The Differences Between Private and Non-Private Training

Musheng Li

ODDS Research Group

Viterbi-Tsinghua Student Research Program

Advised by:

Xinwei Zhang

Prof. Meisam Razaviyayn

Content

I. Background

II. Challenges

III. Experiments

IV. Summary

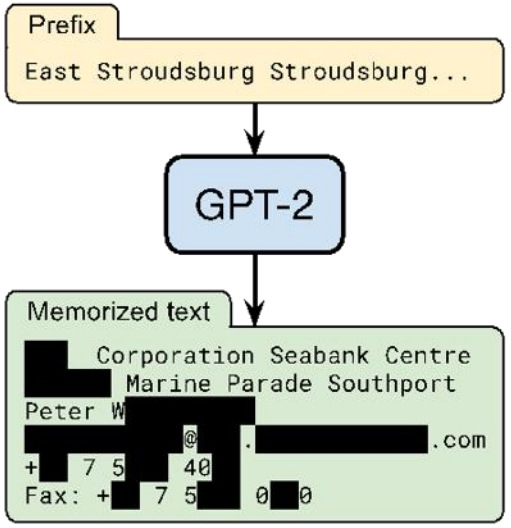
Background

Concerns

- The model weight may contain a lot of **private** information.
  - ➢ Even worse in mordern high-dimension models.
- One can **reconstruct** some sensitive training data with some tricks.



[Fredrikson et al 2015]



[Carlini et al 2020]

We need to protect the privacy!!

Background

What Is DP?

Differential privacy(DP) is a mathematical framework used to **quantify the level of data privacy protection**. Its core idea is to **add noise** to the query results.

Definition

Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

Cynthia Dwork, et al. "The Algorithmic Foundations of Differential Privacy"

ε: The privacy level(lower→stronger)
δ: Probability of violating privacy

DP-SGD Algorithm

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Martín Abadi, et al. "Deep Learning with Differential Privacy", CCS 2016.

Clip Gradient

- Bounded gradient size.
- Quantitative privacy(theorem).

Add Noise

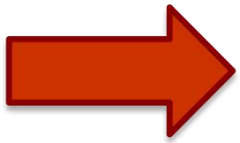
- Noise level $\boldsymbol{\sigma}$: ε, δ, L, C.
- Gaussian mechanism.

Challenges and Goals



Challenges of DP Training

- Worse performance
- More GPU usage
- Slower computing speed
- Harder training
- Balance of accuracy and privacy



Underlying Reasons?

Goals

- Understand DP training.
- Find differences between DP and non-DP.

Experiment Design

Task: Image Classification
Model: ViT-Base-16(Pretrained on ImageNet)
Dataset: CIFAR100

1. Understand DP Training

- Tune hyper-parameters
  - Learning rate, batch size, epochs, epsilon
- Compare GPU occupation
  - Temperature
  - Memory Usage

2. Find differences between DP and non-DP

- Compute gradient trajectory
  - (r, θ) - norm and angle
  - Norm distribution
- Plot loss landscape
  - 2D contour
  - 3D landscape

Experiment 1

Find Out The Best Hyperparameters For DP Training

- Tuning: η - L - epoch - ε, sequentially.
- *log* search space.

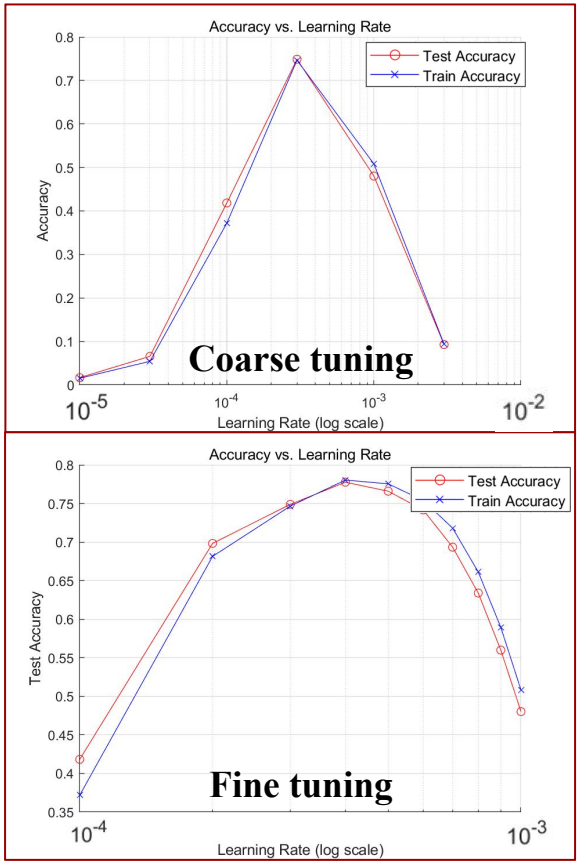
η	1e-5	3e-5	1e-4	3e-4	1e-3	3e-3
Acc(%)	1.7	6.6	41.8	**74.9**	48.0	9.3

