Metric Learning for Adversarial Robustness

Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, Baishakhi Ray
Department of Computer Science, Columbia University

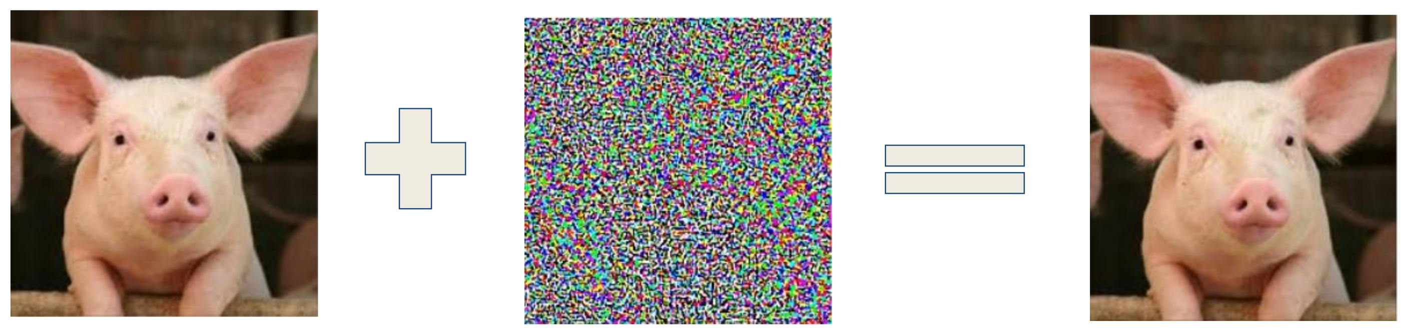
https://papers.nips.cc/paper/8339-metric-learning-for-adversarial-robustness.pdf

COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Code

Adversarial Examples

Neural networks are vulnerable under adversarial attacks.

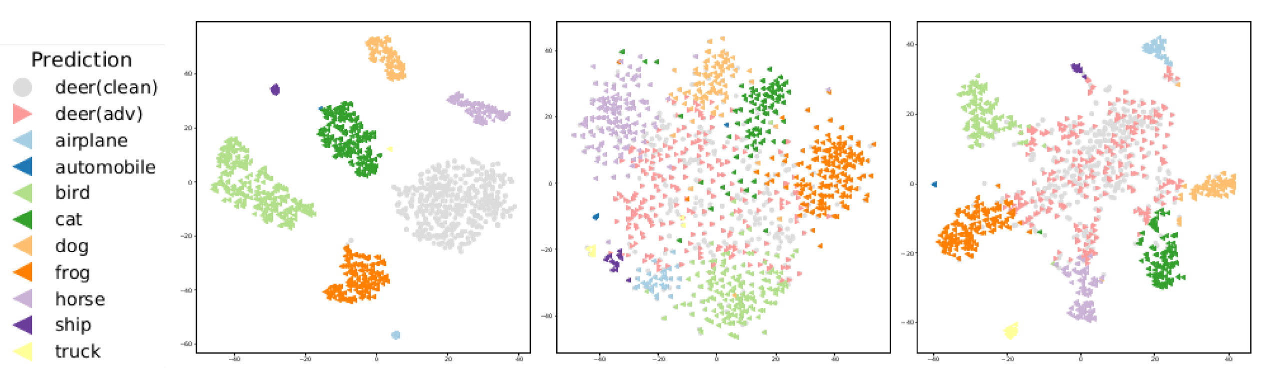


"Pig" Adversarial Noise "Airliner"

Design Motivations

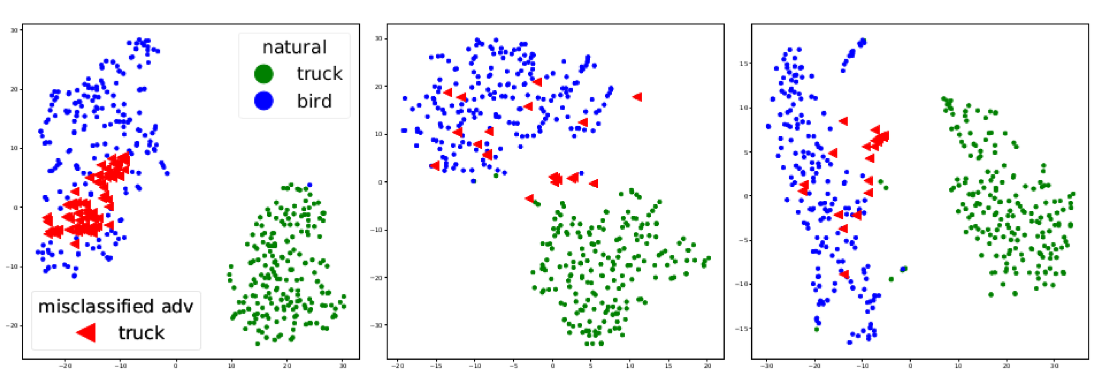
Hidden representations of DNNs under attack have not been systematically studied.

- Representations from the same class are mistakenly classified to different classes.



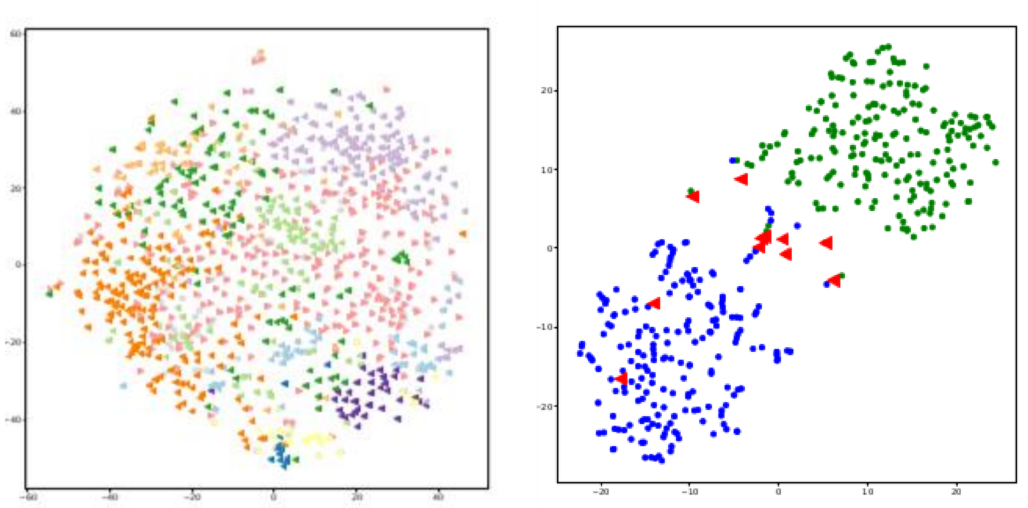
(a) UM (b) AT (c) ALP

- Representations of the adversarial examples are shifted inside the false class.



(a) UM (b) AT (c) ALP

After TLA training



(d) TLA (d) TLA

- The examples from the same class are clustered together.
- The adversarial examples are separated from the corresponding false class by a margin.

Methodology

Triplet Loss

- Pull the representations of the same class together
- Push the representations of different classes away up to a margin
- Anchor $x_a^{(i)}$ and positive $x_p^{(i)}$ are from the same class, negative $x_n^{(i)}$ is from a different class

$$\sum_i^N \mathcal{L}_{trip}(\mathbf{x}_a^{(i)}, \mathbf{x}_p^{(i)}, \mathbf{x}_n^{(i)}) = \sum_i^N [D(h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_p^{(i)})) - D(h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_n^{(i)})) + \alpha]_+$$

Triplet Loss for Adversarial Robustness (TLA)

$$\mathcal{L}_{all} = \sum_i^N \mathcal{L}_{ce}(f(\mathbf{x}_a'^{(i)}), y^{(i)}) + \lambda_1 \mathcal{L}_{trip}(h(\mathbf{x}_a'^{(i)})), h(\mathbf{x}_p^{(i)}), h(\mathbf{x}_n^{(i)})) + \lambda_2 \mathcal{L}_{norm}$$

