CS : Deep Learning Fall 2020

Lecture 1 — September 2, 2020

Prof. Zhao Song *Scribe: Zhao Song*

1 Private Deep Learning

1.1 Two hidden layer neural network

Let $\phi : \mathbb{R} \to \mathbb{R}$ denote the ReLU activation, i.e., $\phi(u) = \max\{u, 0\}$.

We consider a two hidden layer neural network $f : \mathbb{R}^d \to \mathbb{R}$ (which can be decomposed into two functions $h : \mathbb{R}^d \to \mathbb{R}^{m_a}$ and $g : \mathbb{R}^{m_a} \to \mathbb{R}^{m_b}$) as follows:

$$\begin{aligned} h(x) &= \phi(W_A x), \\ g(z) &= W_C^\top \phi(z), \\ f(x) &= g(h(x)) = W_C^\top \phi(W_B \phi(W_A x)), \end{aligned}$$

where $W_A \in \mathbb{R}^{m_a \times d}$, $W_B \in \mathbb{R}^{m_b \times m_a}$ and $W_C \in \mathbb{R}^{m_b}$.

Let $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ denote a set of n input data points. We can think of x_is are images and y_is are the corresponding labels.

In classical deep learning training, the task is to find weights W_A, W_B and W_C such that $\mathcal{L}$ is minimized.

$$\mathcal{L} = \sum_{i=1}^n \| f(x_i) - y_i \|_2^2$$

In this lecture, we describe a slightly different goal. The purpose is to find some W_A, W_B and W_C such that satisfy the following two properties:

1. Utility: $f(x_i) \approx y_i$, $\forall i \in [n]$
2. Privacy: Given $h(x_i)$ and W_A, it is "hard" to recover x_i

There are several ways of modifying deep neural networks to make it more private

1. Modify input data points [HSLA20]
2. Modify weights [HSR+20]

References

[HSLA20] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instace-hiding schemes for private distributed learning. In *ICML*, 2020.

[HSR^{+}20] Yangsibo Huang, Yushan Su, Sachin Ravi, Zhao Song, Sanjeev Arora, and Kai Li. Privacy-preserving learning via deep net pruning. In *arXiv preprint*. https://arxiv.org/pdf/2003.01876.pdf, 2020.