Optimal Param	DP	non-DP
η	4e-4	1e-4
L	1000	1000
epoch	10	5

	SOTA	non-DP	DP(ε=0.5/1/2, δ=1e-5)
Accuracy(%)	94.2	91.1	47.8/70.3/79.6

https://paperswithcode.com/sota/image-classification-on-cifar-100

Accuracy - η

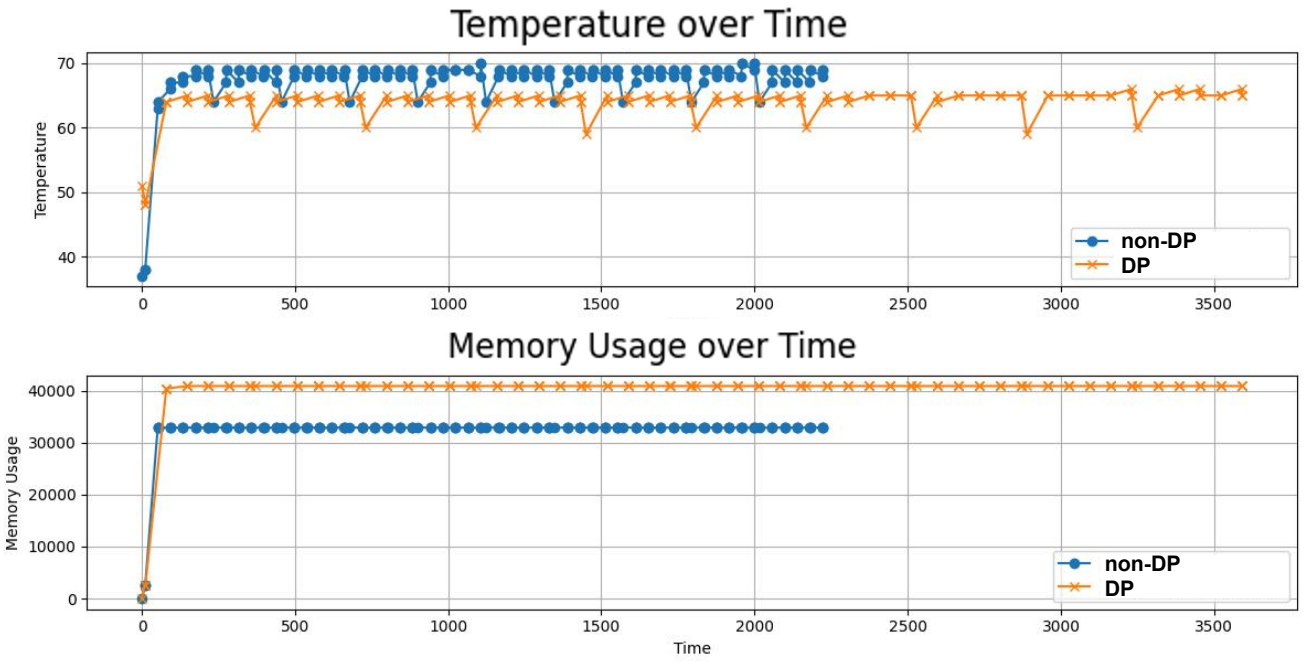


DP has lower accuracy, which is positively correlated with ε.

