Defense Against Adversarial Attacks
Recall the classical detection methods

• Pre-processing the image with different transformation methods
• Train a network to tell the adversarial instances apart
• Leverage spatial/temporal properties to check the consistency – indicating adversarial behaviors
• Map the data to other data manifold by computing meaningful metric to measure differences
Beyond the Min-max Game

• What if we have more knowledge about our learning tasks?
  • Properties of learning tasks and data
  • General understanding about ML models
Characterize Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation

• Attacks against semantic segmentation
  • State-of-the-art attacks against segmentation: Houdini [NIPS2017], DAG [ICCV 2017]
  • We design diverse adversarial targets: hello kitty, pure color, a real scene, ECCV, color shift, strips of even color of classes
• Cityscapes and BDD datasets
Spatial Context Information

• Spatial consistency is a distinct property of image segmentation

• Perturbation at one pixel will potentially affect the prediction of surrounding pixels

\[ \mathcal{H}(m) = - \sum_j V_m[j] \log V_m[j] \]
Perturbation on single patch may loss its adversarial effect

• Spatial consistency: the consistency of segmentation results for randomly selected patches from an image

• Such spatial consistency information from benign and adversarial instances are distinguishable

• We apply mIOU to compare the segmentation results between patches
  • For each class, Intersection over Union (IOU) is calculated as TP/(TP+FP+FN). Here we calculate the relative mIOU for each pair of patches
Pipeline of spatial consistency based detection for adversarial examples on semantic segmentation
We apply mIOU to evaluate the consistency information for patches from benign and adversarial instances quantitatively.

- Detection
Adaptive Attack Against Spatial Consistency Based Detection

• Adaptive attack:
  • Assume the attacker has perfect knowledge of #selected patches: K
  • We generate perturbation that the selected k patches can all be mis-segmented to the corresponding regions within adversarial target

(a) Adaptive attack against image scaling  (b) Convergence analysis of adaptive attack  (c) Adaptive attack against spatial consistency
Detecting adversarial instances based on spatial consistency information

- Both the spatial consistency based detection and the scaling based baseline achieve promising detection rate on different attacks
- The scaling based baseline fails to detect strong adaptive attacks while the spatial based method can

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>mIOU</th>
<th>Detection</th>
<th>Detection</th>
<th>Houdini</th>
<th>Houdini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale (std)</td>
<td>0.5</td>
<td>DRN (16.4M)</td>
<td>66.7</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td></td>
<td></td>
<td>95%</td>
<td>100%</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td></td>
<td></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Spatial (K)</td>
<td>1</td>
<td>DRN (16.4M)</td>
<td>66.7</td>
<td>91%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td>91%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td>94%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td>92%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Takeaways

• Spatial consistency information can be potentially applied to help distinguish benign and adversarial instances against segmentation models.

• Strong adaptive attacker can hardly succeed when large randomness is incorporated into the model
Adversarial Frames In Videos

Attacks on segmentation

Attacks on pose estimation

Attacks on object detection
Defensing Adversarial behaviors in Videos – Temporal Dependency
The results show that choosing more random patches can improve detection rate while $k=5$ is enough to achieve AUC 100%.

The spatial consistency based detection is robust against strong adaptive attackers due to the randomness in patch selection.

<table>
<thead>
<tr>
<th>Task</th>
<th>Attack Method</th>
<th>Target</th>
<th>Previous Frames</th>
<th>Detection</th>
<th>Detection Adap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Semantic Segmentation</td>
<td>Houdini</td>
<td></td>
<td></td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remapping</td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stripe</td>
<td></td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>DAG</td>
<td>CVPR</td>
<td></td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remapping</td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stripe</td>
<td></td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td>Human Pose Estimation</td>
<td>Houdini</td>
<td>shuffle</td>
<td></td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transpose</td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td>Object Detection</td>
<td>DAG</td>
<td>all</td>
<td>Person</td>
<td>Benign</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adversarial</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Benign</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adversarial</td>
<td>98%</td>
</tr>
</tbody>
</table>
Temporal Consistency Based Analysis

• “Yanny” or “Laurel”? – adversarial audio
PeerNets: Exploiting Peer wisdom against adversarial attacks

• Design robust neural networks that are robust to adversarial attacks
• Defense: recover the ground truth instead of just tell adversarial instance apart
• Necessary step: design novel and advanced architectures built on new computational paradigms

• PeerNets:
  • Euclidean convolutions -> graph convolutions
  • Non-local forward propagation: Capture global structure induced by the data graph
  • Design a peer regularization layer
PeerNets: Exploiting Peer wisdom against adversarial attacks

