Learning Adversarially Robust and Rich Image Transformations for Object Detection
Matthew Tan, Kimberly Te, and Nicholas Lai
mratan@stanford.edu, kimte@stanford.edu, nicklai@stanford.edu

Motivation
- Need:
  - State-of-the-art object classification algorithms are incredibly susceptible to adversarial perturbations. These adversarial attacks undermine the effectiveness of neural network models, and pose ethical concerns and safety risks in real systems, such as facial recognition and autonomous driving. Thus, robust and safe systems need adversarial defense strategies.
  - Simple input transformations can help defend against adversarial attacks (Dziugaite et al. and Guo et al.).
- Objective: To learn adversarially robust image transformations as defenses for object classification tasks.

Data and Defenses
- Datasets:
  - MNIST: Black and white handwritten digits split up into 60,000 training examples and 10,000 test examples.
  - CIFAR-10: 32-by-32 color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.
- Defenses:
  - JPEG Compression: Quantization method that removes small perturbations based on JPEG subspace (90%)
  - Image Augmentation: Smoothing with Gaussian blur filter
  - K-Means Compression: Assign clusters with randomly initialized centroids to remove artefacts (centroids = 16, centroids = 50)
  - Vector Quantized-Variational AutoEncoder (VQ-VAE): Variant of variational encoder and decoder which uses discrete latent variables.

Attacks and Models
- Attacks: White-box attacks implemented using the FoolBox frameworks to test robustness of the image transformations
  - Fast Gradient Sign Method (FGSM): \[ x' = x + \varepsilon \text{sgn} (\nabla_x L(\theta, x, y)) \]
  - Projected Gradient Descent (PGD): \[ x^{t+1} = \Pi_B(x^t - \varepsilon \text{sgn}(\nabla_x L(\theta, x^t, y))) \]
  - CarliniWagnerL2Attack: \[ \min_{\mathbf{w} \in \mathbb{R}^d} \left[ \frac{1}{2} \| \mathbf{x} - \mathbf{w}_0 \|_2^2 + \frac{1}{2} \varepsilon^2 \frac{1}{2} \max_{1 \leq i \leq d} \left| w_i - w_{i,0} \right| \right] \]
  - DeepFool Attack: \[ w_{max} = [x, \| x \|_2 \mathbf{1} + \| x \|_2 (x - x_0)] \]
- MNIST Model: 2 Convolutional + 2 Fully-Connected Layer Neural Network
- CIFAR-10 Model: Transfer learning with modified pre-trained DenseNet model.

Results
- Prediction Accuracies for Attacks vs. Defenses on (MNIST, CIFAR)

<table>
<thead>
<tr>
<th>Attacks</th>
<th>No Attack</th>
<th>JPEG</th>
<th>Image Aug</th>
<th>K-Means</th>
<th>VQ-VAE 0.05</th>
<th>VQ-VAE 0.25</th>
<th>VQ-VAE 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGSM</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
<td>0.99, 0.85</td>
</tr>
<tr>
<td>PGD</td>
<td>0.00, 0.00</td>
<td>0.99, 0.65</td>
<td>0.52, 0.24</td>
<td>0.79, 0.545</td>
<td>0.13, 0.28</td>
<td>0.2, 0.26</td>
<td>0.71, 0.23</td>
</tr>
<tr>
<td>Carlini Wagner2</td>
<td>0.00, 0.00</td>
<td>0.99, 0.67</td>
<td>0.67, 0.28</td>
<td>0.84, 0.612</td>
<td>0.14, 0.28</td>
<td>0.22, 0.27</td>
<td>0.81, 0.23</td>
</tr>
<tr>
<td>Deep Fool</td>
<td>0.00, 0.00</td>
<td>0.02, 0.05</td>
<td>0.01, 0.11</td>
<td>0.023, 0.057</td>
<td>0.03, 0.23</td>
<td>0.02, 0.22</td>
<td>0.03, 0.2</td>
</tr>
</tbody>
</table>

Conclusion/Future Work
- Image compression and neural network-based techniques show potential as defensive image transformations in removing adversarial noise.
- Future Work:
  - Fine-tune VQ-VAE and adjust hyperparameters.
  - Apply techniques to larger and more complex datasets (eg. CELEB-A, ImageNet)
  - Explore and combine other techniques in image compression and feature learning to remove (eg. Total Variance Minimization, other denoising autoencoders, GANs)

Selected Works: