AIRI × Передовая Инженерная Школа ИТМО

№25: Состязательная атака на защищенную модель

Свирщевский Юрий, Шуклин Максим,
Гришина Елена

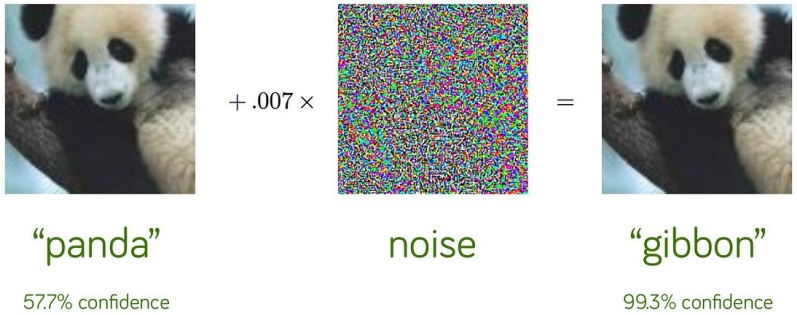
Состязательные атаки

Состязательная атака на модель — способ подбора входных данных, на которых модель выдает неправильный результат.

Варианты постановки:

White-Box — есть доступ к весам и градиентам модели

Black-Box — модель доступна только как оракул



Случайное сглаживание

Случайное сглаживание – замена исходной модели суррогатом:

$$f(x) \Rightarrow g(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)} f(x+\epsilon)$$

В общем случае матожидание не вычислимо, поэтому аппроксимируется:

$$g(x) \approx \hat{g}(x) = \frac{1}{n} \sum_{j=1}^{n} f(x+\epsilon_i),\ \epsilon_i \sim \mathcal{N}(0, \Sigma^2)$$

Случайное сглаживание

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}[f(x+\epsilon) = c], \;\; \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Можно дать теоретическую оценку на радиус устойчивости $g(x)$ внутри которого классификатор не меняет своего предсказания.

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

p_A и p_B обозначают вероятности двух самых популярных классов.

Цель проекта

Проблема: в теории классификатор g устойчив к состязательным атакам, но на практике реализуемо только его приближение, поэтому гарантии неизвестны.

$$\hat{g}(x) = \frac{1}{N}\sum_{i=1}^{N} f(x+\epsilon),\ \ \epsilon_1, \ldots, \epsilon_N \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Цели:

- Исследовать реальную устойчивость приближения к атакам
- Получить теоретические гарантии устойчивости для приближения
- Предложить атаку на случайное сглаживание

Предлагаемая атака на случайное сглаживание

Положим $g_i(x) = f(x + \varepsilon_i)$, где значения ε_i фиксированы.

Оценим градиента случайной функции:

$$\hat{\nabla}_x \hat{g}(x) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_x g_i(x)$$

Двигаемся в направлении этой оценки градиента для фиксированной выборки ε для нахождения состязательных примеров.

Предлагаемая атака на случайное сглаживание

$$\hat{\nabla}_x \hat{g}(x) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_x g_i(x)$$

По оценке градиента можно реализовывать атаку по аналогии с FGSM:

$$x' = x + \varepsilon \cdot \hat{\nabla}_x \hat{g}(x)$$

При рассматривании окрестности исходного изображения:

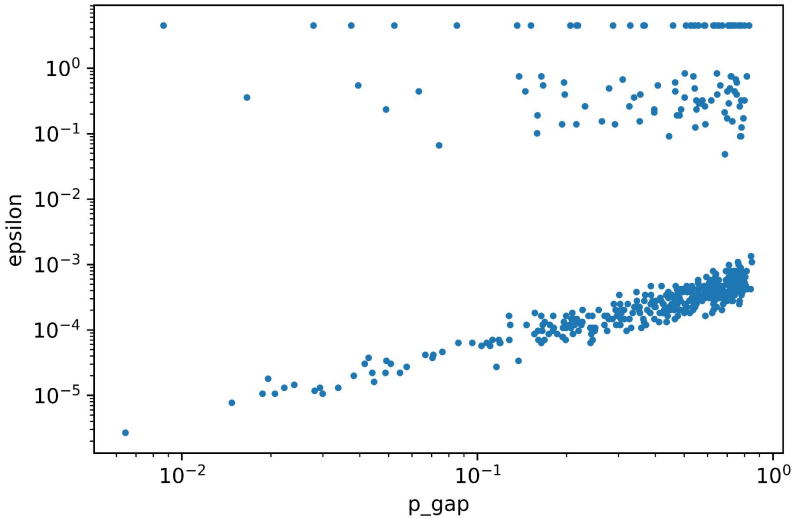
$$x_{k+1} = \mathrm{proj}(x_k + \alpha_k \cdot \hat{\nabla}_{x_k} \hat{g}(x_k))$$