Experiment 2

Compare GPU Occupation of DP and Non-DP Training

- Record the GPU usage(memory and temperature) in different stages:
  - loss.backward()
  - optimizer.step()
  - optimizer.zero_grad()

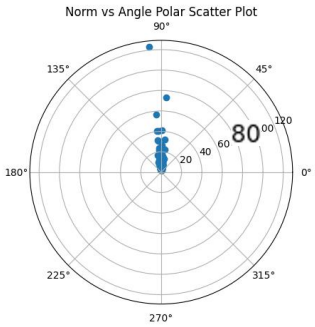
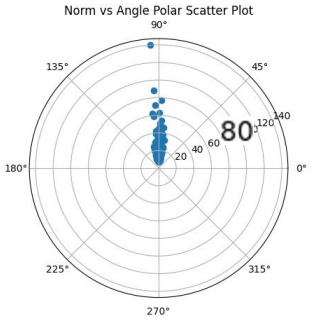
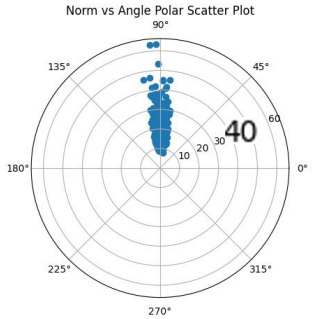
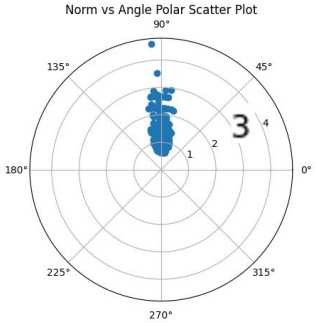
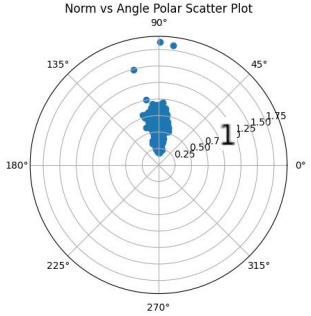
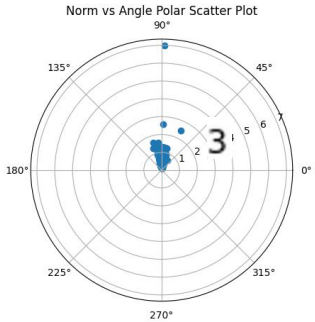


DP requires longer time and higher memory usage(24%). Non-DP has higher temperature.

Experiment 3

Compare Gradient Trajectory of DP and Non-DP Training

- Compute the norm and angle between true gradient and per-sample gradient.
  - **True gradient**: gradient of a batch
  - **Per-sample gradient**: gradient of a sample

