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 Engineering, Georgia Institute of Technology
[3]Adobe Inc.
[4]Care Research and Technology Centre, The UK Dementia Research Institute

Introduction

- Studying the causes of adversarial examples is important but difficult
  - Requires thorough examination of the entire proximity of an input sample in a high-dimensional image space

- Thoughtful strategies are proposed to indirectly justify hypotheses and observe the geometry of input spaces [2, 3]

- This paper: verifying hypotheses on the causes of adversarial examples via carefully-designed controlled experiments

Popular explanations

- Low-probability “pockets” in the manifold [1]
  - Input spaces are not dense
- Model linearity [4]
  - $w^T(x + \Delta x)$ differs significantly from $w^T x$
- Test-error in additive noise [5]
- Non-robust features [6]
- Other geometric interpretations
  - Boundary-tilting perspective [7], geometry of $\mathbb{R}^n$ with the L_0 metric [8], no category for “don’t know” and data distribution is not concentrated [9]

• Low-probability "pockets" in the manifold [1]
  • Input spaces are not dense
• Model linearity [4]
  • $w^T(x + \Delta x)$ differs significantly from $w^T x$
• Test-error in additive noise [5]
• Non-robust features [6]
• Other geometric interpretations
  • Boundary-tilting perspective [7], geometry of $\mathbb{R}^n$ with the L_0 metric [8], no category for "don't know" and data distribution is not concentrated [9]

Contributions

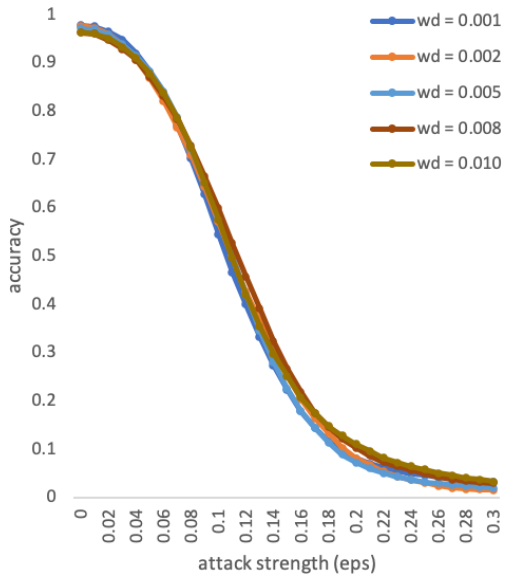
- Verify (or partially verify) several hypothesis on the causes of adversarial examples through carefully-designed controlled experiments.

- Review and collection of explanations on adversarial examples

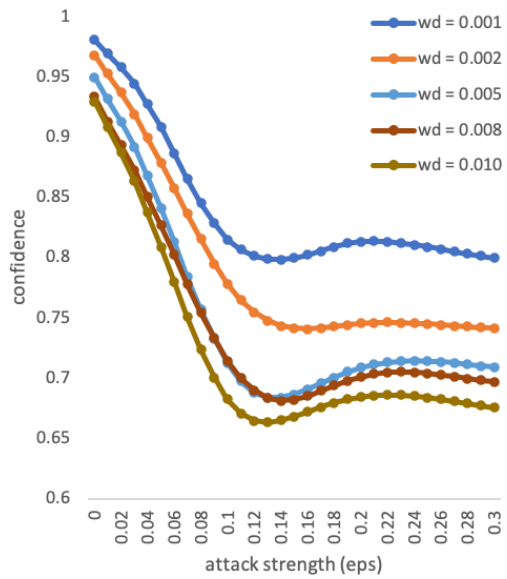
(no SOTA attacks or defenses)

Hypothesis A: model linearity

- $w^T(x + \Delta x)$ can differ significantly from $w^T x$
- Reduce linearity with L_2 norm (i.e., weight decay)



