Progress for March 12

Adam Spindler

March 10, 2021

1 Translation Invariance Experiment

The standard approach for creating adversarial images is to iteratively update an image so that the loss of the target model increases. The image is updated by computing the derivative of the loss function with respect to the input image, which gives you a tensor that points in the direction of increasing the loss function. This gradient is multiplied by a small constant (like 0.01) and added to the input image to slightly move it in the direction of increasing loss. This process is repeated iteratively until the image is modified enough to cause a mis-detection. For my experiments, I go several iterations beyond that to improve transferability.

In another attempt to increase the transferability of my adversarial examples I implemented the translation invariance technique as described in [1]. The idea behind this approach is to shift the adversarial image during its creation so that the adversarial image could be shifted after creation and still retain its adversarialness. This was also shown to increase transferability between models in their paper.

The way that this works specifically is that each iteration of the adversarial image creation, several copies of the current version of the image are made where each one is translated by a small amount. Then each image is passed through the model and its gradient is computed for the loss function. Each of these gradients are then translated back to where they would be on the original input image and averaged together. This combined gradient is then added to the input image. A graphical representation is visible in figure 1, and this idea is expressed mathematically below:

Original:

$$X := X + \epsilon \nabla_X f_{\text{ensemble}}(X) \tag{1}$$

With Translation Invariance:

$$X := X + \epsilon \frac{1}{N} \sum_{i=1}^{N} T_i^{-1}(\nabla_X f_{\text{ensemble}}(T_i(X))) \tag{2}$$

Where:

X	Current version of input image
ϵ	Learning Rate
f_{ensemble}	Ensemble of models which outputs a scalar loss, which is basically the sum of confidence scores
N	Number of translations
T	Translation function
T^{-1}	Inverse of translation function

For this experiment the following set of translations were used each iteration:

```
[(0, 0), (2, 2), (4, 0), (0, 4), (-4, -4), (-8, 0), (0, -8), (0, 8), (0, 8), (8, 8), (-8, -8)]
```

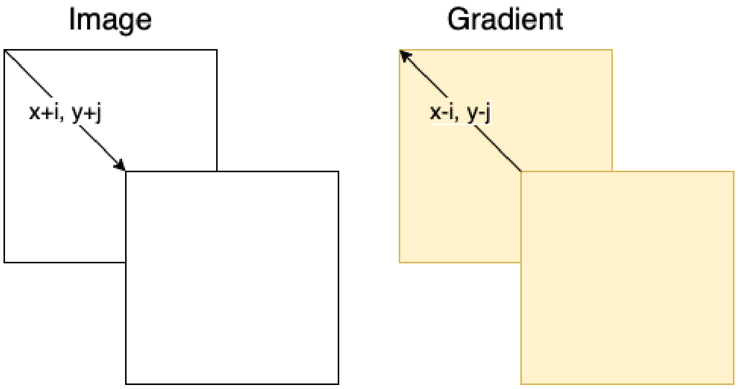

Figure 1: The image is translated in the X and Y directions, shifting it before being passed into the network. Then the gradient is computed and shifted back to align its values with the original image.

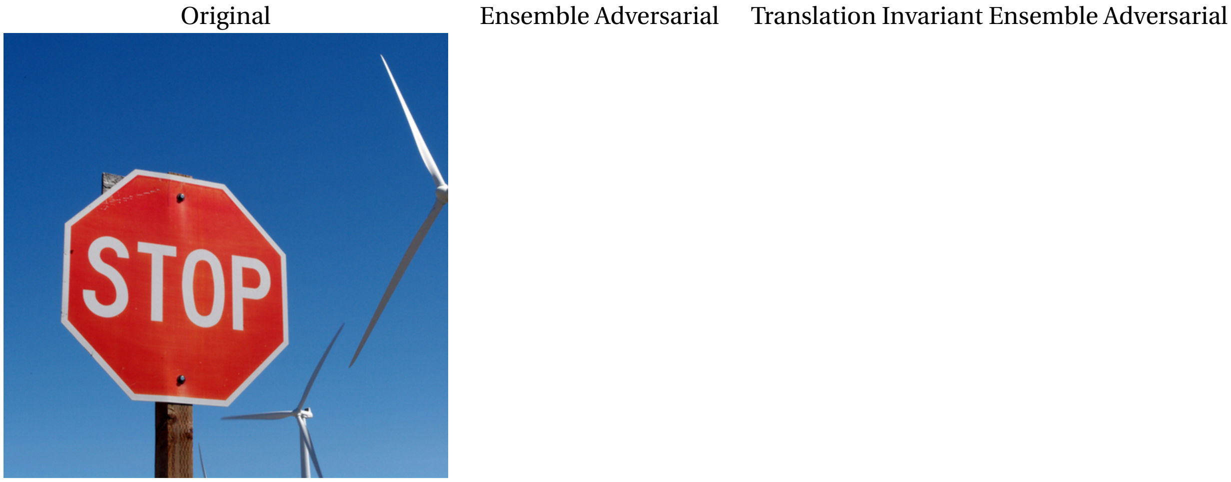

Figure 2: Visual Comparison of Adversarial Image Techniques

References

[1] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks, 2019.