On the Efficacy of Differentially Private Few-shot Image Classification

Anonymous authors
Paper under double-blind review

Abstract

There has been significant recent progress in training differentially private (DP) models which achieve accuracy that approaches the best non-private models. These DP models are typically pretrained on large public datasets and then fine-tuned on private downstream datasets that are relatively large and similar in distribution to the pretraining data. However, in many applications including personalization and federated learning, it is crucial to perform well (i) in the few-shot setting, as obtaining large amounts of labeled data may be problematic; and (ii) on datasets from a wide variety of domains for use in various specialist settings. To understand under which conditions few-shot DP can be effective, we perform an exhaustive set of experiments that reveals how the accuracy and vulnerability to attack of few-shot DP image classification models are affected as the number of shots per class, privacy level, model architecture, downstream dataset, and subset of learnable parameters in the model vary. We show that to achieve DP accuracy on par with non-private models, the shots per class must be increased as the privacy level increases by as much as $20 - 35\times$ at $\epsilon = 1$. We also show that learning parameter-efficient FiLM adapters under DP is competitive with and often superior to learning just the final classifier layer or learning all of the network parameters. Finally, we evaluate DP federated learning systems and establish state-of-the-art performance on the challenging FLAIR benchmark.

1 Introduction

It is well known that neural networks trained without formal privacy guarantees can be attacked to expose a subset of the training data (Carlini et al., 2021; Balle et al., 2022). 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 the privacy of individual training examples. Training with DP stochastic gradient descent (DP-SGD) (Rajkumar & Agarwal, 2012; Song et al., 2013; Abadi et al., 2016), which adapts SGD to guarantee DP, typically impairs model performance due to gradient clipping and the addition of noise during training in order to mask the contribution of individual examples to model updates. However, there has been significant recent progress in training DP models which achieve accuracy that approaches the best non-private models in both NLP (Li et al., 2022b; Yu et al., 2022) and computer vision (Kurakin et al., 2022; De et al., 2022; Mehta et al., 2022; Cattan et al., 2022).

The majority of these approaches are based on transfer learning where the models have been pretrained on large public datasets and then fine-tuned (Yosinski et al., 2014) on a private downstream dataset with DP-SGD, as transfer learning has been shown to be highly effective on non-private data (Kolesnikov et al., 2020; Shysheya et al., 2022). In the non-private setting, the subset of model parameters to fine-tune ranges from all model parameters (Kolesnikov et al., 2020) to only the final layer, with the tuning of parameter-efficient adapters (Perez et al., 2018; Houlsby et al., 2019; Mahabadi et al., 2021) becoming increasingly prevalent. Transfer learning has also proven successful in the DP setting with (Yu et al., 2022) and without (Mehta et al., 2022) adapters.

However, strong DP results have only been demonstrated with relatively large datasets, with no extensive DP few-shot studies performed. The few-shot setting is crucial to any application where obtaining large amounts of labeled data is problematic. It is particularly significant in federated learning, where a global model is trained using data from multiple distributed users, and personalized federated learning, which involves customizing a federated learning model with a specific user's data. In such scenarios, each user's data may be sensitive and of limited size, such as medical images (Sheller et al., 2020), personal photos (Massiceti et al., 2021), or confidential personal data or actions entered on a mobile device (Differential Privacy Team, 2017; Ding et al., 2017).

In addition, the strong DP transfer learning results that have recently been reported have largely considered the case where the data distribution of the downstream dataset is similar to the pretraining data distribution (Tramèr et al., 2022). A more demanding test is out-of-domain transfer where more information needs to be extracted from the downstream dataset, making private learning more challenging. Support for differing data distributions is essential for frequently encountered specialist settings such as medical imaging, Earth imaging, or personalized object recognition.

In this work, we answer the question: *Under what conditions is differentially private few-shot image classification effective?* Our contributions are:

- We provide the first comprehensive study on the efficacy of DP few-shot image classification. In particular, in the centralized setting we perform an exhaustive set of experiments that reveals how the accuracy of DP and non-private models are affected as the number of shots per class, privacy level, downstream dataset, model architecture, and the subset of learnable parameters in the model vary. We also investigate whether the trends observed in the centralized setting apply to federated learning. Novel insights include:

  Amount of data required: It is known that classification accuracy under DP decreases as the level of privacy increases and the amount of data decreases, however: (i) we quantify how much more data is required under various levels of DP to match non-private accuracy. In particular, we found that the number of shots per class must be increased by roughly $20 - 35\times$ at $\epsilon = 1$ to match non-private performance, depending on the subset of learnable parameters; and (ii) we show that accuracy under DP is strongly related to the difficultly of the transfer learning task.

  Model parameterization: We show that fine-tuning parameter-efficient FiLM adapters in addition to the final linear classifier layer performs close to or better than fine-tuning all parameters in the model or fine-tuning only the final layer under few-shot DP. This is demonstrated by superior accuracy for the FiLM configuration on the challenging VTAB-1k benchmark and establishing state-of-the-art in terms of accuracy (macro average precision increased from 44.3% to 51.9%) and communication efficiency (cost reduced from 11.9M to 0.017M parameters per round) on the large-scale FLAIR federated learning benchmark.

  Characterization of few-shot DP learning dynamics: We show that non-private few-shot transfer learners are generally in the interpolating regime where they achieve 100% training accuracy. Under strong DP, trained networks are generally in the regularization regime where test and train accuracies are comparable.
- We assess the vulnerability of DP few-shot models with a strong membership inference attack (MIA) and find that non-private models are highly susceptible and the privacy level must be increased to a high level to mitigate them.
- Finally, we establish recommended practice guidelines for training DP few-shot models.

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 backbones pretrained on non-private data.

Preliminaries We denote input images $\boldsymbol{x}$ and image labels $y \in \{1, \ldots, C\}$ where C is the number of image classes indexed by c. Assume that we have access to a model $f(\boldsymbol{x}) = h_{\boldsymbol{\phi}}(b_{\boldsymbol{\theta}}(\boldsymbol{x}))$ that outputs class-probabilities for an image $p(y|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ and comprises a feature extractor backbone $b_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^{d_b}$ 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 $h_{\boldsymbol{\phi}} : \mathbb{R}^{d_b} \to \mathbb{R}^C$ with weights $\boldsymbol{\phi}$. Let $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ be the private 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 $\boldsymbol{\phi}$ 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 $\boldsymbol{\phi}$ 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 myriad of adaptors for both 2D convolutional and transformer networks including FiLM, Adapter (Houlsby et al., 2019), LoRA (Hu et al., 2022a), 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., 2022), though another adapter could be used. A FiLM layer scales and shifts the activations $\boldsymbol{a}_{ij}$ arising from the jth output of a layer in the ith 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. GroupNorm, LayerNorm, etc., 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 $\boldsymbol{\theta}$.

Transfer Difficulty (TD) The overlap between the distributions of the pretraining data and the downstream dataset as well other factors such as the number of classes in the downstream dataset are key determinants of the ease and success of transfer learning. We measure the *transfer difficulty* (TD) as the relative difference between the accuracy of the *All* and *Head* learnable parameter configurations for a non-private model: $TD = 100\left(Acc_{All} - Acc_{Head}\right)/Acc_{All}$. This simple metric captures how different the downstream dataset is from the pretraining data as well as other factors that complicate transfer learning such as the number of classes C in the downstream dataset and its size $|\mathcal{D}|$. If transfer learning is easy (i.e. TD is low), then only adapting the head of the network is sufficient. If transfer learning is more difficult (i.e. TD is high), then the backbone must also be adapted. Table 1 provides the TD 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. We use *add/remove* adjacency, where two datasets are adjacent if one can be obtained from the other by adding or removing one data record, which could be a single datapoint in case of example-level privacy or data belonging to a single user in case of user-level privacy.The additive error is typically chosen such that $\delta < 1/|\mathcal{D}|$. We refer to Dwork & Roth (2014) for a thorough introduction to DP.

DP-SGD (Rajkumar & Agarwal, 2012; Song et al., 2013; Abadi et al., 2016) adapts stochastic gradient descent (SGD) to guarantee DP. DP-SGD selects mini-batches using Poisson sampling, clips the ℓ_2 norm of per-example gradients, and adds isotropic Gaussian noise to the sum of mini-batch gradients. The level of privacy ((ϵ, δ)-DP) is controlled by the noise multiplier σ^2 which scales the variance of the added noise, the number of steps, and the sampling ratio (the Poisson sampling probability, i.e., expected batch size/$|\mathcal{D}|$).

Membership Inference Attacks (MIAs) MIAs aim to determine if a particular example was used in the training set of a model (Shokri et al., 2017). MIAs can be used to derive lower bounds to complement the theoretical upper bounds of (ϵ, δ)-DP for trained models. While there are many types of MIA (Hu et al., 2022b), in this work we consider attacks that operate in the black-box mode (i.e. only model outputs can be observed) and can evaluate the loss on particular training or test examples (Carlini et al., 2022; Ye et al., 2022). In addition, we assume that attacks have access to images from the training data distribution and know the training algorithm used and its hyperparameters. To evaluate the effectiveness of a MIA, we examine the Receiver Operating Characteristic (ROC) curve which plots the attack true positive rate (TPR) against its false positive rate (FPR). We focus on the TPR at low FPR regime since a MIA is harmful if it can infer membership of even a small number of training examples with high confidence (Carlini et al., 2022).

3 Related Work

DP Transfer Learning Section 1 describes various works where DP transfer learning using models pretrained on large public datasets achieve accuracy close to non-private approaches. However, to the best of our knowledge, there are no comprehensive studies on few-shot transfer learning under DP. The closest work to ours is Luo et al. (2021) where the authors evaluate DP fine-tuning of a sparse subset of the parameters of models pretrained on public data on a small number of few-shot downstream datasets. Their work employs a relatively small backbone (ResNet18), pretrained on a small public dataset (miniImageNet), with limited analysis. In contrast, our work utilizes large backbones, a large public pretraining set, a wider range of privacy levels and downstream datasets, in addition to assessing vulnerability to attacks and the federated learning setting. Tramèr et al. (2022) point out that current DP benchmarks rely excessively on downstream datasets with a high level of overlap with the pretraining data. Our work addresses this issue by evaluating on datasets with a wide range of TD.

Federated Learning (FL) and Transfer Learning There has been a recent surge of interest in using large pretrained models as initialization for training 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 partially mitigate the former issue, while to fully tackle it as well as 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., 2017), which provides user-level DP guarantees by applying the Gaussian mechanism to parameter updates sent to the server. Recently, a few studies have investigated 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 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 transfer-learned image classification. In this work we aim to fill this gap and evaluate these methods on real-world datasets.

4 Centralized Learning Experiments

In our experiments, we endeavor to answer the question: "Under what conditions is differentially private few-shot image classification effective?" We focus on transfer learning approaches that utilize large backbones pretrained on public data. We do this empirically by varying the: (i) number of shots S; (ii) set of learnable parameters in f (*All*, *Head*, *FiLM*); (iii) downstream dataset $\mathcal{D}$ (with varying TD); and (iv) network architecture: BiT-M-R50x1 (R-50) (Kolesnikov et al., 2020) with 23.5M parameters, Vision Transformer VIT-Base-16 (VIT-B) (Dosovitskiy et al., 2021) with 85.8M parameters, both pretrained on the ImageNet-21K dataset. In all experiments, we assume that the pretraining data is public and the downstream data is private. Source code for all experiments is included in the supplementary material.

Datasets For the experiments where S is varied, we use the CIFAR-10 (low TD) and CIFAR-100 (medium TD) datasets (Krizhevsky, 2009) which are commonly used in DP transfer learning, and SVHN (Netzer et al., 2011) which has a high transfer difficulty and hence requires a greater degree of adaptation of the pretrained backbone. We also evaluate on the challenging VTAB-1k transfer learning benchmark (Zhai et al., 2019) that consists of 19 datasets grouped into three distinct categories (natural, specialized, and structured) with training set size fixed at $|\mathcal{D}| = 1000$ and widely varying TD.

Training Protocol For all centralized experiments, we first draw $\mathcal{D}$ of the required size ($|\mathcal{D}| = CS$ (i.e. the number of classes C multiplied by shot S) for varying shot or $|\mathcal{D}| = 1000$ for VTAB-1k) from the entire training split of the current dataset under evaluation. For the purposes of hyperparameter tuning, we then split $\mathcal{D}$ into 70% train and 30% validation. We then perform 20 iterations of Bayesian optimization based hyperparameter tuning (Bergstra et al., 2011) with Optuna Akiba et al. (2019) to derive a set of hyperparameters that yield the highest accuracy on the validation data. This set of parameters is subsequently used to train a final model on all of $\mathcal{D}$. We evaluate the final, tuned model on the entire test split of the current dataset. Details on the set of hyperparameters that are tuned and their ranges can be found in Appendix A.3.2. For DP fine-tuning on $\mathcal{D}$, we use Opacus (Yousefpour et al., 2021) and compute the required noise multiplier depending on the targeted (ϵ, δ). We report the results over three runs. We set $\delta = 1/|\mathcal{D}|$ and report (ϵ, δ)-DP computed with the RDP accountant (Mironov, 2017). Similarly to previous work (De et al., 2022; Mehta et al., 2022; Sander et al., 2022) we do not account for privacy loss originating from the tuning of the hyperparameters. See Appendix A.3 for additional training details.

