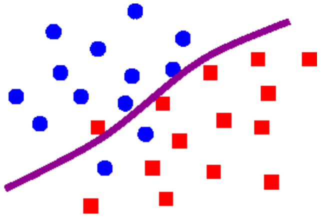
Understanding Adversarial Examples Through Deep Neural Network's Classification Boundary and Uncertainty Regions

Bowei Xi
Department of Statistics
Purdue University
xbw@purdue.edu

Many Unanswered Questions about DNN

What is the shape of DNN classification boundary? A popular description –



Discrepancy between the established generalization error bounds for DNN, $O(\frac{c(depth,width)}{\sqrt{n}})$, and the existence of adversarial examples.

Where are the regions containing adversarial examples?

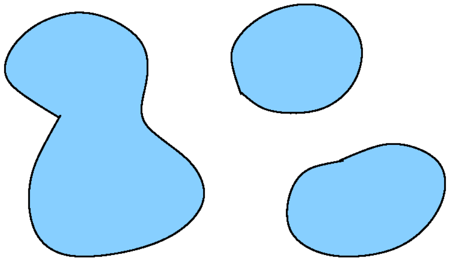
Adaptive attacks can successfully break a defense strategy.

Classification Boundary for One DNN Model

DNN function $M(W) = c$. Need to train multiple DNN models to establish the classification boundary for the attacked model.

The DNN model structure must strictly remain the same, using different intial random seeds.

DNN Uncertainty Regions: A bounded region where at least two DNN models disagree on the hard labels of the points inside the region.

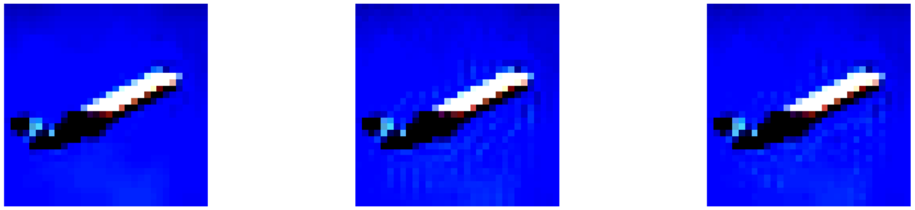


CIFAR10 MobileNet Experiment

Uncertainty Region Construction: Generate sufficient amount of adversarial examples using an attack algorithm; the region spanned by the perturbed dimensions with the interval size not close to 0.

