1 定义

1.1 符号说明

- $\mathcal{W} \subseteq \mathbb{R}^d$: 神经网络参数空间
- $W \in \mathcal{W}$: 模型权重向量
- $\mathcal{X}$: 输入空间
- $\mathcal{Y}$: 输出空间
- $M_W : \mathcal{X} \to \mathcal{Y}$: 由 W 参数化的模型
- $\mathcal{A}$: 攻击算法
- Q: 攻击者可用的查询次数上限
- ϵ, δ: 隐私参数
- $\sim$: 相邻关系符号

1.2 定义

定义 1 (参数相邻)**.** 设 $W, W' \in \mathcal{W}$ 为两个神经网络权重向量, 若两者的 L_2-范数距离满足 $\|W - W'\|_2 \leq 1$, 则称 W 与 W' 是相邻的, 即 $W \sim W'$。

定义 2 (防御机制)**.** 对于查询 x, 防御机制 $\mathcal{M}$ 返回:

$$\mathcal{M}(x; W) = f_W(x) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_m)$$

其中 $f_W(x) \in \mathbb{R}^m$ 为模型在权重 W 下的 Logits 输出, η 为各维度独立的高斯噪声。

定义 3 (模型提取攻击)**.** 模型提取攻击是一个概率算法 $\mathcal{A}$, 给定:

(1) 通过查询接口 $\mathcal{O}$ 对目标模型 M_W 的预言机访问

(2) 自适应查询预算 $Q \in \mathbb{N}$

输出一个替代模型 $\tilde{M}$, 使得对于某个相似度度量 sim 和阈值 τ:

$$\Pr\left[\mathrm{sim}(M_W, \tilde{M}) \geq \tau\right] \geq \beta$$

其中 $\beta \in (0, 1]$ 是成功概率。

定义 4 (模型提取攻击抵抗性)**.** 一个具有查询接口 $\mathcal{O}$ 的模型 M_W 是 (α, β)-抵抗提取攻击的, 如果对于任何概率多项式时间攻击算法 $\mathcal{A}$, 最多对 $\mathcal{O}$ 进行 Q 次查询, 都有:

$$\Pr\left[\|W_{\mathrm{ext}} - W\|_2 \leq \alpha\right] \leq \beta$$

其中 W_{ext} 是提取模型的参数, 概率取遍 $\mathcal{O}$ 和 $\mathcal{A}$ 的随机性。

定义 5 (具有差分隐私的查询接口)**.** 一个查询接口 $\mathcal{O}_{\epsilon,\delta}$ 满足 (ϵ,δ)-差分隐私，如果对于所有相邻权重向量 $W \sim W'$，所有查询序列 $\mathbf{x} = (x_1, \ldots, x_n)$，以及所有可测集 $S \subseteq \mathcal{Y}^n$：

$$\Pr[\mathcal{O}_{\epsilon,\delta}(M_W, \mathbf{x}) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{O}_{\epsilon,\delta}(M_{W'}, \mathbf{x}) \in S] + \delta$$

其中概率取遍 $\mathcal{O}_{\epsilon,\delta}$ 的随机性。

2 Proof

2.1 定义

攻击者限制访问次数为 Q，在第 i 次访问中，攻击者向 $\mathcal{O}$ 访问 $\{x_i\}$，然后 $\mathcal{O}$ 向攻击者返回 $\{y_i = M_W(x_i) + \eta\}$，其中 $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$；然后攻击者根据 $\{x_i, y_i\}$ 输出 W_{ext}，需要证明模型可以抵抗提取，即证明 $\Pr\left[\|W_{\text{ext}} - W\|_2 < \alpha\right] < \beta$。