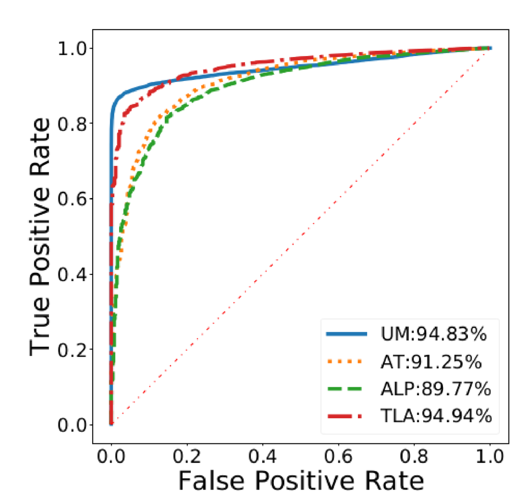
Semi-hard negative example retrieval

- Negative examples randomly sampled from another class are too easy.
- Retrieving the closest examples from the total training set is slow and too hard as a negative example.
- We propose to retrieve the closest negative example from a mini-batch of training data.

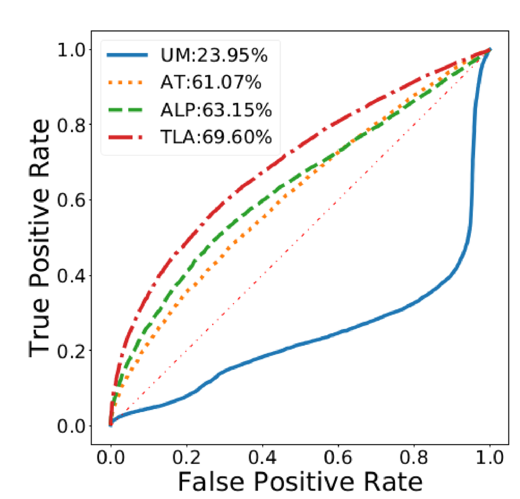
- We propose to regularize the distorted representation under attack with metric learning, to produce more robust classifier.
- Our method, TLA, produces desirable internal representation that improves adversarial robust accuracy and misclassification detection efficiency.

