On the Robustness of Deep k-Nearest Neighbor

Chawin Sitawarin
EECS, UC Berkeley
chawins@berkeley.edu

David Wagner
EECS, UC Berkeley
daw@cs.berkeley.edu

2nd Deep Learning and Security Workshop (IEEE S&P 2019)

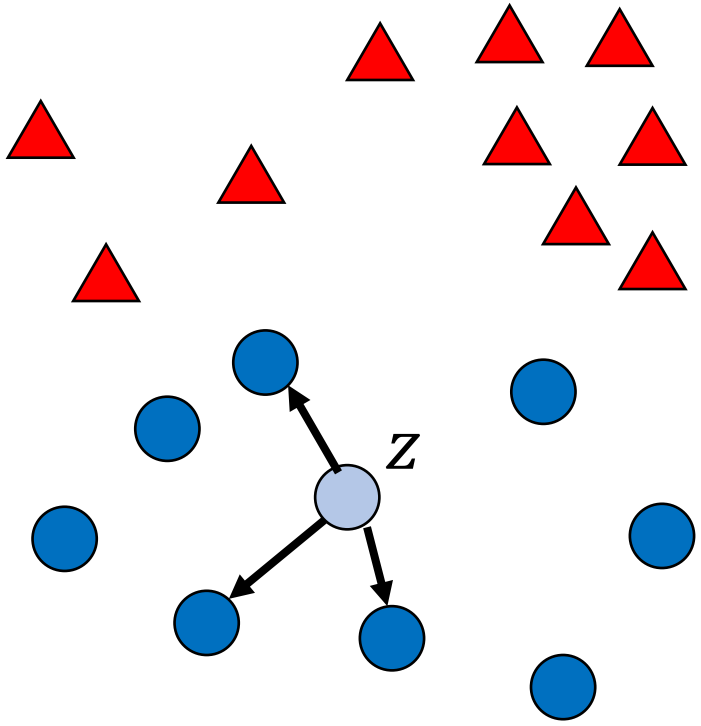
Motivation

" Adversarial examples for kNN and DkNN "

- No previous work attacks kNN directly
- Deep k-Nearest Neighbor (DkNN) shows a possibility for detecting adversarial examples but it is difficult to evaluate
- kNN is not differentiable so most existing attacks don't work
- To measure how robust they really are, we need a white-box attack (no security through obscurity)

k-Nearest Neighbor

k = 3



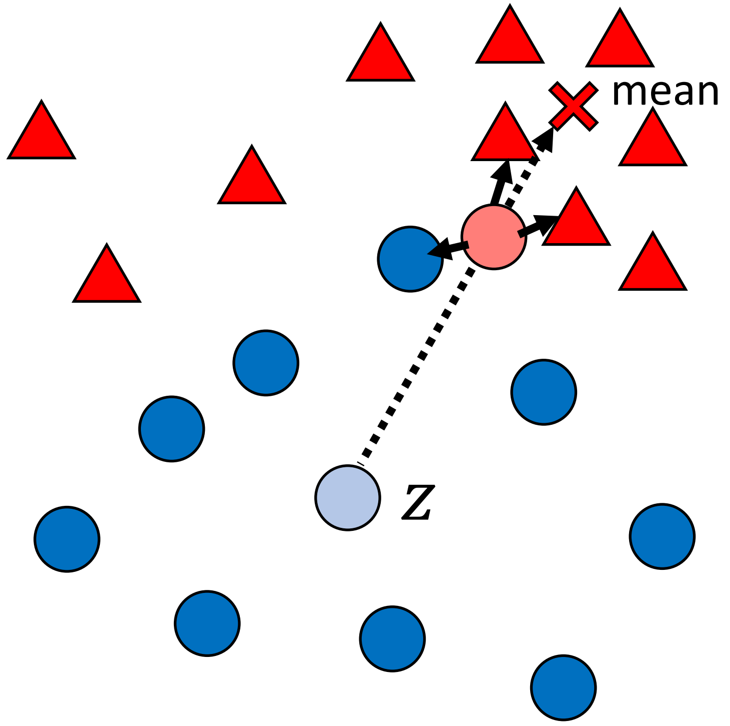
Threat model: white-box, untargeted, Lp norm-ball adversarial examples

- All training samples are known to the attacker

Attack on kNN

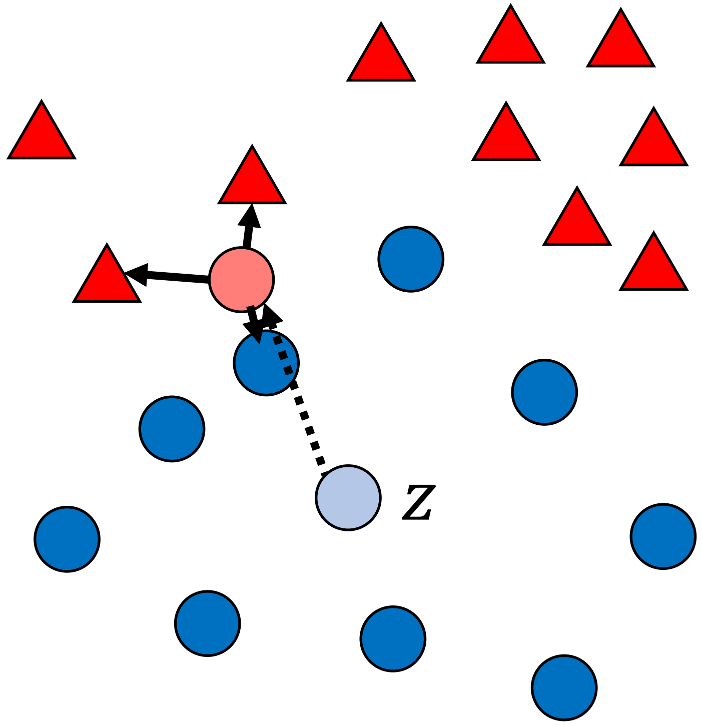
- Baseline: mean attack
  - Move z towards mean of the nearest class
  - Use binary search to determine the distance
- But this is not optimal

Mean Attack



Attack on kNN

- Our gradient-based attack
  - Main idea: move z towards a set of m nearest neighbors from a different class, $\{x_i\}$

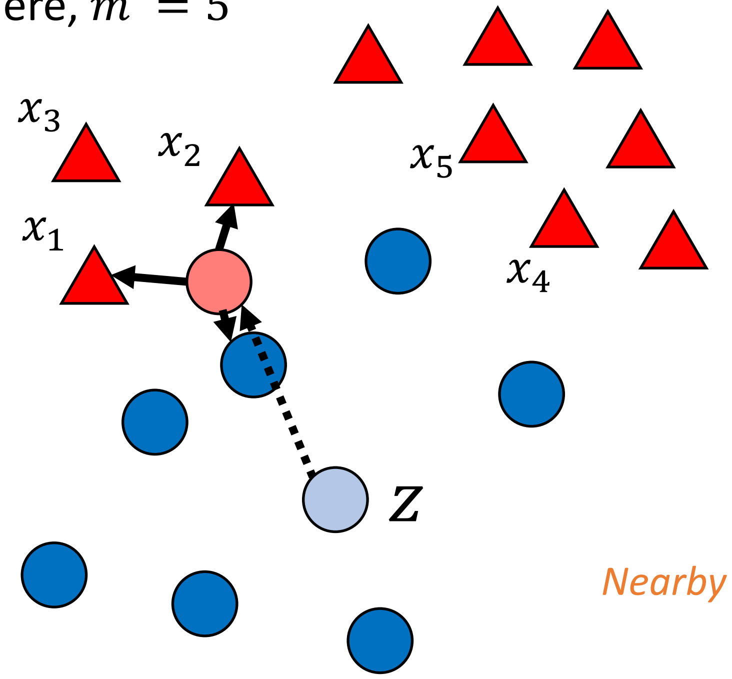


Attack on kNN

- Our gradient-based attack
  - Main idea: move z towards a set of m nearest neighbors from a different class, $\{x_i\}$
  - Set up as a constrained optimization problem

*We use Euclidean distance here, but it can be directly substituted with cosine distance

