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. The problem is motivated by pointing out that a DNN can be tricked into doing a wrong classification by introducing perturbations that the human eye cannot observe. The perturbations are obtained using the *fast gradient sign* method [Goodfellow et al., 2014] where a cost function $\mathcal{L}(\theta, X, Y)$ is differentiated with respect to X denoted by

$$\mathcal{J}_{\mathcal{L}}(X) = \frac{\partial \mathcal{L}(\theta, X, Y)}{\partial X}$$

and the perturbation is calculated by the formula

$$\delta X = \phi \cdot sign(\mathcal{J}_{\mathcal{L}}(X))$$

where ϕ is some coefficient that controls the scale of the gradient to be added. There are constrains on the value of ϕ as a too large value will cause the synthesised adversarial sample to be distorted enough to be detected.

Current works to tackle the problem of adversarial samples approach the problem in two main ways: adversarial training and increasing the model complexity. In the former, [Szegedy et al., 2013] performed back-feeding of adversarial examples to the training set; this method of data augmentation does not train the DNN on adversarial samples that they have not been exposed which makes this technique vulnerable. The strategy of increasing model complexity was not discussed in detail but the paper points out that any attempt to increase the model complexity will be futile should the architecture is known to the attacker. As such, the paper's idea is to design a technique such that even if the full knowledge of the architecture is known to the attacker, its model will still be resistant to adversarial attacks.

With this motivation, the paper explains its RFN technique which is different from the earlier two techniques as it introduces a random variable I_p- a matrix with dimensions same as the real sample and each i, j component to be Bernoulli distributed with parameter p that nullifies the real sample $X \odot I_p$ where $\odot$ is the Hardmard-Product. It is somewhat similar to Hinton's dropout technique, but the nullification takes place in both training and classification (the paper used the term testing) phase. The argument that the paper uses to justify its resistance against adversarial attacks is as follows: the usage of I_p makes finding the derivative for backpropagation impossible[1] and attackers need to approximate I_p to generate an adversarial perturbation δX. An adversarial sample produced using its model this has the form in (1) and due to the $I_p \odot I_p^*$ it weakens δX as the most impactful adversarial perturbation.

The paper then completes with some numbers on its classification accuracy and resiliency; there is a lack of a clear definition of resiliency. It then extends its RFN technique to be used together with other adversarial technique; which from Table 2 of the paper shows that a combination of RFN and adversarial training gives the most desirable results.

In the appendix it gives information on the hyper parameter: network structure, activation function, learning rate, batch size and epoch. It is interesting to see that for some reason the sigmoid activation function was used on the solely RFN learning technique whereas the rest utilised the tanh function. Also for the sole adversarial training technique, a learning rate of 1 was used. No information was given on these deviating choices of hyper parameters from the norm used in their experiments.

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 a tall order for the attackers, through the addition of random variables. Making it hard to construct adversarial samples might be a better approach as the doing adversarial training has its limitations; the augmentation of adversarial samples to the training set to increase the resistance of the model from adversarial samples only trains the model well on adversarial samples that it has been exposed to. Also if it is hard to craft effective adversarial samples, the problem of adversarial samples [Szegedy et al., 2013] able to generalise, in the form of *cross model generalisation* and *cross training set generalisation* will not impactful. The RFN method seems to be more of a deterrence technique than a resistive technique as compared to the other methods.

Negativity

In its adversarial resistant analysis on why its RFN method is more robust than the known methods of adversarial resistance, its argument only assumed that the attackers used the same method to obtain perturbations which are then added to the real sample to form the adversarial samples. The discussion on how attackers might tweak their

[1] I do not agree with it as $\frac{\partial f(\tilde{X}, I_p))}{\partial \tilde{X}} = I_p$

methodology to generate adversarial samples effective against them is little, not every clear and not convincing. Based on my naive thinking, the attacker might chose to not draw a sample I_p^* which then the nullified adversarial sample looks like

$$(X + \delta X) \odot I_p = (X \odot I_p) + \delta X \odot I_p \tag{1}$$

thus there is no distortion of the δX. In the first place how much different is $\delta X \odot I_p$ from δX is still not really known based on my point in the third paragraph. But the lack of analysis on the next point makes the RFN method less desirable as a technique to foil adversarial attacks.

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 say anything about 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, which leads to the next point.

As seen in the examples from [Szegedy et al., 2013], we see that the perturbations introduced are not visible to the human eye but are classified by the trained model as classes that are very far from its original class. A lack of showing how their RFN technique to datasets of colour images like those mentioned above makes this paper less complete. Perhaps the RFN technique cannot be applied to colour images or the paper has not considered extending to a more complex dataset with more classes in a bigger space.

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*.

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