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:** Input image transformations by applying black-box compression and deep learning techniques
  - **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.

Figure 1: Examples of VQ-VAE reconstruction on MNIST (top), CIFAR-10 (bottom) without any adversarial perturbations.

Attacks and Models

- **Attacks**: White-box attacks implemented using the FoolBox frameworks to test robustness of the image transformations

Fast Gradient Sign Method (FGSM)	$x + \varepsilon\,\mathrm{sgn}(\nabla_x L(\theta, x, y)).$
Projected Gradient Descent (PGD)	$x^{t+1} = \Pi_{x+\mathcal{S}}\left(x^t + \alpha\,\mathrm{sgn}(\nabla_x L(\theta, x, y))\right).$
CarliniWagnerL2Attack	$\text{minimize}\ \ \lVert\frac{1}{2}(\tanh(w)+1) - x\rVert_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)$
DeepFool Attack	$\arg\min_{\boldsymbol{r}} \lVert\boldsymbol{r}\rVert_2$ s.t. $\exists k : \boldsymbol{w}_k^\top(\boldsymbol{x}_0+\boldsymbol{r}) + b_k \geq \boldsymbol{w}_{\hat{k}(\boldsymbol{x}_0)}^\top(\boldsymbol{x}_0+\boldsymbol{r}) + b_{\hat{k}(\boldsymbol{x}_0)},$

- **MNIST Model:** 2 Convolutional + 2 Fully-Connected Layer Neural Network
- **CIFAR-10 Model:** Transfer learning with modified pre-trained DenseNet model.

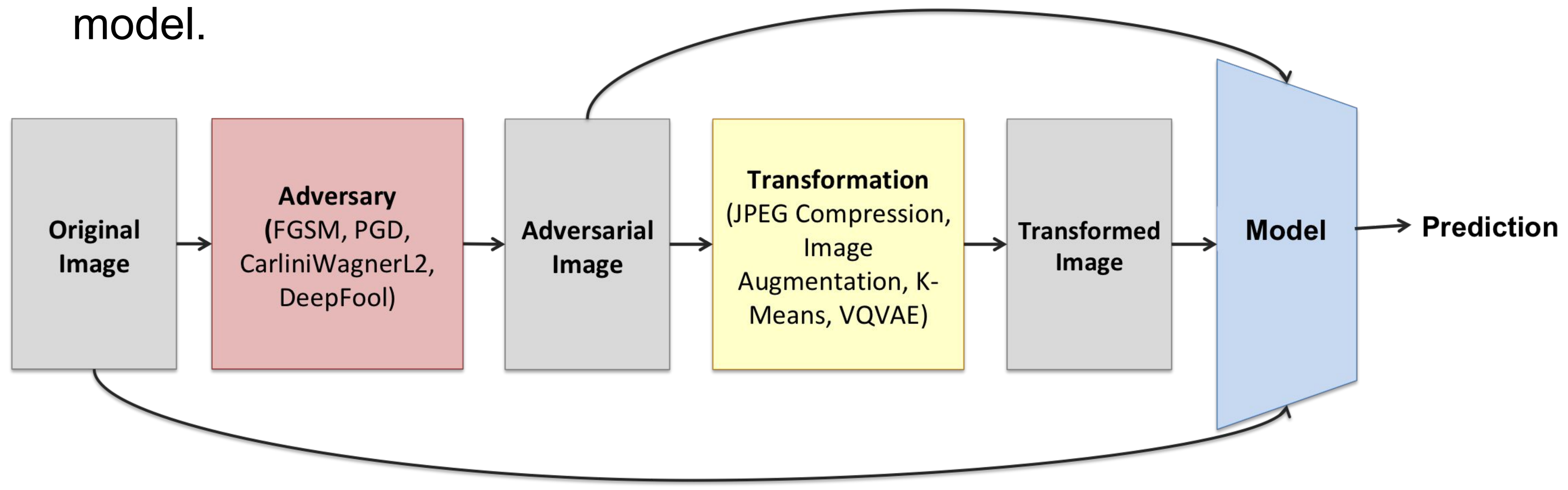

Figure 2: Flow chart diagram of the experimental process. Original images, adversarial images, and their transformed images were run through the models with various attacks and defenses.

Results

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

(MNIST, CIFAR)	No Defense	JPEG	Image Aug	K-Means	VQVAE 0	VQVAE 0.25	VQVAE 0.75
No Attack	0.99, 0.85	0.99, 0.85	0.99, 0.85	0.99,0.85	0.99, 0.85	0.99, 0.85	0.99, 0.85
FGSM	0.00, 0.00	**0.99, 0.65**	0.52, 0.24	0.701, 0.545	0.13, 0.28	0.2, 0.26	0.71, 0.23
PGD	0.00, 0.00	**0.99, 0.73**	0.67, 0.28	0.844, 0.612	0.14, 0.28	0.22, 0.27	0.81, 0.23
Carlini Wagner2	0.00, 0.00	**0.99, 0.67**	0.7, 0.26	0.82, 0.555	0.15, 0.28	0.47, 0.27	0.86, 0.23
Deep Fool	0.00, 0.00	0.02, 0.05	0.01, 0.11	0.023, **0.057**	0.03, 0.23	0.02, 0.22	**0.03**, 0.2

Figure 3: Highest accuracies against a given attack are bolded. JPEG compression gave the highest accuracies for the most attacks on MNIST and CIFAR. For VQ-VAE variations indicates standard deviation of noise added.

