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

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)
Threat model: white-box, untargeted, $L_p$ 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
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

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

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

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

such that $\|\delta\|_p \leq \epsilon$ and $x + \delta \in [0, 1]^d$
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
  o Main idea: move $z$ towards a set of $m$ nearest neighbors from a different class, $\{x_i\}$
  o Set up as a constrained optimization problem
  o Use sigmoid as a soft threshold
  o Choose $\eta$ to be mean distance to k-th neighbor

Here, $m = 5$

Approximate hard threshold with a soft, differentiable one

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

such that $\|\delta\|_p \leq \epsilon$ and $x + \delta \in [0, 1]^d$
Results on $k$NN

• $k$NN uses cosine distance with $k = 75$ on MNIST dataset

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Accuracy (%)</th>
<th>Mean Perturbation ($L_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Attack</td>
<td>95.74</td>
<td>-</td>
</tr>
<tr>
<td>Mean Attack</td>
<td>5.89</td>
<td>8.611</td>
</tr>
<tr>
<td><strong>Our Gradient Attack</strong></td>
<td>9.89</td>
<td><strong>6.565</strong></td>
</tr>
</tbody>
</table>

Most have perceptible / semantic perturbation

![Image of attack examples]
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

Chawin Sitawarin
DLS '19 (IEEE S&P)
On the Robustness of Deep k-Nearest Neighbor
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

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

such that \(\|\delta\|_p \leq \epsilon\) 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

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

such that \( \| \delta \|_p \leq \epsilon \) and \( x + \delta \in [0, 1]^d \)
Results on DkNN

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

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Accuracy (%)</th>
<th>Mean Perturbation (L₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Attack</td>
<td>98.83</td>
<td>-</td>
</tr>
<tr>
<td>Mean Attack</td>
<td>13.13</td>
<td>4.408</td>
</tr>
<tr>
<td>P&amp;M’18 Attack</td>
<td>16.02</td>
<td>3.459</td>
</tr>
<tr>
<td>Our Gradient Attack</td>
<td>0.00</td>
<td>2.164</td>
</tr>
</tbody>
</table>
Results on DkNN

Prediction

Neighbors at layer 4

Neighbors at layer 3

Neighbors at layer 2

Neighbors at layer 1

Input

Mean Attack

Our Gradient Attack
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 given input
• It can be used to filter out adversarial examples and out-of-distribution samples
• Promising but not very effective currently
  o Some adversarial examples have a high credibility score
  o 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
  o Requires larger perturbation
  o Some perturbation also has semantic meaning

• Improving the DkNN
  o Ongoing work: DkNN on representations of a robust network
    (e.g. adversarially trained networks)
  o More robust variants of kNN (e.g. weighted voting)
Extra Slides
**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 $\eta$ to be mean distance to k-th neighbor

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

such that $\|\delta\|_p \leq \epsilon$ 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

Distribution of credibility

- Grad Attack ($\ell_\infty = 0.2$)
- Grad Attack ($\ell_\infty = 0.3$)
- Clean Test Samples

Gradient attack
Baseline

Mean Credibility

$\ell_\infty$-Norm of Adversarial Perturbation

Number of Samples

Credibility