Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality

Xingjun Ma\(^1\), Bo Li\(^2\), Yisen Wang\(^3\), Sarah M. Erfani\(^4\), Sudanthi Wijewickrema\(^1\),
Grant Schoenebeck\(^1\), Dawn Song\(^2\), Michael E. Houle\(^4\), James Bailey\(^1\)

\(^1\)The University of Melbourne, \(^2\)UC Berkeley, \(^3\)Tsinghua University, \(^4\)University of Michigan, \(^5\)National Institute of Informatics

Why

- Adversarial attack is a major security threat to deep networks (DNNs).
- Better methods are needed for adversarial detection and defense.
- Adversarial subspaces need to be characterized for better understanding of adversarial attack.

What

- We characterize the dimensional properties of adversarial subspaces using Local Intrinsic Dimensionality (LID).
- We show that adversarial subspaces possess higher intrinsic dimensionality.
- We demonstrate how LID can be used to discriminate adversarial examples.

Adversarial Examples and Adversarial Subspaces

- Small perturbations on inputs can easily fool a deep neural network.
- Perturbations are small, imperceptible to human eyes.
- Open issues: All networks are vulnerable to adversarial attack.
- Adversarial examples transfer across models.

Adversarial Examples

- Given input \((x, y)\) and a target class \(\tilde{y}\), the attack generates a new example \(x_{\text{adv}}\), so as to:
  - \[\minimize_{\tilde{x}} g(x_{\text{adv}}) \text{ subject to } f(x_{\text{adv}}) = \tilde{y}(x_{\text{adv}}) = y \]

Adversarial Attack

- Current attacks:
  - Fast Gradient Method (FGM).
  - Basic Iterative Method (BIM).
  - Jacobian-based Saliency Map Attack (JSMA).
  - Optimization Based Attack (Opt.)

Adversarial Defense/Detection

- Defense methods:
  - Adversarial training.
  - Defensive distillation.
  - Gradient masking.
  - Feature squeezing.

Adversarial Subspaces

- Adversarial subspace is the local subspace that immediately surrounding an adversarial example.

Local Intrinsic Dimensionality of Adversarial Subspaces

- Expansion Dimension:
  - Two balls of differing radii \(r_1\) and \(r_2\) dimension \(m\) can be deduced from ratios of volumes:
  - \[V_m = \left(\frac{r_1}{r_2}\right)^m \Rightarrow m = \frac{\ln(V_m/V_2)}{\ln(r_2/r_1)}\]

- \(V_1\) and \(V_2\) are estimated by the numbers of points contained in the two balls.

LID of Adversarial Subspaces

- Higher dimensionality: Adversarial subspaces are of higher dimensionality (LID).

- Consistency: Adversarial subspaces generated by different attacks share similar dimensional properties.

LID of Different Layers


- Deeper layers: LID difference is more pronounced at deeper layers.

Potential for Detection

- Maximum Likelihood Estimator (Hill 1975, Amaleg et al. 2015):
  - \[\text{LID}(r) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{\text{density of } X_{\text{adv}}}{\text{density of } X} \]

- Extreme Value Theory:
  - Nearest distances are extreme events.
  - Lower tail distribution follows Generalized Pareto Distribution.

- Efficient estimation within a random minibatch.

Potential of LID

- LID characteristics of adversarial examples from five current attacks can be easily discriminated from those of normal examples.

- New experiments with batch normalization shows better and more consistent results on new attacks.

Contact: Xingjun (Daniel) Ma
Email: xingjunm@student.unimelb.edu.au