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

Summary

The paper presents the random feature nullification (RFN) to preserve the DNN's integrity against adversarial samples. Based on the discussion on the currently related work to increase the DNN's resistance to adversarial samples, the inclusion of a non-deterministic feature is a very 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 that makes it impossible to generate adversarial samples even when it is disclosed. However, [Goodfellow et al., 2014]

Positivity

The discussed approaches to tackle the problem of adversarial attacks have so far been trying to train the model on adversarial samples or to increase the non-linearity of the model. The RFN technique adopts another approach which is to make the construction of effective adversarial samples[1] a tall order for the attackers, through the addition of random variables. Making it hard to construct adversarial samples might be an avenue that future efforts of tackling adversarial attacks might want to explore as the doing adversarial training has its limitations;

- 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

From [Goodfellow et al., 2014] adversarial examples are known to generalize, i.e. an adversarial example generated for one model can be misclassified by another model, although they were trained on different learning architectures using disjoint training sets. Hence although the experimental results shown by the paper does numerically points to better resilience of their RFN method, it is only resilient to adversarial samples constructed using their RFN learning technique; it does not show resilience against adversarial samples generated from other learning techniques.

The paper worked with only the MNIST dataset where each real sample is a 28×28 grayscale image. And the nullification of MNIST data to 0 would simply mean it changed the pixel to a white pixel and since an image from MNIST is predominantly white[2], it is possible that the RFN technique does not really "null" much of the information from the images. A reason for this suspicion is even with their choice of nullification rate to be a high 50%, the classification error is 0.0170. Even with the highest reported nullification rate of 70% the classification error is still a decent 0.0266. Such results are counter-intuitively, as one might feel that by nulling such a high proportion of the image could be though of as the image being occluded and occlusions in images limit the information that could be obtained (might need to find some papers that support my claim here). Such a problem might also cause a problem in trying to extend the RFN technique to colour images like CIFAR-10, CIFAR-100 and ImageNet.

A lack of showing how their RFN technique to datasets of colour images like those mentioned above makes this paper slightly less attractive as compared to if it did some work on the extension of their technique to those datasets. As from the examples shared in [Szegedy et al., 2013] on how adversarial samples are generated from images from ImageNet shows that discussion on adversarial samples are not restricted to MNIST datasets and the complexity

My Thoughts

- how does the technique 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

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

[1] adversarial samples that is able to successful deceive the DNN to make wrong classifications.

[2] no preprocessing of the dataset is assumed here as the paper does not say anything about it

- 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.

- does not talk about the confidence of the predictions.

References

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

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.