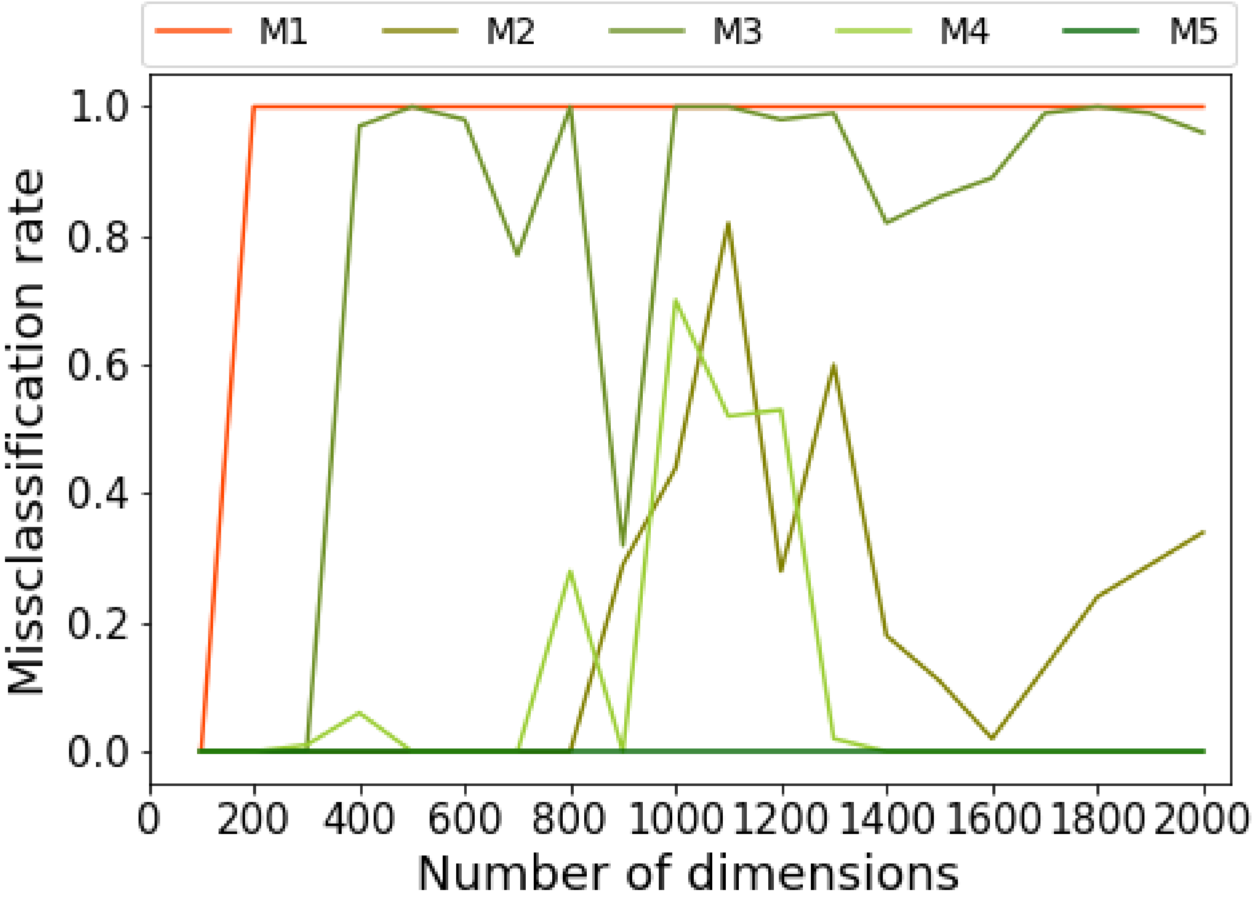
Re-train 5 MobileNet on CIFAR10, the misclassification rates range from 0.0727 to 0.0767 on clean test data.

Attack an airplane image using BIM L_2 attack. Left to right: Original clean; misclassified as deer under BIM attack; misclassified as deer through sampling.



CIFAR10 MobileNet Experiment

BIM L_2 attack perturbed 3017 out of 3072 dimensions.



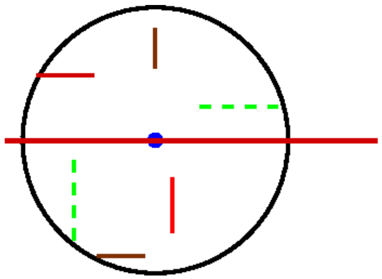
DNN Boundary Surrounding A Clean Image

Type 1 regions (solid): Adversarial examples mostly mis-classified by the attacked model, but correctly classified by some other DNN.

Type 2 regions (solid): Adversarial examples mis-classified by all the trained DNN models.

Type 3 regions (dashed): Adversarial examples correctly classified by the attacked model but misclassified by some other DNN model.

Type 1 and 2 regions are part of the attacked model's classification boundary; type 1 and 3 regions are its uncertainty regions; type 2 regions are the transferable adversarial regions.



Conclusion

DNN classification boundary is highly fractured, unlike other classifiers.

Transferability of adversarial examples is not universal.

We perturb far fewer pixels and generate a lot more adversarial examples through sampling.

Conjecture 1: The union of the uncertainty and transferable adversarial regions containing adversarial examples has zero probability mass.

Conjecture 2: DNN function is discontinuous at the boundary of these regions, and may be discontinuous inside some of these regions.