Trustworthy ML - writeup 1

Eden Lumbroso 208996587

April 2023

1 White-box vs. query-based black-box attack

1.1 Benign accuracy

Model benign accuracy is: 87.50%

1.2 White-box attack

- untargeted success rate: 98.50%
- targeted success rate: 93.50%

1.3 Black-box attack

- **Untargeted black-box attack (momentum=0.00):**
  - success rate: 93.50%
  - median(# queries): 3600
- **Targeted black-box attack (momentum=0.00):**
  - success rate: 76.50%
  - median(# queries): 6400
- **Untargeted black-box attack (momentum=0.90):**
  - success rate: 96.50%
  - median(# queries): 2800
- **Targeted black-box attack (momentum=0.90):**
  - success rate: 86.00%
  - median(# queries): 4400

Analysis: It seems from the experiments, that black-box attacks are weaker than white-box attacks (which is not at all surprising) but not by much. Black-box attacks are much more computationally expensive than white-box attacks. With momentum, we see improved success rate, and also reduced number of queries. This is probably because there is a correlation between consecutive gradients, which means that by moving in the direction of average pass gradients, we are probably moving in a good direction.

2 Transferability-based black-box attack

- **Test accuracies:**
  - Model 0: 0.8750
  - Model 1: 0.8250
  - Model 2: 0.7900

Table 1: Untargeted attacks' transferability

Source → Target	**Model 0**	**Model 1**	**Model 2**
Model 0	0.985	0.57	0.535
Model 1	0.685	0.965	0.59
Model 2	0.59	0.545	0.95

Table 2: Targeted attacks' transferability

Source → Target	**Model 0**	**Model 1**	**Model 2**
Model 0	0.96	0.295	0.275
Model 1	0.395	0.895	0.275
Model 2	0.355	0.25	0.86

- **Ensemble attacks' transferability from models 1+2 to model 0:**
  - untargeted attack: 0.7450
  - targeted attack: 0.4900

The transferability of the targeted and untargeted attacks improved while using ensemble attacks. In the untageted case, the success rate improved from 68.5% and 59% when transferring from model 1 and 2 respectively, to 74.5% when transferring from their ensemble. In the targeted case, the success rate improved from 39.5% and 35.5% to 49% on the ensemble. One possible explanation to this improvement is that the adversarial attack is forced to find more generic perturbations, that do not rely on specifics of one model, because we use the expected loss of 2 models.

3 Bit-flip attacks

- The maximum RAD is: 0.7273
- Only 2.62% of bits lead to $> 15\%$ RAD when flipped.
- The second bit has the highest RAD by far. Flipping it from 1 to 0 reduces the exponent by 128. This means the exponents flips from positive to negative when the bit is flipped, which in turn flips the value of the weight from above 1 to below 1 (is positive). The reason this bit has such high RAD is most weights are small (close to 0) and all of them are betweem -0.5 and 0.5 approximately (on the checked model). This means that the second bit is always 0. flipping any other bit will keep the value between 2 and -2, and thus won't change it by much. Flipping the second bit on the other hand will make the weight explode to values of 10^30.