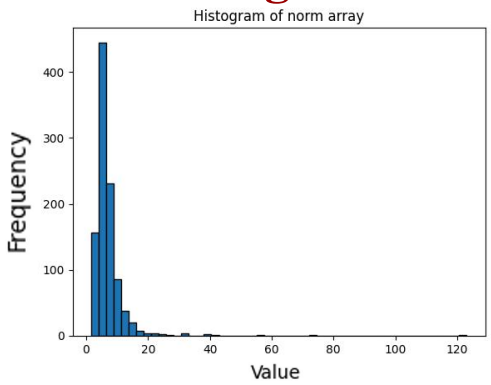
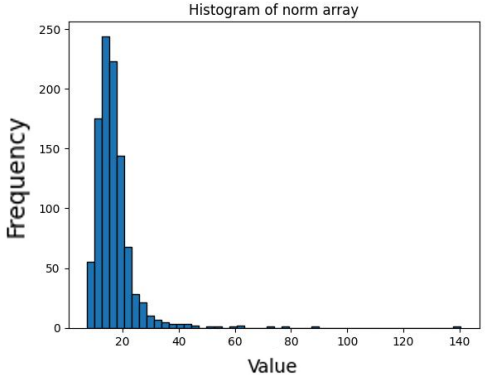
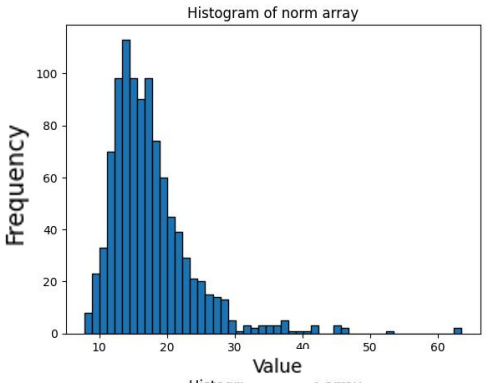
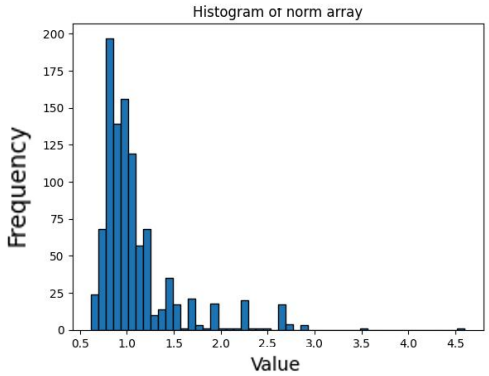
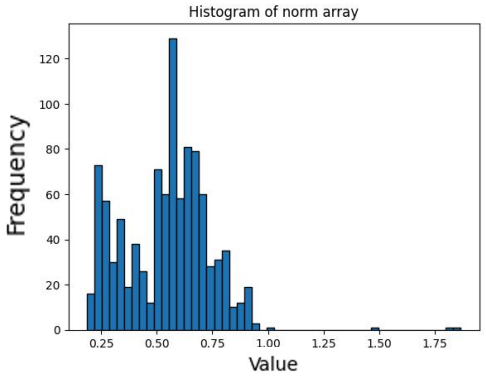
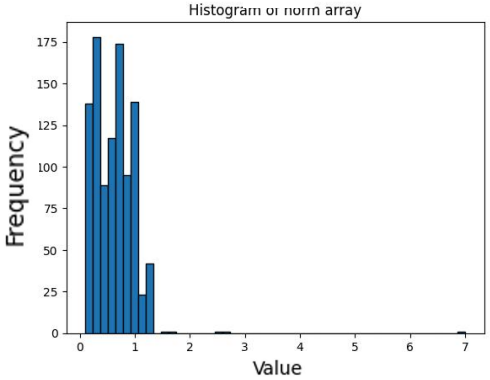
	Begin	Mid	End
DP			
non-DP			

DP has much larger norm and smaller angle variance.

Experiment 3

Compare Gradient Trajectory of DP and Non-DP Training

- Plot the norm distribution.

	Begin	Mid	End
DP			
non-DP			

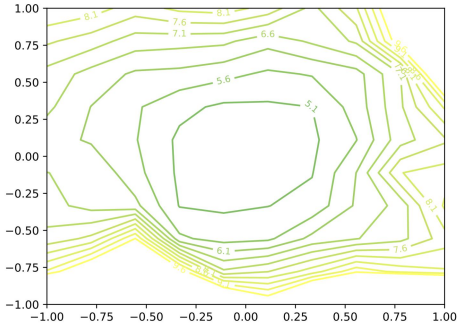
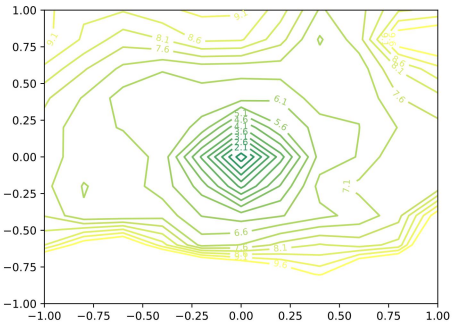
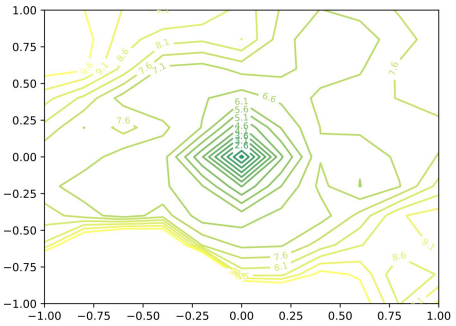
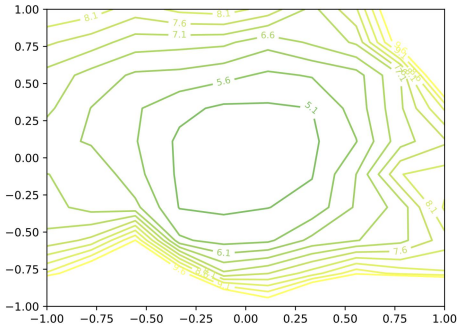
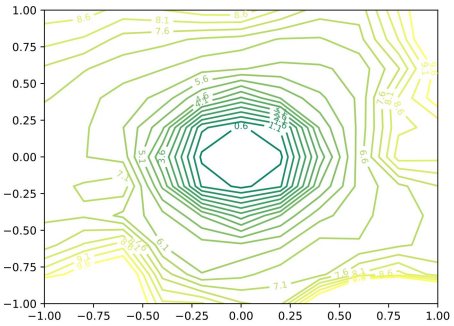
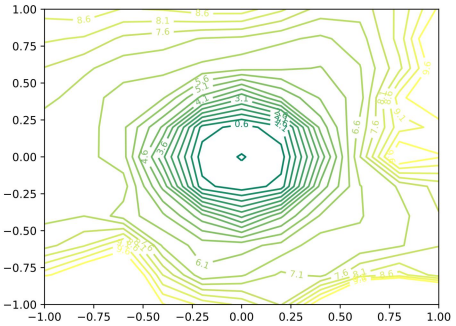
The norm of DP gradient follows a chi-squared distribution(n↑). Complex for non-DP case.