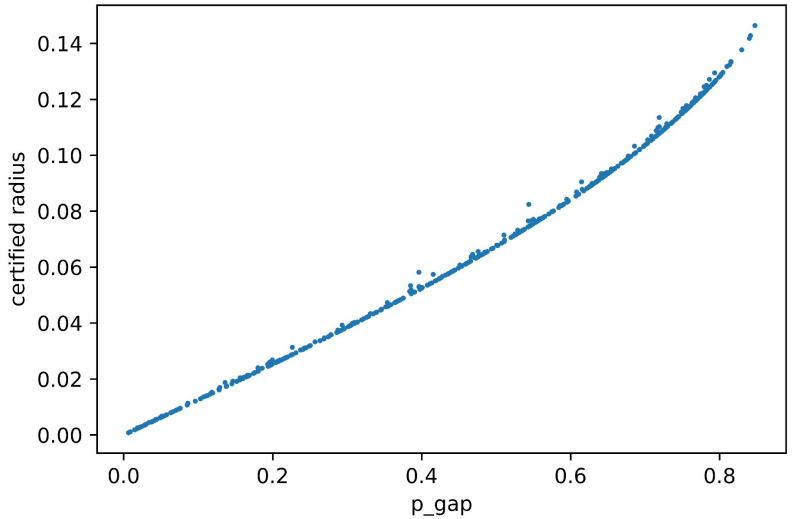
$\mathrm{proj}(x')$ — проекция x′ на нужную окрестность x

