Differentially Private Federated Few-shot Image Classification

Aliaksandra Shysheya[*]
University of Cambridge

Marlon Tobaben[*]
University of Helsinki

John Bronskill
University of Cambridge

Andrew Paverd
Microsoft Research

Shruti Tople
Microsoft Research

Santiago Zanella-Beguelin
Microsoft Research

Richard E. Turner
University of Cambridge

Antti Honkela[†]
University of Helsinki

Abstract

In Federated Learning (FL), the role of a central server is to simply aggregate the gradient or parameter updates sent by an array of remote clients, which perform local model training using their individual data. Even though the server in FL does not have access to raw user data, the privacy of users may still be compromised through model parameters. To mitigate this and provide guaranteed level of privacy, user-level differentially private (DP) FL aggregation methods can be employed which are able to achieve accuracy approaching that of non-private training when there is a sufficient number of remote clients. In most practical distributed learning scenarios, the amount of labelled data each client has is usually limited, necessitating few-shot learning approaches. An effective approach to few-shot learning is transfer learning where the model employs a backbone pretrained on large public datasets and then fine-tunes it on a downstream dataset. A key advantage of transfer learning systems is that they can be made extremely parameter efficient by updating only a small subset of model parameters during fine-tuning. This advantage is extremely beneficial in the FL setting, as it helps minimize the communication cost spent on each client-server communication during training by transferring only those model parameters that need to be updated. To understand in which settings DP FL few-shot transfer learning can be effective, we perform a set of experiments that reveals how the accuracy of DP FL image classification systems is affected as the model architecture, dataset, and subset of learnable parameters in the model varies. We evaluate on three FL datasets, establishing state-of-the-art performance on the challenging FLAIR federated learning benchmark.

1 Introduction

In Federated Learning (FL) (McMahan et al., 2016), training data are split between remote clients, e.g. mobile phones or personal laptops, and the goal is to train a global (McMahan et al., 2016) or personalized (Fallah et al., 2020) model using the distributed data without transferring it to a central server. Model training is typically performed via numerous aggregation steps (communication rounds) between the server and clients. On each step, a subset of clients is selected, and each member performs local computation on their personal data, returning information to the server. Then the server aggregates all information from the chosen clients, updates the shared model parameters, and proceeds to the next round until convergence. Perhaps the most commonly used FL aggregation algorithm is FedAvg (McMahan et al., 2016), which makes several optimization steps over model parameters using stochastic gradient descent (SGD) locally and then computes a weighted sum of the updates from each client (based on the number of samples per client) to get a new shared model

[*] Equal contribution

[†] Correspondence to Antti Honkela <antti.honkela@helsinki.fi>

state. Even though client data is not directly shared with the server in FL aggregation, the privacy of clients may still be compromised via model parameters leaking some subset of training data (Yin et al., 2021; Huang et al., 2021; Geiping et al., 2020). This necessitates FL aggregation algorithms that ensure formal user-level privacy guarantees.

For applications where training data are sensitive (Abowd, 2018; Cormode et al., 2018), it has become increasingly common to train under Differential Privacy (DP) (Dwork et al., 2006) which is considered to be the gold standard for protecting individual training examples from discovery. To incorporate DP guarantees into standard centralized training DP-SGD (Rajkumar & Agarwal, 2012; Song et al., 2013; Abadi et al., 2016), a DP version of standard SGD, was introduced. Similarly to DP-SGD, DP-FedAvg (McMahan et al., 2018) is an adaptation of the baseline FL algorithm FedAvg (McMahan et al., 2016), which provides user-level DP guarantees by clipping the norm of local parameter updates and applying the Gaussian mechanism to the aggregated global update. If there are a sufficient number of remote clients, DP FL systems are able to achieve accuracy approaching that of non-private training (McMahan et al., 2018).

In addition to client data being sensitive, in most practical distributed learning scenarios, the amount of labeled data each client has is usually limited (e.g. medical images (Sheller et al., 2020), personal photos (Massiceti et al., 2021), or personal data or actions entered on a mobile device (Differential Privacy Team, 2017; Ding et al., 2017)), necessitating few-shot learning approaches. An effective approach to few-shot learning is transfer learning, which utilizes a backbone pretrained on large public datasets and then fine-tunes it on a downstream dataset (Yosinski et al., 2014; Kolesnikov et al., 2019). This approach is especially useful for private learning, as non-private pretraining provides a rich feature representation and reduces the amount of information required from the private downstream data to build a performant model. Another key advantage of transfer learning systems is that they can be made extremely parameter efficient by updating only a small subset of model parameters during fine-tuning (Shysheya et al., 2023). This advantage is extremely beneficial in the FL setting, as models with a smaller number of updateable parameters are preferred in order to reduce the client-server communication cost which is typically bandwidth-limited.

