Differentially Private and Certified Robust Deep Learning

Anonymous

***Abstract*—**
***Index Terms*—**

I. Introduction

Contributions.

Organization.

II. Related Work

III. Methodology

There is a connection between the input perturbation and the gradient perturbation. A transformation from the input perturbation to the gradient perturbation will be provided in this section.

A. Perturbation Transformation

Given a clean example x_i and an element-wise perturbation b with $b_j \sim N(0, \sigma^2)$, a perturbed example is denoted as $z_i = x_i + b$. A model trains on the perturbed data can preserve its parameters from inference attack. Its loss function can be formulated as follows,

$$L_{priv}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(z_i, \theta). \tag{1}$$

This loss function is leveraged to calculate the gradient as follows,

$$\nabla_\theta L_{priv} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta l(z_i, \theta_t) \tag{2}$$

We can leverage Taylor expansion to approximate $\nabla_\theta l(z_i, \theta_t)$ at the data point x_i as follows,

$$\begin{aligned}\nabla_\theta l(z_i, \theta_t) &= \nabla_\theta l(x_i, \theta_t) + \nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t)(z_i - x_i) \\ &= \nabla_\theta l(x_i, \theta_t) + \nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t) b\end{aligned} \tag{3}$$

Assume the gradient perturbation p is defined as $p = \nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t) b$. As can be seen from equation 3, the input perturbation b is transformed to the gradient perturbation p through $\nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t)$.

Since $b_j \sim N(0, \sigma^2)$, $\nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t)$ determines the scale of p. To analyze the statistics of $\nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t)$, we vectorize b and ~~θ_t~~, and let $A = \nabla_{x_i}\nabla_{\theta_t} l(x_i, \theta_t)$. Therefore, the problem becomes that given $p = Ab$, where $b \in \mathbb{R}^v$, $b_j \sim N(0, \sigma^2), \theta_t \in \mathbb{R}^w, A \in \mathbb{R}^{w \times v}$, what is the scale of p?

x_i, gradient

We first let A_k denotes k-th row of A, then $A_i \in \mathbb{R}^{1\times v}$, $A_k b$ is a summation of b_j, and $A_k b \sim N(0, A_k A_k^\intercal \sigma^2)$. Therefore, $p = Ab$ transforms b from $\mathbb{R}^{v\times 1}$ tp $\mathbb{R}^{w\times 1}$, and each row element $p_k \sim N(0, A_k A_k^\intercal \sigma^2)$. Choosing the minimum value M from $A_k A_k^\intercal$, we can guarantee that 1) for $k \in [1, w]$, $A_k A_k^\intercal \geq M$, and 2) p can be approximated by $\hat{p} \sim N(0, M\sigma^2)$. Then the traditional DP analysis technique, e.g., MA, can be leveraged to analyze the privacy cost by taking the gradient perturbation $\hat{p}$.

Hetergeneous

DP-SGD brings some noises, if

Looking at the definition of the transformation matrix A, it requires both the clean example x_i and the perturbed example z_i. However, the calculation of A will not incur the privacy issue. Because the clean example x_i does not actually contribute to the gradient $\nabla_\theta l(z_i, \theta_t)$, the transformation process and the calculation of A is actually a theoretical analysis.

IV. Experiments

V. Conclusions and Future Work

References: