Supplementary Material for "Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation"

Alexander Levine and Soheil Feizi
University of Maryland, College Park
{alevine0, sfeizi}@cs.umd.edu

Architecture and Training Parameters for MNIST

See Tables 1 and 2.

Layer	Output Shape
(Input)	$2 \times 28 \times 28$
2D Convolution + ReLU	$64 \times 14 \times 14$
2D Convolution + ReLU	$128 \times 7 \times 7$
Flatten	6272
Fully Connected + ReLU	500
Fully Connected + ReLU	100
Fully Connected + SoftMax	10

Table 1: Model Architecture of the Base Classifier for MNIST Experiments. 2D Convolution layers both have a kernel size of 4-by-4 pixels, stride of 2 pixels, and padding of 1 pixel.

Training Epochs	400
Batch Size	128
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	.01 (Epochs 1-200) .001 (Epochs 201-400)
Momentum	0.9
L_2 Weight Penalty	0

Table 2: Training Parameters for MNIST Experiments

Training Parameters for CIFAR-10

As discussed in the main text, we used a standard ResNet18 architecture for our base classifier: the only modification made was to increase the number of input channels from 3 to 6. See Table 3 for training parameters.

Training Parameters for ImageNet

As with CIFAR-10, we used a standard ResNet50 architecture for our base classifier: the only modification made was to increase the number of input channels from 3 to 6. See Table 4 for training parameters.

Training Epochs	400
Batch Size	128
Training Set Preprocessing	Random Cropping (Padding:4) and Random Horizontal Flip
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	.01 (Epochs 1-200) .001 (Epochs 201-400)
Momentum	0.9
L_2 Weight Penalty	0.0005

Table 3: Training Parameters for CIFAR-10 Experiments

Training Epochs	36
Batch Size	256
Training Set Preprocessing	Random Resizing and Cropping, Random Horizontal Flip
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	.1 (21 Epochs) .01 (10 Epochs) .001 (5 Epochs)
Momentum	0.9
L_2 Weight Penalty	0.0001

Table 4: Training Parameters for ImageNet Experiments

Mutual information derivation for Lee et al. 2019

Here we present a derivation of the expression given in Equation 21 in the main text. Let $\mathbf{X}$ be a random variable representing the original image: in this derivation, we assume that $\mathbf{X}$ is distributed uniformly in $\mathcal{S}^d$. Let $\mathbf{Y}$ be a random variable representing the image, after replacing each pixel with a random, different value with probability $(1-\kappa)$. By the definition of mutual information, we have:

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \tag{1}$$

Note that, with $\mathbf{X}$ distributed uniformly, it consists of d i.i.d. instances of a random variable $X_\circ$, itself uniformly distributed in $\mathcal{S}$. Similarly, each component of $\mathbf{Y}$ is an instance of a random variable defined by:

$$Y_\circ = \begin{cases} X_\circ & \text{with probability } \kappa \\ \text{Uniform on } \mathcal{S} - \{X_\circ\} & \text{with probability } 1-\kappa \end{cases} \tag{2}$$

We can then factorize the expression for mutual information, using the fact that each instance of $(X_\circ, Y_\circ)$ is independent:

$$I_{\text{Lee et al.}} = I(\mathbf{X}, \mathbf{Y}) = d(H(X_\circ) - H(X_\circ|Y_\circ)) \tag{3}$$

By the definitions of entropy and mutual entropy, we have:

$$\begin{aligned} I_{\text{Lee et al.}} = -d\Bigg(&\sum_{s\in\mathcal{S}} \Pr(X_\circ = s)\log_2 \Pr(X_\circ = s) \\ &- \sum_{(s,s')} \Pr(X_\circ = s, Y_\circ = s')\log_2 \frac{\Pr(X_\circ = s, Y_\circ = s')}{\Pr(Y_\circ = s')}\Bigg) \end{aligned} \tag{4}$$

Note that, by symmetry, $Y_\circ$ is itself uniformly distributed on $\mathcal{S}$. Then we have:

$$\begin{aligned} I_{\text{Lee et al.}} = -d\Bigg(&\sum_{s\in\mathcal{S}} |\mathcal{S}|^{-1}\log_2 |\mathcal{S}|^{-1} \\ &- \sum_{(s,s')} \Pr(X_\circ = s, Y_\circ = s')\log_2 \frac{\Pr(X_\circ = s, Y_\circ = s')}{|\mathcal{S}|^{-1}}\Bigg) \end{aligned} \tag{5}$$

Splitting (s, s') into cases for $(s = s')$ and $(s \neq s')$:

$$\begin{aligned} I_{\text{Lee et al.}} = -d\Bigg(&\sum_{s} |\mathcal{S}|^{-1}\log_2 |\mathcal{S}|^{-1} \\ &- \sum_{s} \Pr(X_\circ = Y_\circ = s)\log_2 \frac{\Pr(X_\circ = Y_\circ = s)}{|\mathcal{S}|^{-1}} \\ &- \sum_{s\neq s'} \Pr(X_\circ = s, Y_\circ = s')\log_2 \frac{\Pr(X_\circ = s, Y_\circ = s')}{|\mathcal{S}|^{-1}}\Bigg) \end{aligned} \tag{6}$$

