WHY DOES PRIVATE FINE-TUNING RESIST DIFFERENTIAL PRIVACY NOISE? A REPRESENTATION LEARNING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we investigate the impact of differential privacy (DP) on the fine-tuning of publicly pre-trained models, focusing on Vision Transformers (ViTs). We introduce an approach for analyzing the DP fine-tuning process by leveraging a representation learning law to measure the separability of features across intermediate layers of the model. Through a series of experiments with ViTs pre-trained on ImageNet and fine-tuned on a subset of CIFAR-10, we explore the effects of DP noise on the learned representations. Our results show that, without proper hyperparameter tuning, DP noise can significantly degrade feature quality, particularly in high-privacy regimes. However, when hyperparameters are optimized, the impact of DP noise on the learned representations is limited, leading to high model accuracy even in high-privacy settings. These findings provide insight into how pre-training on public datasets can help mitigate the privacy-utility trade-off in private deep learning applications.

1 INTRODUCTION

Recently, privately fine-tuning publicly pre-trained models with differential privacy (DP) has drawn much attention in private deep learning. For example, De et al. (2022) demonstrated that fine-tuning an ImageNet-pretrained Wide-ResNet achieves 95.4% accuracy on CIFAR-10 under $(\epsilon = 2.0, \delta = 10^{-5})$-DP, surpassing the 67.0% accuracy from training a three-layer convolutional neural network from scratch with private training (Abadi et al., 2016). Furthermore, Li et al. (2021); Yu et al. (2021) showed that pre-trained models like BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2018) achieve near-no-privacy-utility trade-offs when fine-tuned for sentence classification and generation tasks.

However, the empirical success of privately fine-tuning large, pre-trained models seems to contradict the worst-case dimensionality dependence seen in private learning problems. Specifically, noisy stochastic gradient descent (NoisySGD) requires adding noise scaled to $\sqrt{p}$ to each coordinate of the gradient in a model with p parameters, making it impractical for models with millions of parameters. This suggests that the benefits of pre-training may help mitigate the dimension dependence and privacy-utility trade-off inherent in NoisySGD. Recent works (Tang et al., 2023; Wang et al., 2024) have explored this issue, showing that if the last-layer features are well-learned, fine-tuning only the last layer—also known as linear probing—can achieve high accuracy with nearly no privacy-utility trade-offs. However, when the features are not sufficiently extracted, fine-tuning part or all of the model's parameters becomes essential (De et al., 2022). In such cases, the precise behavior of representations across intermediate layers remains difficult to analyze.

In this work, we investigate the private fine-tuning behavior of intermediate layers by leveraging representation learning tools in DP fine-tuning. Specifically, we adopt a representation law introduced by He & Su (2023; 2024), which quantifies the separability of representations across intermediate layers.

Based on experiments with Vision Transformers pretrained on ImageNet, our key observations are as follows:

- If hyperparameters are not chosen appropriately, the injected DP noise can destroy the extracted features.
- With a well-tuned selection of hyperparameters, a representation law shows that the impact of DP noise on representation learning is limited. As the representations are not significantly affected by DP noise, this explains why public pretraining helps mitigate the privacy-utility trade-off.

2 PRELIMINARIES

2.1 DIFFERENTIAL PRIVACY

The focus of this study is on *differentially private learning*, where the objective is to train a model while adhering to the mathematical definition of privacy known as differential privacy (Dwork, 2006). Differential privacy ensures that no individual training data point can be identified from the trained model, even when additional side information is available. The most commonly used DP notion, (ϵ, δ)-DP, is formally defined as follows.

Definition 2.1 (Differential privacy, DP)**.** A randomized mechanism M satisfies (ϵ, δ)-DP for $\epsilon \geq 0$ and $0 \leq \delta \leq 1$ if, for any pair of neighboring datasets D and D'—where one can be obtained from the other by adding or removing a single individual record—and any event S, we have

$$\mathbb{P}(M(D) \in S) \leq e^{\epsilon} \cdot \mathbb{P}(M(D') \in S) + \delta. \tag{1}$$

When ϵ and δ are small, it is harder to distinguish between D and D' just based on the outputs. Thus, ϵ and δ are called the privacy budget and a smaller value of ϵ means the algorithm is more private.

