Adversarial attacks and robust neural networks

Data Science Lab, project 3
Meriem DEHMAS - Julien SENTUC - Aurèle GOETZ

Double adversarial attack

- Focus on FGSM targeted attacks
- General ideas:
  - The adversarial of an adversarial is surprisingly good
  - Attacking cannot change the groundbase of an image, so the “adversarial signature” of an example will remain
    - environment, colors

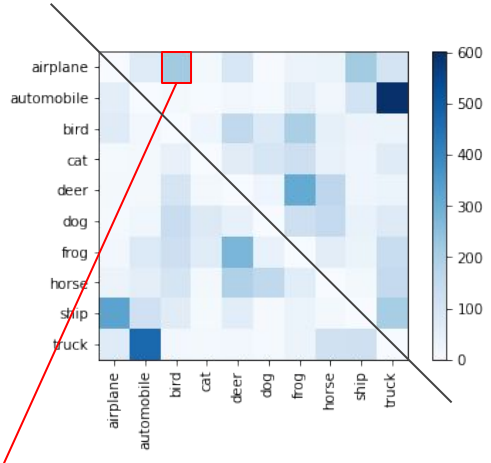


Figure 1: best adversarial class (y) for each original class (x)

Results (attack airplane -> birds)

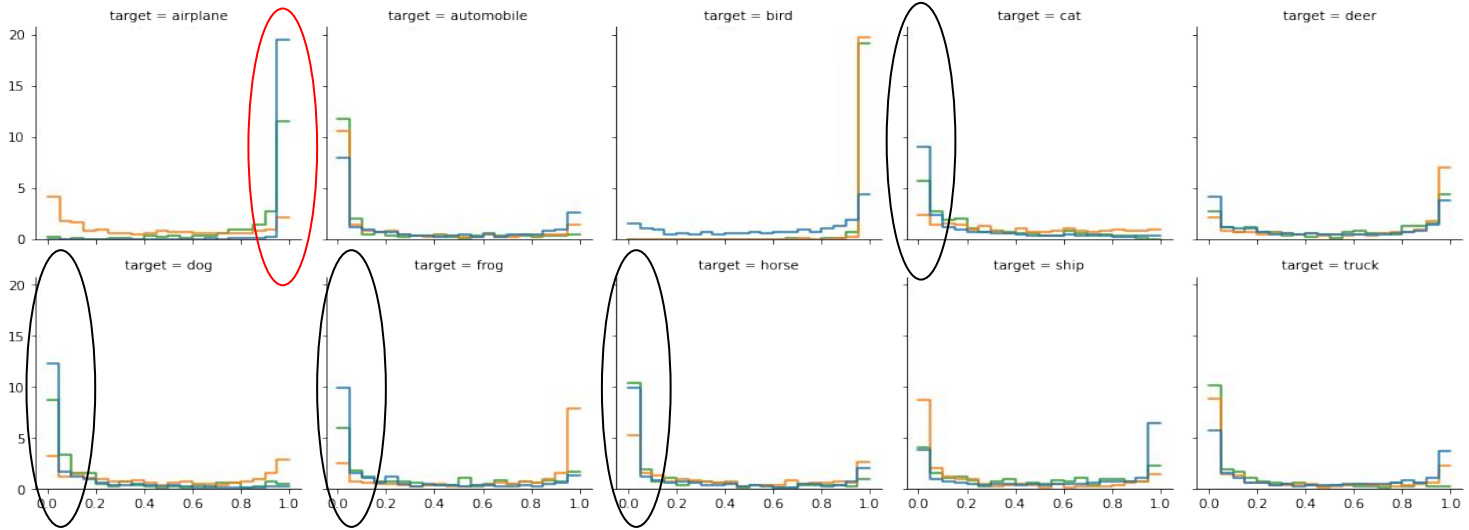


Figure 2: Confidence distribution in all possible targeted class after a targeted attack

original class examples (airplane)

target class examples (bird)

adversarial examples of original class airplane turned into birds

Results (attack automobile -> truck)

