Review: Building Adversary Resistant Deep Neural Networks with Random Feature Nullification

Summary

The paper presents the random feature nullification to preserve the DNN's integrity against adversarial samples. Based on the discussion on the related work to increase the DNN's resistance to adversarial samples, the non-deterministic technique introduced is a nature progression from the current available techniques of adversarial training and model complexity enhancement. (weakness of adversarial training and model complexity enhancement here?) The paper points out that to be able to put up a strong resistance against adversarial attacks, the model architecture has to be one makes it impossible to generate adversarial samples even when it is disclosed.

Positivity

- the introduced method is more efficient than the method of adversarial training; there is no need to train on both the real sample and a crafted adversarial sample which is an important aspect as the amout of real samples we potentially have is inexhaustive.
-

Negativity

My Thoughts

- how does the techique to create adversarial samples work? How do we add the perturbation to make the DNN believe that it is of a particular class A or B when it is actually of class C?
- why compute $\frac{\partial \mathcal{L}(\theta, f(\tilde{X}, I_p), Y)}{\partial \tilde{X}}$ instead of $\frac{\partial \mathcal{L}(\theta, f(\tilde{X}, I_p), Y)}{\partial f(\tilde{X}, I_p)}$? Would seeing the data as $f(\tilde{X}, I_p)$ be better than seeing it as $\tilde{X}$? Or perhaps I should first myself if that derivative can be computed. (my tentative answer is yes). Although I_p is a random variable, what it does is that it nullifies every sample to 0 with proportion p. This can be seen as a form of occlusion with the image being occluded at many small spots.
- is $\delta X \odot I_p^*$ a notation or the exact representation of the perturbation?
- can the attacker not draw a sample I_p^* which then the nullified adversarial sample looks like

$$(X + \delta X \odot I_p^*) \odot I_p = (X \odot I_p) + \delta X \odot I_p^* \odot I_p$$

- could explore deeper on how attackers can bypass the random feature nullification technique other than the traditional method used by them to produce adversarial samples.
- could have tested it on CIFAR dataset since MNIST dataset is a smaller space than the CIFAR dataset and this technique might not work as well on images with more colors and variety. There are papers that the author cited from that discussed about adversarial examples in other datasets.
- why is there a choice of different activation function and learning rate to train the models which makes the comparison not fair.
- game theory aspects of the paper: every sample is seen as an agent and they are competing to have their cost reduced, i.e. to be classified correctly. However in the midst of this competition, the social cost of the DNN (which is the error made by the DNN) is increased as it is susceptible to adversarial attacks. Hence we want to construct the DNN such that the NE of the game is such that it is not susceptible to the adversarial samples.

References

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.