World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:10, 2021

Adversarial Attacks and Defenses on Deep Neural Networks

Authors : Jonathan Sohn

Abstract : Deep neural networks (DNNs) have shown state-of-the-art performance for many applications, including computer vision, natural language processing, and speech recognition. Recently, adversarial attacks have been studied in the context of deep neural networks, which aim to alter the results of deep neural networks by modifying the inputs slightly. For example, an adversarial attack on a DNN used for object detection can cause the DNN to miss certain objects. As a result, the reliability of DNNs is undermined by their lack of robustness against adversarial attacks, raising concerns about their use in safety-critical applications such as autonomous driving. In this paper, we focus on studying the adversarial attacks and defenses on DNNs for image classification. There are two types of adversarial attacks studied which are fast gradient sign method (FGSM) attack and projected gradient descent (PGD) attack. A DNN forms decision boundaries that separate the input images into different categories. The adversarial attack slightly alters the image to move over the decision boundary, causing the DNN to misclassify the image. FGSM attack obtains the gradient with respect to the image and updates the image once based on the gradients to cross the decision boundary. PGD attack, instead of taking one big step, repeatedly modifies the input image with multiple small steps. There is also another type of attack called the target attack. This adversarial attack is designed to make the machine classify an image to a class chosen by the attacker. We can defend against adversarial attacks by incorporating adversarial examples in training. Specifically, instead of training the neural network with clean examples, we can explicitly let the neural network learn from the adversarial examples. In our experiments, the digit recognition accuracy on the MNIST dataset drops from 97.81% to 39.50% and 34.01% when the DNN is attacked by FGSM and PGD attacks, respectively. If we utilize FGSM training as a defense method, the classification accuracy greatly improves from 39.50% to 92.31% for FGSM attacks and from 34.01% to 75.63% for PGD attacks. To further improve the classification accuracy under adversarial attacks, we can also use a stronger PGD training method. PGD training improves the accuracy by 2.7% under FGSM attacks and 18.4% under PGD attacks over FGSM training. It is worth mentioning that both FGSM and PGD training do not affect the accuracy of clean images. In summary, we find that PGD attacks can greatly degrade the performance of DNNs, and PGD training is a very effective way to defend against such attacks. PGD attacks and defence are overall significantly more effective than FGSM methods.

Keywords : deep neural network, adversarial attack, adversarial defense, adversarial machine learning
Conference Title : ICADLPR 2021 : International Conference on Advances in Deep Learning for Pattern Recognition
Conference Location : New York, United States
Conference Dates : October 07-08, 2021