



Introduction

This defense is implemented using a two-step procedure.

- 1) Apply several transformations to the input image to make the attack perturbations less effective.
- 2) Classify using networks trained with adversarial examples. This kind of training helps the networks generalize better to images that are not just outside the training set, but do not even occur in “nature”.






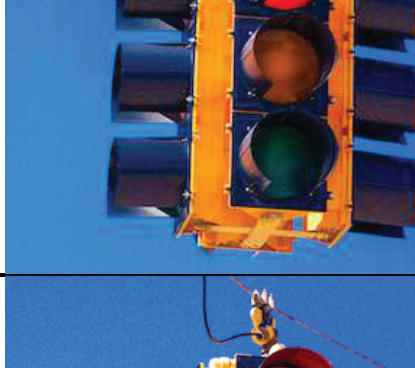
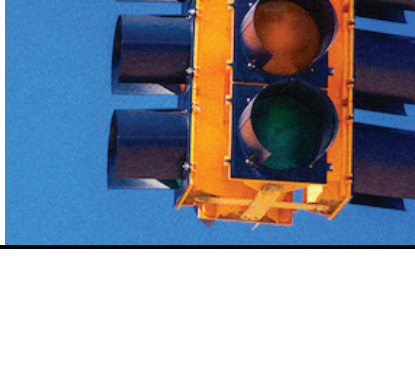
These steps improve the robustness of a classifier. In this context, the robustness of a classifier \mathcal{F} can be defined as $\mathbb{E}[\Delta_{adv}(\mathcal{X}, \mathcal{F})]$ where $\Delta_{adv}(\mathcal{X}, \mathcal{F})$ is the minimum distortion required to misclassify a sample.

$$\Delta_{adv}(\mathcal{X}, \mathcal{F}) = \arg \min_{\delta \mathcal{X}} \|\delta \mathcal{X}\|_{\infty} : \mathcal{F}(\mathcal{X} + \delta \mathcal{X}) \neq \mathcal{F}(\mathcal{X})$$

Source code

Available at <https://github.com/anlthms/nips-2017>

Image transformations

Transformation	Adversarial Image	P(traffic light)	P(hummingbird)
None		2.2e-05	0.9998
Rotate by 5 degrees		0.9976	9.5e-06
Shift by 5%		0.9887	5.4e-05
Shear by 5 degrees		0.9983	1.7e-05
Zoom in by 5%		0.9988	4.9e-06
JPEG compress with quality=50		0.9981	3.4e-06
Sprinkle gaussian noise		0.9856	2.8e-04

