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

Huanrui Yang¹, Jingchi Zhang¹, Hsin-pai Cheng¹, Wenhan Wang², Yiran Chen¹, Hai Li¹

¹Duke University ²Microsoft
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 → 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


Results

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


- **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