Note that $\Pr(X_\circ = Y_\circ = s) = |\mathcal{S}|^{-1}\kappa$, because $X_\circ = s$ with probability $|\mathcal{S}|^{-1}$, and then $Y_\circ$ is assigned to $X_\circ$ with probability κ. Also, for $s \neq s'$, we have

$$\Pr(X_\circ = s, Y_\circ = s') = |\mathcal{S}|^{-1}(1-\kappa)(|\mathcal{S}|-1)^{-1}, \tag{7}$$

because $X_\circ = s$ with probability $|\mathcal{S}|^{-1}$, $Y_\circ$ is not equal to $X_\circ$ with probability $(1-\kappa)$, and then $Y_\circ$ assumes each value in $\mathcal{S} - \{X_\circ\}$ with uniform probability. Plugging these expressions into Equation 6 gives:

$$\begin{aligned} I_{\text{Lee et al.}} = -d\Bigg(&\sum_{s} \frac{\log_2 |\mathcal{S}|^{-1}}{|\mathcal{S}|} - \sum_{s} \frac{\kappa}{|\mathcal{S}|}\log_2 \kappa \\ &- \sum_{s\neq s'} \frac{(1-\kappa)}{(|\mathcal{S}|-1)|\mathcal{S}|}\log_2\left[(1-\kappa)(|\mathcal{S}|-1)^{-1}\right]\Bigg) \end{aligned} \tag{8}$$

Now all summands are constants: we note that summing over all $s \in \mathcal{S}$ is now equivalent to multiplying by $|\mathcal{S}|$ and summing over $(s, s') \in \mathcal{S}^2$ with $s \neq s'$ is equivalent to multiplying by $|\mathcal{S}|(|\mathcal{S}|-1)$:

$$\begin{aligned} I_{\text{Lee et al.}} = -d\big(&\log_2 |\mathcal{S}|^{-1} - \kappa\log_2\kappa \\ &- (1-\kappa)\log_2\left[(1-\kappa)(|\mathcal{S}|-1)^{-1}\right]\big) \end{aligned} \tag{9}$$

This simplifies to the expression given in the text.

Additional Adversarial Examples

See Figure 1.

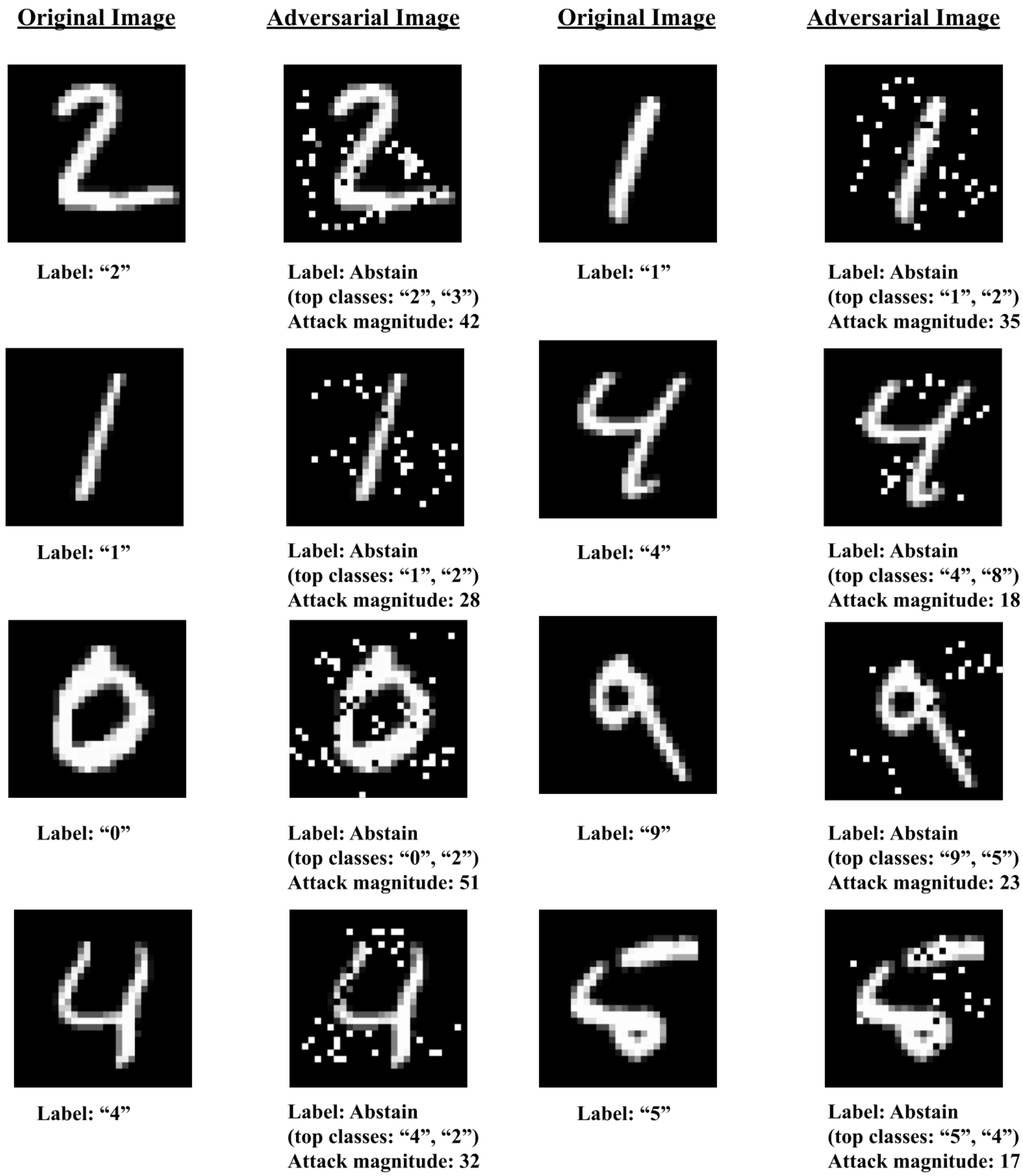

Figure 1: Additional adversarial examples generated on MNIST by the Pointwise attack on our robust classifier, with $k = 45$.