AIP Conference Proceedings Volume 2608 12 June 2023 Article number 020028 2nd International Recent Trends in Engineering, Advanced Computing and Technology Conference, RETREAT 2021 Virtual, Online 1 December 2021 through 3 December 2021 Code 189566

A defense and detection against adversarial attack using De-noising auto-encoder and super resolution GAN

Abstract

Neural networks have flourished in heterogeneous industries to automate tasks that evince it being an utmost priority for the adopters. The adversarial attack poses a threat for Deep Neural Networks and their variants. This attack is designed such that it adds adversarial noise to an image. Several such techniques can be found in contemporary research capable of corrupting neural networks leading to misclassification. Various defense mechanisms have been purported and built with Deep Neural Networks to defend and increase the robustness of the primary classifier neural network model. However, models accommodating high-resolution image data and pre-trained neural network classifiers are sparse. This research develops a model that can be integrated with any existing trained neural network, establishing a generic line of defense against adversarial attacks. The proposed model detects highly distorted images, repudiating to avoid misclassification. Additionally, this work is intended to work with high-resolution adversarial image samples. The ADDA-Adv. tool restores the adversarial samples and provides better accuracy as compared to recent works.