4.1 Few-shot DP Data Requirements

Fig. 1 depicts the performance of transfer learning under DP when varying S, ϵ, and TD. Tabular results can be found in Tables 2 to 7. We see that accuracy decreases as S and ϵ decrease and as TD increases. For $S \leq 10$, accuracy is poor under DP. However, if the TD is low or medium, a moderate number of shots ($S \approx 100$) is sufficient to approach the accuracy of the non-private setting. For example, at $S = 100$, the model achieves better than 90% accuracy on CIFAR-10 using only 2% of the full training split at $\epsilon = 1$. On the other hand, if TD is high, learning is more challenging and more shots are required to approach non-private accuracy. For example, for $S = 100$ and $\epsilon = 2$, SVHN achieves just over 20% accuracy and falls well short of non-private levels even at $S = 500$.

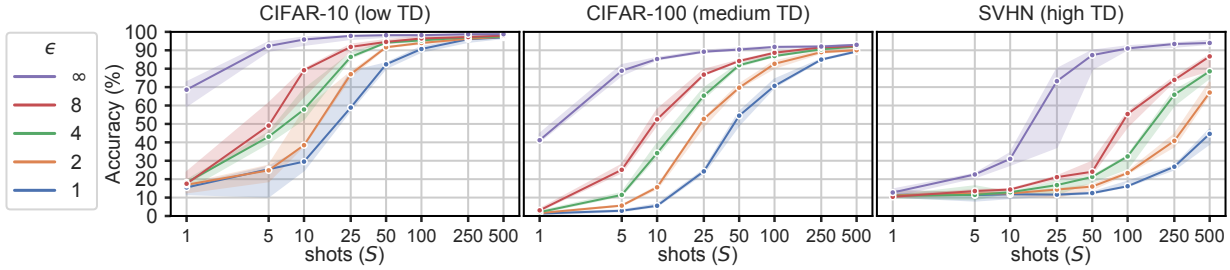

Figure 1: Classification accuracy as a function of shots and ϵ for CIFAR-10, CIFAR-100 and SVHN. Backbone is VIT-B and the best performing configuration out of *All*, *FiLM* and *Head* is used for each combination of ϵ and S, with $\delta = 1/|\mathcal{D}|$. The accuracy is reported over three seeds with the line showing the median and the band reporting the lowest and highest accuracy. **Analysis: Classification accuracy decreases as S and ϵ decrease and TD increases.**

Fig. 2 shows the multiplier on the number of DP shots to match non-private accuracy (see Appendix A.2.2 for additional figures and details). On the left we average over $S \in \{5, 10\}$, datasets, and network architectures. For all configurations, at $\epsilon = 8$, S must be increased by approximately $4 - 8\times$ to meet non-private accuracy and $20 - 35\times$ at $\epsilon = 1$. In effect, as the privacy level increases, the required multiplier increases in an exponential manner. The multipliers are lower for simpler forms of adaptation (e.g. *Head* requires $20 \times S$ at $\epsilon = 1$) than for more complex forms (e.g. *All* requires $35 \times S$ at $\epsilon = 1$). On the right we average over $S \in \{5, 10\}$, network architectures, and learnable parameters. Even though high TD datasets require more data for good accuracy, the multiplier values are similar and independent of the TD of the dataset (around $30\times$ at $\epsilon = 1$ and $6\times$ at $\epsilon = 8$). **Fig. 3 shows the classification accuracy as a function of TD at $S = 100$. The accuracy gap between non-private and private training increases as TD increases.**

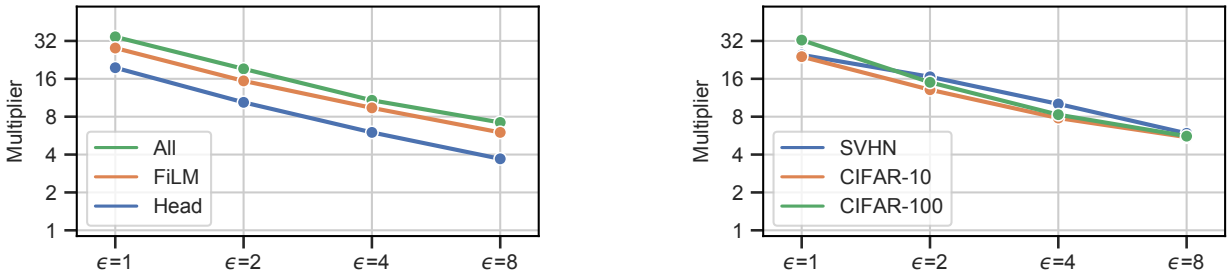

Figure 2: Multiplier of shots required to reach non-private accuracy. Left: Average over $S \in \{5, 10\}$, datasets, and network architectures. Right: Average over $S \in \{5, 10\}$, network architectures, and learnable parameters. The data is obtained using linear interpolation. See Appendix A.2.2. $\delta = 1/|\mathcal{D}|$. **Analysis: S must be increased by approximately $20 - 35\times$ to meet non-private accuracy at $\epsilon = 1$ and $4 - 8\times$ at $\epsilon = 8$**

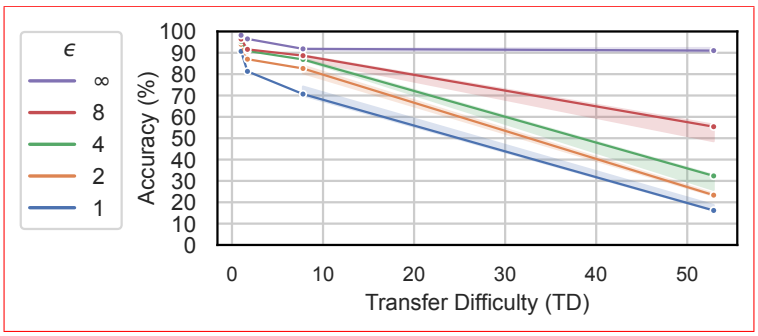

Figure 3: **Classification accuracy as a function of transfer difficulty (TD) and ϵ for CIFAR-10 (TD$=1.0$), EuroSAT (TD$=1.7$), CIFAR-100 (TD$=7.8$) and SVHN (TD$=52.9$) at $S = 100$. EuroSAT has been chosen because the result for $S = 100$ can be easily taken from the VTAB results (Tables 8 to 13) due to $C = 10$. Backbone is VIT-B and the best performing configuration out of *All*, *FiLM* and *Head* is used for each ϵ, with $\delta = 1/|\mathcal{D}|$. The accuracy is reported over three seeds with the line showing the median and the band reporting the lowest and highest accuracy. Analysis: The accuracy gap between non-private and private training increases as TD increases.**

