《Certified Adversarial Robustness via Randomized Smoothing》-ICML2019 文章阅读

[ICML2019]https://arxiv.org/pdf/1902.02918.pdf
Designing a smoothed classifier against adversarial samples using Gaussian noise

1. 背景介绍

本文提出了"randomized smoothing"，并对其进行了理论分析。即对分类器高斯噪声处理，使得新分类器对对抗攻击足够鲁棒。

Certified Defense:

训练一个对输入样本x的ℓ_2或ℓ_∞邻域内所有样本都具有鲁棒性的分类器。具体的方法分为exact method 和 conservative method。

目的在于寻找一个扰动δ，满足$g(x) \neq g(x+\delta)$，如果存在这样的扰动，则 decline to make a certification；如果找不到这样的扰动，则假设成立。然而，没有一种 exact method适用于中型复杂度的神经网络。Conservative certification 可扩展到任意大小的神经网络，但是其得到的鲁棒性 guarantee比较 loose。

given a model $h \circ f$, and a new test point (x, y), we would like to prove $h \circ f\left(x'\right) = h \circ f(x)$, for all x' in the allowed oerturbation set. That is, to provide certification about the optimality of the following equation

$$\max_{x' \in \Delta(x_{test})} Loss\left(h\left(f\left(x'\right)\right), y_{test}\right)$$

Randomized Smoothing:

给定一个neural network f(base classifier)，并且$f(x) = y$。将f转变成smoothed classifier g.

When queried at x, the smoothed classifier g returns whichever class the base classifier f is most likely to return when x is perturbed by isotropic Gaussian noise:

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\varepsilon) = c)$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Certification guarantee:

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$$

where p_A is the prob of the top class, p_B is the prob of the runner-up class. The smoothed classifier g will return the top class, where Φ^{-1} denotes the inverse Gaussion CDF

Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (2)$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (3)$$

Pseudocode for certification and prediction

```
# evaluate g at x
function PREDICT(f, σ, x, n, α)
    counts ← SAMPLEUNDERNOISE(f, x, n, σ)
    ĉ_A, ĉ_B ← top two indices in counts
    n_A, n_B ← counts[ĉ_A], counts[ĉ_B]
    if BINOMPVALUE(n_A, n_A + n_B, 0.5) ≤ α return ĉ_A
    else return ABSTAIN

# certify the robustness of g around x
function CERTIFY(f, σ, x, n_0, n, α)
    counts0 ← SAMPLEUNDERNOISE(f, x, n_0, σ)
    ĉ_A ← top index in counts0
    counts ← SAMPLEUNDERNOISE(f, x, n, σ)
    p_A ← LOWERCONFBOUND(counts[ĉ_A], n, 1 − α)
    if p_A > 1/2 return prediction ĉ_A and radius σ Φ^{-1}(p_A)
    else return ABSTAIN
```

Prediction: 创建n个带有高斯噪声的x样本，用base classifier $f(x+\eta)$对其分类，得到top two classes $\hat{c}_A$和$\hat{c}_B$, If $n_A \gg n_B$, then predict the top class; otherwise, abstain.

Certification: Use small #samples to identify c_A; Use large #samples to estimate p_A; Set $p_B = 1 - p_A$ and compute.

证明思路：https://zhuanlan.zhihu.com/p/463037691