2.2 证明过程

定理 1 (基于高斯输出扰动的模型提取抵抗性)**.** 设 $M_W : \mathcal{X} \to \mathbb{R}^m$ 为一个参数化模型，其参数为 $W \in \mathbb{R}^d$。假设对任意固定输入 $x \in \mathcal{X}$，模型输出关于参数 W 是 *L-Lipschitz* 的，即

$$\|M_W(x) - M_{W'}(x)\|_2 \leq L\|W - W'\|_2, \quad \forall W, W'.$$

攻击者至多进行 Q 次（可自适应）查询，每次通过如下带噪声的预言机获得输出：

$$\mathcal{O}(x) = M_W(x) + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2 I_m),$$

并最终输出参数估计 W_{ext}。

则对任意 $\alpha > 0$，存在常数 $c > 0$，使得

$$\Pr\big(\|W_{\text{ext}} - W\|_2 < \alpha\big) \;\leq\; \frac{2QL^2R^2/\sigma^2 + \log 2}{c\, d\, log\frac{R}{\alpha}}.$$

证明. 证明思路是将连续参数估计问题归约为有限多假设判别问题，并应用 Fano 不等式。

1. 参数空间的打包构造 在 $B(0, R)$ 中构造一个 2α-打包集合

$$\mathcal{W} = \{W^{(1)}, \ldots, W^{(N)}\},$$

满足任意 $i \neq j$ 都有 $\|W^{(i)} - W^{(j)}\|_2 \geq 2\alpha$，且其规模满足 $\log N \geq c\, d\, log\frac{R}{\alpha}$，其中 $c > 0$ 为常数。假设真实参数 W 从集合 $\mathcal{W}$ 中均匀随机选取，并用随机变量 $V \in \{1, \ldots, N\}$ 表示其索引。

2. 从参数估计到分类问题的归约 攻击者给出的估计为 W_{ext}。定义判决规则

$$\hat{V} = \arg\min_j \|W_{\text{ext}} - W^{(j)}\|_2.$$

由打包集合的分离性可知，若 $\|W_{\text{ext}} - W\|_2 \leq \alpha$，则必然有 $\hat{V} = V$。因此，

$$\Pr(\|W_{\text{ext}} - W\|_2 \leq \alpha) \;\leq\; \Pr(\hat{V} = V).$$

3. 观测分布及 KL 散度 记攻击者观测到的全部输出为 $Y = (y_1, \ldots, y_Q)$。在条件 $V = i$ 下，有

$$Y \sim \mathcal{N}(\mu_i, \sigma^2 I_{Qm}), \quad \mu_i = (M_{W^{(i)}}(x_1), \ldots, M_{W^{(i)}}(x_Q)).$$

对任意 $i \neq j$，对应分布的 KL 散度为

$$\mathrm{KL}(P_{W^{(i)}} \| P_{W^{(j)}}) = \frac{1}{2\sigma^2} \sum_{q=1}^{Q} \| M_{W^{(i)}}(x_q) - M_{W^{(j)}}(x_q) \|_2^2.$$

由 Lipschitz 条件及 $\|W^{(i)} - W^{(j)}\|_2 \leq 2R$，可得

$$\mathrm{KL}(P_{W^{(i)}} \| P_{W^{(j)}}) \leq \frac{2QL^2R^2}{\sigma^2}.$$

4. 互信息上界 设 $P_Y = \frac{1}{N} \sum_{j=1}^{N} P_{W^{(j)}}$ 为 Y 的边缘分布。由于 V 服从均匀分布，有

$$I(V; Y) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}(P_{W^{(i)}} \| P_Y).$$

利用 KL 散度对第二个参数的凸性，可得

$$I(V; Y) \leq \frac{1}{N^2} \sum_{i,j} \mathrm{KL}(P_{W^{(i)}} \| P_{W^{(j)}}) \leq \frac{2QL^2R^2}{\sigma^2}.$$

5. 应用 Fano 不等式 由 Fano 不等式，

$$\Pr(\hat{V} \neq V) \geq 1 - \frac{I(V; Y) + \log 2}{\log N},$$

等价地，

$$\Pr(\hat{V} = V) \leq \frac{I(V; Y) + \log 2}{\log N}.$$

结合上述结论即可得到定理结论。 □