2.2 A REPRESENTATION LAW

The representation learning tool we adopt in this paper is a law introduced by He & Su (2023; 2024).

For vision models, this law measures the separability of the outputs at each layer of a deep learning model.

For a classification problem with K classes, let x_{ik} be the intermediate output of the deep neural network for the i-th image in the k-th class, with sample size n_k. The total sample size is $n = \sum_{i=1}^{k} n_k$. Denote $\overline{x}_k$ the sample mean of x_{ik} and $\overline{x}$ the sample mean of the outputs across all K classes. Define the between-class sum of squares SS_b and within-class sum of squares SS_w as

$$\mathrm{SS}_b = \frac{1}{n} \sum_{k=1}^{K} n_k (\overline{x}_k - \overline{x})(\overline{x}_k - \overline{x})^T,$$

$$\mathrm{SS}_w = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_k - \overline{x})(\overline{x}_k - \overline{x})^T.$$

The separability of the output can then be measured by the following:

$$D = \mathrm{Tr}\left[\mathrm{SS}_b \mathrm{SS}_w^{\dagger}\right],$$

where $\mathrm{SS}_w^{\dagger}$ is the Moore–Penrose inverse of SS_w. He & Su (2023) demonstrated that the separability of each layer decays linearly as the depth of the network increases. Specifically, let D_l denote the separability of the l-th layer of a deep neural network. A law of data separation states:

$$D_l \approx \rho^l D_0,$$

for some coefficient $0 < \rho < 1$.

In He & Su (2024), a law was introduced that applies to next-token prediction, where the prediction residual (PR) in the form of regression exhibits a linear decay across layers. To assess how well a large language model (LLM) predicts the next token, they fit a linear regression model to the dataset

D. Specifically, they treat each token x as its index in the token vocabulary and use a least-squares regression approach to predict the next token. The resulting model is expressed as:

$$\hat{x}_{\text{next}} = w \cdot h + b,$$

where w and b are the learned parameters, and h is the hidden representation from the model. This framework allows us to quantify the LLM's ability to predict the next token. To evaluate the performance of this prediction, the PR is defined as:

$$\text{PR} := \frac{\sum (x_{\text{next}} - \hat{x}_{\text{next}})^2}{\sum (x_{\text{next}} - \bar{x}_{\text{next}})^2},$$

where x_{next} represents the true next token, $\hat{x}_{\text{next}}$ is the predicted next token, and $\bar{x}_{\text{next}}$ is the mean of all true next tokens. The PR metric quantifies the proportion of variance in the true next token that is not explained by the model's prediction. A high PR value indicates that the token embeddings have limited predictive power, while a low PR value suggests that the embeddings are more effective at predicting the next token.

Specifically, the PR for the l-th layer in a model with depth L follows the relationship:

$$PR_l \approx \rho^{l-1} \times PR_1,$$

where $0 < \rho < 1$. This indicates that the predictive power of each layer decays linearly as the depth of the network increases.

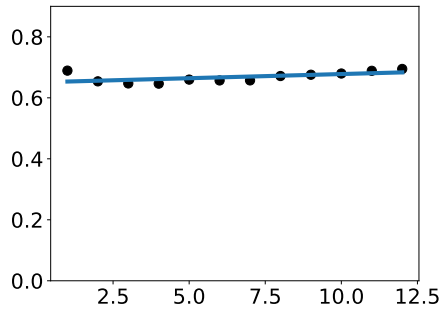
To apply this law to ViTs, we focus on the [CLS] (stands for classification) token, which is a special learnable token added to the input sequence of a ViT. The [CLS] token aggregates information from all patches of the image and is used for classification or regression tasks. We treat the representation of this [CLS] token as the input x in linear regression, with the actual classification label y as the target. By performing regression on the [CLS] token's embedding, we obtain the PR for each layer, similar to the next-token prediction task. This allows us to quantify how the predictive power of the model changes across its layers, following the same linear decay as observed in language models.

