Adversarial attacks are a large problem for security sensitive applications!

Evaluation of adversarial attacks

Previous work:
Overly simplified evaluation of adversarial attacks
Focus is only on 100% error rates. Even 50% error rate is problematic in the real world.

We propose:
Evaluate attacks with different hyperparameters to obtain accuracy-perturbation curves

Resulting accuracy-perturbation curves show how the classifiers relative accuracy drops with larger perturbations.

Experimental setup

Adversarial example methods:
- FGSM
- BIM
- DeepFool
- AutoPGD

Image classifiers
- RBFN
- Logistic regression
- 2x CNN

Datasets
- MNIST
- CIFAR-10

Human evaluation survey

Accuracy-perturbation curves

Increased adversarial perturbation is more likely to confuse.

"Efficiency" curves of classifiers response to adversaries of varying perturbation.

Classifier B beats A!
Its accuracy dropped at higher perturbation.

Classifier D beats C!
Scattered plots are compared using min-max wrap.

Min wrap ➔ worst-case

Results

Adversarial training

CNN1BIM is adversarially trained.
Curves show the increased robustness.

Conclusion

Accuracy-Perturbation curves give stronger insight into the efficiency of the attack or defence.

A useful tool for adversarial attack evaluation!