4.2 Characterization of Learning Under Few-Shot DP

In this section, we provide empirical evidence to highlight the different traits of private and non-private learning. Fig. 4 shows snapshots at $\epsilon \in \{1, 8, \infty\}$ of the train and test accuracies as a function of S for CIFAR-100 (medium TD) and SVHN (high TD) (see Figs. 14 and 15 for versions with additional values of ϵ). The three snapshots for each dataset can be viewed as discrete points on a continuum from low to high ϵ. We see that learning under DP is fundamentally different from non-private. Non-private models with sufficient capacity operate in the interpolating regime and attain close to 100% training accuracy at all values of S, but have substantially lower test accuracy when S is low. In contrast, models that are are trained with DP-SGD are learning under heavy regularization and thus the training and test accuracies are significantly lower, but similar in value. When S is low, test accuracy is relatively poor and as S increases, test accuracy steadily improves. The point at which accuracy begins to improve varies with TD (CIFAR-100 test accuracy improvement starts much earlier than for SVHN). Independent of S, for low ϵ, the train-test gap is very small, with the train accuracy indicative of the test performance. As ϵ increases, the train accuracy grows as the amount of regularization pressure from DP is reduced, ultimately entering the interpolating regime. For SVHN with $\epsilon = \infty$, *Head* leaves the interpolating regime for $S > 100$, as there is not enough capacity to adapt to a high TD dataset. As ϵ increases and S remains low, the test accuracy does not increase as quickly as the train accuracy and the accuracy gap grows. However, as S increases, test accuracy starts to catch up with train accuracy, reducing the gap. Fig. 21 shows the train and test accuracies for all 19 VTAB datasets as a function of ϵ, where the general trends noted in Fig. 4 are also evident. While the results of Section 4.1

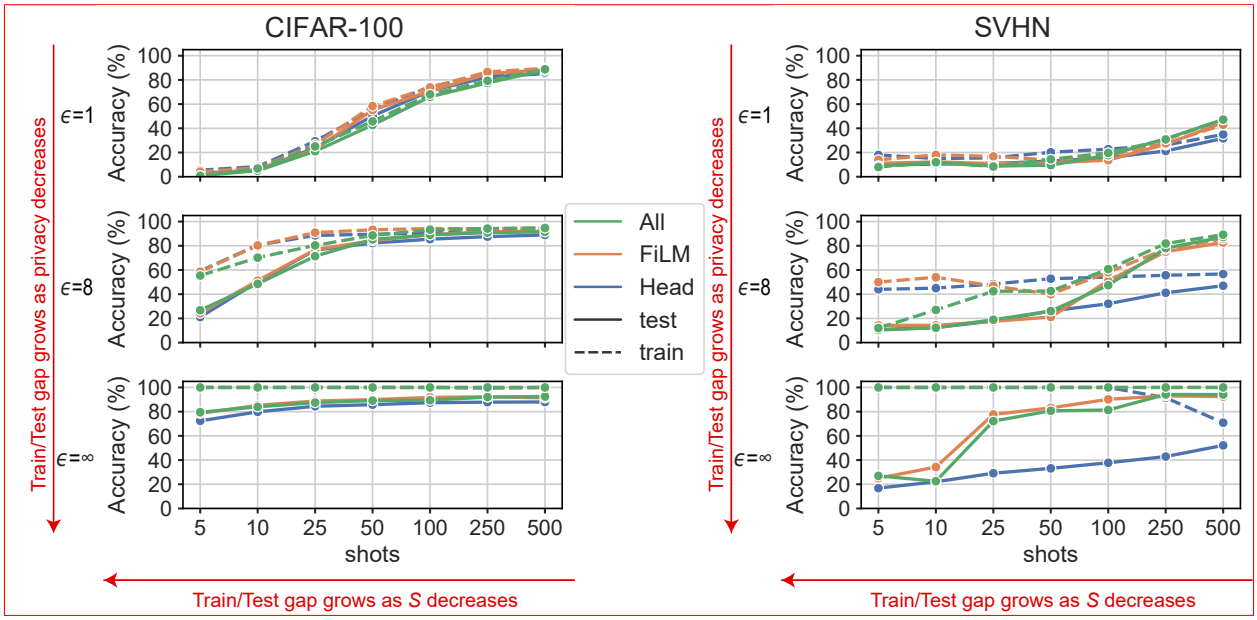

Figure 4: Snapshots at $\epsilon \in \{1, 8, \infty\}$, $\delta = 1/|\mathcal{D}|$ of the train/test accuracies as a function of S for CIFAR-100 and SVHN. The trends in accuracy gap are shown with red arrows. At low S, the gap grows as ϵ increases, at high S gap decreases to nearly zero, and the gap grows as S decreases. **Analysis: In the non-private setting ($\epsilon = \infty$), learning operates in the interpolation mode (i.e. train accuracy is 100%, yet accuracy continues to increase as S increases). As the privacy level increases, learning operates learn under heavy regularization and the gap between train and test accuracy reduces.**

indicate that both private and non-private test accuracy benefit from additional training data, it is evident that their learning behavior is significantly different.

4.3 Few-shot DP Model Parameterization

Fig. 5 depicts classification accuracy as a function of S, two different values of ϵ, and learnable parameters. *FiLM* is comparable to or better than *All* and *Head* in terms of accuracy despite fine-tuning fewer than 0.05%

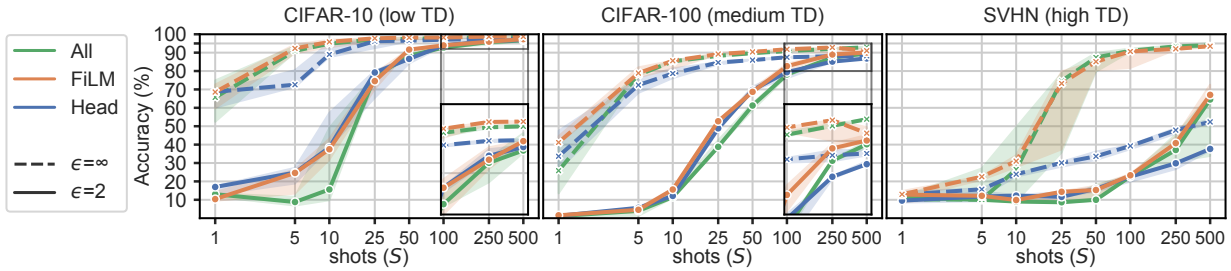

Figure 5: Classification accuracy as a function of shots and learnable parameters on VIT-B for CIFAR-10, CIFAR-100 and SVHN for $\epsilon \in \{2, \infty\}$ with $\delta = 1/|\mathcal{D}|$. The accuracy is reported over three seeds with the line showing the median and the band reporting the lowest and highest accuracy. **Analysis: *FiLM* is comparable to or better than *All* and *Head* in terms of accuracy despite fine-tuning fewer than 0.05% of the parameters in the backbone.**

