Abstract

In the past few years, it has been shown that most Deep Neural Networks (DNNs) exhibit inherent vulnerability to maliciously crafted inputs, called Adversarial Examples (AEs). These are usually created using inputs from the dataset by adding a small but very specific noise, which causes large errors on the DNN output.

Despite intense research in this area, the problem still pertains and has not yet been satisfactorily solved. Numerous defence mechanisms against AEs, as well as methods for detecting such modified inputs, have been suggested, yet, none of them provides complete robustness against adversarial attacks. Due to that, the focus of research has slowly shifted towards analysing the AEs and searching for the exact reason why and how they cause DNNs to fail.

In our work, we perform numerous experiments within the image classification domain with the aim of better understanding the properties of adversarial examples and various ways in which they cause faulty outputs. More specifically, our main contributions concern two parts: 1) we analyse the inherent robustness of networks with sigmoidal activations, and how it relates to saturating of the individual neurons. Moreover, we examine the robustness of the newly proposed Recurrent Vision Transformer (RecViT) architecture and explore the possibilities of different training strategies to make RecViT more robust; and 2) we assess the manifold disentanglement theorem using multiple methods and follow with the proposal of a novel algorithm for investigating the hidden-layer representations. This algorithm is based on analysing the proximity of certain inputs to the original data manifolds, and in our experiments, we use it to compare the behaviour of multiple distinct types of AEs.

Keywords: adversarial attacks, robustness, deep neural networks, image classification, explainability