To date, we are only aware of one study (Song et al., 2022) that has used large pretrained models fine-tuned via FL aggregation algorithms for the task of transfer-learned image classification. However, this study evaluated DP FL systems using a single model architecture and on a single dataset with the data distribution significantly overlapping with the pretraining data distribution. In this work, we aim to understand under which conditions DP FL few-shot transfer learning can be effective as well as complement the existing study by evaluating on broader range of model architectures and datasets, including those with different degrees of distribution overlap. Our contributions are:

- We perform a comprehensive set of experiments that reveals how the accuracy of DP and non-private FL models are affected by dataset distribution overlap, model architecture, and the subset of learnable parameters in the model vary.
- We establish state-of-the-art performance under DP on the challenging FLAIR (Song et al., 2022) few-shot federated learning benchmark in terms of both classification metrics (macro average precision increased from 44.3% to 51.9%) and communication efficiency (cost reduced from 11.9M to 0.017M parameters per round) using the backbone from the original paper of Song et al. (2022).
- We further improve the state-of-the-art results under DP on FLAIR using a larger architecture, without any loss in communication efficiency compared to a smaller backbone. Macro average precision increased from 51.9% to 59%, while communication cost is only 0.013M parameters per round.

2 Background

In this section, we provide background information, definitions, and nomenclature required for subsequent sections. We focus our analysis on few-shot transfer learning based image classifiers that rely on large pretrained backbones.

Preliminaries We denote input images $\boldsymbol{x}$ and image labels $y \in \{1, \dots, C\}$ where C is the number of image classes indexed by c. Assume that we have access to a model $f(\boldsymbol{x}) = h_(b_{\boldsymbol{\theta}}(\boldsymbol{x}))$ that outputs class-probabilities for an image $p(y = c|\boldsymbol{x}, \boldsymbol{\theta},)$ for $c = 1, \dots, C$ and comprises a feature extractor backbone $b_{\boldsymbol{\theta}} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_b}$ with parameters $\boldsymbol{\theta}$ pretrained on a large upstream public dataset such as Imagenet-21K (Russakovsky et al., 2015) where d is the input image dimension and d_b is the output feature dimension, and a linear layer classifier or head $h{:}\mathbb{R}^{d_b} \rightarrow \mathbb{R}^C$ with weights . Let $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ be the downstream dataset that we wish to fine-tune the model f to. We denote the number of training examples per class or *shot* as S.

Learnable Parameters In all experiments, the head parameters are initialized to zero and are always learned when fine-tuning on $\mathcal{D}$. For the backbone weights $\boldsymbol{\theta}$, we consider three options: (i) *Head*: $\boldsymbol{\theta}$ are fixed at their pretrained values and do not change during fine-tuning, only the head parameters are updated; (ii) *All*: $\boldsymbol{\theta}$ are initialized with pretrained values, but can be updated during fine-tuning in addition to the head; and (iii) *FiLM*: using FiLM (Perez et al., 2018) layers. There exists a myriad of adaptors for both 2D convolutional and transformer networks including FiLM, Adapter (Houlsby et al., 2019), LoRA (Hu et al., 2021), VPT (Jia et al., 2022), AdaptFormer (Chen et al., 2022c), NOAH (Zhang et al., 2022), Convpass (Jie & Deng, 2022), Model Patch (Mudrakarta et al., 2019), and CaSE (Patacchiola et al., 2022) that enable a pretrained network to adapt to a downstream dataset in a parameter efficient manner. In this work, we use FiLM due to its simplicity, high performance, and low parameter count (Shysheya et al., 2023), though another adapter could be used. A FiLM layer scales and shifts the activations $\boldsymbol{a}_{ij}$ arising from the j^{th} output of a layer in the i^{th} block of the backbone as $\texttt{FiLM}(\boldsymbol{a}_{ij}, \gamma_{ij}, \beta_{ij}) = \gamma_{ij}\boldsymbol{a}_{ij} + \beta_{ij}$, where γ_{ij} and β_{ij} are scalars. We implement FiLM by fixing $\boldsymbol{\theta}$ at their pretrained values except for a subset of the scale and offset parameters utilized in the backbone normalization layers (e.g. BatchNorm, GroupNorm, or LayerNorm — see Appendix A.3.1 for details), which can update during fine-tuning. For example, in a ResNet50, there are only 11 648 learnable FiLM parameters, which is fewer than 0.05% of $\boldsymbol{\theta}$.