(a) accuracy

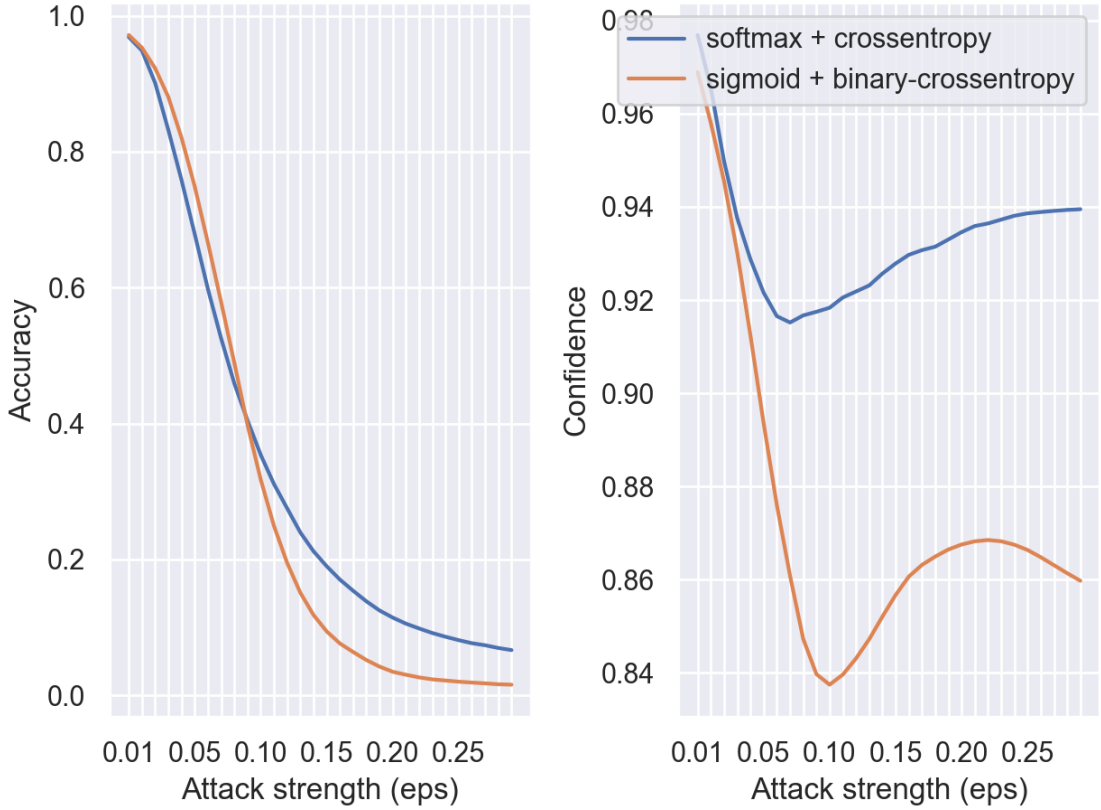


(b) confidence

- Not conclusive at the accuracy level
- Clear correlation at the confidence level

Hypothesis B: one-sum probability constraint

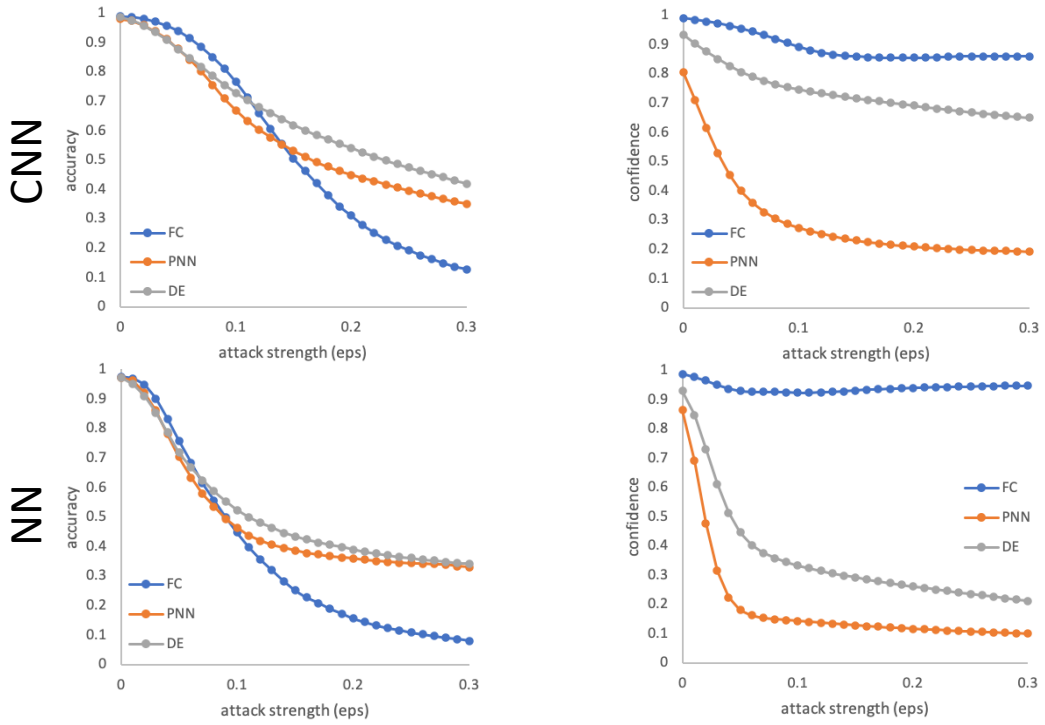
- High confidence results from the constraint that all probabilities must add up to 1



