Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality

Xingjun Ma[1], Bo Li[2], Yisen Wang[3], Sarah M. Erfani[1], Sudanthi Wijewickrema[1], Grant Schoenebeck[4], Dawn Song[2], Michael E. Houle[5], 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

Adversarial Examples

- 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 Attack

Given input (x, y) and a target class l, the attack generates a new example x_adv, so as to:

$$\text{minimize } \|x - x_{adv}\|_p \quad \text{subject to } f(x_{adv}) \neq f(x) \text{ or } f(x_{adv}) = l$$

- 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.
- Detection methods:
  - Deep feature based detectors.
  - Artifacts based detectors: Kernel Density (KD) and Bayesian Uncertainty (BU).

Adversarial Subspaces

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

- *Nonlinear view:*
  - Densely scattered.
  - Low probability regions.
  - Close to data submanifold.
- Linear view:
  - Small changes at individual dimensions can sum up to significant change in final output.

Local Intrinsic Dimensionality of Adversarial Subspaces

Intuition

- Adversarial subspace is close to, yet semantically far from original data subspace.
- Adversarial examples can "escape" to adversarial subspace with only a small perturbation.
- Dimensional Escape.
- Adversarial subspaces have higher dimensionality.

Expansion Dimension

- Expansion Dimension:
- Two balls of differing radii r_1 and r_2, dimension m can be deduced from ratios of volumes:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \Rightarrow m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}$$

- V_1 and V_1 are estimated by the numbers of points contained in the two balls.

Local Intrinsic Dimensionality

Given a data sample $x \in X$, let $r > 0$ be a random variable denoting the distance from x to other data samples. The local intrinsic dimension of x at distance r is

$$\mathrm{LID}_F(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln(F((1+\epsilon)\cdot r)/F(r))}{\ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)},$$

wherever the limit exists.

- $F(r)$: cumulative distribution function.

Estimation of LID

- Maximum Likelihood Estimator (Hill 1975, Amsaleg et al. 2015):

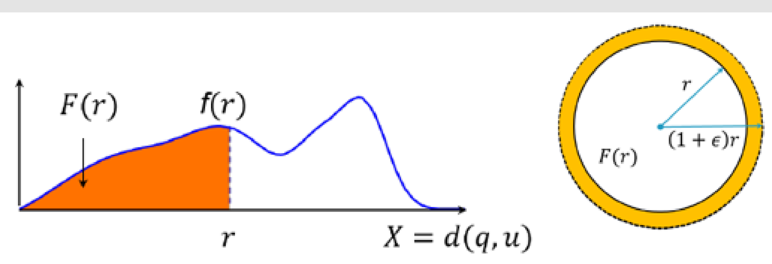
$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)}\right)^{-1}$$

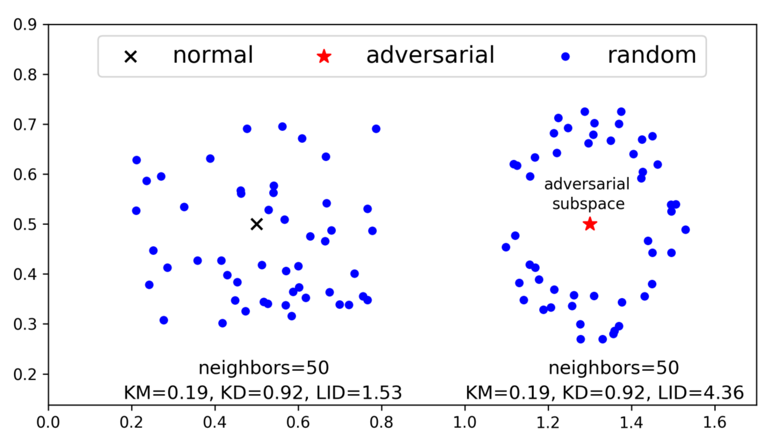
- Extreme Value Theory:
  - Nearest distances are extreme events.
  - Lower tail distribution follows Generalized Pareto Distribution.
- Efficient estimation within a random minibatch.

