ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation

Yuzhe Yang

Guo Zhang, Dina Katabi, Zhi Xu
New defense: ME-Net

emphasizes *global structures* in images
Adversarial Examples

“pig”
Adversarial Examples

“pig”

Adversarial noise
Adversarial Examples

“pig” + Adversarial noise → “airliner”
Adversarial Examples

“pig” + Adversarial noise → “airliner”

Highly structured
Idea: Destroy the Structure of Adversarial Noise

“pig” + Adversarial noise → “airliner”

Highly structured
Idea: Destroy the Structure of Adversarial Noise

“airliner”  Subsample  →  texture image
Idea: Destroy the Structure of Adversarial Noise

“airliner” Subsample

But, subsampling destroys the structure of both adversarial noise and image
Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!
Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

“airliner”

Subsample

Images are known to be low rank!
Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

"airliner"

Subsample

Images are known to be low rank!

Matrix Estimation can be used to recover global structures in an image
Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!
ME-Net

Preprocessing

Subsample

Matrix Estimation

Neural nets
Combine with adversarial training!
# CIFAR-10 Black-Box Attacks

<table>
<thead>
<tr>
<th></th>
<th>CW</th>
<th>FGSM</th>
<th>PGD</th>
<th>Boundary</th>
<th>SPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vanilla</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Madry et al.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ME-Net</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## CIFAR-10 Black-Box Attacks

<table>
<thead>
<tr>
<th>Method</th>
<th>CW</th>
<th>FGSM</th>
<th>PGD</th>
<th>Boundary</th>
<th>SPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>8.9%</td>
<td>24.8%</td>
<td>7.6%</td>
<td>3.5%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Madry et al.</td>
<td>78.7%</td>
<td>67%</td>
<td>64.2%</td>
<td>61.9%</td>
<td>47.0%</td>
</tr>
<tr>
<td>ME-Net</td>
<td>93.6%</td>
<td>92.2%</td>
<td>91.8%</td>
<td>87.4%</td>
<td>93.0%</td>
</tr>
</tbody>
</table>

High robustness against black-box attacks!
# White-box Attacks

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>SVHN</th>
<th>Tiny-ImageNet (Top-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madry et al.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME-Net + SGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME-Net + Adv.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# White-box Attacks

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>SVHN</th>
<th>Tiny-ImageNet (Top-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madry et al.</td>
<td>91.6%</td>
<td>45.0%</td>
<td>47.1%</td>
<td>22.1%</td>
</tr>
<tr>
<td>ME-Net + SGD</td>
<td>82.6%</td>
<td>40.8%</td>
<td>43.4%</td>
<td>18.9%</td>
</tr>
<tr>
<td>ME-Net + Adv.</td>
<td>91.0%</td>
<td>52.8%</td>
<td>69.4%</td>
<td>28.5%</td>
</tr>
</tbody>
</table>

*Improve white-box robustness when combined with AT!*
Customized Adaptive Attacks

Neural nets
Also Robust to Customized Adaptive Attacks [Results in Paper]
Also Robust to Customized Adaptive Attacks [Results in Paper]

Attack low-rank space

Attack processed images

Neural nets
Please visit our poster #63