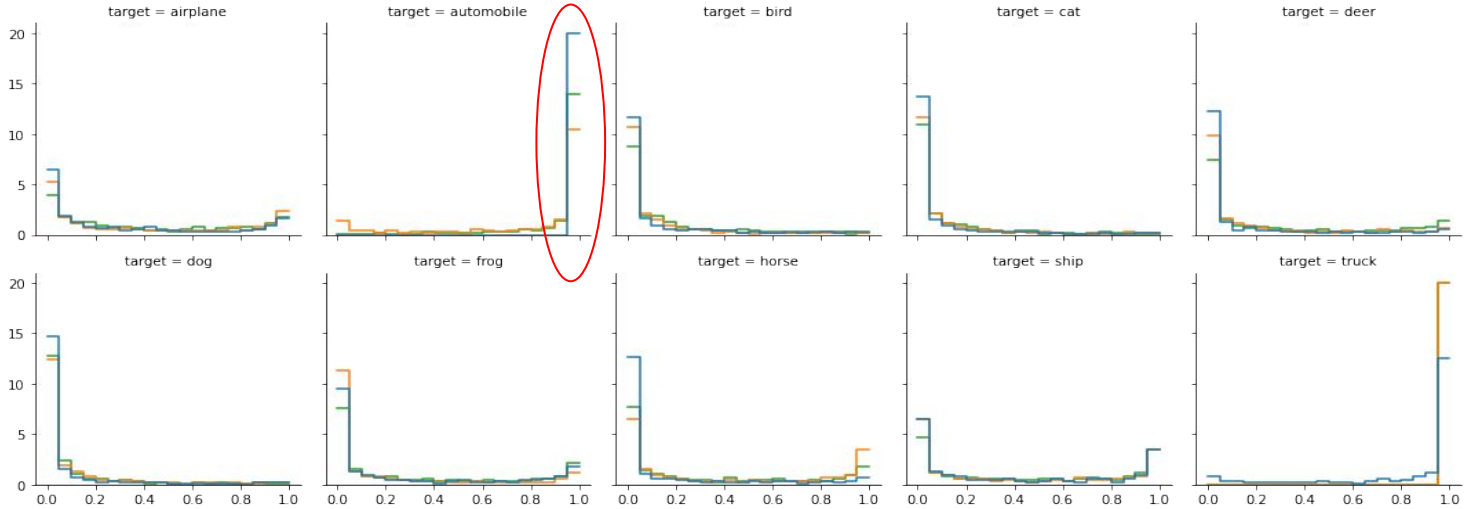


Figure 3: Confidence distribution in all possible targeted class after a targeted attack

original class examples (automobile)

target class examples (truck)

adversarial examples of original class automobile turned into trucks

Inspecting the behavior of examples in latent spaces - KNN

start class : “cat”

end class : “deer”

-------- **after pool 1** --------

[3, 0, 21, **19**, **32**, 13, 10, 0, 2, 0]

[2, 0, 21, **16**, **35**, 11, 12, 1, 2, 0]

-------- **after pool 2** --------

[0, 0, 8, **13**, **62**, 3, 10, 2, 2, 0]

[0, 0, 8, **5**, **71**, 1, 14, 1, 0, 0]

-------- **after pool 3** --------

[2, 0, 5, **27**, **43**, 11, 12, 0, 0, 0]

[0, 0, 8, **6**, **60**, 8, 18, 0, 0, 0]

-------- **after fc1 + relu** --------

[0, 0, 0, **76**, **19**, 5, 0, 0, 0, 0]

[0, 0, 0, **2**, **90**, 2, 1, 5, 0, 0]

-------- **after fc2** --------

[0, 0, 0, **83**, **11**, 5, 0, 1, 0, 0]

[0, 0, 0, **0**, **95**, 2, 0, 3, 0, 0]

Inspecting the behavior of examples in latent spaces

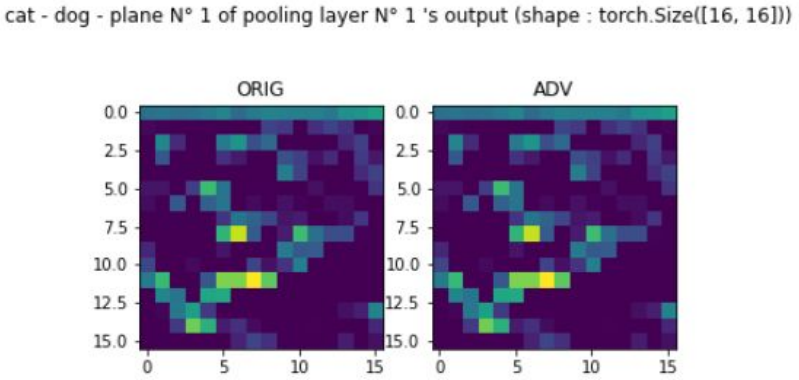
Meaning: The 1st fully connected layer has high weights for the inputs' entries which are precisely different for adv. and orig. vectors.

Try to see to what correspond these entries.

```
7789-cat-horse
[ 1 29]
7795-deer-horse
[ 1 29]
7820-cat-ship
[ 1  9 13 29]
7813-airplane-ship
[ 0  1  4  5  8  9 13 18 24 29 36 37]
```

but:

```
7792-automobile-automobile
[ 1 29]
```



Distance between adv.’ and predictions’ latent vectors through the network