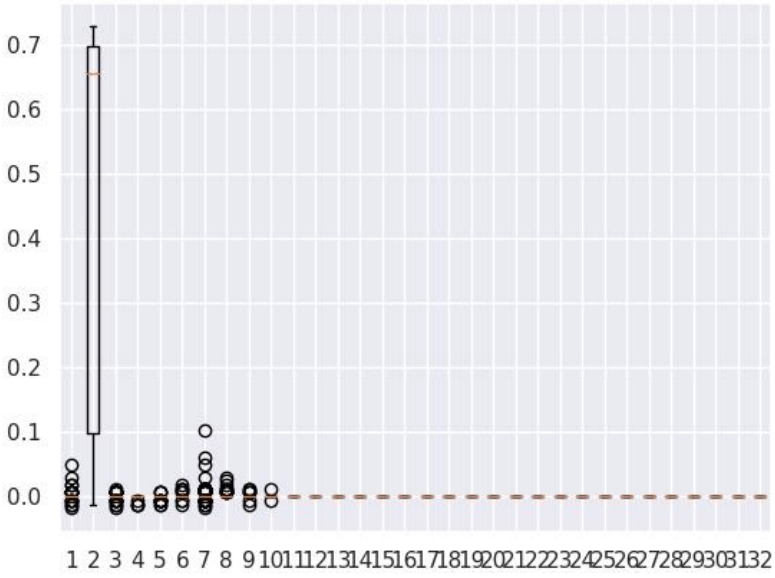

Figure 1: Bit index versus RAD

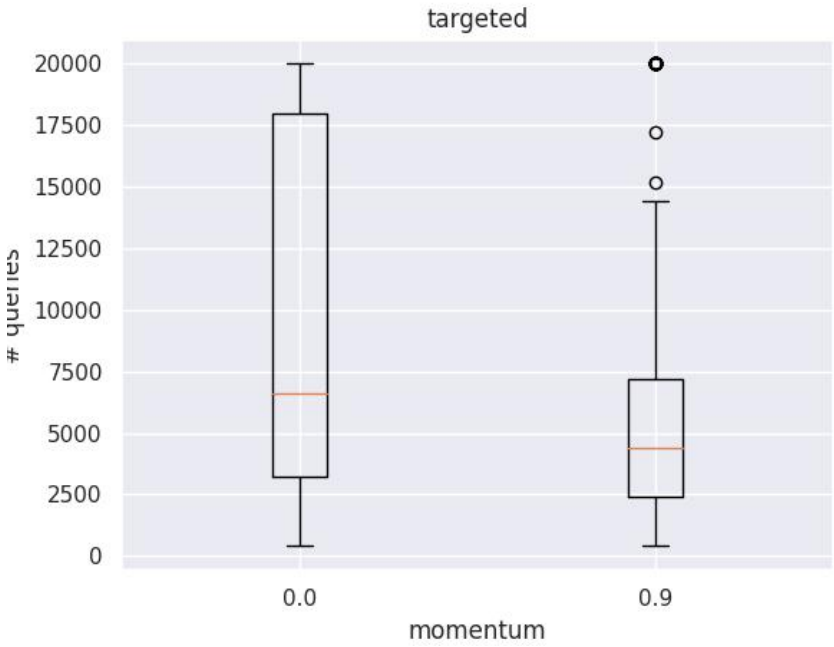

Figure 2: Targeted black-box attack: Number of queries

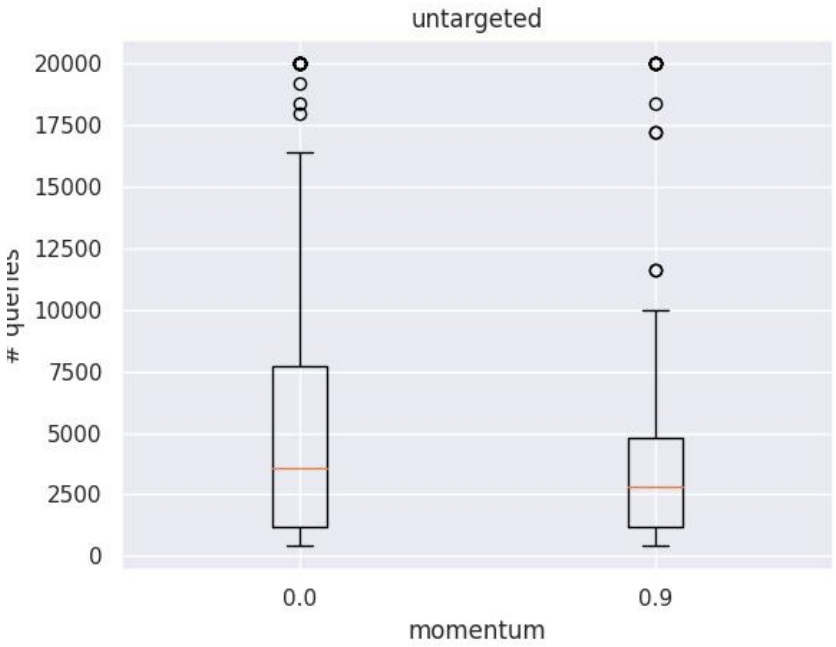

Figure 3: Untargeted black-box attack: Number of queries