3 Main Results

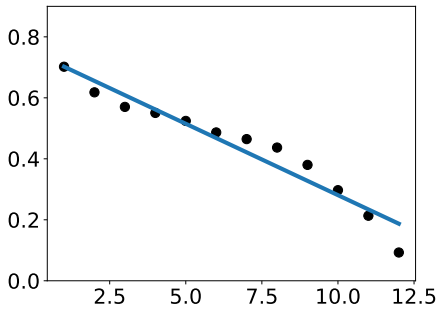
Experimental settings. We consider privately fine-tuning a Vision Transformer (ViT) with 12 blocks, publicly pretrained on ImageNet, on a subset of CIFAR-10. For each of the 10 classes in CIFAR-10, we randomly select 100 images, resulting in a total sample size of 10^3. With a fixed value of $\delta = 10^{-3}$ (the reciprocal of the total sample size), we examine both a high-privacy regime with $\epsilon = 1$ and a low privacy regime with $\epsilon = 8$. The tunable parameters range from 1 to 12 blocks of the ViT.

DP noise may destroy the features. As noted by Wang et al. (2024), when fine-tuning only the last layer, any constant learning rate is sufficient to ensure model convergence, regardless of the privacy budget. When fine-tuning all 12 blocks, we also considered using a constant learning rate, and the results are presented in Figure 1. In the low-privacy regime (Figure 1(b)), the derived law closely resembles the case with optimized hyperparameters, and the corresponding accuracy remains high at 95.8%. However, in the high-privacy regime with $\epsilon = 1$, as shown in Figure 1(a), the behavior significantly deviates from the case with optimized hyperparameters in Figure 2. Additionally, the accuracy drops to 11.5%, which is nearly equivalent to random guessing. This suggests that, without appropriate hyperparameter tuning, a larger noise can degrade the features without a more fine-grained hyperparameter tuning.

Optimized hyperparameters mitigate the privacy-utility trade-off. By choosing the hyperparameters using Optuna (Akiba et al., 2019), we display the representation law for the high privacy regime in Figure 2 and the low privacy regime in Figure 3. As we can see, with a fixed privacy budget ϵ, the number of tunable blocks has a limited impact on the learned representations, as the laws are nearly identical. Moreover, for different privacy budgets ($\epsilon = 1$ and $\epsilon = 8$), the representation laws are nearly the same. This means that despite the larger noise in the high privacy regime, the learned representations remain of high quality if the hyperparameters are well-tuned.

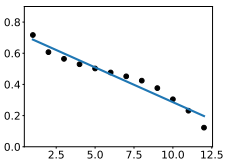


(a) $\epsilon = 1$ with learning rate 0.003. The accuracy is 11.5%.

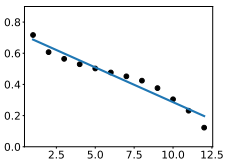


(b) $\epsilon = 8$ with learning rate 0.003. The accuracy is 95.8%.

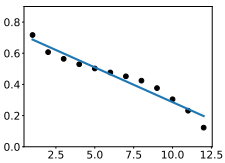
Figure 1: The representation law when privately finetuning vision transformers.



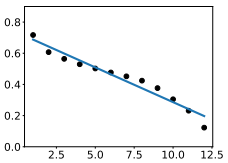
(a) Tune 1 block.



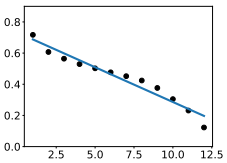
(b) Tune 2 blocks.



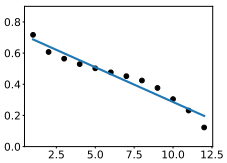
(c) Tune 3 blocks.



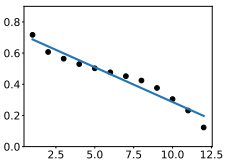
(d) Tune 4 blocks.



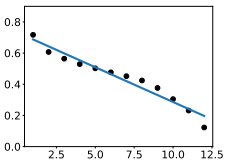
(e) Tune 5 blocks.



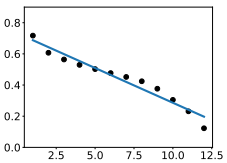
(f) Tune 6 blocks.



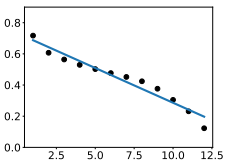
(g) Tune 7 blocks.



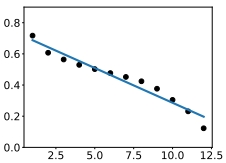
(h) Tune 8 blocks.