Here, $m = 5$



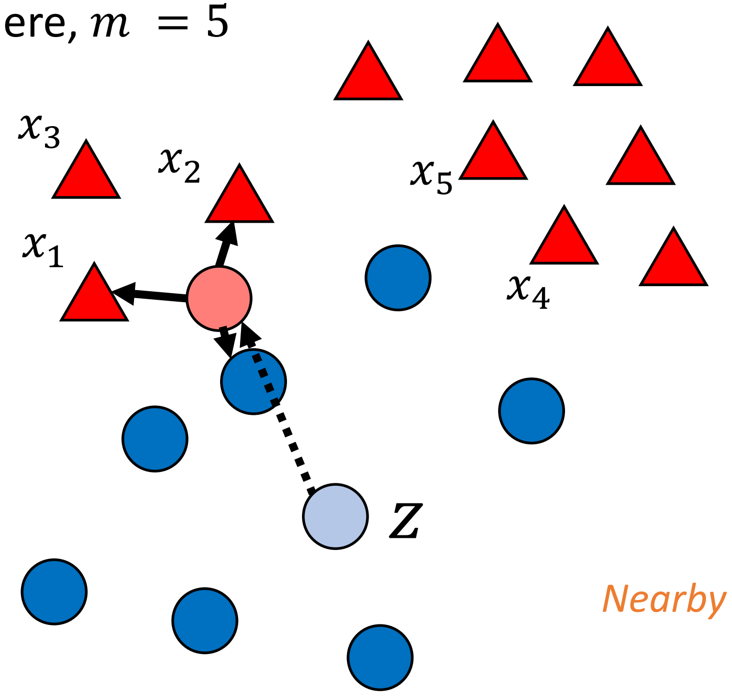
$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \|\overset{\textit{Nearby}\text{ training sample}}{x_i} - (\overset{\text{Input}}{z} + \overset{\text{Perturbation}}{\delta})\|_2$$

$$\text{such that } \underbrace{\|\delta\|_p \leq \epsilon}_{L_p\text{-norm constraint}} \text{ and } \underbrace{x + \delta \in [0,1]^d}_{\text{Constraint on input domain}}$$

Attack on kNN

- Our gradient-based attack
  - Main idea: move z towards a set of m nearest neighbors from a different class, $\{x_i\}$
  - Set up as a constrained optimization problem

Here, $m = 5$



- It's sufficient to be *close* to only x_1 and x_2!
- But it is difficult to know which x_i ahead of time

Nearby training sample Input Perturbation

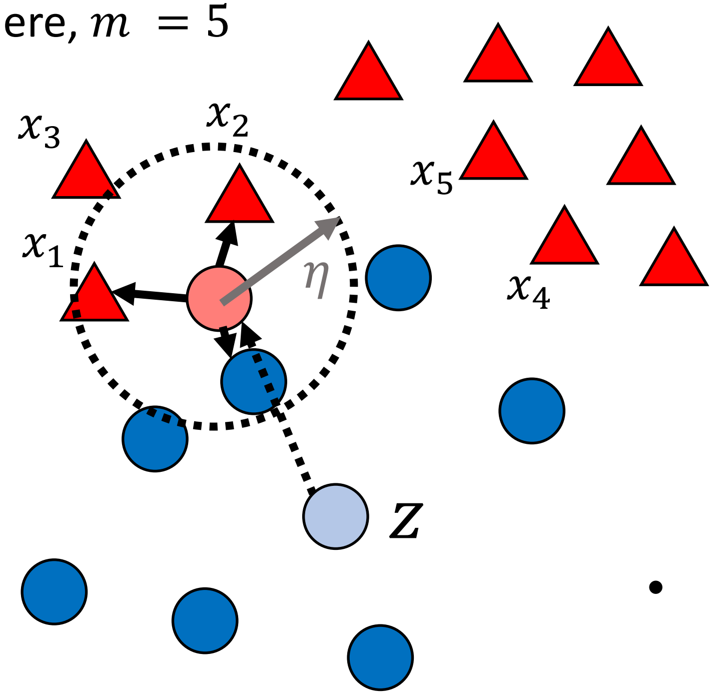
$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \|x_i - (z + \delta)\|_2$$

$$\text{such that } \underbrace{\|\delta\|_p \leq \epsilon}_{\text{L}_\text{p}\text{-norm constraint}} \text{ and } \underbrace{x + \delta \in [0,1]^d}_{\text{Constraint on input domain}}$$

Attack on kNN

- Our gradient-based attack
  - Main idea: move z towards a set of m nearest neighbors from a different class, $\{x_i\}$
  - Set up as a constrained optimization problem

Here, $m = 5$

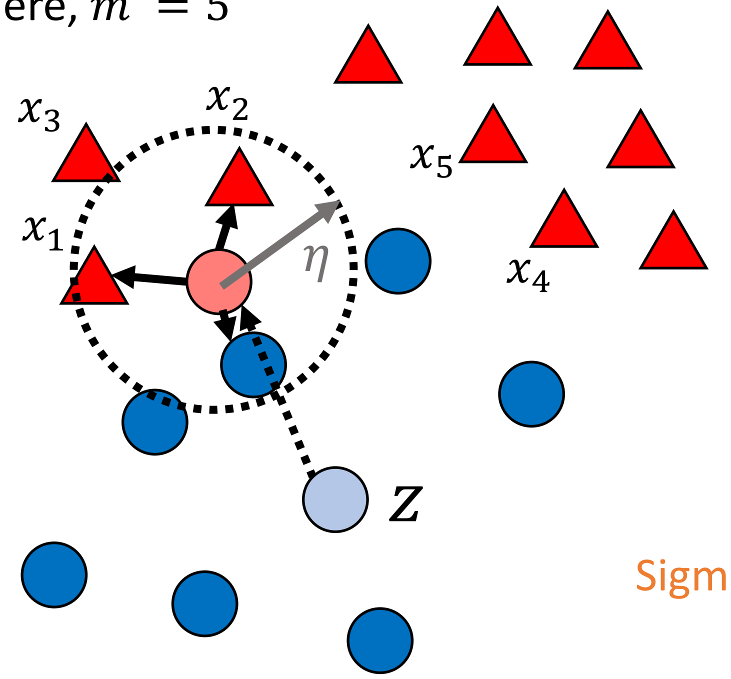


- We want to ignore samples that are too far away by setting a threshold
- But hard threshold is not differentiable

Attack on kNN

- Our gradient-based attack
  - Main idea: move z towards a set of m nearest neighbors from a different class, $\{x_i\}$
  - Set up as a constrained optimization problem
  - Use sigmoid as a soft threshold
  - Choose η to be mean distance to k-th neighbor

Here, $m = 5$



Approximate hard threshold with a soft, differentiable one

Sigmoid function

Threshold

$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \sigma\left(\|x_i - (z+\delta)\|_2 - \eta \right)$$

$$\text{such that } \|\delta\|_p \le \epsilon \text{ and } x + \delta \in [0,1]^d$$

Results on kNN

- kNN uses cosine distance with k = 75 on MNIST dataset

Attacks	Accuracy (%)	Mean Perturbation (L_2)
No Attack	95.74	-
Mean Attack	**5.89**	8.611
Our Gradient Attack	9.89	**6.565**

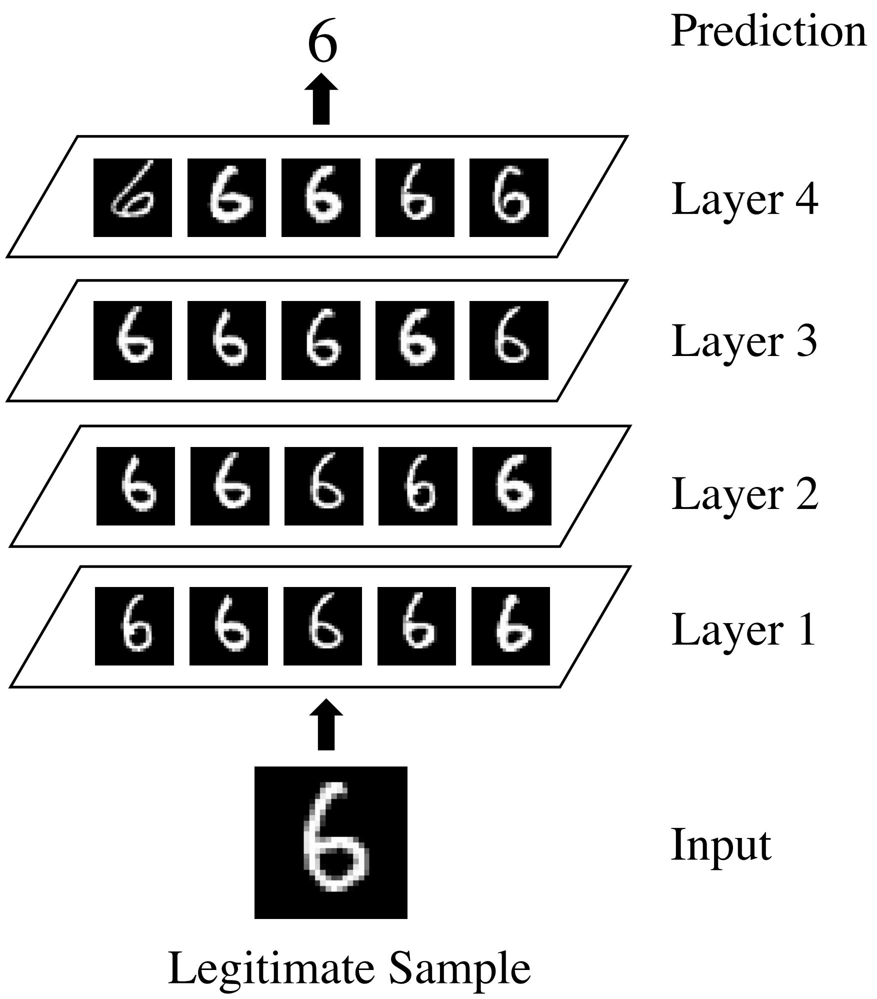
Most have perceptible / semantic perturbation

