World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:18, No:06, 2024

Resisting Adversarial Assaults: A Model-Agnostic Autoencoder Solution

Authors : Massimo Miccoli, Luca Marangoni, Alberto Aniello Scaringi, Alessandro Marceddu, Alessandro Amicone

Abstract : The susceptibility of deep neural networks (DNNs) to adversarial manipulations is a recognized challenge within the computer vision domain. Adversarial examples, crafted by adding subtle yet malicious alterations to benign images, exploit this vulnerability. Various defense strategies have been proposed to safeguard DNNs against such attacks, stemming from diverse research hypotheses. Building upon prior work, our approach involves the utilization of autoencoder models. Autoencoders, a type of neural network, are trained to learn representations of training data and reconstruct inputs from these representations, typically minimizing reconstruction errors like mean squared error (MSE). Our autoencoder was trained on a dataset of benign examples; learning features specific to them. Consequently, when presented with significantly perturbed adversarial examples, the autoencoder exhibited high reconstruction errors. The architecture of the autoencoder was tailored to the dimensions of the images under evaluation. We considered various image sizes, constructing models differently for 256x256 and 512x512 images. Moreover, the choice of the computer vision model is crucial, as most adversarial attacks are designed with specific AI structures in mind. To mitigate this, we proposed a method to replace image-specific dimensions with a structure independent of both dimensions and neural network models, thereby enhancing robustness. Our multi-modal autoencoder reconstructs the spectral representation of images across the red-green-blue (RGB) color channels. To validate our approach, we conducted experiments using diverse datasets and subjected them to adversarial attacks using models such as ResNet50 and ViT_L_16 from the torch vision library. The autoencoder extracted features used in a classification model, resulting in an MSE (RGB) of 0.014, a classification accuracy of 97.33%, and a precision of 99%.

Keywords : adversarial attacks, malicious images detector, binary classifier, multimodal transformer autoencoder
Conference Title : ICCV 2024 : International Conference on Computational Vision
Conference Location : Venice, Italy
Conference Dates : June 20-21, 2024

Open Science Index, Computer and Systems Engineering Vol:18, No:06, 2024 publications.waset.org/abstracts/174687.pdf