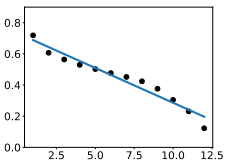
(i) Tune 9 blocks.



(j) Tune 10 blocks.



(k) Tune 11 blocks.

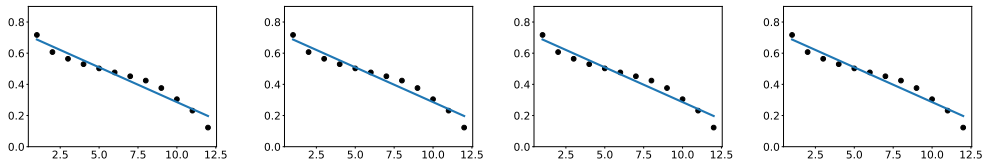


(l) Tune 12 blocks.

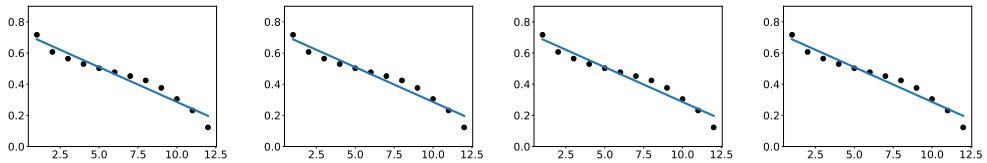
Figure 2: The representation law when privately fine-tuning vision transformers by tuning 1–12 blocks in the high privacy regime ($\epsilon = 1$). The accuracy in each case ranges from 92.83%-93.57%.

4 Conclusions and Future Study

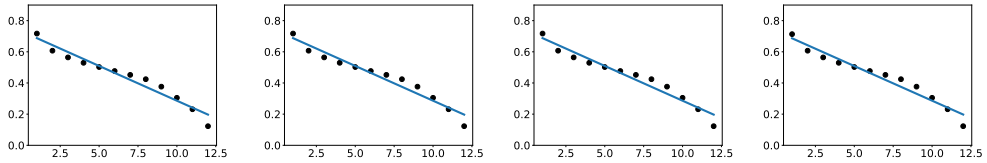
This study presents a detailed examination of the private fine-tuning process for large pre-trained models under differential privacy. Our key findings demonstrate that while DP noise can severely degrade model performance if hyperparameters are poorly chosen, careful tuning of hyperparameters can effectively mitigate this issue. Specifically, our analysis shows that the separability of intermediate layer representations remains largely unaffected by DP noise when appropriate tuning is applied, even in high-privacy regimes. This highlights the importance of hyperparameter optimization in private fine-tuning tasks and suggests that pre-training on large public datasets can help alleviate the inherent privacy-utility trade-off. Future work should focus on further exploring the dynamics of intermediate layer representations in private learning and developing more robust



(a) Tune 1 block. (b) Tune 2 blocks. (c) Tune 3 blocks. (d) Tune 4 blocks.



(e) Tune 5 blocks. (f) Tune 6 blocks. (g) Tune 7 blocks. (h) Tune 8 blocks.



(i) Tune 9 blocks. (j) Tune 10 blocks. (k) Tune 11 blocks. (l) Tune 12 blocks.

Figure 3: The representation law when privately fine-tuning vision transformers by tuning 1–12 blocks in the low privacy regime ($\epsilon = 8$). The accuracy in each case ranges from 92.83%-93.56%.

methods for hyperparameter selection. An important future topic is to extend our experiments on ViT to language models. The next token prediction law proposed by He & Su (2024) may provide some insights on the effects of learned representations for language models.

REFERENCES

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (eds.), *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer, 2006.

Hangfeng He and Weijie J. Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023. doi: 10.1073/pnas.2221704120.

Hangfeng He and Weijie J. Su. A law of next-token prediction in large language models. *arXiv preprint arXiv:2408.13442*, 2024.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

Xinyu Tang, Ashwinee Panda, Vikash Sehwag, and Prateek Mittal. Differentially private image classification by learning priors from random processes. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/7058bc192a37f5e5a57398887b05f6f6-Abstract-Conference.html.

Chendi Wang, Yuqing Zhu, Weijie J. Su, and Yu-Xiang Wang. Neural collapse meets differential privacy: Curious behaviors of noisygd with near-perfect representation learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=7rrN6E4KU0.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.