Clean

Adv.

Deep k-Nearest Neighbor

- Proposed by Papernot & McDaniel ’18
- Essentially, kNN on outputs of multiple layers of a neural network
- Simple scheme that offers some interpretability
- Can detect out-of-distribution samples and adversarial examples to some degree



Attack on DkNN

- Baseline: mean attack
  - Same as kNN
- Our gradient-based attack
  - Similar to our gradient-based attack on kNN

Gradient-based attack on kNN

$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \sigma\left(\|x_i - (z+\delta)\|_2 - \eta \right)$$

$$\text{such that } \|\delta\|_p \leq \epsilon \text{ and } x + \delta \in [0,1]^d$$

Attack on DkNN

- Our gradient-based attack
  - Similar to our gradient-based attack on kNN
  - Instead of distance in the pixel space, we consider distance in the representation space
  - And sum over all the layers

Sum over all layers

Representation at λ-th layer

$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \sum_{\lambda=1}^{l} \sigma\left(\|f_\lambda(x_i) - f_\lambda(z+\delta)\|_2 - \eta_\lambda \right)$$

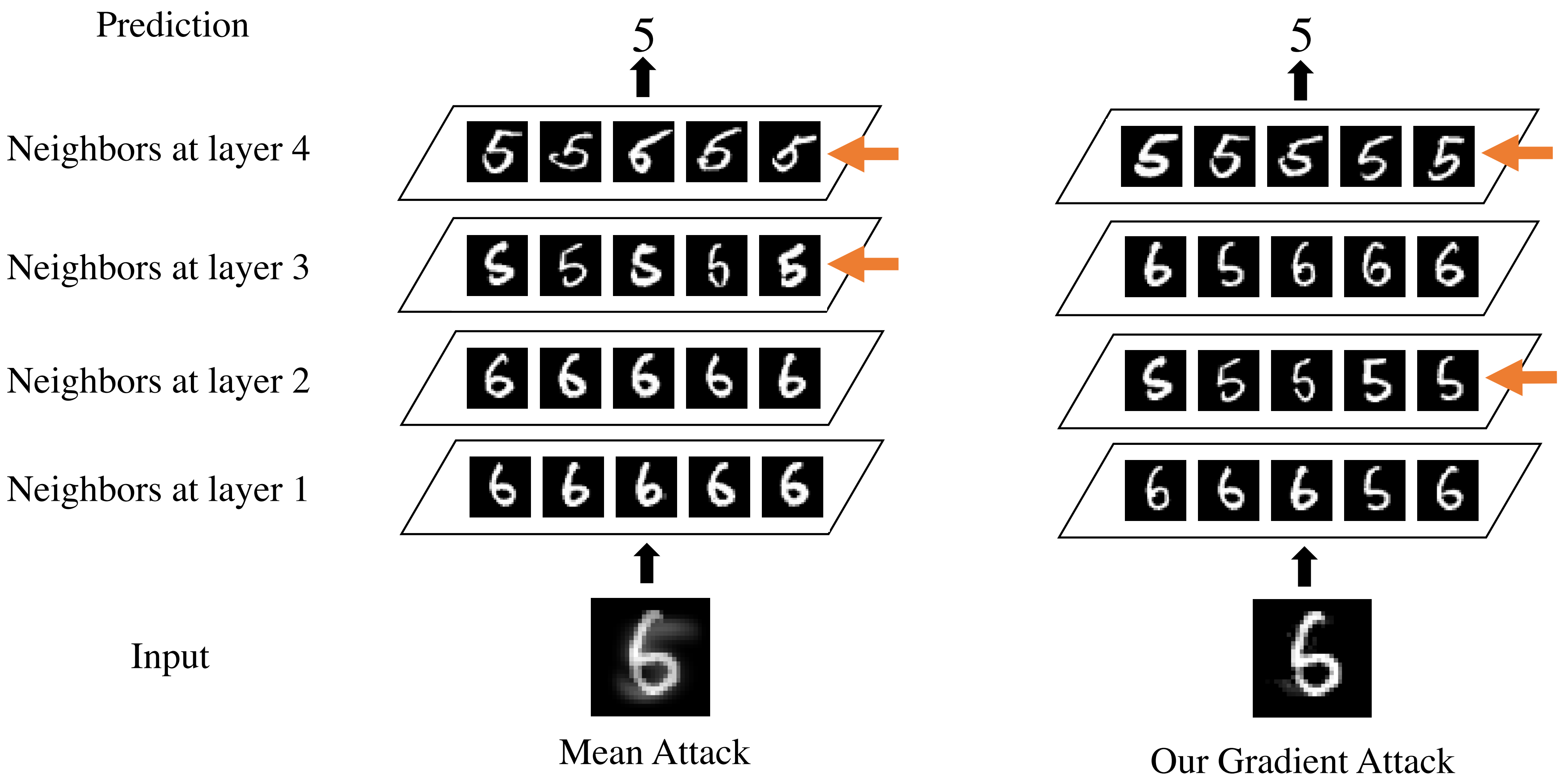
$$\text{such that } \|\delta\|_p \leq \epsilon \text{ and } x + \delta \in [0,1]^d$$

