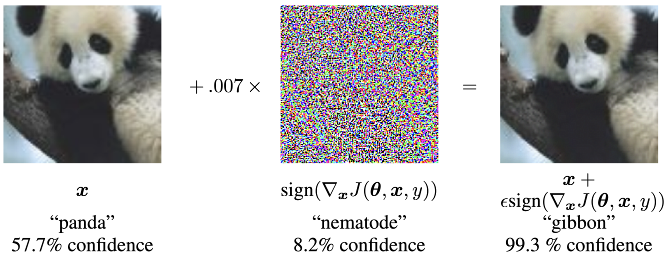
VERIFYING THE CAUSES OF ADVERSARIAL EXAMPLES

Honglin Li,[1,4] Yifei Fan,[2,3] Frieder Ganz,[3] Anthony Yezzi,[2] and Payam Barnaghi[1,4]

[1]Department of Brain Sciences, Imperial College London, [2]School of Electrical and Computer Engineering, Georgia Institute of Technology, [3]Adobe, and [4]Care Research and Technology Centre, The UK Dementia Research Institute.

INTRODUCTION

Adversarial examples [1] remain a critical issue in computer vision, which hinders the industry from building robust explainable real-world applications.



CONTRIBUTION

Verification of several hypotheses regarding the **causes** of adversarial examples through carefully-designed **controlled experiments**.

- geometric factors: direct causes
- statistical factors: magnifier for high confidence

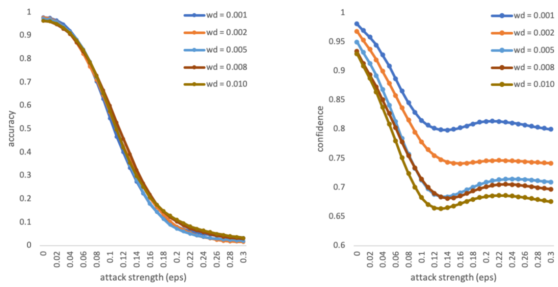
HYPOTHESES AND VERIFICATION

General evaluation metric:

- Evaluation at both the accuracy and confidence levels
- Weak untargeted FGSM attack for better illustration

Hypothesis A: Linearity of the classifier

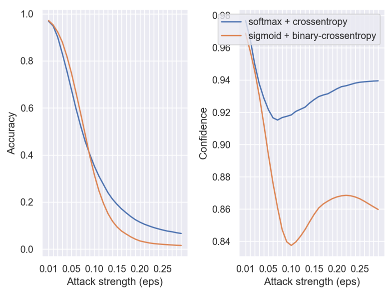
Linear coefficients can lead to high prediction confidence.



Higher weight decay (L_2 normalization) means smaller absolute values in linear coefficients.

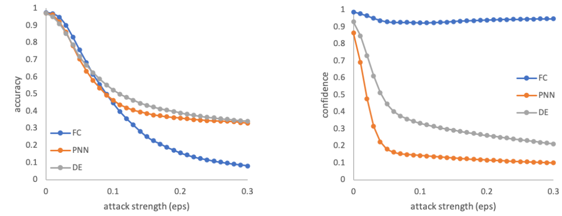
Hypothesis B: One-sum probability space

High confidence is assigned once all other possibilities are ruled out.



Sigmoid + binary-crossentropy break the one-sum output constraint.

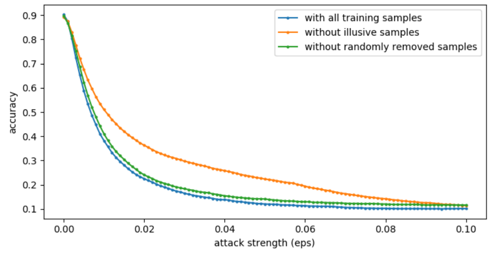
Hypothesis C: Combination of linearity and one-sum



The proposed PNN and DE heads [2] can simultaneously remove linearity and break the one-sum constraint (elaborated in Section IV).

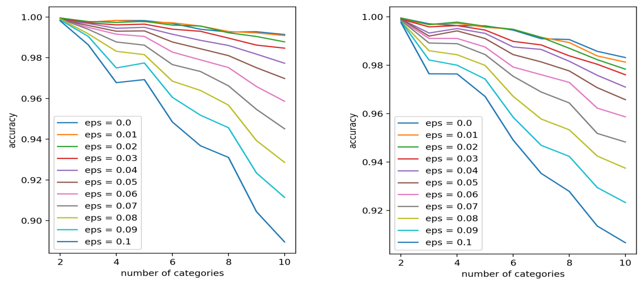
Hypothesis D: Path-connected regions

Uncertain "bridges" are created to connect samples of the same category in a path-connected manner.



Fewer hard illusive samples results in fewer uncertain "bridges."

Hypothesis E: Excessive number of target categories

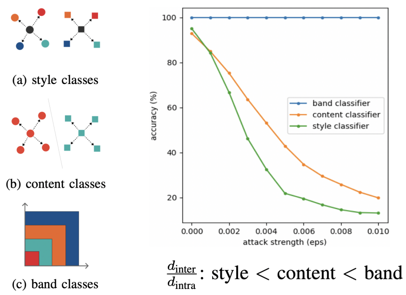


additive mode: including all available training samples

constant mode: 10,000 training samples (balanced)

More target categories result in less robust classifiers.

Hypothesis F: Geometry/entropy of input spaces



Robustness positively correlated to the ratio $d_{\text{inter}}/d_{\text{intra}}$.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] H. Li, P. Barnaghi, S. Enshaeifar, and F. Ganz, "Continual learning using task conditional neural networks," *arXiv preprint arXiv:2005.05080*, 2020.