Dataset Distribution Overlap (DDO) The overlap between the distributions of the pretraining data and the downstream dataset is a key determinant of the ease and success of transfer learning. We measure the overlap as the relative difference between the accuracy of the *All* and *Head* learnable parameter configurations for a non-private model trained centrally, i.e. with all training data on one server. If two domains overlap substantially, then only adapting the head of the network is sufficient. If the overlap is small, then the backbone must also be adapted. Table A.1 provides the DDO values for all of the datasets used in the paper.

Differential Privacy (DP) DP (Dwork et al., 2006) is the gold standard for protecting sensitive data against privacy attacks. A stochastic algorithm is differentially private if it produces similar output distributions on similar datasets. More formally, (ϵ, δ)-DP with privacy budget $\epsilon \geq 0$ (lower means more private) and additive error $\delta \in [0, 1]$ bounds how much the output distribution can diverge on adjacent datasets. The additive error is typically chosen such that $\delta < 1/|\mathcal{D}|$. We refer to Dwork & Roth (2014) for a comprehensive introduction to DP. Centralized training with DP guarantees commonly assumes example-level privacy, which defines dataset adjacency on an examplar level, i.e. two datasets are adjacent if one can be obtained from the other by adding or removing one datapoint. In contrast, in DP FL, aggregation schemes that guarantee user-level privacy are considered, where two federated datasets are adjacent if one can be obtained from the other by adding or removing the data of a single user or client. Being stronger than example-level privacy, user-level privacy is constructed to prevent information about any training client from being leaked by a published model.

3 Related Work

Non-private FL and Transfer Learning There has been a recent surge of interest in using large pretrained models as initialization for training non-private decentralized models in both NLP (Lin et al., 2022; Stremmel & Singh, 2021; Weller et al., 2022; Tian et al., 2022) and computer vision (Chen et al., 2022b; Tan et al., 2022; Qu et al., 2021; Chen et al., 2022a; Nguyen et al., 2022; Liu et al., 2022). Most of these works were able to improve upon state-of-the-art results under different tasks and settings within FL as well as showing that the client data heterogeneity problem often seen in FL can be partially mitigated with pretrained networks.

FL and DP Even though the server in FL does not have access to raw user data, the privacy of users may still be compromised if i) the server is untrusted (Huang et al., 2021) or ii) a third party has access to the model after training (Geiping et al., 2020; Carlini et al., 2022). Cryptographic techniques like secure aggregation Goryczka et al. (2013) can protect against the former, while to tackle the latter, DP adaptations of the FL aggregation algorithms are needed McMahan et al.

(2018). Similarly to DP-SGD, DP-FedAvg (McMahan et al., 2018) is an adaptation of the baseline FL algorithm FedAvg (McMahan et al., 2016), which provides user-level DP guarantees by applying the Gaussian mechanism to parameter updates sent to the server.

Recently, a few studies have been conducted investigating the use of large pretrained models for FL under DP constraints in NLP (Basu et al., 2021), representation learning (Xu et al., 2022), and image classification (Song et al., 2022). The closest work to ours is the work of Song et al. (2022) who introduce FLAIR, a few-shot federated learning image classification dataset, which they use to perform a relatively small evaluation of pretrained models (only ResNet18 was used) fine-tuned using FL under DP. However, to the best of our knowledge, there are no other studies on how large pretrained models fine-tuned via FL aggregation algorithms behave under DP constraints for the task of transfer-learned image classification. In this work we aim to close this gap and provide experimental evaluation of these methods on real-world datasets.

4 Experiments

In our experiments, we endeavor to answer the question: "Under what conditions is differentially private few-shot federated image classification effective?" We focus on transfer learning approaches that utilize large pretrained backbones. We do this empirically by varying the: (i) set of learnable parameters in f (*All*, *Head*, *FiLM*); (ii) downstream dataset $\mathcal{D}$ (with varying DDO); and (iii) network architecture: ResNet18 (R-18) (He et al., 2016) pretrained on ImageNet-1K with 11.2M parameters, BiT-M-R50x1 (R-50) (Kolesnikov et al., 2019) pretrained on ImageNet-21K with 23.5M parameters, Vision Transformer VIT-Base-16 (VIT-B) (Dosovitskiy et al., 2020) pretrained on ImageNet-21K with 85.8M parameters. Source code for all experiments can be found at: `https://github.com/cambridge-mlg/dp-few-shot`.