of the parameters in the backbone. When the TD is low, training only *Head* is competitive with *FiLM* and *All*, but when TD is medium or high, *Head* falls short as it cannot adapt the backbone to a dataset that has a different data distribution. These observations have two implications: (i) *FiLM* is able to adapt to differing downstream datasets under DP and serves as a computationally efficient alternative to *All*; (ii) The result provides empirical support for the observations of Li et al. (2022a) that the number of parameters has little effect on the privacy utility trade-off when fine-tuning large pretrained models. Prior theory (Chaudhuri et al., 2011; Bassily et al., 2014) suggested that *All* should perform worse under DP compared to configurations with fewer parameters.

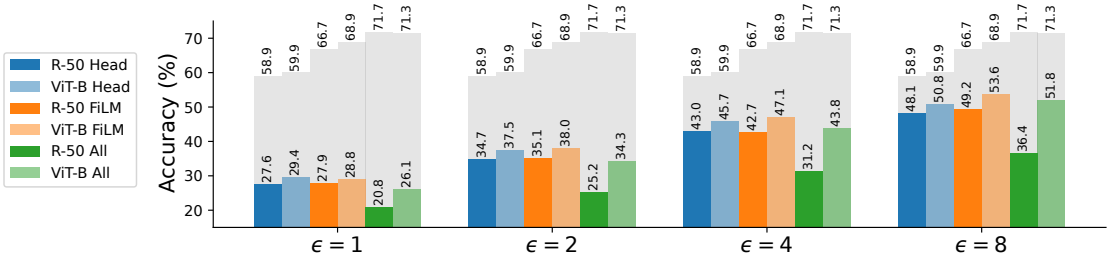

Figure 6: Average classification accuracy over all VTAB-1k datasets as a function of backbone, learnable parameters, and privacy level (ϵ) at $\delta = 10^{-3}$. Colored columns indicate results under DP, light gray indicates non-private accuracy for the corresponding configuration. **Analysis: DP classification accuracy (colored columns) decreases significantly as ϵ is decreased and always falls short of non-private accuracy (gray columns). For non-private settings, the *All* learnable parameters setting outperforms *FiLM* which outperforms *Head*. In contrast, for DP settings, *All* performs worst, *FiLM* and *Head* perform similarly, though *FiLM* is better in the majority of cases.**

Fig. 6 shows average classification accuracy over all of the datasets in the VTAB-1k benchmark (tabular results are in Tables 8 to 13 and comprehensive graphical results are in Figs. 19 and 20). We see that DP classification accuracy decreases significantly as ϵ is decreased and always falls short of non-private accuracy. For non-private settings, the *All* learnable parameters setting outperforms *FiLM* which outperforms *Head*. In contrast, for DP settings, *All* performs worst, *FiLM* and *Head* perform similarly, though *FiLM* is better in the majority of cases. One explanation for this is that under DP at low S, *All* requires more data compared to *Head* and *FiLM* for accuracy to progress beyond random chance as can be seen in Fig. 5.

Fig. 7 shows the difference between the accuracy of *FiLM* and *Head* for VTAB-1k datasets as a function of ϵ. The datasets are ordered from low to high TD (see Table 1). At $\epsilon = 1$, *Head* has an advantage over *FiLM* on several datasets. *FiLM* shows a significant advantage when the TD increases and as ϵ increases. Refer to Appendix A.2.5 for additional heat maps.

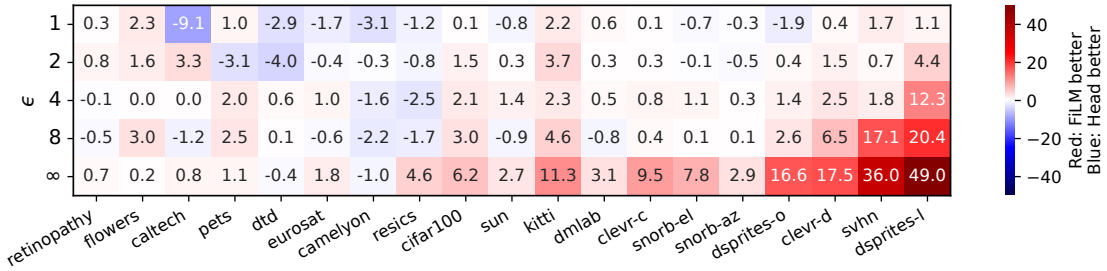

Figure 7: Heat map showing the accuracy difference between *FiLM* and *Head* for the VTAB-1k datasets as a function of ϵ. Backbone is VIT-B. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from low to high TD. $\delta = 10^{-3}$. **Analysis: At $\epsilon = 1$, *Head* has an advantage over *FiLM* on several datasets. *FiLM* shows a significant advantage when the TD increases and as ϵ increases.**

4.4 Membership Inference Attacks

We use the state-of-the-art Likelihood Ratio Attack (LiRA) (Carlini et al., 2022) to attack models trained on CIFAR-100 with varying S and privacy level ϵ using 256 shadow models. Refer to Appendix A.3.5 for additional detail. Excerpts from attack results are shown in Fig. 8. The complete set of attack ROC curves are shown in Figs. 22 and 23, while Table 14 reports TPR at several low FPR values, AUC score, and the maximum membership inference advantage (defined as TPR - FPR by Yeom et al. (2018)) achieved over the curve. Key observations are:

- Non-private ($\epsilon = \infty$) models are extremely vulnerable to MIAs (see Fig. 8, middle). For example, in the case of $\epsilon = \infty$, $S = 10$, *Head* configuration, 84.5% of the examples can be successfully identified with a false positive rate of only 0.1%.
- Vulnerability of non-private ($\epsilon = \infty$) models decreases as S increases. Also, the *FiLM* configuration is consistently less vulnerable than *Head* (see Fig. 8, middle). We hypothesize that *FiLM* generalizes better, so training examples do not stand out as much as in the *Head* configuration.
- When S is fixed, vulnerability to MIAs greatly decreases with decreasing ϵ (see Fig. 8, right). However, when $S = 10$ with $\epsilon = 2$ and *FiLM*, 6.5% of the examples can be successfully identified with a FPR of 1% and 0.8% of the examples with 0.1% FPR (see Table 14), which may still be unacceptable for some applications.
- Under DP, there appears to be little or no difference between the vulnerability of the *FiLM* and *Head* configurations at the same ϵ (see Fig. 8, right).
- Under DP with small ϵ, the vulnerability to MIA decreases as S increases (see Fig. 8, left).

