Towards Information Theoretic Adversarial Examples

Chia-Yi Hsu*, Pin-Yu Chen† and Chia-Mu Yu*
*National Chung Hsing University, Taiwan
†IBM Research

***Abstract*— Deep learning has shown impressive performance on wide applications. However, recent research shows that deep neural networks are vulnerable to well-crafted input samples, called adversarial examples. Adversarial examples are indistinguishable to humans but can easily fool deep neural networks. Nowadays, most of attacks measure human's perception of the image quality with L_2-norm or L_∞-norm perturbation constraints. In this paper, we introduce mutual information (MI) to evaluate image quality of adversarial examples instead of L_p-norm measures. With MI as an information theoretic metric, our quantitative and qualitative results show that the resulting adversarial examples are more similar to unperturbed data samples.**

I. INTRODUCTION

In recent years, deep learning has made significant progress in many tasks of machine learning such as image classification, language translation, and object detection. However, recent studies demonstrated that well-trained deep neural networks (DNNs) are vulnerable to adversarial examples [1]. The purpose of the attacker is to find adversarial examples and minimize perturbations at the same time.

There have been many efforts to attack DNNs. Carlini and Wanger [2] proposed an optimization-based framework for targeted and untargeted attacks, abbreviated C&W attack. They design a L_2 norm regularized loss function apart from the model prediction loss defined by the logit layer representations in DNNs.

There have been many attacks using L_p-norm to restrict perturbations so that adversarial examples are imperceptible for humans. In this paper, our main contribution is that we are the first replacing statistical distance by information-theoretic metrics. MI is most used in generative adversarial networks (GANs) which measures the similarity between random variables and images.

II. OUR METHOD

Belghazi et al. [3] proposed a method using neural network to estimate mutual information called MINE. The concept is to select $\mathcal{F}$ to be the family of functions $T_\Theta : \mathcal{X} \times \mathcal{Z} \to \mathrm{R}$ parametrized by a deep neural network with parameters $\theta \in \Theta$. We formalize the bound:

$$I(X, Z) \geq I_\Theta(X, Z),$$

where $I_\Theta(X, Z)$ is the neural information quantity defined as

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]).$$

In our case, we need to compute MI of a single pair images. However, MINE must use batch version so that we make an image to batch version by gaussian random projection.

We use the C&W attack loss without L_2-norm regularized loss in addition to negative MI maximized by gradient ascent. We formalize our attack as the following optimization problem:

$$\underset{\delta, \Theta}{\text{minimize}} \quad c \cdot f(x + \delta) - \alpha \cdot I_\Theta(x, x + \delta)$$

$$\text{such that} \quad x + \delta \in [0, 1]^n \text{ and } \delta \in [-\epsilon, \epsilon]^n.$$

with f defined as

$$f(x') = \max\{\mathrm{Z}(x')_{l_x} - \max(\mathrm{Z}(x')_i : i \neq l_x), -\kappa\}.$$

The loss f is the best objection function found earlier, modified slightly so that we can control the confidence with which the misclassification occurs by adjusting κ.