On Detecting Adversarial Inputs with Entropy of Saliency Maps

Dian Ang Yap
Stanford University
dayap@stanford.edu

Joyce Xu
Stanford University
jexu@stanford.edu

Vinay Uday Prabhu
UnifyID AI Labs
vinay@unify.id

Abstract

Adversarial attacks pose serious security concerns in a wide range of real-life machine learning applications. An increasingly important component of building robust systems is detecting intentionally adversarial examples before classification. However, many current methods of detection are computationally inefficient or broadly ineffective. In this paper, we propose a gradient-based technique of detecting adversarial samples that relies on calculating the entropy of the Jacobian saliency map of the input. We demonstrate that quantitative and qualitative evaluation of adversarial saliency maps through Shannon entropy can be an efficient, effective way of detecting adversarial attacks, especially in deep neural networks with a linear nature.

1. Introduction

Adversarial attacks highlight the security vulnerabilities of machine learning models, especially in convolutional neural networks with locally linear nature and high-dimensional input space[3]. An image indistinguishable from the original to the human eye can be interpreted very differently and misclassified by deep neural networks, which poses security concerns in a variety of real-life applications from robotics to autonomous driving.

Existing literature has shown multiple ways of detecting adversarial examples, such as image transformation [8], network mutations [9], finding trajectory of internal representations and convolutional layer outputs across all layers [1] [5]. However, most defense techniques often require modifying the target model or depend on the prior knowledge of attacks.

Here, we propose a method to detect and visualize gradient-based adversarial attacks through entropy of saliency maps, which can be introduced in real time during inference or training time, and which does not require prior knowledge of the attack or modifications of target model beforehand.

2. Methods and Experiments

Given an image $\boldsymbol{x}$ with true label y, we experiment with two adversarial attacks. The first is a targeted adversarial attack where given $\boldsymbol{x}$, y and a target label $\tilde{y}$ where $y \neq \tilde{y}$, we perform gradient ascent over the image to maximize $\tilde{y}$ and stop when the network classifies the image as the $\tilde{y}$ instead of y [7].

The second attack is Fast Gradient Sign Attack (FGSM) [3] where the attack updates the input data to maximize the loss based on the backpropagated gradients, which does not require a target label $\tilde{y}$. Formally, given $\boldsymbol{x}$, FGSM creates a perturbed image $\tilde{\boldsymbol{x}}$ such that

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

By introducing an imperceptible non-random perturbation $\eta = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ to the image such that $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \eta$, the network prediction could be misclassified. Since deep models behave linearly, a large number of small perturbations in high dimensional input spaces can yield significant change in the model's output.

A saliency map [6] presents the heatmap of how significant each pixel contributes to the classification score by taking the maximum absolute over 3 input channels of the gradient. In a non-attacked image, the saliency map focuses on the core subjects in an image, which is specific and of high intensity; under perturbed attacks, the saliency map generally attends to wider, less focused regions.

We run experiments of SqueezeNet [4] on the validation set of ImageNet [2], with the examples unobserved during training, and measure the entropy of the saliency map using Shannon entropy where p_i is the probability of pixels of value i as

$$Q = -\sum_{i=0}^{n-1} p_i \log_2 p_i \tag{2}$$

3. Results and Discussion

Since the gradient-based attacks introduce a large number of small variations in a high-dimensional input space to the original image, the Shannon entropy of the perturbed image