5 Federated Learning Experiments

In this section, we investigate how imposing user-level DP influences the performance of large pretrained models fine-tuned via federated aggregation. In our evaluation, we use FLAIR Song et al. (2022), which is a recently proposed real-world dataset for multi-label image classification. It has around 50k users (overall around 400k images) 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 low to medium TD dataset. As in (Song et al., 2022), δ is set to $N^{-1.1}$, where $N = 41131$ is the number of training clients, and $\epsilon = 2$. We also perform experiments on CIFAR-100 and Federated EMNIST, which have many fewer training users, but are widely used for benchmarking federated learning. Those results are in Appendix A.2.8.

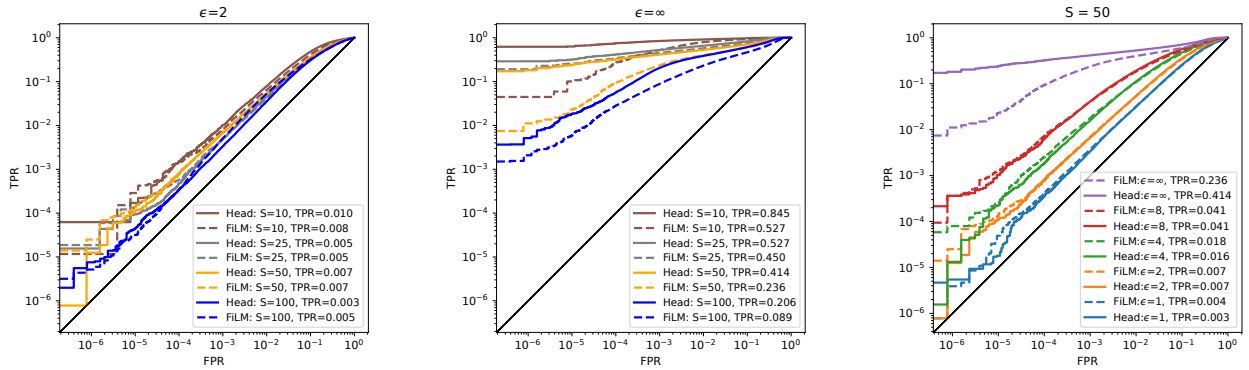

Figure 8: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for two values of ϵ (2 and ∞) where S varies and for $S = 50$ where ϵ varies. TPR values in legends are measured at FPR=0.001. Complete results in Table 14 and Figs. 22 and 23. $\delta = 1/(100S)$. **Analysis: Middle - Non-private models are extremely vulnerable to MIAs. For $\epsilon = \infty$, $S = 10$, *Head* configuration, 84.5% of the examples can be successfully identified with FPR = 0.1%. Also, vulnerability decreases as S increases. Right: Increasing the privacy level reduces the vunerability of the model as expectd, however when $S = 10$ with $\epsilon = 2$ and *FiLM*, 6.5% of the examples can be successfully identified with a FPR of 1%.**

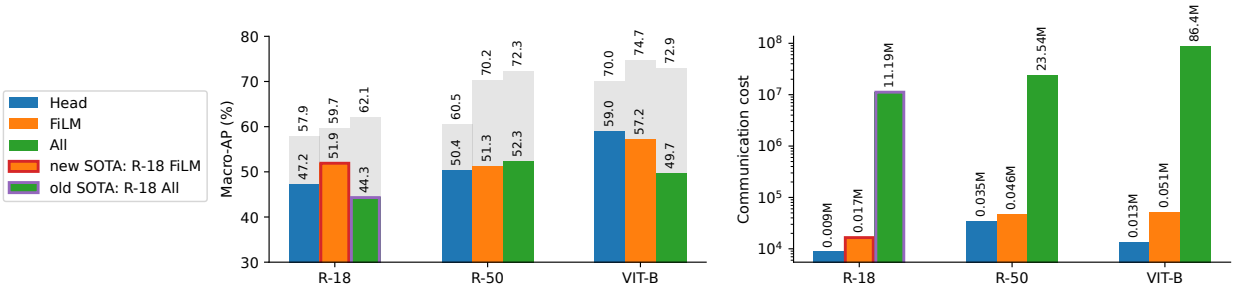

Figure 9: Left: Private (colored) and non-private (gray) FL performance on FLAIR as a function of backbone and learnable parameters. $\epsilon = 2, \delta = 41131^{-1.1}$. We use Macro-AP as the primary metric to report accuracy for FLAIR. The R-18 *All* result on FLAIR is taken from Song et al. (2022). Our FLAIR results set a new state-of-the-art. Right: FLAIR communication cost – the number of parameters sent at every user-server communication round. **Analysis: We set a new state-of-the-art result on the FLAIR benchmark by using *FiLM* on R-18 by increasing Macro-AP from 47.2% to 51.9% and drasticly reducing the communication cost from 11.9M parameters per round to 17K parameters. We further improve those results by using bigger backbones (R-50, VIT-B) with corresponding decreases in communication cost.**

As in the centralized experiments, we use R-50 and VIT-B, both pretrained on ImageNet-21K. We also perform experiments on a smaller architecture, ResNet18 (R-18) (He et al., 2016) pretrained on ImageNet-1K with 11.2M parameters, as it was initially used to achieve SOTA results on FLAIR.

For FL experiments, user-level DP is considered. We use FedADAM (Reddi et al., 2021) aggregation, which was shown to have better empirical performance than standard FedAvg (McMahan et al., 2017). We do not use Bayesian optimization for hyperparameter tuning, as each FL run is prohibitively expensive. Instead, we perform a small grid search over the server and client learning rates. Refer to Appendix A.3.6 for the hyperparameter ranges searched. For fair comparison on FLAIR, we fixed the other training hyperparameters to the values in the original paper Song et al. (2022).

Fig. 9 (left) shows the performance of different model configurations on FLAIR with (color) and without (grey) DP. We report macro average precision (Macro-AP) results here, while additional metrics are shown in Tables 15 and 16. As communication cost is important in FL, in Fig. 9 (right) we report the number of

parameters required to be transmitted for each model configuration in one user-server interaction. Summarizing Fig. 9, key observations are:

- With R-18 as 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 on state-of-the-art result on FLAIR in both DP and non-private settings. Under DP, the Macro-AP increases to 59%, while for non-private, the Macro-AP increases 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% for any model configuration.
- Although *FiLM* is outperforming *Head* and *All* on R-18, it is not always the case for other backbones. However, taking into account both test performance and communication cost, we can clearly see that either *FiLM* or *Head* is preferred. *FiLM* performs better for smaller backbones (R-18 and R-50), while *Head* is slightly better for VIT-B.