- Peer Regularization layer
- For N images, each image will look for its K nearest neighbors based on cosine similarity
  - For each image, there is a \( n \times d \) feature map

\[
\tilde{x}_p^i = \sum_{k=1}^{K} \alpha_{ijk}^p q_k x_q^k, \quad \alpha_{ijk}^p q_k = \frac{\text{LeakyReLU}(a(x_p^i, x_q^k))}{\sum_{k'=1}^{K} \text{LeakyReLU}(a(x_p^i, x_q^{k'}))}
\]
PeerNets: Exploiting Peer wisdom against adversarial attacks

- Randomized approximation
- Monte Carlo approximation
  - Select smaller batch and sample the nearest neighbor from each batch
    \[ \{l_{m1}, \ldots, l_{mN}\} \subset \{1, \ldots, N'\} \]
    \[ \tilde{x}_p^i = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{ijm,kpqmk} x_{jmk} \]
- Other optimization method?
PeerNets: Exploiting Peer wisdom against adversarial attacks

• Select \( M = 1 \) during training and large \( M \) during inference
  • Limitations?
Results for PeerNets

Original

Reconstructed
Visualization of perturbation
Takeaways

• Alternate Euclidean Graph convolution to harness information from peers can provide global information

• Can be added to any models as regularized layer –> good principle

• Not affect the benign accuracy -> important

• How to scale up?

• How to consider more peer images instead of pixels?

• Temporal information?
Similar reading

• Countering adversarial images using input transformations
  • Image quilting – nearest patches
  • Computationally expensive
Interesting reading

- A simple neural network module for relational reasoning
Interesting reading

• Deformable Convolutional Networks
  • Deformable convolution and deformable RoI pooling
  • Augment the spatial sampling locations with additional offset which can be learned
Towards Deep Learning Models Resistant to Adversarial Attacks

\[
\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right]
\]

- Use a natural saddle point (min-max) formulation to capture the notion of security against adversarial attacks in a principled manner.
- The formulation casts both attacks and defenses into a common theoretical framework.
- Motivate projected gradient descent (PGD) as a universal “first-order adversary”.

Model Capacity
Towards Deep Learning Models Resistant to Adversarial Attacks
Decision Boundary Based Detection

MNIST Test image 3153

No defense | Adv. training
---|---

CIFAR-10 Test image 5415

No defense | Adv. training
---|---

Benign

**OptBrittle**

**OptMargin** (ours)

**FGSM** (unsuccesful)
Decision Boundary Analysis of Adversarial Examples
<table>
<thead>
<tr>
<th>Training attack</th>
<th>False pos.</th>
<th>False neg.</th>
<th>Accuracy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benign</td>
<td>OPTBRITTLE</td>
<td>OPTMARGIN</td>
<td>Our approach</td>
</tr>
<tr>
<td>MNIST, normal training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPTBRITTLE</td>
<td>1.0%</td>
<td>1.0%</td>
<td>74.1%</td>
<td>90.4%</td>
</tr>
<tr>
<td>OPTMARGIN</td>
<td>9.6%</td>
<td>0.6%</td>
<td>7.2%</td>
<td></td>
</tr>
<tr>
<td>MNIST, PGD adversarial training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPTBRITTLE</td>
<td>2.6%</td>
<td>2.0%</td>
<td>39.8%</td>
<td></td>
</tr>
<tr>
<td>OPTMARGIN</td>
<td>10.3%</td>
<td>0.4%</td>
<td>14.5%</td>
<td></td>
</tr>
<tr>
<td>CIFAR-10, normal training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPTBRITTLE</td>
<td>5.3%</td>
<td>3.2%</td>
<td>56.8%</td>
<td>96.4%</td>
</tr>
<tr>
<td>OPTMARGIN</td>
<td>8.4%</td>
<td>7.4%</td>
<td>5.3%</td>
<td></td>
</tr>
<tr>
<td>CIFAR-10, PGD adversarial training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPTBRITTLE</td>
<td>0.0%</td>
<td>2.4%</td>
<td>51.8%</td>
<td></td>
</tr>
<tr>
<td>OPTMARGIN</td>
<td>3.6%</td>
<td>0.0%</td>
<td>1.2%</td>
<td></td>
</tr>
</tbody>
</table>
Takeaways

• Decision boundaries of DNNs are important towards improving learning robustness
• Isolated islands in the data manifold would lead to harder detected/defensed adversarial behaviors