Interpretation of LID

$$\mathrm{LID}_F(r) = \frac{r \cdot F'(r)}{F(r)}$$

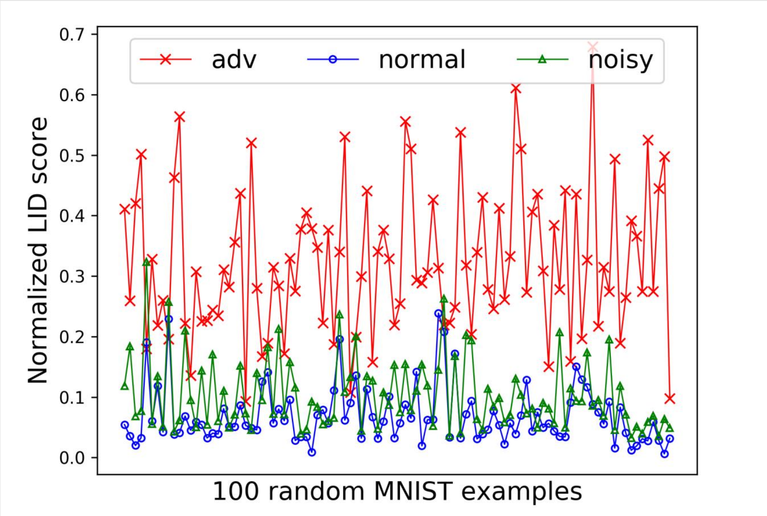
- Characterizes local spatial expansion rate.
- More sensitive than KD and BU.





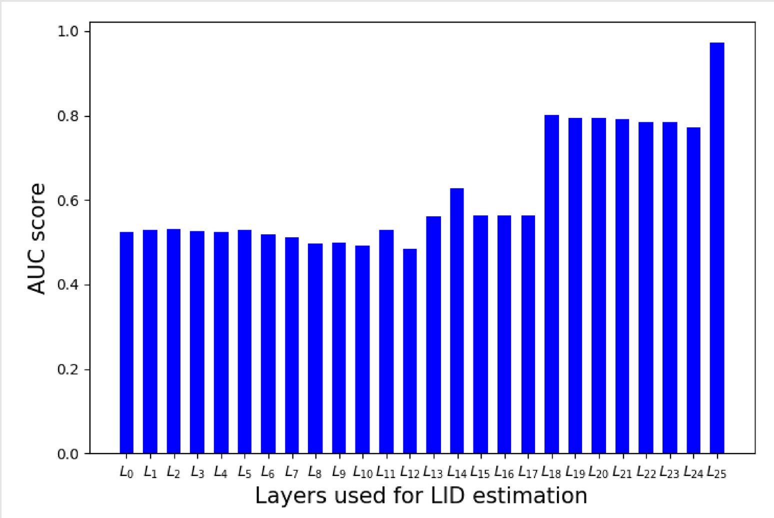
LID of Adversarial Subspaces

- **Higher dimensionality:** Adversarial subspaces are of higher dimensionality (LID).
- **Consistency**: Adversarial subspaces generated by different attacks share similar dimensional properties.



LID of Different Layers

- **Intermediate layers:** Adversarial subspaces already begin to appear.
- **Deeper layers:** LID difference is more pronounced at deeper layers.



Potential for Detection

- 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.

Table 1: A comparison of the discrimination power (AUC score (%) of a logistic regression classifier) among LID, KD, BU, and KD+BU. The AUC score is computed for each attack strategy on each dataset, and the best results are highlighted in **bold**.

Dataset	Feature	FGM	BIM-a	BIM-b	JSMA	Opt
MNIST	KD	78.12	98.14	98.61	68.77	95.15
	BU	32.37	91.55	25.46	88.74	71.30
	KD+BU	82.43	99.20	98.81	90.12	95.35
	LID	**96.89**	**99.60**	**99.83**	**92.24**	**99.24**
CIFAR-10	KD	64.92	68.38	98.70	85.77	91.35
	BU	70.53	81.60	97.32	87.36	91.39
	KD+BU	70.40	81.33	98.90	88.91	93.77
	LID	**82.38**	**82.51**	**99.78**	**95.87**	**98.94**
SVHN	KD	70.39	77.18	99.57	86.46	87.41
	BU	86.78	84.07	86.93	91.33	87.13
	KD+BU	86.86	83.63	99.52	93.19	90.66
	LID	**97.61**	**87.55**	**99.72**	**95.07**	**97.60**

Dataset	%	FGM	BIM	PGD	Deepfool	EAD-0	EAD-40	Opt-0	Opt-40
CIFAR-10	AUC	88.55	95.28	94.45	98.78	98.85	98.82	98.75	98.45
	Accuracy	80.89	87.74	86.80	95.98	93.23	94.58	95.61	94.02
	Precision	82.21	77.55	77.10	95.98	94.25	95.45	95.75	94.42
	Recall	80.10	88.98	85.92	96.20	92.45	93.91	95.70	96.48

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