March 21, 2023

Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks

Croce et al.

1 Summary

1.1 Auto-PGD: A budget-aware step size-free variant of PGD

a APGD algorithm (only free variable is the budget N_{iter})

- Gradient step: classic algorithm + momentum term
- Step size selection: For every checkpoint of iteration, step size is halved if {increasing update steps less than 0.75 fraction after the last checkpoint} or {step size was not reduced at the last checkpoint and no improvement}
- Restarts from the best point: restart at the point attaining the highest objective f_{max} so fa at a checkpoint
- Exploration vs exploitation: period length is reduced in each step.

b Comparison of APGD to usual PGD

- Attack models on MNIST and CIFAR-10
- APGD shows better CE loss and robust accuracy for all iteration budgets

1.2 An alternative loss

a Difference of Logits Ratio Loss: shift and rescaling invariant

$$\mathrm{DLR}(x, y) = -\frac{z_y - \max_{i \neq y} z_i}{z_{\pi_1} - z_{\pi_3}}$$

b APGD versus PGD on different losses

- APGD outperforms PGD/PGD with momentum regardless of the employed loss
- DLR loss improves upon CE loss, comparable to CW loss with less severe failure

1.3 AutoAttack: an ensemble of parameter-free attacks

Combine $APGD_{CE}$, $APGD_{DLR}$ with FAB and Square Attack to form the ensemble AutoAttack. AutoAttack has diversity in l_p-ball, FAB and Square Attack have few parameters which generalize wellacross models and datasets.

a Benefits

- There exist classifiers for which some attack fails while at least one of them works
- On the same model diverse attack may succeed on different point

1.4 Experiments

a Deterministic defenses

In all but one case AutoAttack achieves a lower robust accuracy than reported in the original papers, and the improvement is larger than 10% in 13 out of 49 cases, larger than 30% in 8 cases

b Randomized defenses

AutoAttack achieves always lower robust accuracy than reported in the respective papers, with APGDCE being the best performing attack

c Analysis of SOTA of adversarial defenses

2 Pros

2.1 Auto Attack is tested with many(50) classifiers.

2.2 The Author provided very detail numeric evaluation results for their target models.

3 Cons

3.1 Explanation of setting various hyper parameters such as step size reducing rate, check points, and momentum is weak.

3.2 It would be better if the author mentioned about whether each idea for Auto-PGD is from others' work or their own new idea.

3.3 The author argue that the different attack method can succeed on different points even if they have similar robust accuracy, but there are no evaluation on this part.

4 Meaning of the paper

Propose AutoAttack which can be used as minimal test for any new defense.

AutoAttack would almost always have provided a better estimate of the robustness of the models than in the original evaluation

AutoAttack is also suitable for the evaluation of randomized defenses

Propose APGD algorithm which automatically adjust learning steps.

Propose Difference of Logits Ratio loss.

5 Discussion and Questions

It seems that the ideas(Gradient step, Step size selection, Restarts from the best point, Exploration vs exploitation) adopted in Auto-PGD is applicable in generous model training. Do these ideas come from model training? If not, these ideas can be adopted to NN model training?