Bamboo: Ball-Shape Data Augmentation Against Attacks from All Directions

Huanrui Yang[1], Jingchi Zhang[1], Hsin-pai Cheng[1],
Wenhan Wang[2], Yiran Chen[1], Hai Li[1]

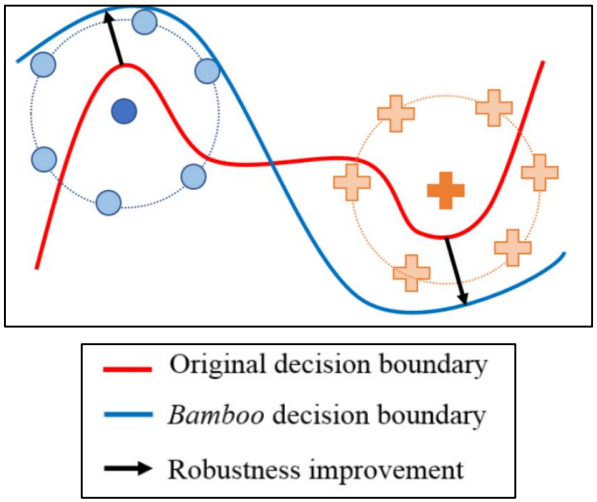
[1]Duke University [2]Microsoft

Motivation

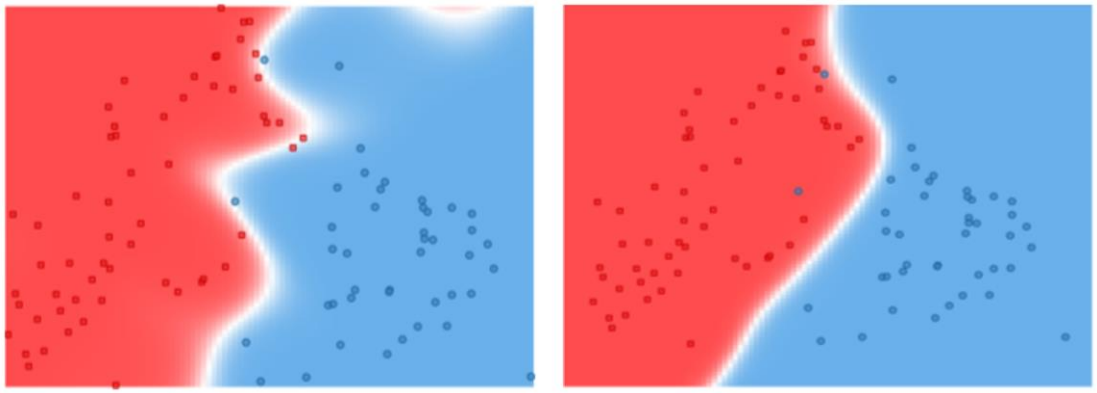
- DNN models are vulnerable to adversarial attacks
  - Small perturbation in the input can ruin output result
- Adversarial training
  - Training with the adversarial example generated from a **known** attack
  - May not work under unknown attacks
- Optimization based method
  - Optimizing a **min-max problem** to generate "worst" adversarial example and train model simultaneously
  - Costly and unstable to optimize
- Need a method that can **efficiently** improve the overall robustness **without knowing** the attack to be faced
- Can be considered as a special case of increasing model generalizability $\longrightarrow$ Data augmentation

Method and Intuition

- Increasing robustness against perturbation
  - → Moving the decision boundary away from data points
- Considering the low-curvature property of DNN's decision boundary*, we propose to **uniformly** sample the augmented data **on the surface** of a fixed-radius ball



Toy example: behave as we expect

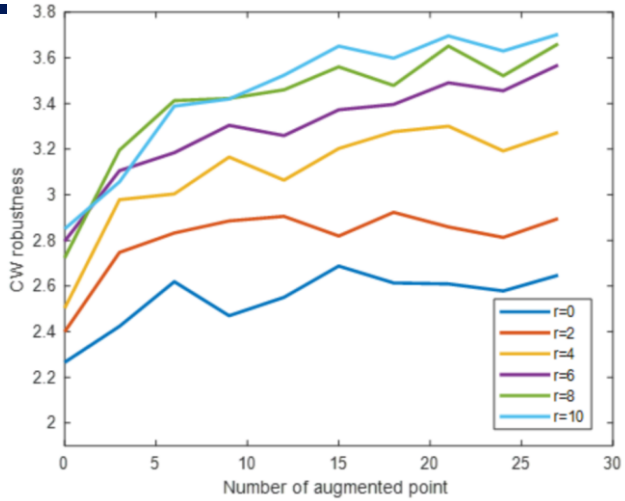


(a) Without data augmentation (b) *Bamboo* data augmentation

*Alhussein Fawzi, Seyed Mohsen Moosavi Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

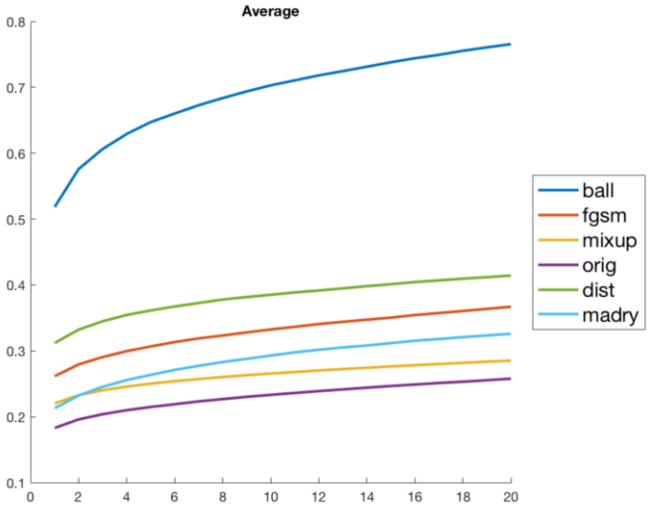
Results

- Effect on model robustness
  - Larger ball radius and larger amount of augmented points leads to higher robustness against CW attack*



*Nicholas Carlini and David Wagner.Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017. .

- Effect on distance to decision boundary
  - Empirically evaluate the distance between data points and decision boundary along random orthogonal directions
  - Figure shows the top-20 smallest distances averaged across MNIST test set
  - Our method achieves the largest distance on both MNIST and CIFAR-10



- Achieve better performance comparing to previous defending methods against multiple types of attack, see paper for details and more results