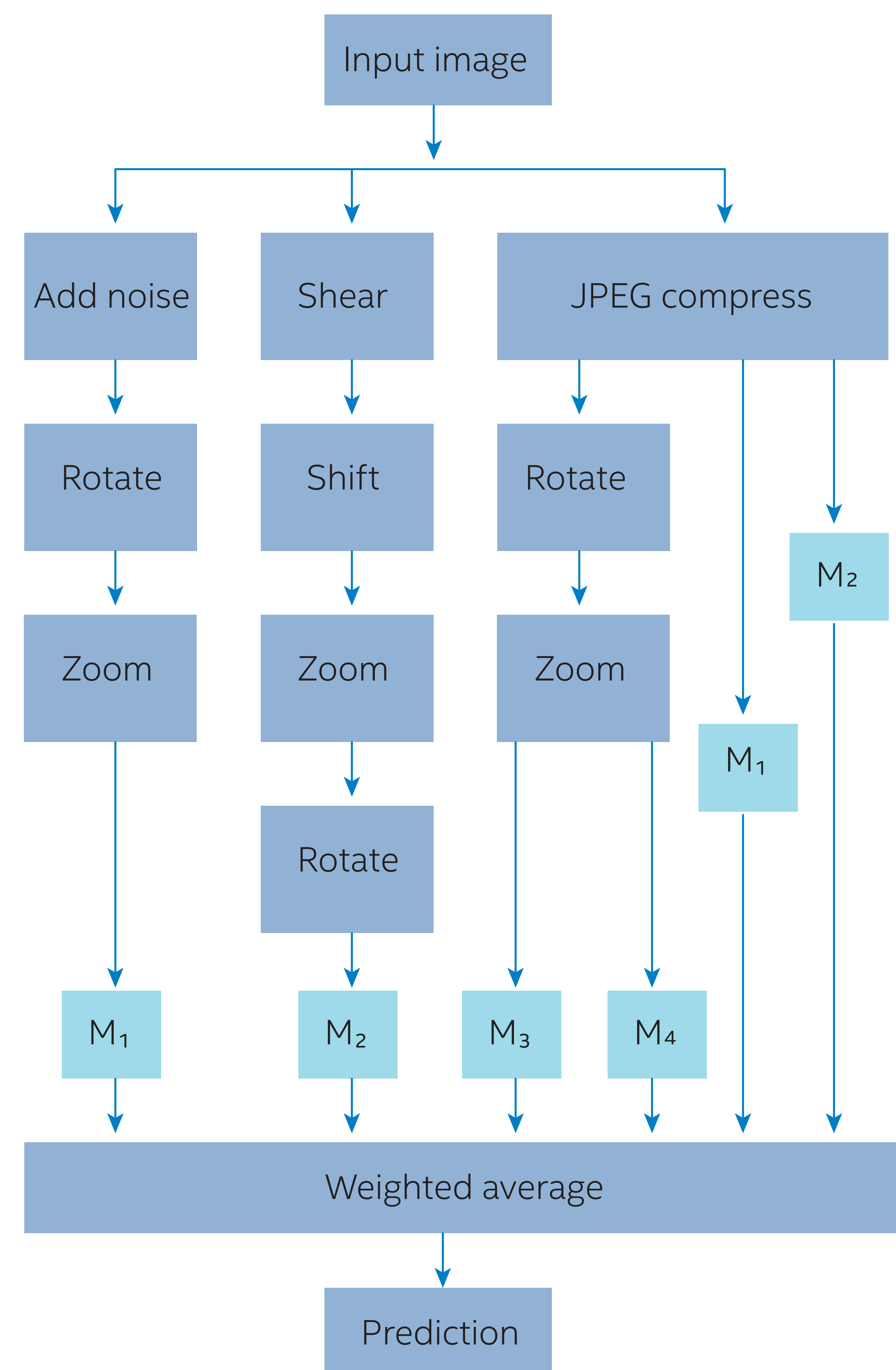
The original version of this image is classified by Inception V3 as a “traffic light” with high confidence. An adversarial version was produced by iteratively perturbing the image until it was misclassified by the network as a “hummingbird” with high confidence.

Transforming the adversarial image in any of the different ways shown on the left renders the adversarial attack ineffective.

However, it has been shown that it is possible to craft adversarial images that defeat this defense. This can be achieved by using the same transformations while adversarial gradients are computed.

The next panel shows how such image transformations were employed in the contest entry.

Complete pipeline



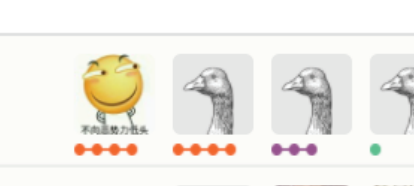
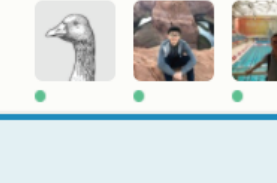
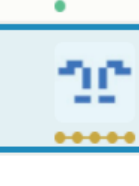
Models:

- M₁ - Adversarially trained Inception V3*
- M₂ - Adversarially trained Inception V3 - Resnet V2
- M₃ - VGG16
- M₄ - Resnet V2 152

*Adversarially trained models were provided by the contest organizers.

Conclusions

Private Leaderboard

#	Team Name	Team Members	Score
1	TsAIL		0.95316
2	iy swim		0.92352
3	Anil Thomas		0.91483

The defense outlined here was able to do well against the attacks featured in this contest by using a combination of image transformations and adversarially trained models. However, it should be noted that it would be harder to defend against a newly created attack that has intimate knowledge of this defense.

The defense could be strengthened by using it to produce gradients that can be used in new attacks and then fine-tuning the model parameters by training with new adversarial examples thus created.