World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:12, No:03, 2018

Non-Targeted Adversarial Image Classification Attack-Region Modification Methods

Authors : Bandar Alahmadi, Lethia Jackson

Abstract : Machine Learning model is used today in many real-life applications. The safety and security of such model is important, so the results of the model are as accurate as possible. One challenge of machine learning model security is the adversarial examples attack. Adversarial examples are designed by the attacker to cause the machine learning model to misclassify the input. We propose a method to generate adversarial examples to attack image classifiers. We are modifying the successfully classified images, so a classifier misclassifies them after the modification. In our method, we do not update the whole image, but instead we detect the important region, modify it, place it back to the original image, and then run it through a classifier. The algorithm modifies the detected region using two methods. First, it will add abstract image matrix on back of the detected image matrix. Then, it will perform a rotation attack to rotate the detected region around its axes, and embed the trace of image in image background. Finally, the attacked region is placed in its original position, from where it was removed, and a smoothing filter is applied to smooth the background with foreground. We test our method in cascade classifier, and the algorithm is efficient, the classifier confident has dropped to almost zero. We also try it in CNN (Convolutional neural network) with higher setting and the algorithm was successfully worked.

Keywords : adversarial examples, attack, computer vision, image processing
Conference Title : ICIPACV 2018 : International Conference on Image Processing, Analysis and Computer Vision
Conference Location : London, United Kingdom
Conference Dates : March 15-16, 2018