University of Ljubljana
Faculty of Computer and Information Science

Accuracy-perturbation Curves for Evaluation of Adversarial Attack and Defence Methods

Jaka Šircelj and Danijel Skočaj
jaka.sircelj,danijel.skocaj@fri.uni-lj.si

Adversarial attacks are a large problem for security sensitive applications!

Classified as Brad Pitt

M. Sharif et al., in ACM TOPS, 2019

Classified as Speed limit 45

K. Eykholt et al, in CVPR, 2018

Attacks call not only for defence but also for strong evaluation!

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-peturbation 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
- MLP
- 2x CNN

Datasets

- MNIST
- CIFAR-10

Human evaluation survey

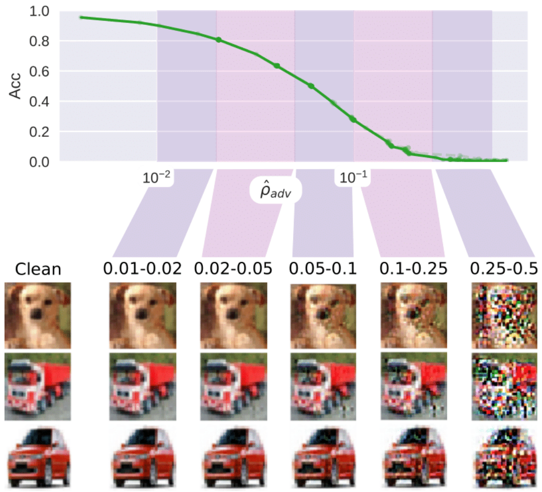
Accuracy-perturbation curves

Increased adversarial perturbation is more likely to confuse.

Accuracy-perturbation curves

↓

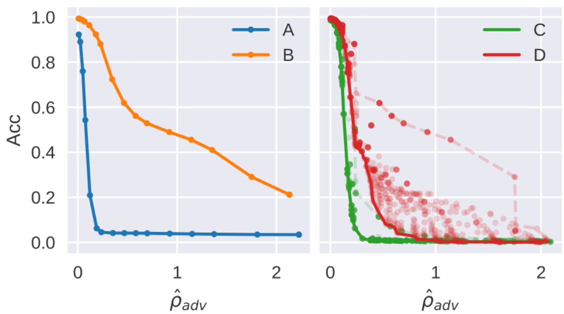
"Efficiency" curves of classifiers response to adversaries of varying perturbation.



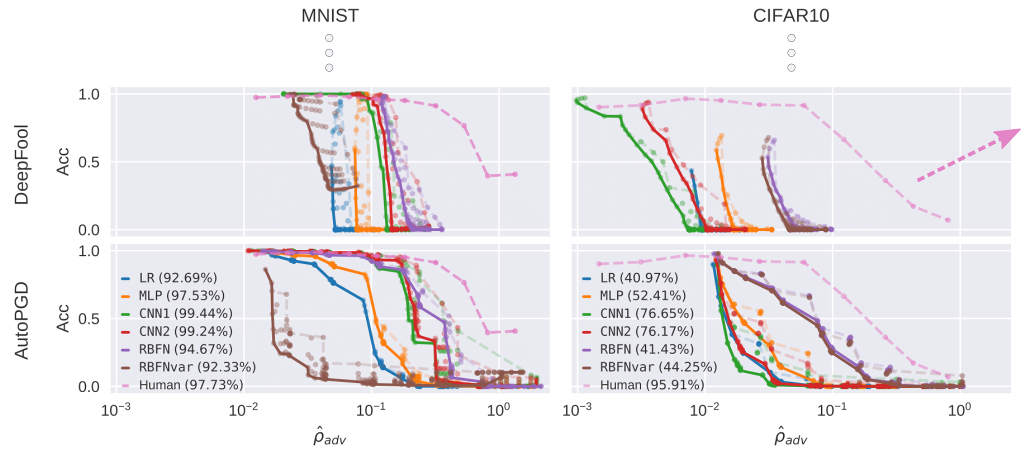
Classifier B beats A!
Its accuracy dropped at higher perturbation.

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

Min wrap → worst-case



Results

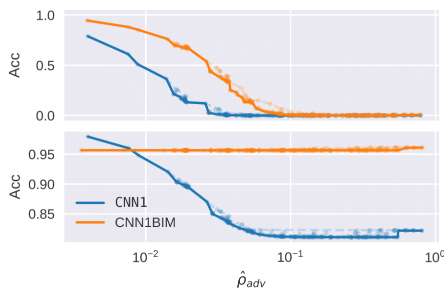


Humans win!

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 **usefull tool** for adversarial attack evaluation!