Experiment 4

Compare Loss Landscape of DP and Non-DP Training

- Visualize the loss contour of models on different stages.

Hao Li, et al. "Visualizing the Loss Landscape of Neural Nets", NeurIPS 2018.

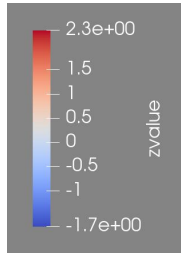
	Begin	Mid	End
DP			
non-DP			

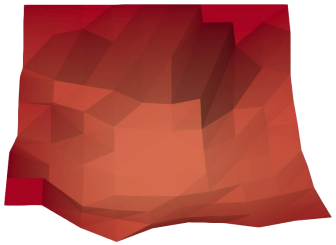
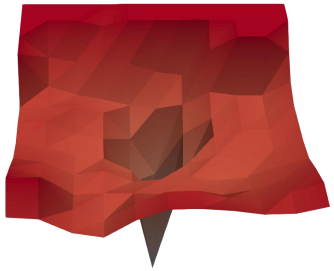
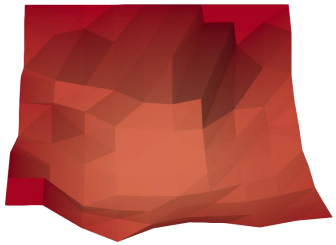
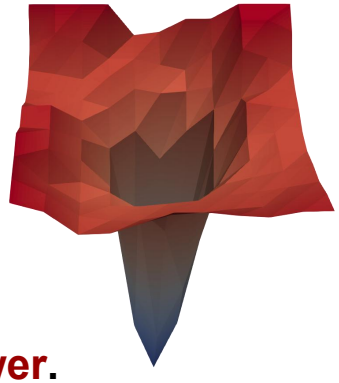
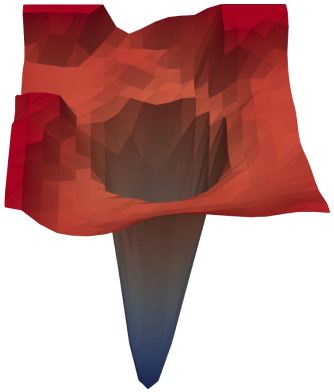
The loss landscape of DP training is more irregular, the concave is narrower.

Experiment 4

Compare Loss Landscape of DP and Non-DP Training

- 3D visualize the loss landscape of models on different stages.



	Begin	Mid	End
DP			
non-DP			

The loss landscape of DP models is shallower.

Summary



The Differences of DP and non-DP Training

1. **Clip gradient**
2. **Add noise**

Experiments Show that DP Training:

1. **Has worse performance**(low accuracy: 91.1%→79.6%)
   - The DP noise results in slower and less accurate convergence.
2. **Is harder to train**(more epochs, slower speed)
   - The loss landscape is more irregular, and the gradient has noise.
3. **Requires more GPU memory**(24% more)
   - Need to store more gradients: per-sample, noisy/clean, clipped.
4. **Has larger gradient norm and smaller angle variance**
   - The DP noise. The partial fine-tuning of projection layer.(further research)



Thank You!

Musheng Li

ml_361@usc.edu

Q & A