grad}(y)\|$$

Ordre de grandeur :

$$\|\text{grad}(x) - \text{grad}(y)\|_{\infty} \leq .02$$

$$\|x - y\|_{\infty} \leq .1$$

## Une meilleure norme ?

$$\|x - y\| \rightarrow \|\text{grad}(x) - \text{grad}(y)\|$$

Ordre de grandeur :

$$\|\text{grad}(x) - \text{grad}(y)\|_{\infty} \leq .02$$

$$\|x - y\|_{\infty} \leq .1$$

Projection ?