6 Discussion and Recommendations

Our work shows that DP few-shot learning works surprisingly well in the low TD setting, while the high TD setting is more difficult. Alternative strategies may include side-stepping privacy costs by leveraging the zero-shot capabilities of large pretrained models such as CLIP (Radford et al., 2021) or utilizing public data in addition to private data in the fine-tuning process as well (Golatkar et al., 2022) in order to improve utility. In summary, our experiments show that:

- **How much additional data is required under few-shot DP?** Image classification accuracy decreases as ϵ and S decrease, and as TD increases. As a result, one should expect to use roughly $4-8\times$ larger S for $\epsilon = 8$ and $20-35\times$ larger S for $\epsilon = 1$ under DP to achieve accuracy comparable to non-private. The multipliers are surprisingly similar across different TD levels.
- **Transfer learning dynamics under DP are fundamentally different from non-private** Non-private models with sufficient capacity operate in the interpolating regime and attain close to 100% training accuracy at all values of S, but have substantially lower test accuracy when S is low. In contrast, models that are are trained with DP-SGD are learning under heavy regularization and thus the training and test accuracies are significantly lower, but similar in value.
- **Parameter-efficient *FiLM* adapters perform well under DP** *FiLM* is comparable to or better than *All* and *Head* in terms of accuracy, demonstrating its ability to adapt to differing downstream datasets despite fine-tuning fewer than 0.05% of the parameters in the backbone. When the TD is easy, *Head* is competitive with *FiLM* and *All*, but when TD is difficult, *Head* falls short as it cannot adapt the backbone to a downstream dataset that has a different data distribution. *FiLM* is also effective in the DP FL setting, achieving state-of-the-art accuracy on the FLAIR benchmark while reducing communication cost by orders of magnitude.
- **Non-private Few-Shot Models Are Particularly Vulnerable to MIAs** The vulnerability of non-private few-shot models increases as S decreases. DP significantly mitigates the effectiveness of MIAs, however, we found that DP few-shot models can expose 6.5% of the examples with a 1% FPR even when $\epsilon = 2$ (on CIFAR-100 with S=10, *FiLM* on R-50).

Limitations We identify the following limitations in this work: (i) We focused exclusively on few-shot transfer learning from relatively large pretrained models and did not consider meta-learning approaches or training from scratch. (ii) We used FiLM adapters exclusively and did not consider other parameter-efficient adapters. Based on our experience, adapters do not have a large effect on the overall trends that we observed (in comparison to the items that we did vary) and making a fair comparison on a reasonable set of adapters would have exceeded our computational resources. (iii) The Transfer Difficulty (TD) metric is not ideal as it

depends on the network architecture and training hyperparameters, but in practice it aligns extremely well with the empirical difficulty of adapting to a downstream dataset. (iv) We always set $\delta = 1/\mathcal{D}$. While this is standard practice, in some experiments in the paper where $\mathcal{D}$ varies, fair comparisons of results can be difficult and in that case an alternative would have been to choose a small constant value of δ. (v) For each experiment, we used hyperparameter tuning to set a constant learning rate, but this learning rate was not annealed as is often done in non-private training.

Broader Impact Statement Few-shot learning systems hold much positive potential – from personalizing object recognizers for people who are blind (Massiceti et al., 2021) to rendering personalized avatars (Zakharov et al., 2019) (see (Hospedales et al., 2020) for a full review). These systems, however, also have the potential to be used in adverse ways – for example, in few-shot recognition in military/surveillance applications. We demonstrate how to execute highly successful membership inference attacks against few-shot leaning models which could be employed in a harmful manner. However, we also show how to effectively mitigate such attacks by training with DP.

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 U.S. census bureau adopts differential privacy. In Yike Guo and Faisal Farooq (eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 2867. ACM, 2018. doi: 10.1145/3219819.3226070. URL https://doi.org/10.1145/3219819.3226070.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (eds.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 2623–2631. ACM, 2019. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17455–17466, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/91cff01af640a24e7f9f7a5ab407889f-Abstract.html.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pp. 1138–1156. IEEE, 2022. doi: 10.1109/SP46214.2022.9833677. URL https://doi.org/10.1109/SP46214.2022.9833677.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 464–473. IEEE Computer Society, 2014. doi: 10.1109/FOCS.2014.56. URL https://doi.org/10.1109/FOCS.2014.56.

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

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *ArXiv preprint*, abs/1612.03801, 2016. URL https://arxiv.org/abs/1612.03801.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 2546–2554, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

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

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In Michael Bailey and Rachel Greenstadt (eds.), *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

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.

Yannis Cattan, Christopher A. Choquette-Choo, Nicolas Papernot, and Abhradeep Thakurta. Fine-tuning with differential privacy necessitates an additional hyperparameter search. *CoRR*, abs/2210.02156, 2022. doi: 10.48550/arXiv.2210.02156. URL https://doi.org/10.48550/arXiv.2210.02156.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011. doi: 10.5555/1953048.2021036. URL https://dl.acm.org/doi/10.5555/1953048.2021036.

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*, abs/2205.13535, 2022c. URL https://arxiv.org/abs/2205.13535.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein (eds.), *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pp. 1655–1658. ACM, 2018. doi: 10.1145/3183713.3197390. URL https://doi.org/10.1145/3183713.3197390.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *CoRR*, abs/2204.13650, 2022. doi: 10.48550/arXiv.2204.13650. URL https://doi.org/10.48550/arXiv.2204.13650.

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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3571–3580, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/253614bbac999b38b5b60cae531c4969-Abstract.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8366–8376. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00819. URL https://doi.org/10.1109/CVPR52688.2022.00819.

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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *ArXiv preprint*, abs/2004.05439, 2020. URL https://arxiv.org/abs/2004.05439.

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 Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

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. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022b.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 7232–7241, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3b3fff6463464959dcd1b68d0320f781-Abstract.html.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41. URL https://doi.org/10.1007/978-3-031-19827-4_41.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *ArXiv preprint*, abs/2207.07039, 2022. URL https://arxiv.org/abs/2207.07039.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.215. URL https://doi.org/10.1109/CVPR.2017.215.

Kaggle and EyePacs. Kaggle diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection/data, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pp. 491–507. Springer, 2020. doi: 10.1007/978-3-030-58558-7_29. URL https://doi.org/10.1007/978-3-030-58558-7_29.