In our evaluation, we use three datasets with different DDO. The first is FLAIR Song et al. (2022), which is a recently proposed real-world dataset for multi-label image classification. It has more than 50k users with heterogeneous data as well as a long-tailed label distribution, making it particularly appealing for benchmarking federated learning both in non-private and private settings. Comprising mainly natural image data, FLAIR is a high DDO dataset. The second dataset is CIFAR-100, which is medium DDO. For CIFAR-100, we use 500 training clients and 100 test clients, with each client having 100 samples and no clients sharing any data. To introduce more client heterogeneity, the data are distributed using the Pachinko Allocation Method (Li & McCallum, 2006) as in Reddi et al. (2021). The allocation method is described in detail in Appendix A.2. The third dataset is Federated EMNIST Caldas et al. (2018), a dataset of black-and-white handwritten symbols from 62 classes grouped according to the writer. EMNIST is a highly out-of-distribution dataset (i.e. low DDO) with respect to the ImageNet-21K pretraining data. As the number of training users in CIFAR-100 (500 users) and Federated EMNIST (3400 users) is relatively low, we need to increase ϵ from 2 to 8, such that the amount of added noise during aggregation is not excessive. δ is set to $N^{-1.1}$, where N is the number of training clients.

We use FedADAM (Reddi et al., 2021) aggregation, which was shown to have a better empirical performance than standard FedAvg (McMahan et al., 2016). For hyper-parameter tuning, we perform a small grid search over the server and client learning rates. The hyper-parameter ranges searched are provided in Appendix A.3.2.

Figure 1 shows the performance of different model configurations on all three datasets with and without DP. Tabular results are shown in Tables 1 to 3. For FLAIR, we report macro average precision (Macro-AP) results in the figure, while other metrics are in Tables 1 and 2. For a fair comparison on FLAIR, we fixed all of the training hyperparameters to the values from the original paper (Song et al., 2022), except for local and server learning rates. For CIFAR-100 and Federated EMNIST, we report standard test classification accuracy. All training details and hyperparameters are in Appendix A.3.2.

As communication cost is important in FL, in Figure 2 we report the number of parameters required to be transmitted for each model configuration in one user-server interaction. The results are shown for just FLAIR. The results for the other datasets are of the same magnitude, varying only because the number of classes is different.

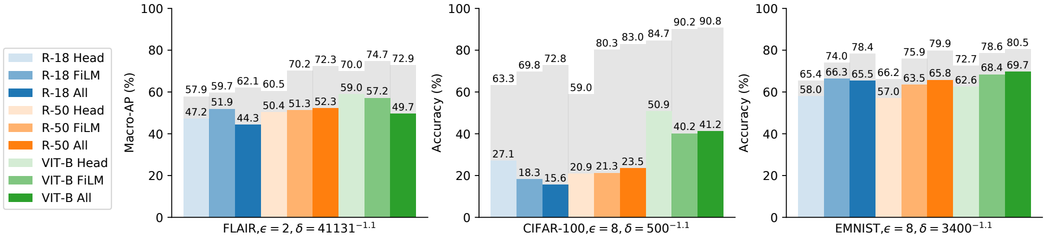

Figure 1: Private (colored) and non-private (gray) FL performance on FLAIR (left), CIFAR-100 (middle) and EMNIST (right) as a function of backbone and learnable parameters. We use Macro-AP as the primary metric to report accuracy for FLAIR, and standard accuracy on other datasets. The R-18 *All* result on FLAIR is taken from Song et al. (2022). Our FLAIR results set a new state-of-the-art.

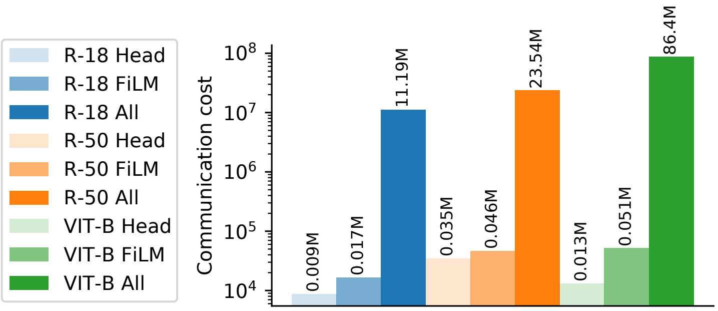

Figure 2: FLAIR communication cost – the number of parameters sent at every user-server communication round.

