Assignment 9

Ana Araújo n 59457

Pedro Almeida nº 58844

Problem 2:

"How may adversarial attack methods be adapted to produce more robust networks?"

Deep Neural Networks (DNNs) have achieved great successes in a variety of applications. However, recent work has demonstrated that DNNs are vulnerable to adversarial perturbations. In adversarial attacks, very small perturbations in the network's input data are performed, which lead to classifying an input erroneously with a high confidence. Such maliciously perturbed instances can cause the learning system to misclassify them into either a maliciously chosen target class (in a targeted attack) or classes that are different from the ground truth (in an untargeted attack). [1], [2]

An adversarial example for the picture classification problem is a natural image that has been purposefully altered to include a visually unnoticeable alteration that can significantly lower classification accuracy.

The following are a few examples of adversarial attack methods that have been modified to create more reliable networks:

Adversarial training is by far the most well-liked and uninterrupted protection strategy for enhancing network robustness. The fundamental principle of adversarial training is to use an adversarial example as training material to prepare a DNN for an adversarial attack. [3]

- **Fast Gradient Sign Method (FGSM):** FGSM algorithm proposed by [4]can deceive neural networks by adding noise linearly to generate countermeasure samples in high-dimensional space.
  The FGSM approach, which is suggested, creates adversarial examples by using the gradients of the loss function with respect to the input. The strategy produces adversarial examples that can trick the model even with modest perturbations by modifying the input in the direction of the gradient's sign. The effectiveness of the FGSM approach versus a variety of models, including deep neural networks, was further demonstrated by the authors. [4]
- **Projected Gradient Descent (PGD):** PGD is an attack method that includes projecting the disturbed input back onto a legal area after iteratively employing the FGSM attack with a small step size. Using adversarial training and adversarial regularization, PGD has been modified to generate networks that are more resilient. For example, in [5], by maximizing the original loss function of the targeted model, the adversarial example with backpropagation was update and the perturbation was optimized by accumulating small updates on perturbed images consecutively. The attack was generated for several modern, CNN classifiers using ImageNet and compare the attack performance with other universal adversarial attack methods. The adversarial attack method showed that can achieve a higher fooling rate and can realize good generalization on cross-model evaluation. [5]
- **DeepFool:** DeepFool is an attack method that involves iteratively finding the nearest decision boundary and perturbing the input to cross that boundary. DeepFool has been adapted to produce more robust networks through adversarial training and adversarial regularization. For example, in [6]proposed an algorithm, DeepFool, to compute adversarial examples that fool other classifiers. It is based on an iterative linearization of the classifier to generate minimal perturbations that are sufficient to change classification labels. They provided extensive experimental evidence on three datasets and eight classifiers, showing the superiority of the proposed method to compute adversarial perturbations, as well as the efficiency of the proposed approach. Due to its accurate estimation of the adversarial perturbations, the proposed DeepFool algorithm provides an efficient and accurate way to evaluate the robustness of classifier. The proposed approach can be used as a reliable tool to accurately estimate the minimal perturbation vectors, and build more robust classifiers. [6]

These techniques have potential for strengthening the defenses of deep learning models against hostile attacks. It is important to keep in mind that new attack strategies that can get past existing defenses are continually being created as part of the ongoing arms race between adversary attacks and defenses.

References:

[1] D. Jakubovitz and R. Giryes, “Improving dnn robustness to adversarial attacks using jacobian regularization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 514–529.

[2] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, “Generating adversarial examples with adversarial networks,” *arXiv preprint arXiv:1801.02610*, 2018.

[3] Z. He, A. S. Rakin, and D. Fan, “Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 588–597.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” Dec. 2014, Accessed: May 13, 2023. [Online]. Available: https://arxiv.org/abs/1412.6572

[5] Y. Deng and L. J. Karam, “Universal Adversarial Attack Via Enhanced Projected Gradient Descent,” in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1241–1245. doi: 10.1109/ICIP40778.2020.9191288.

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.