Homework

1023040801 董奥

2024 年 5 月 28 日

摘要

it is an abstract.

1 Introduction

In recent years, neural networks have achieved great success in many fields, including image processing and natural language processing [1,2]. The progress of neural networks cannot be separated from large datasets. However, some datasets are crowdsourced and contain a lot of sensitive information. While using them to train models meets demand, it is also necessary to provide strict privacy guarantees.

Deep learning models are inherently opaque, with a vast number of parameters making it nearly impossible to describe the details of the data features they discover. However, this does not mean that the datasets are immune to leakage. Competent attackers might be able to extract parts of the training data. For example, Fredrikson et al. demonstrated a model inversion attack that could recover images from a facial recognition system [3].

Although model inversion attacks only require "black-box" access to the model, we assume a very powerful attacker who fully understands the model and has access to its parameters. Differential privacy methods can provide effective protection against such powerful adversaries.

In this paper, I will introduce methods for training models using differential privacy, discuss the current issues, and demonstrate these methods with experiments.

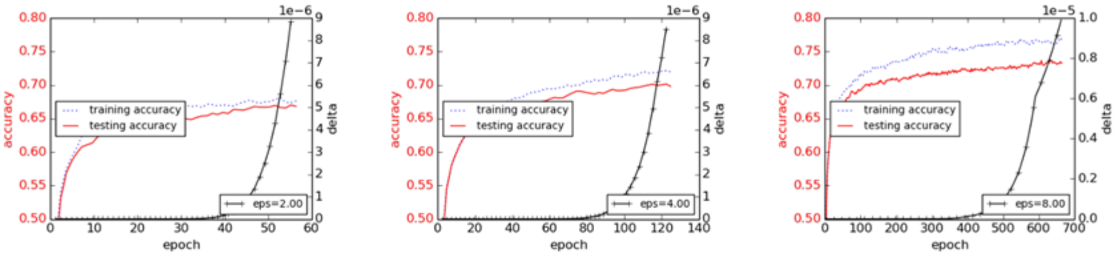
2 Some examples

2.1 Include a Figure

It can be observed that the stronger the privacy guarantees, the lower the model's accuracy. In the worst-case scenario, the accuracy might even drop to as low as 65%. However, these results are from earlier experiments, and the impact of differential privacy training on model performance has become increasingly smaller in recent times.

2.2 Add a Table

I tested the time consumed for each step of a batch, and the data is as follows:



(1) $\varepsilon = 2$ (2) $\varepsilon = 4$ (3) $\varepsilon = 8$

图 1: a image

Item	Time
Forward	0.03
Backward	5.24

表 1: An example table.

2.3 Some Lists

In fact, there are still many issues that need to be addressed before differential privacy algorithms can be widely applied.

1. Time issue
2. model performance issue

2.4 A Mathematic

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.