Зависимость ε и $R_{certified}$ от p_A-p_B

Атаковали модель ResNet32 на датасете CIFAR-10

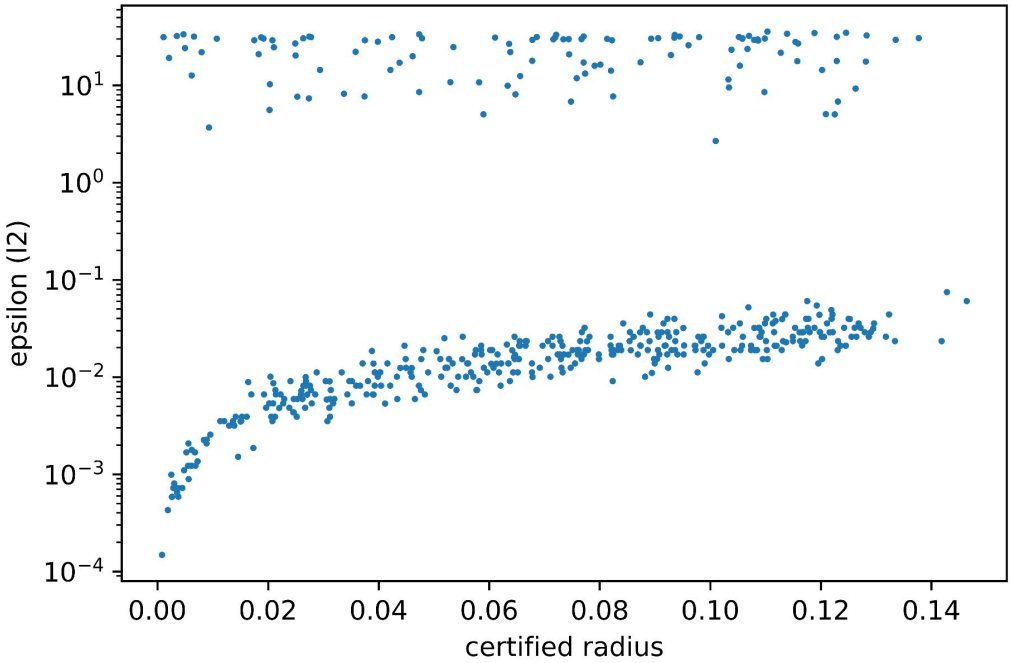
p_gap = p_A-p_B





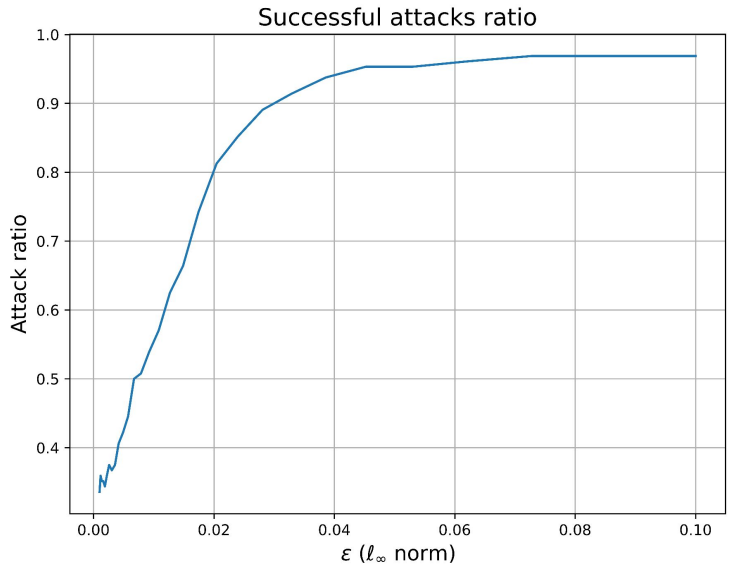
AIRI × Передовая Инженерная Школа ИТМО

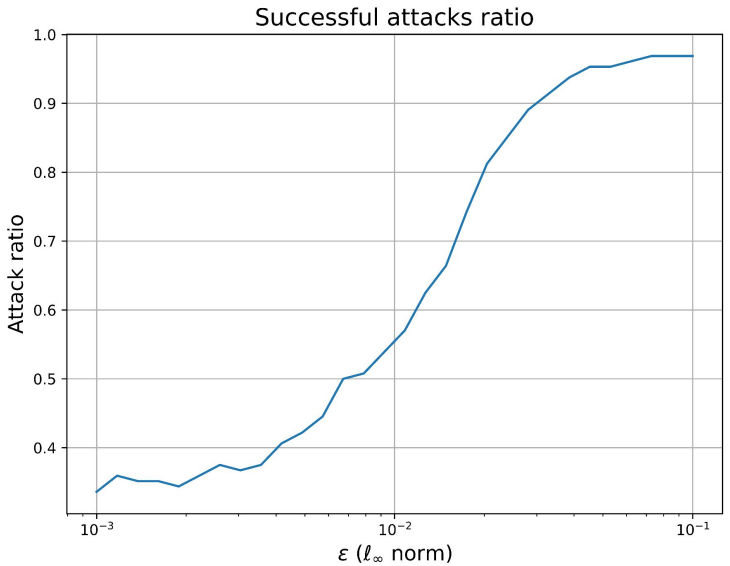
Зависимость ε в l2 норме и $R_{certified}$



Шаг вдоль случайного градиента

$$x' = x + \varepsilon \cdot \hat{\nabla}_x \hat{g}(x)$$





Теоретическая оценка

Мы показали, что для $\forall x,x',t$ верно неравенство:

$$\mathbb{P}(\hat{g}(x) - \hat{g}(x') \geq t) \leq 2e^{-\frac{N(t-g(x)+g(x'))^2}{2}}$$

Вероятность успешной атаки для $\forall x'$ в радиусе устойчивости точки x убывает экспоненциально.

Оценка доказывает: невозможно провести на случайное сглаживание атаку, вероятность успешности которой убывала бы “не слишком быстро”

Результаты

+ Показана теоретическая устойчивость оценки сглаженного классификатора к атакам
+ Предложили метод атаки случайного сглаживания, незаметный для человеческого глаза
– Не удалось провести атаку, которая бы попала в радиус устойчивости

Направления дальнейших исследований

- Проведение экспериментов для разных архитектур нейросетей
- Получение более точных оценок устойчивости $\hat{g}$ к атакам
- Разработка новых видов защищенных классификаторов

Вопросы

Спасибо за внимание!