Results

Effect of TLA on mis-classification detection: ROC Curve



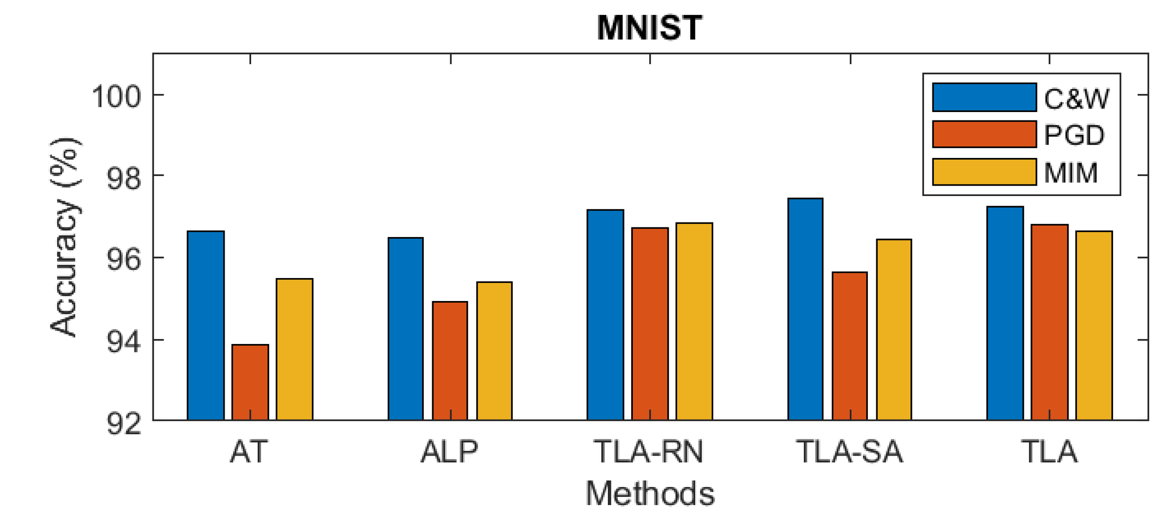
(a) MNIST



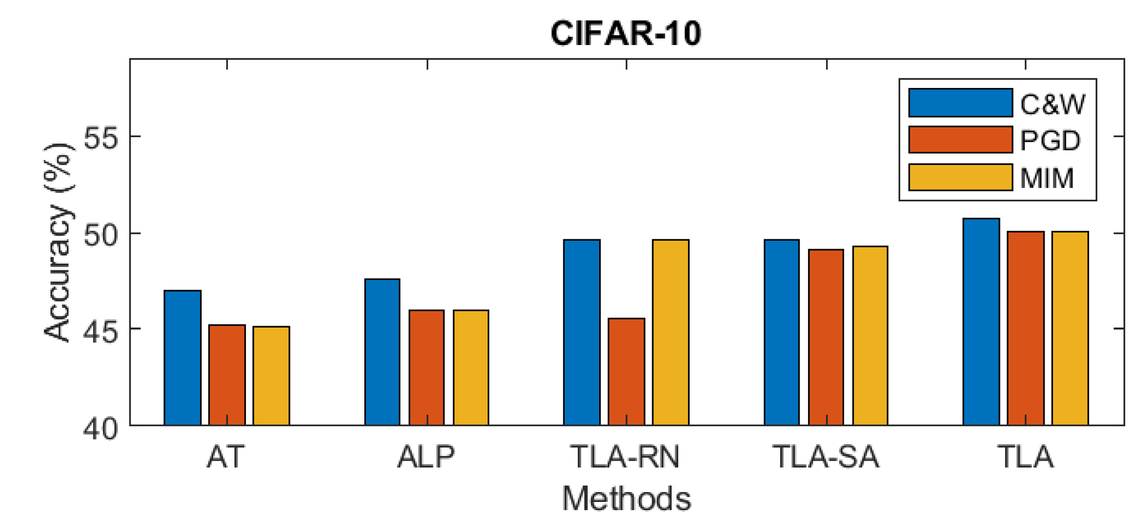
(b) CIFAR-10

- PGD attack under larger perturbation norm bound compared with training time
- TLA Improves the **AUC** score for detection by
  - **3.69%** on MNIST
  - **6.46%** on CIFAR-10.

Effect of TLA on classification accuracy under attacks

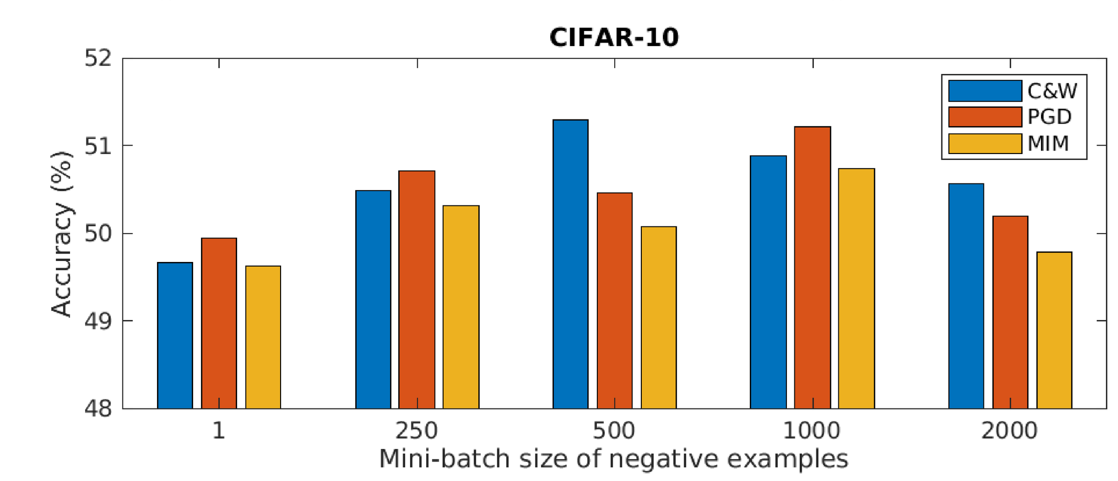


- With 2000 steps of PGD
- Improve the accuracy by
  - **1.86%** compared with Adversarial Logit Pairing
  - **2.92%** compared with Adversarial Training.



- With 400 steps of PGD
- Improve the accuracy by
  - **4.05%** compared with Adversarial Logit Pairing
  - **4.82%** compared with Adversarial Training.

Semi-hard negative example retrieval using mini-batch



- By choosing the proper mini-batch size, performance is improved by up to **1.64%**

Nearest Neighbors Retrieval Using Learned Metric