Summarizing Figures 1 and 2, key observations are:

FLAIR and CIFAR-100 (high-moderate DDO)

- With R-18 used in the original paper, we achieve state-of-the-art performance under DP with *FiLM*, improving Macro-AP from 44.3 to 51.9. This improvement comes with a reduction in communication cost from 11.2M parameters per each user-server interaction to only 17k.
- With VIT-B we further improve the state-of-the-art results on FLAIR in both DP and non-private settings. Under DP, we improved the Macro-AP to 59.0, while for non-private, the Macro-AP increased from 62.1 to 74.7.
- *Head* is more robust under DP than *All* or *FiLM*. *Head* has the smallest relative drop in performance of around 10% on FLAIR and 33.8% on CIFAR-100 using VIT-B.
- Under DP, in the case of a large number of training clients as in FLAIR, tuning either *FiLM* or *Head* is the most performant in terms of both accuracy and communication cost. *Head* is preferable in the case of a large output feature dimension (as in R-50 with $d_b = 2048$), while *FiLM* adaptation is better when the output feature dimension is smaller (as in R-18 with $d_b = 512$).
- In the case of a small number of training clients as in CIFAR-100, the performance deteriorates significantly under DP and *Head* performs the best (50.9% for VIT-B). Interestingly, VIT-B achieves far better accuracy than the other backbones.

Federated EMNIST (low DDO)

- Although *All* has the largest relative drop in accuracy under DP (14.1% for R-50), it achieves the best accuracy for $\epsilon = 8$ and $\epsilon = \infty$.
- *Head* under-performs regardless of the backbone used, providing empirical evidence that adapting the head only is insufficient for datasets with low DDO.

- Under DP, *FiLM* reaches accuracy on par with *All* (68.4% vs. 69.7% for VIT-B). *FiLM* transfers far fewer parameters than *All* during training (Figure 2), making it the preferred approach for FL on low DDO datasets.
- VIT-B outperforms all other backbones regardless of configuration. Surprisingly, R-18 pretrained on ImageNet-1k performs similarly to R-50 pretrained on ImageNet-21k, perhaps indicating that the learned fine-grained features of a larger backbone are not that useful in transfering to a low DDO dataset.

Table 1: Non-private Federated Learning performance on FLAIR as a function of backbone b_θ and learnable parameters. C stands for averaged per-class metrics (Macro) and O denotes overall metrics (Micro). P, R and AP denote precision, recall, and average precision, respectively. The R-18 *All* result is taken from the original paper Song et al. (2022). Due to the significant computational requirements, only a single random seed was used in all experiments on FLAIR.

b_θ		ϵ	C-P	O-P	C-R	O-R	C-F1	O-F1	C-AP	O-AP
R-18	ALL	∞	71.8	83.5	48.6	76.0	58.0	79.5	62.1	88.8
	FILM	∞	73.8	82.0	44.8	74.4	55.7	78.0	59.7	87.7
	HEAD	∞	71.0	79.9	43.8	72.9	54.1	76.2	57.9	85.8
R-50	ALL	∞	76.9	85.2	**62.0**	82	68.6	83.6	72.3	91.9
	FILM	∞	78.3	83.8	57.9	80.0	66.6	81.9	70.2	90.6
	HEAD	∞	76	82.3	42.7	71.3	54.6	76.4	60.5	86.7
VIT-B	ALL	∞	79.6	**86.8**	57.4	**82.9**	66.7	**84.8**	72.9	**93.1**
	FILM	∞	**81.9**	**86.8**	59.3	81.6	**68.8**	84.1	**74.7**	92.7
	HEAD	∞	81.6	83.7	52	72.2	63.4	77.5	70.0	87.6

Table 2: Federated Learning performance on FLAIR under DP with $\epsilon = 2$ as a function of backbone b_θ and learnable parameters. C stands for averaged per-class metrics (Macro) and O denotes overall metrics (Micro). P, R and AP denote precision, recall, and average precision, respectively. The R-18 *All* result is taken from the original paper Song et al. (2022). Due to significant computational requirements, only single random seed was used in all experiments with FLAIR.

b_θ		ϵ	C-P	O-P	C-R	O-R	C-F1	O-F1	C-AP	O-AP
R-18	ALL	2	47.3	77.5	32.3	64.3	38.4	70.3	44.3	80.2
	FILM	2	59.0	81.0	39.1	70.3	47.0	75.3	51.9	85.2
	HEAD	2	47.6	81.4	34.2	66.4	39.8	73.1	47.2	83.4
R-50	ALL	2	56.2	83.1	38.1	70.9	45.4	76.6	52.3	86.6
	FILM	2	59.7	79.3	39.4	69.9	47.5	74.3	51.3	84.2
	HEAD	2	57.0	79.8	38.0	68.5	45.6	73.7	50.4	83.8
VIT-B	ALL	2	47.8	82.3	37.5	71.0	42.1	76.2	49.7	86.1
	FILM	2	58.1	**84.2**	**42.5**	**76**	49.1	**79.9**	57.2	**89.2**
	HEAD	2	**67.1**	83.4	39.8	68.9	**50.0**	75.5	**59.0**	85.9

Table 3: Federated Learning performance on CIFAR-100 and EMNIST with ($\epsilon = 8$) and without ($\epsilon = \infty$) DP as a function of backbone b_θ and learnable parameters. Accuracy (in %) is reported. R-18 backbone is pretrained on ImageNet-1k, VIT-B and R-50 are pretrained on ImageNet-21k. The ± sign indicates the 95% confidence interval over 3 runs with different seeds.