Discussion

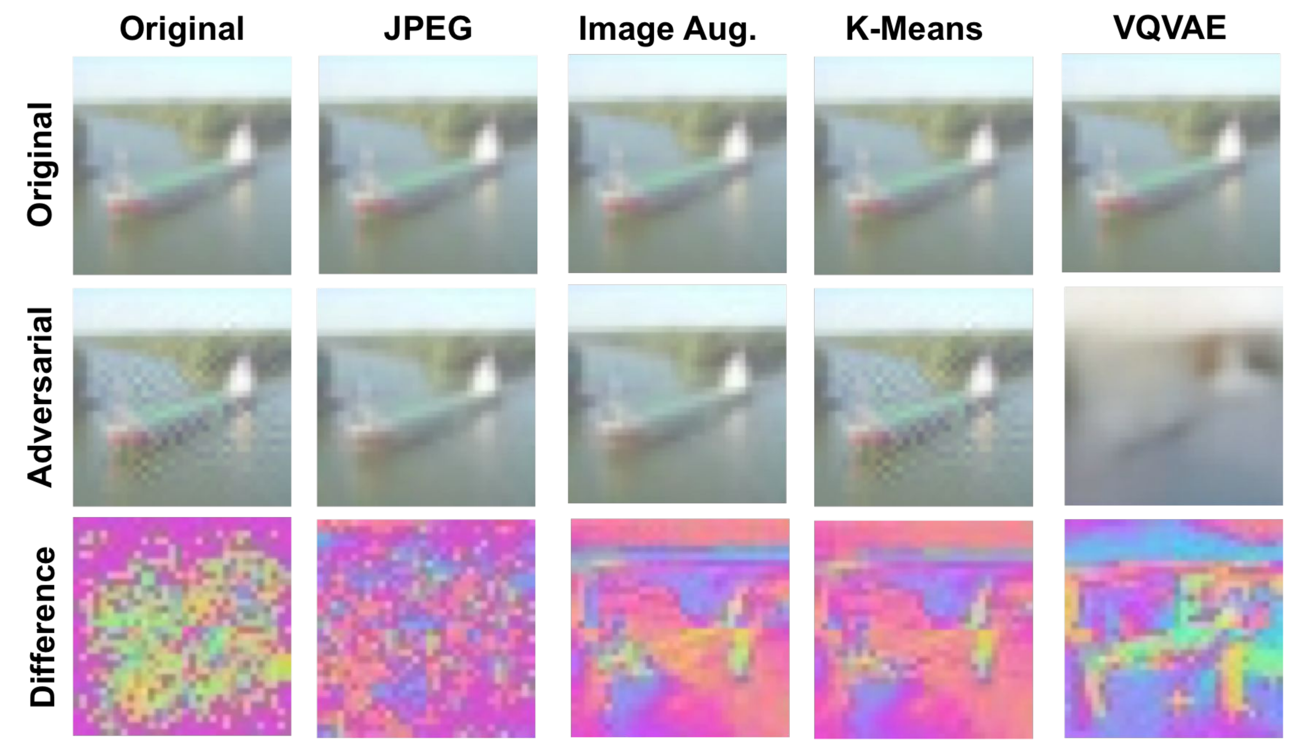

Figure 4: Examples of image transformations on the original image, DeepFool, and their noise differences.

- Adversarial attacks were effective against undefended object classification methods with 0.00% accuracy on both datasets. DeepFool was the most harmful, yet produced the least realistic images.
- Despite large differences between original and adversarial examples after transformation, the defenses improved accuracies. They appeared to minimize perturbations and preserved useful features. JPEG defense showed highest accuracies on both the MNIST and CIFAR datasets. K-Means had the next highest performance.
- VQ-VAE was also investigated as a possible defense. Adding noise to the VQ-VAE training was effective in the MNIST dataset, but less so in the CIFAR dataset.
- Image transformations were more effective on the MNIST dataset over CIFAR-10 dataset, possibly due to its complexity of images.

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:

Guo, Chuan, et al. "Countering adversarial images using input transformations." arXiv preprint arXiv:1711.00117 (2017).
Dziugaite, Gintare Karolina, Zoubin Ghahramani, and Daniel M. Roy. "A study of the effect of jpg compression on adversarial images." arXiv preprint arXiv:1608.00853 (2016).