HKUST MSBD 5014: Independent Project - Proposal

Muhammad Adam Usmani : 20729858
Supervisor : Nevin L. Zhang, Professor
Industry Partner : Huawei

1 Introduction

Adversarial examples have attracted significant attention in machine learning. It has been shown that deep neural network (DNN) models can be easily fooled by adding unnoticeable adversarial perturbations to the input. The most common reason for this is to cause a malfunction in a machine learning model. To overcome this issue, researchers have developed XAI methods to explain the output of a DNN by identifying the relevant evidence in the input. The aim of this paper is to study the effect of several adversarial attack methods on a variety of XAI algorithms and analyze the differences between adversarial examples and their corresponding benign examples. This will hopefully lead to some new ways to detect adversarial examples.

2 Dataset

In this paper we will be using natural images from the ImageNet dataset. The images in the dataset are all RGB and resized to 224x224 pixels.

3 Methodology

The adversarial attack methods will include both untargeted adversarial attacks and targeted adversarial attacks. Untargeted adversarial attacks are attacks that just want the model to be confused and to predict a wrong class. Targeted adversarial attacks are attacks which compel the model to predict a (wrong) desired output.

Untargeted Adversarial Attacks:

Fast Gradient Sign Method (FGSM)
This is a single step attack, ie.. the perturbation is added in a single step instead of adding it over a loop (Iterative attack)

Basic Iterative Method(BIM)
This method applies perturbations in several small steps rather than in a single step

Targeted Adversarial Attacks:

Projected Gradient Descent (PGD)
This attack is a white-box attack which means the attacker has access to the model gradients

The XAI methods include activation based methods and gradient based methods. Activation based methods involve deciphering the activations of the individual neurons or a group of neurons to get an intuition of what they are doing. Gradient based methods tend to manipulate the gradients that are formed from a forward and backward pass while training a model.

Activation Based Methods:

Image Occlusion
This is a model-agnostic explanation method: it involves systematically occluding different portions of the input image and monitoring the output of the classifier.

Gradient Based Methods:

Saliency Maps
This is a way to visualize the gradients wrt all the

pixels. A saliency map highlights the pixels that have the largest impact on class score if perturbed.

Gradient based Class Activations Maps (Grad CAM, Guided Grad CAM, Grad Cam ++)

Grad-CAM uses the gradients of any target concept (say logits for “dog” or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Guided Gradient Weighted Class Activation Map(Guided Grad CAM) is class discriminative as well as highlights fine-grained important regions of an image for prediction in high resolution for any CNN Architecture

Grad-CAM++, built on Grad-CAM, provides better visual explanations of CNN model predictions, in terms of better object localization as well as explaining occurrences of multiple object instances in a single image

There is a strong possibility that additional methods will be used throughout this project

4 Evaluation

The evaluation process will involve viewing the results of a variety of different adversarial attack methods and comparing the results of adversarial inputs against benign inputs. Comparison metrics will be used to compare the explained benign output with the explained adversarial example output. Two such methods include using algorithms such as Mean Squared Error (MSE) or the Structural Similarity Index (SSIM).

5 Expected Outcome

It is expected that explanations will deviate between adversarial inputs created using different adversarial attack methods and benign inputs.

1. Performance evaluation of Explainable AI methods against adversarial noise https://projekter.aau.dk/projekter/files/334478864/MasterThesis.pdf

2. A Guide to Understanding Convolutional Neural Networks (CNNs) using Visualization https://www.analyticsvidhya.com/blog/2019/05/understanding-visualizing-neural-networks/

3. Torchattacks : A PyTorch Repository for Adversarial Attacks https://www.arxiv-vanity.com/papers/2010.01950/

4. Welcome to TorchRay https://facebookresearch.github.io/TorchRay/index.html