		R-18			R-50			VIT-B		
DATASET	ϵ	HEAD	FILM	ALL	HEAD	FILM	ALL	HEAD	FILM	ALL
CIFAR-100	∞	63.3±0.2	69.8±0.3	72.8±0.7	59.1±0.5	79.8±0.5	83.0±0.1	84.6±0.1	90.2±0.3	**90.8±0.3**
	8	27.1±1.4	18.3±0.9	15.6±1.0	20.9±0.6	21.3±1.0	23.5±1.3	**50.8±0.1**	40.2±2.3	41.2±3.4
EMNIST	∞	65.4±0.1	74.0±0.9	78.4±1.1	66.2±0.4	75.9±0.4	79.9±0.4	72.7±0.2	78.6±0.1	**80.5±0.1**
	8	58.0±0.4	66.3±0.5	65.5±0.2	57.0±0.3	63.5±0.1	65.8±0.3	62.6±0.1	68.4±0.2	**69.7±0.3**

5 Discussion

In summary, our experiments show that DP few-shot FL can be effective in terms of accuracy and can lower communication cost per round in comparison to *All* by orders of magnitude using *FiLM* in the case of low DDO and *Head* in the case of high DDO. To achieve a high level of privacy (i.e. low ϵ) the number of clients needs to be in the tens of thousands in order to minimize the noise added during parameter updates.

Acknowledgements

Marlon Tobaben and Antti Honkela are supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; and grant 325573), the Strategic Research Council at the Academy of Finland (Grant 336032) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. Aliaksandra Shysheya, John Bronskill, and Richard E. Turner are supported by an EPSRC Prosperity Partnership EP/T005386/1 between the EPSRC, Microsoft Research and the University of Cambridge. This work has been performed using resources provided by the CSC – IT Center for Science, Finland, and the Finnish Computing Competence Infrastructure (FCCI), as well as the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service `https://www.hpc.cam.ac.uk` funded by EPSRC Tier-2 capital grant EP/P020259/1. We thank Joonas Jälkö, Stratis Markou, Massimiliano Patacchiola and Runa Eschenhagen for helpful comments and suggestions.

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. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018.

Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RUQ1zwZR8_.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zümrüt Müftüoglu, Sahib Singh, and Fatemehsadat Mireshghallah. Benchmarking differential privacy and federated learning for BERT models. *CoRR*, abs/2106.13973, 2021. URL https://arxiv.org/abs/2106.13973.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018. URL http://arxiv.org/abs/1812.01097.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han-Wei Shen, and Wei-Lun Chao. On the importance and applicability of pre-training for federated learning, 2022a. URL https://arxiv.org/abs/2206.11488.

Jinyu Chen, Wenchao Xu, Song Guo, Junxiao Wang, Jie Zhang, and Haozhao Wang. Fedtune: A deep dive into efficient federated fine-tuning with pre-trained transformers, 2022b. URL https://arxiv.org/abs/2211.08025.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022c.

Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 1655–1658, 2018.

Apple Differential Privacy Team. Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, 2017.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/24389bfe4fe2eba8bf9aa9203a44cdad-Paper.pdf.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, December 2020.

Google. Tensorflow federated: Machine learning on decentralized data. https://www.tensorflow.org/federated, 2019a.

Google. Tensorflow privacy: Library for training machine learning models with privacy for training data". https://github.com/tensorflow/privacy/, 2019b.

Slawomir Goryczka, Li Xiong, and Vaidy Sunderam. Secure multiparty aggregation with differential privacy: A comparative study. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pp. 155–163, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450315999. doi: 10.1145/2457317.2457343. URL https://doi.org/10.1145/2457317.2457343.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0CDKgyYaxC8.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 577–584, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143917. URL https://doi.org/10.1145/1143844.1143917.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 157–175, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-naacl.13.

