Towards Evaluating the Robustness of Neural Networks

Ravi Raju

12-21-2018

1 Motivation

This paper was written to dismantle defense distillation which until had been the prominent defense aginst adverserial attacks. This is know as the CW attack with 3 varients based on different distance metrics: L_2, L_0, and L_∞.

2 Solution

The way to formulate the optimization problem goes as follows:

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & \mathcal{C}(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

where $\mathcal{D}$ is some distance metric(L_2, L_0, and L_∞), x is the image, and δ is the perturbation to the image to alter the classification. $\mathcal{C}$ is the classifier and t is the target classification label we desire to fool the classifier.

This is generally a difficult problem to solve so instead we try solve:

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ \text{such that} \quad & x+\delta \in [0,1]^n \end{aligned}$$

where f is some objective function and c > 0 is a constant. The best way to choose c is evalute the function until it is less than zero. They use binary search to find this constant in their experiments. We use these three different optimization substitutions so we can use other optimizers besides SGD: Projected Gradient Descent (PGD), Clipped gradient descent, and change of variables. PGD seemed to yield the best results.

Now we finally reveal the actual attacks being formed beginning with the L_2 attack:

$$\text{minimize} \quad \|\frac{1}{2}(\tanh(w)+1) - x\|_2^2 + c \cdot f(\frac{1}{2}(\tanh(w))+1).$$

We choose a target class, t and the search for a w that minimzes the above equation.

The L_0 version of this problem is non-differentiable so instead we solve the L_2 version to eliminate pixels that do not contribute to adverserial examples. This is different from JSMA in that it grows a set of pixels. For L_∞, the solution is also not fully differentiable.

$$\text{minimize} \quad c \cdot f(x+\delta) + \sum_i [(\delta_i - \tau)^+].$$

More details are in the paper if you need context.

3 Results

This is most apparent if you take a look at Table IV. The mean distance of the perturbation is much smaller. JSMA cannot calculate for L_0. For results against defense distillation look at Table VI.

4 Discussion/Takeaway

Two important details are that you need to: 1) use a powerful attack and 2) demonstrate that transferability fails. By transferability, we mean that adverserial examples from an easy-to-attack model should not perturb a model with defense that claims to provide robustness. In the paper, three different methods were evaluated to find adverserial examples: L_2, L_0, and L_∞.