Results on DkNN

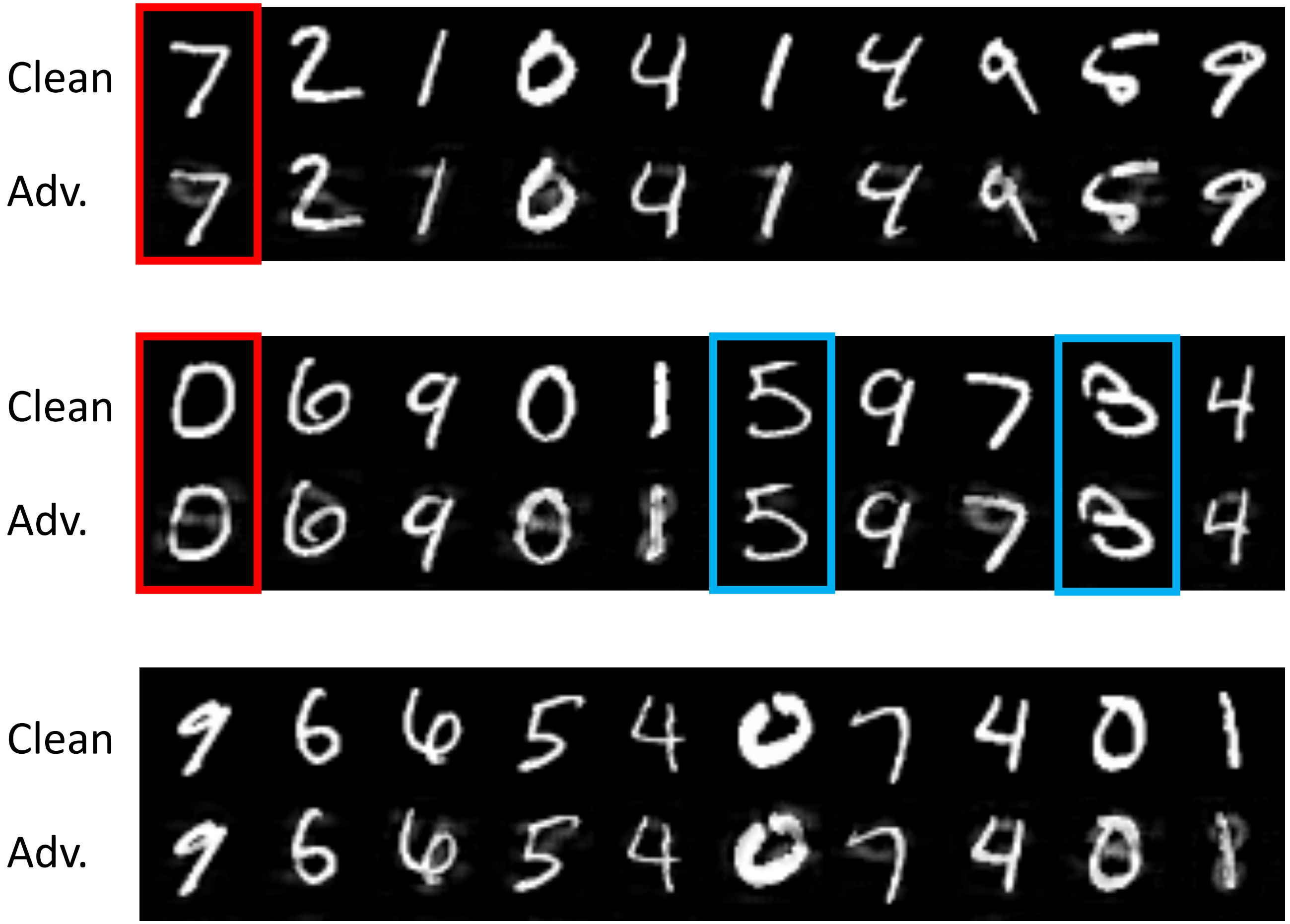
- We use the same network and hyperparameters suggested by Papernot & McDaniel ’18