Zicheng Liu, Da Li, Javier Fernandez-Marques, Stefanos Laskaridis, Yan Gao, Łukasz Dudziak, Stan Z. Li, Shell Xu Hu, and Timothy Hospedales. Federated learning for inference at anytime and anywhere, 2022. URL https://arxiv.org/abs/2212.04084.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BJ0hF1Z0b.

Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, and Andrew Howard. K for the price of 1: Parameter efficient multi-task and transfer learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJxvEh0cFQ.

John Nguyen, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat. Where to begin? exploring the impact of pre-training and initialization in federated learning, 2022. URL https://arxiv.org/abs/2206.15387.

Massimiliano Patacchiola, John Bronskill, Aliaksandra Shysheya, Katja Hofmann, Sebastian Nowozin, and Richard E Turner. Contextual squeeze-and-excitation for efficient few-shot image classification. *arXiv preprint arXiv:2206.09843*, 2022.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Li Fei-Fei, Ehsan Adeli, and Daniel L. Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10051–10061, 2021.

Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In Neil D. Lawrence and Mark A. Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pp. 933–941. JMLR.org, 2012. URL http://proceedings.mlr.press/v22/rajkumar12.html.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Aliaksandra Shysheya, John F Bronskill, Massimiliano Patacchiola, Sebastian Nowozin, and Richard E Turner. Fit: Parameter efficient few-shot transfer learning for personalized and federated image classification. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=9aokcgBVIj1`.

Congzheng Song, Filip Granqvist, and Kunal Talwar. Flair: Federated learning annotated image repository. *arXiv preprint arXiv:2207.08869*, 2022.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013, Austin, TX, USA, December 3-5, 2013*, pp. 245–248. IEEE, 2013. doi: 10.1109/GlobalSIP.2013.6736861. URL `https://doi.org/10.1109/GlobalSIP.2013.6736861`.

Joel Stremmel and Arjun Singh. Pretraining federated text models for next word prediction. In Kohei Arai (ed.), *Advances in Information and Communication*, pp. 477–488, Cham, 2021. Springer International Publishing. ISBN 978-3-030-73103-8.

Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pretrained models: A contrastive learning approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yuanyishu Tian, Yao Wan, Lingjuan Lyu, Dezhong Yao, Hai Jin, and Lichao Sun. Fedbert: When federated learning meets pre-training. *ACM Trans. Intell. Syst. Technol.*, 13(4), aug 2022. ISSN 2157-6904. doi: 10.1145/3510033. URL `https://doi.org/10.1145/3510033`.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. Pretrained models for multilingual federated learning, 2022. URL `https://arxiv.org/abs/2206.02291`.

Zheng Xu, Maxwell Collins, Yuxiao Wang, Liviu Panait, Sewoong Oh, Sean Augenstein, Ting Liu, Florian Schroff, and H. Brendan McMahan. Learning to generate image embeddings with user-level differential privacy, 2022. URL `https://arxiv.org/abs/2211.10844`.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16337–16346, June 2021.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3320–3328, 2014.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022.

A Appendix

A.1 Dataset Distribution Overlap

Table A.1 shows the data distribution overlap between the pretraining data (ImageNet-1K for FLAIR and ImageNet-21K for CIFAR-100 and EMNIST) and each of the downstream datasets used in our experiments.

Table A.1: Amount of distribution overlap between each of the 3 downstream datasets used in our experiments and the pretraining data. The Score column is computed as the difference between the non-private accuracy of the *All* learnable parameter configuration and the *Head* configuration, normalized by the *All* accuracy, and then scaled by 100. The lower the score, the more the data distribution overlap between the pretraining data and the downstream dataset. In the Distribution Overlap column, we map the score into three buckets: a score of 0-5 is High, 5-10 is Medium, and 10 or greater is Low. To compute the scores, we use the VIT-B backbone and use accuracies (Macro-AP for FLAIR) from the non-private FL experiments.

Dataset	**Score**	**Distribution Overlap**
FLAIR (Song et al., 2022)	4	High
CIFAR100 (Krizhevsky, 2009)	7	Medium
EMNIST (Caldas et al., 2018)	10	Low

A.2 Federated CIFAR-100 allocation method

Pachinko Allocation Method (PAM) is a topic modeling method that structures topics as a rooted directed acyclic graph (DAG), with leaves corresponding to the vocabulary words. Each interior node models a correlation among its children (vocabulary words or other internal nodes) with a Dirichlet distribution. To generate a document, each interior node samples a multinomial distribution from the corresponding Dirichlet distribution. Then to sample a word from the document, starting from the root, a child is sampled from the corresponding generated multinomial distribution and the chosen child node is used as the next sampling point. This procedure is repeated until the leaf node is reached.