- Lift one-sum constraint
  - Sigmoid + binary-crossentropy
  - Lower confidence
- Two stages
  - Probability decrease at correct classes (boosted by one-sum constraint)
  - Probability increase at incorrect classes (hindered by one-sum constraint)

Hypothesis C: linearity + one-sum constraint

- Combination of the previous two causes leads to a stronger impact
- The proposed MLP-PNN (i.e., PNN) and DE can lift both constraints



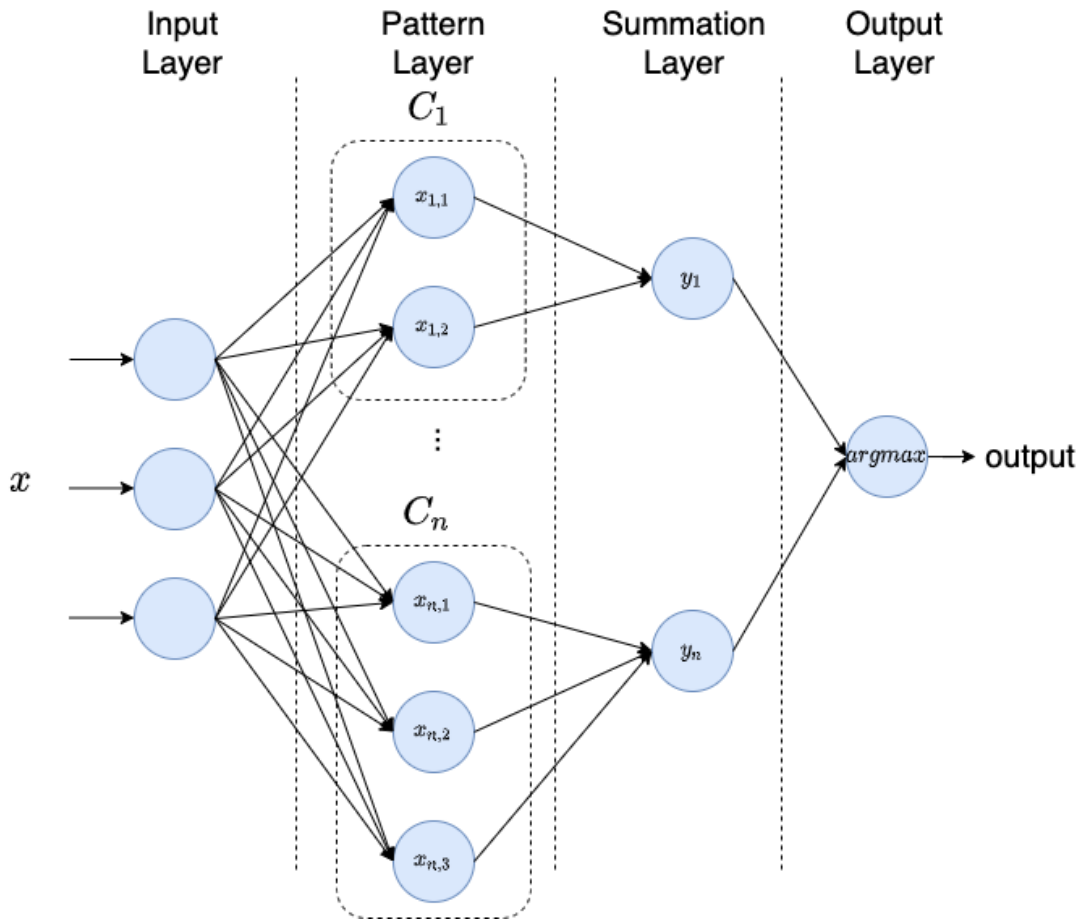
(a) accuracy (b) confidence

More robust under stronger attacks when the two constraints lifted by MLP-PNN and DE:

- Higher prediction accuracy
- Lower prediction confidence

Section IV elaborate the technical details of MLP-PNN and DE

Difference between PNN and DE



In the pattern layer:

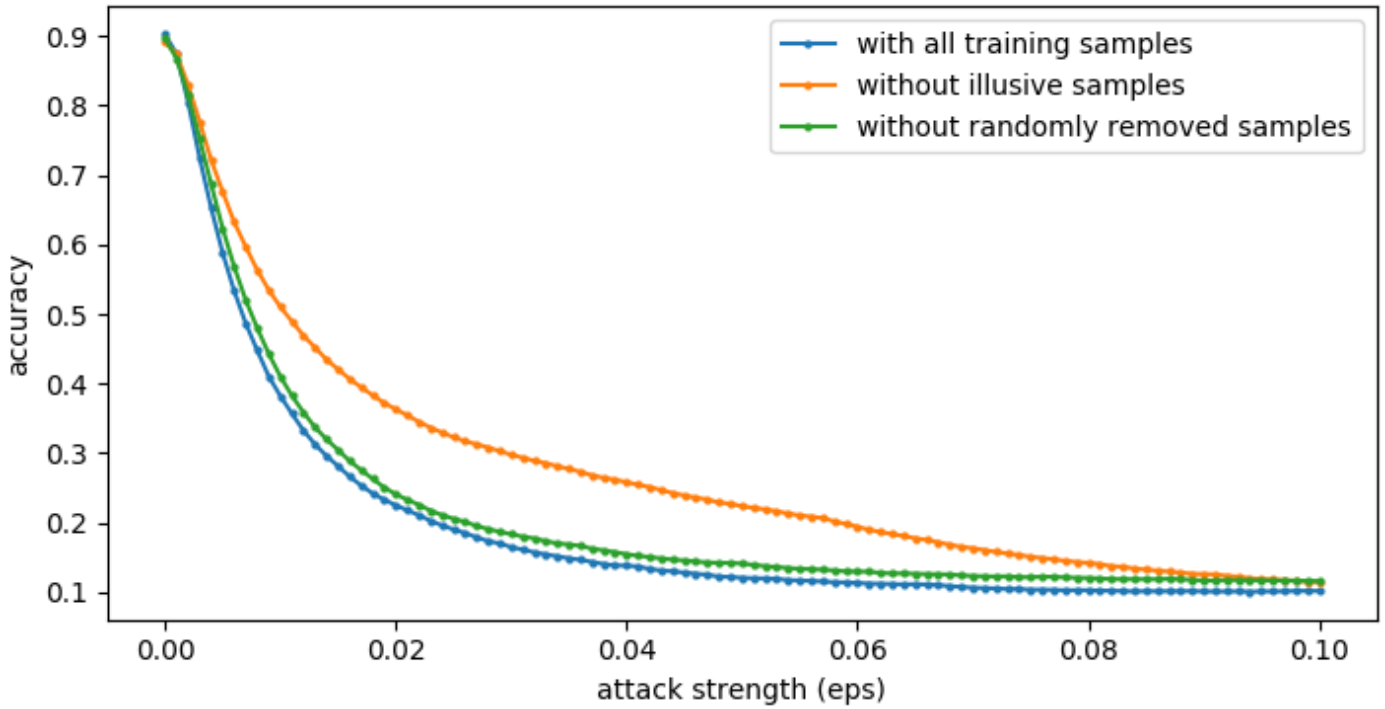
$$K(f(\mathbf{x}), \mathbf{x}_k) = \exp[-\frac{1}{2}\frac{||f(\mathbf{x}) - \mathbf{x}_k||}{2\sigma^2}]$$

σ is :

- a pre-defined constant (.5) in PNN
- a trainable variable in DE

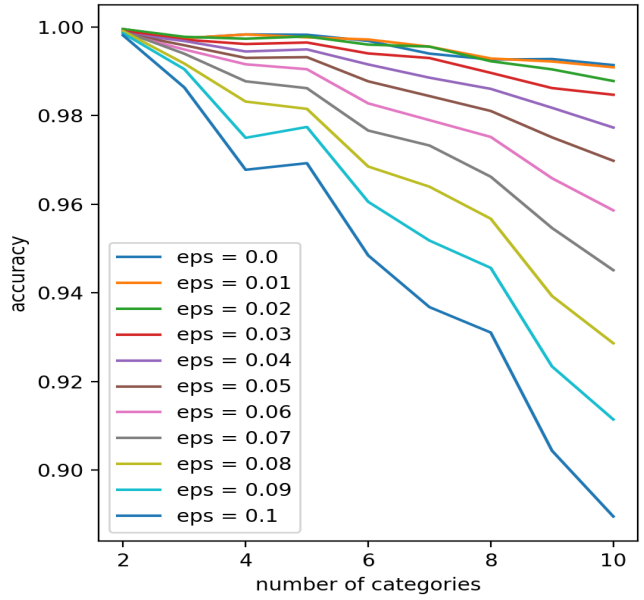
Hypothesis D: path-connected regions

- Uncertain “bridges” for connecting illusive (hard) samples in path-connected regions
- Training without illusive (hard) samples enhance robustness

