README

October 10, 2020

1 EECS127_final_project_1

Author: Yunxuan Mao, Xuan Jiang, Shuai Liu

> This Project is mainly based on the technique developed by Wong and Kolter: J. Z. Kolter and E. Wong. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In ICML. 5283–5292.

1.1 1. Background

For this project, we will be considering a three-layer feedforward neural network with ReLU non-linearity; That is, the network $f_\theta : \mathbb{R}^{n_1} \to \mathbb{R}^{n_3}$ consists of the layers:

$$\vec{z_1} \in \mathbb{R}^{n_1}, \vec{\hat{z}_2} \in \mathbb{R}^{n_2}, \vec{z_2} \in \mathbb{R}^{n_2}, \vec{\hat{z}_3} \in \mathbb{R}^{n_3}$$

Where $\vec{z_1} = \vec{x}$ is the input for the network and $\vec{z_3}$ is the output of f_θ and the parameters are:

$$W_1 \in \mathbb{R}^{n_2 \times n_1}, W_2 \in \mathbb{R}^{n_3 \times n_2}, \vec{b_1} \in \mathbb{R}^{n_2}, \vec{b_2} \in \mathbb{R}^{n_3}$$

1.2 2. Adversarial Samples

In recent years, it has been found that classifier trained using standard methods are not very robust. That is, although a classifier may perform well when the inputs are sampled from real-world processes (e.g., images of real-world handwritten images), if they are artificially perturbed by even a small amount that is imperceptible to humans, they can easily be mislead. For example, if we have trained a classifier for detecting handwritten digits, an adversary might be able to take an image of a four and slightly change some pixel values such that our classifier now thinks the image is of a two.

More formally, the goal of an adversary is to find an example $\vec{x'}$ that is close to a real input $\vec{x}$, but is classified incorrectly. (The classifier is assumed to have correct output on$\vec{x}$.) That is, the adversary wishes to solve the following optimization problem:

$$\max_{\vec{x'}} L(f_\theta(\vec{x'}), y_{\vec{true}})$$
$$s.t. \ \|\vec{x} - \vec{x'}\|_\infty \leq \epsilon$$

1.3 3. Technological Process

We first showed that **Fast Gradient Signed Method** could be the solution to a **first-order approximation** of the adversarial optimization problem(With code in the notebook).

Then we formed a **convex relaxation** of the original problem above.

$$
\begin{aligned}
&\min_{z} \vec{c}^T \vec{\hat{z}}_3 \\
&s.t. \ \|\vec{z}_1 - \vec{x}\|_\infty \leq \epsilon \\
&\quad \vec{\hat{z}}_2 = W_1 \vec{z}_1 + \vec{b}_1 \\
&\quad \vec{z}_2 = ReLU(\vec{\hat{z}}_2) \\
&\quad \vec{\hat{z}}_3 = W_2 \vec{z}_2 + \vec{b}_2 \\
&\quad \vec{\hat{c}} = \vec{y_{true}} - \vec{y_{target}}
\end{aligned}
$$

After that, we derived the **dual optimization problem** in terms of the **Fenchel conjugates** and solve it by finding expressions for these conjugates. Finally we formed the expression of **Dual Network**, found the bounds of the dual problem and proved its robustness(with code in the notebook).

For more information please refer to **main.pdf** and **project.pdf**