To distribute CIFAR-100 across clients, we use the coarse and fine labels provided for each image. In particular, in CIFAR-100, there are 20 coarse labels and 5 fine labels for each coarse label. This label structure is represented by a DAG, where root is connected to coarse labels and each coarse label is connected to the corresponding fine labels. The root node models the distribution over its children with a symmentic Dirichlet distribution with parameter α ($Dir(\alpha)$), while each coarse label node has a symmetric Dirichlet distribution with parameter β ($Dir(\beta)$). We sample a subset of N images per each client as we would sample N words from a new document using PAM. In more detail, for each client we sample a multinomial distribution from $Dir(\alpha)$ over the coarse labels and a set of multinomial distributions from $Dir(\beta)$, one for each coarse label node. After that, to add a new image to the client dataset, we first sample a leaf node using PAM, and then randomly sample (without replacement) an example with the generated fine label. This example is added to the client dataset. If there are no more examples with a particular label, we remove the label node from the coarse label node multinomial and renormalize the resulting distribution to obtain a new multinomial. In our experiments, we used 100 images per client, 500 training clients, $\alpha = 0.1$ and $\beta = 10$. We used the implementation provided in tensorflow-federated (Google, 2019a).

A.3 Training and Evaluation Details

A.3.1 FiLM Layer Implementation

Table A.2 details the locations and count of the parameters that are updateable for the *FiLM* configuration in each of the backbones used in the experiments.

Table A.2: Backbone parameter count, FiLM parameter count, FiLM parameter count as a percentage of the backbone parameter count, and FiLM parameter locations within the backbone for each of the backbones used in the experiments.

BACKBONE	BACKBONE COUNT	FILM COUNT	FILM (%)	LOCATIONS
R-18	11.2M	7808	0.07	GROUPNORM SCALE AND BIAS THAT FOLLOWS EACH 3X3CONV LAYER
R-50	23.5M	11648	0.05	GROUPNORM SCALE AND BIAS THAT FOLLOWS EACH 3X3CONV LAYER FINAL GROUPNORM SCALE AND BIAS BEFORE HEAD
VIT-B	85.8M	38400	0.04	ALL LAYERNORM SCALE AND BIAS

A.3.2 FEDERATED LEARNING EXPERIMENTS

All experiments were performed in TensorFlow using tensorflow-federated Google (2019a) for federated aggregation and tensorflow-privacy Google (2019b) for privacy accounting and the adaptive clipping algorithm Andrew et al. (2021). CIFAR-100 and Federated EMNIST datasets were taken from tensorflow-federated.

FLAIR Each model configuration is trained for 5000 rounds with a cohort size of 200. Each sampled user trains the model locally with SGD for 2 epochs with local batch size set to 16. The maximum number of images for each user is set to 512. For DP, $\epsilon = 2, \delta = N^{-1.1}$, where N is the number of training users. As in the original paper, we set L2 norm quantile to 0.1 for adaptive clipping and we use 200 users sampled uniformly per round to simulate the noise-level with a cohort size of 5000.

For the non-private setting we perform the grid search over:

- server learning rate $\in \{0.01, 0.05, 0.1\}$
- client learning rate $\in \{0.01, 0.05, 0.1\}$

For the private setting ($\epsilon = 2$) we fixed the client learning rate to the optimal value found for the non-private run and a perform grid search over the server learning rate in the set $\{a/2, a/10, a/50, a/100\}$, where a is the optimal server learning rate found for the non-private setting.

CIFAR-100 and Federated EMNIST Each model configuration is trained for 500 rounds with a cohort size of 20. Each sampled user trains the model locally with SGD for 5 epochs with local batch size set to 100. The maximum number of images for each user is set to 512. For DP, $\epsilon = 8, \delta = N^{-1.1}$, where N is the number of training users ($N = 500$ for CIFAR-100, $N = 3400$ for Federated EMNIST). As in the original paper, we set L2 norm quantile to 0.1 for adaptive clipping and we use 20 users sampled uniformly per round to simulate the noise-level with a cohort size of 100.

For the non-private setting we perform the grid search over:

- server learning rate $\in \{0.05, 0.1, 0.5\}$
- client learning rate $\in \{0.01, 0.05, 0.1\}$

For the private setting ($\epsilon = 8$) we fixed the client learning rate to the optimal value found for the non-private run and perform a grid search over:

- server learning rate $\in \{a/2, a/10, a/50, a/100\}$, where a is the optimal server learning rate found for the non-private setting.
- quantile for adaptive clipping bound $\in \{0.1, 0.5, 0.8\}$