# Paper

- **Title:** Countering Adversarial Images using Input Transformations
- **Authors:** Chuan Guo, Mayank Rana, Moustapha Cissé, Laurens van der Maaten
- **arXiv link:** https://arxiv.org/abs/1711.00117

# TL;DR

The paper describes five strategies that can be used to counter adversarial images by transforming the inputs before feeding them to the network.

# Motivation

- Recent research (Goodfellow et al. 2015, Cisse et al. 2017) has shown that existing deep learning models are not robust to small, adversarially designed perturbations of the input.
- The paper provides various effective strategies against such adversarial attacks.

# The attacks

The paper considers four attacks, a brief description of them is:

1. Fast Gradient Sign Method (FGSM, Goodfellow et al. 2015) : Let $l(\cdot, \cdot)$ be the loss function that was used to train the classifier. The FGSM adversarial example corresponding to $x$ with label y will be:

   $$x' = x + \epsilon \ sign(\nabla_x l(x, y))$$

   Here, $\epsilon$ governs the perturbation magnitude.

2. Iterative FGSM (I-FGSM, Kurakin et al. 2017) : This is stronger than FGSM, as it iteratively applies the FGSM update:

   $$x^m = x^{m-1} + \epsilon \ sign(\nabla_{x^{m-1}} l(x^{m-1}, y))$$

   where $x^0 = x$ and $x' = x^m$
   Both FGSM and I-FGSM aims to minimize the Chebyshev distance($L_\infty$ norm) between the inputs and the adversarial examples they generate.

3. DeepFool (Seyed-Mohsen et al. 2016) : projects $x$ onto a linearization of the decision boundary defined by $h(\cdot)$ for M iterations, where $h(\cdot)$ is a binary classifier.

$$x^m = x^{m-1} - \epsilon \frac{h(x^{m-1})}{||\nabla_{x^{m-1}} h(x^{m-1})||_2^2} \nabla_{x^{m-1}} h(x^{m-1})$$

where $x^0 = x$ and $x' = x^m$
For a multiclass classifier $h(\cdot)$, we take the projections on the nearest class boundary.

4. Carlini Wagner's $L_2$ attack (CW-L2, Carlini & Wegner 2017) : Let $Z(x)$ be the operation that computes the logits for input $x$, and $Z(x)_k$ be the logit value for the class k. CW-L2 finds the solution to

$$\min_{x'}[||x - x'||_2^2 + \lambda_f max(-\kappa, Z(x')_{h(x)} - max(Z(x')_k :\neq h(x)))]$$

where $\kappa$ denotes the margin parameter and $\lambda_f$ is the trade off between perturbation norm and the hinge loss of predicting a different class.

## The defenses

The paper considers five input transformation methods:

1. Image cropping ans rescaling : Image cropping-rescaling has the effect of altering the spatial positioning of the adversarial perturbation, which is important in making attacks successful.
At training time, crop and resize images as part of data augmentation and at test time, average predictions over random image crops.

2. Bit-depth reduction : This was introduced by Xu et al., 2017. It performs simple quantization that can remove small (adversarial) variations in pixel values from an image.

3. JPEG compression : JPEG compression removes small perturbations from images as it removes high frequencies after converting image to frequency domain.

4. Total variance minimization : Pixel dropout is combined with total variance minimization. A subset of pixels is selected and then the "simplest" image that is consistent with this subset is reconstructed i.e.

   (a) Select a random subset of pixels by sampling a Bernoulli random variable $X(i, j, k)$ for each pixel $(i, j, k)$

(b) Use TV minimization to reconstruct an image $z$ that is similar to input image $x$ for a subset of selected pixels and is simple in terms of total variation by solving:

$$\min_z ||(1-x) \odot (z-x)||_2 + \lambda_{TV} \cdot TV_p(z)$$

where $\odot$ is the $L_p$-total variation of z

$$TV_p(z) = \sum_{k=1}^{K} \left[ \sum_{i=2}^{N} ||z(i,:,k) - z(i-1,:,k)||_p + \sum_{j=2}^{N} ||z(:,j,k) - z(:,j-1,k)||_p \right]$$

Thus, small perturbations that result in very different pixel values among neighbouring pixels are discouraged upon TV minimization.

5. Image quilting : Image quilting (Efros & Freeman 2001) is a non-parametric technique that makes an image by combining small patches that are taken from a database of images. The database is predefined and in order to reduce edge artifacts, it computer minimum graph cuts as described in (Boykov et al. 2001)

## Conclusion

Image quilting and TV minimization turn out to be the most effective strategies. The authors note that this is due to the fact that a strong input transformation defense should have non-differentiability and randomness, both of which are present in Image quilting and TV minimization.