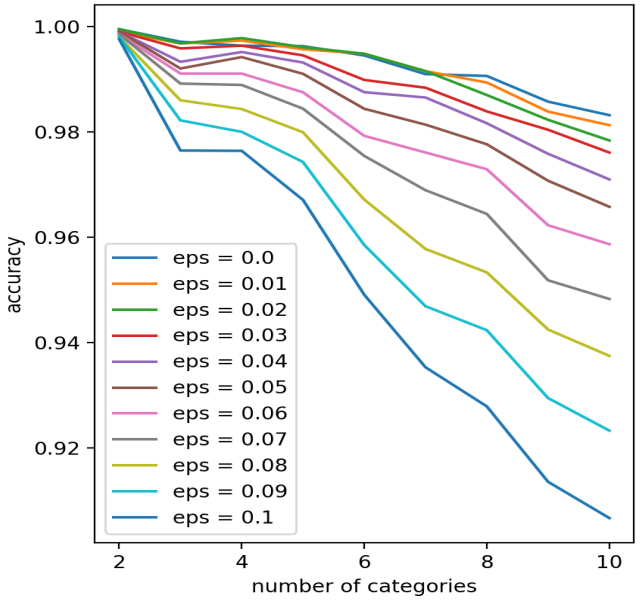


Hypothesis E: excessive number of categories

- Fewer target categories leads to higher adversarial robustness



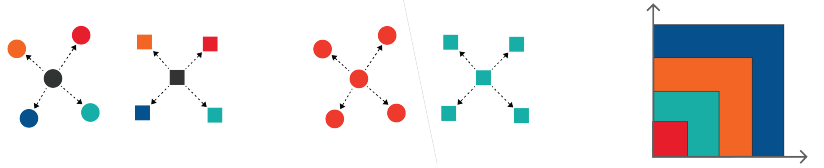
(a) additive mode: including all available training samples



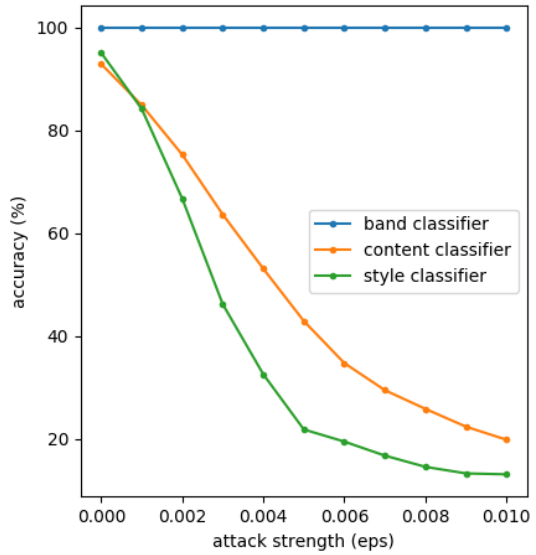
(b) constant mode: 10,000 training samples (balanced)

Hypothesis F: geometry of categories

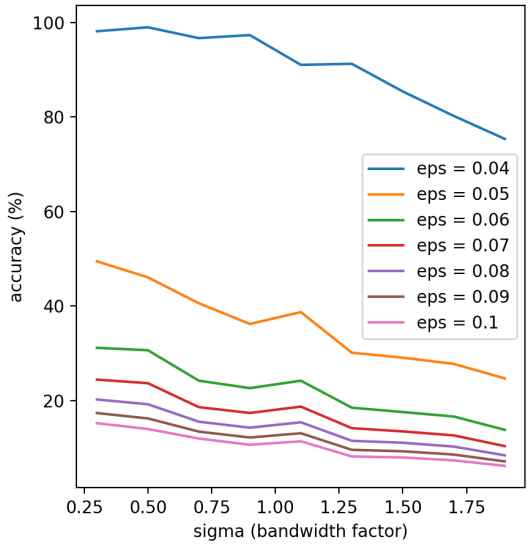
- Adversarial robustness depends on geometry of the input space:
  - entropy of the distribution of categories
  - can be measured by $d_{\text{inter}}/d_{\text{intra}}$



(a) style classes (b) content classes (c) band classes



(a) $\frac{d_{\text{inter}}}{d_{\text{intra}}}$: style $<$ content $<$ band



(b) increasing band overlap

Summary

- Verified hypothesis on causes of adversarial examples
  - Geometric factors: direct causes
  - Statistical factors: magnifier for high confidence

- Future work
  - More rigorous investigation on the root causes of adversarial examples
  - Design on more robust models