Attacks	Accuracy (%)	Mean Perturbation (L_2)
No Attack	98.83	-
Mean Attack	13.13	4.408
P&M’18 Attack	16.02	3.459
Our Gradient Attack	**0.00**	**2.164**

Results on DkNN



Results on DkNN



- Some perturbations have semantic meaning
- But some are imperceptible
- Suggests that L2-norm is not always a good metric
- Suggests that there is some hope for the defense

Credibility

- DkNN can output a *credibility score* for a give input
- It can be used to filter out adversarial examples and out-of-distribution samples
- Promising but not very effective currently
  - Some adversarial examples have a high credibility score
  - Some clean samples have a low credibility score
- We refer to paper for more details

Conclusion

- We propose an attack on kNN and DkNN
- Nonetheless, they appear to be more robust compared to other algorithms out of the box
    - Requires larger perturbation
    - Some perturbation also has semantic meaning
- Improving the DkNN
    - Ongoing work: DkNN on representations of a robust network (e.g. adversarially trained networks)
    - More robust variants of kNN (e.g. weighted voting)

Extra Slides

Sigmoid

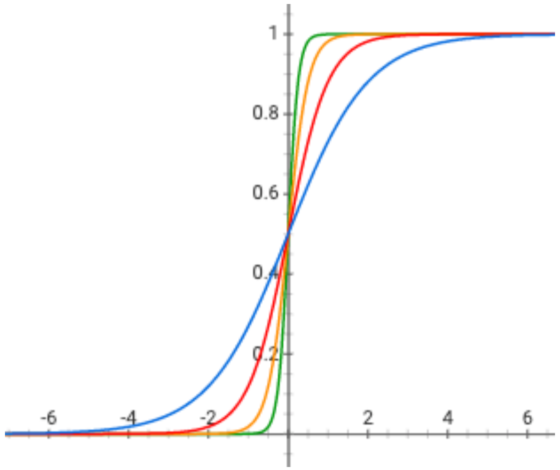
- **Gradient-based attack**
  - Main idea: move z towards *a group of m nearby samples* (x_i) from a different class
  - Set up as a constrained optimization problem
  - Use sigmoid as a soft threshold
  - Choose η to be mean distance to k-th neighbor

Sigmoid function

Threshold

$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{m} \sigma\Big(\overbrace{\|x_i - (z+\delta)\|_2}^{d} - \eta\Big)$$

$$\text{such that } \|\delta\|_p \leq \epsilon \text{ and } x + \delta \in [0,1]^d$$



$$\sigma(d - \eta) \approx \begin{cases} 0 & \text{if } d < \eta \\ 1 & \text{if } d > \eta \end{cases}$$

$$\sigma(x) = \frac{1}{1 + e^{-ax}}$$

Credibility