Spatially Transformed Adversarial Examples

Ravi Raju

12-22-2018

1 Motivation

The point of this paper was to diversify the type of adverserial perturbations against deep models. Previously, we looked optimization based attacks like Carlini-Wagner attack/FGSM. The problem with these methods is that, while producing minimal perturbations, they are not representative of real world adverseries.

2 Solution

The new method, spAdv, is changing some of the pixel positions rather than directly modifying the pixels in the image. They alter the image with a per-pixel flow field f to synthesize the image. (Is this generalizable across all images)? It looks like it is based off of an interpolation. But it also looks like an optimization problem is being solved. This is looking very analgous to the CW attack except the regularizer being used deals with the total variation. (Rudin et al. 1992).

3 Results

This was attacking average pooling along with adverserial training. An advantage of this method is that it smoothly deforms digits (MNIST) so that they look natural. In the CIFAR 10 example and MNIST examples, qualitiatively it's easier to see that these images are more realistic than FGSM, CW. But solving CW would provided a tighter bound so is this paper considering we do not need to solve a problem like CW because it is not realistic?

4 Discussion/Takeaway

I think we should use the CW attacks for a baseline attack. But we should consider this type of transformation as something to be done on top. The class activation mapping (CAM) would be a very interesting idea to check out on the intermediate of DNNs.