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

Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward Training at ImageNet Scale with Differential Privacy. *ArXiv preprint*, abs/2201.12328, 2022. URL https://arxiv.org/abs/2201.12328.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pp. II–104. IEEE, 2004.

Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In William W. Cohen and Andrew W. Moore (eds.), *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pp. 577–584. ACM, 2006. doi: 10.1145/1143844.1143917. URL https://doi.org/10.1145/1143844.1143917.

Xuechen Li, Daogao Liu, Tatsunori Hashimoto, Huseyin A. Inan, Janardhan Kulkarni, Yin Tat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? *CoRR*, abs/2207.00160, 2022a. doi: 10.48550/arXiv.2207.00160. URL https://doi.org/10.48550/arXiv.2207.00160.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=bVuP3ltATMz.

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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.13. 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.

Zelun Luo, Daniel J. Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5059–5068. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00502. URL https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Scalable_Differential_Privacy_With_Sparse_Network_Finetuning_CVPR_2021_paper.html.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 1022–1035, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/081be9fdff07f3bc808f935906ef70c0-Abstract.html.

Daniela Massiceti, Luisa M. 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10798–10808. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01064. URL https://doi.org/10.1109/ICCV48922.2021.01064.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

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.

Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *CoRR*, abs/2205.02973, 2022. doi: 10.48550/arXiv.2205.02973. URL https://doi.org/10.48550/arXiv.2205.02973.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pp. 263–275. IEEE Computer Society, 2017. doi: 10.1109/CSF.2017.11. URL https://doi.org/10.1109/CSF.2017.11.

Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, and Andrew G. Howard. K for the price of 1: Parameter-efficient multi-task and transfer learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJxvEh0cFQ.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

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*, abs/2206.09843, 2022. URL https://arxiv.org/abs/2206.09843.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16528.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL `http://proceedings.mlr.press/v139/radford21a.html`.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 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.

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. TAN without a burn: Scaling laws of DP-SGD. *CoRR*, abs/2210.03403, 2022. doi: 10.48550/arXiv.2210.03403. URL `https://doi.org/10.48550/arXiv.2210.03403`.

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.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Aliaksandra Shysheya, John Bronskill, Massimiliano Patacchiola, Sebastian Nowozin, and Richard E. Turner. Fit: Parameter efficient few-shot transfer learning for personalized and federated image classification. *ArXiv preprint*, abs/2206.08671, 2022. URL `https://arxiv.org/abs/2206.08671`.

Congzheng Song, Filip Granqvist, and Kunal Talwar. Flair: Federated learning annotated image repository. *ArXiv preprint*, abs/2207.08869, 2022. URL `https://arxiv.org/abs/2207.08869`.

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 pre-trained 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), 2022. ISSN 2157-6904. doi: 10.1145/3510033. URL https://doi.org/10.1145/3510033.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *CoRR*, abs/2212.06470, 2022. doi: 10.48550/arXiv.2212.06470. URL https://doi.org/10.48550/arXiv.2212.06470.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. Pretrained models for multilingual federated learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1413–1421, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.101. URL https://aclanthology.org/2022.naacl-main.101.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pp. 3485–3492. IEEE Computer Society, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

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.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 3093–3106, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3560675. URL https://doi.org/10.1145/3548606.3560675.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018*, pp. 268–282. IEEE Computer Society, 2018.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3320–3328, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Gosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. *ArXiv preprint*, abs/2109.12298, 2021. URL https://arxiv.org/abs/2109.12298.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9458–9467. IEEE, 2019. doi: 10.1109/ICCV.2019.00955. URL https://doi.org/10.1109/ICCV.2019.00955.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *ArXiv preprint*, abs/1910.04867, 2019. URL `https://arxiv.org/abs/1910.04867`.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *ArXiv preprint*, abs/2206.04673, 2022. URL `https://arxiv.org/abs/2206.04673`.

A Appendix

A.1 Transfer Difficulty

Table 1 shows the transfer difficulty for each of the datasets used in our experiments.

Table 1: Transfer highy for the 19 VTAB-1k datasets (plus CIFAR-10). The Score column is computed as the difference between the 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 lower the difficulty of the transfer. In the TD column, we map the score into three buckets: a score of 0-5 is low, 5-10 is medium, and greater than 10 is high. To compute the scores, we use the VIT-B backbone and use accuracies from the VTAB-1k experiments for the VTAB-1k datasets and the accuracies from the results of the effect of shots and DP experiments at $S = 100$ for CIFAR-10.

Dataset	**Score**	**TD**
Caltech101 (Fei-Fei et al., 2006)	0.4	low
CIFAR10 (Krizhevsky, 2009)	1.0	low
CIFAR100 (Krizhevsky, 2009)	7.8	medium
Flowers102 (Nilsback & Zisserman, 2008)	0.2	low
Pets (Parkhi et al., 2012)	1.1	low
Sun397 (Xiao et al., 2010)	8.8	medium
SVHN (Netzer et al., 2011)	52.9	high
DTD (Cimpoi et al., 2014)	1.3	low
EuroSAT (Helber et al., 2019)	1.7	low
Resics45 (Cheng et al., 2017)	6.7	medium
Patch Camelyon (Veeling et al., 2018)	3.8	low
Retinopathy (Kaggle & EyePacs, 2015)	0.2	low
CLEVR-count (Johnson et al., 2017)	26.2	high
CLEVR-dist (Johnson et al., 2017)	38.7	high
dSprites-loc (Matthey et al., 2017)	71.4	high
dSprites-ori (Matthey et al., 2017)	37.7	high
SmallNORB-azi (LeCun et al., 2004)	33.3	high
SmallNORB-elev (LeCun et al., 2004)	28.2	high
DMLab (Beattie et al., 2016)	21.9	high
KITTI-dist (Geiger et al., 2013)	13.6	high

A.2 Additional Results

A.2.1 Additional Effect of Shots per Class and ϵ Results

Tables 2 to 7 depict tabular results for different backbones (R-50, VIT-B), different learnable parameter sets (*Head*, *FiLM*, *All*), different numbers of shots per class ($S = 1, 5, 10, 25, 50, 100, 250, 500$) and various privacy levels ($\epsilon = 1, 2, 4, 8, \infty$), all at $\delta = 1/|\mathcal{D}|$).

Table 2: Classification accuracy as a function of ϵ, S and learnable parameters for CIFAR-10. Backbone is R-50 pretrained on ImageNet-21k. Accuracy figures are percentages and the ± sign indicates the 95% confidence interval over 3 runs with different seeds.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
All	$\epsilon = 1$	10.0±0.0	9.8±0.3	9.9±0.2	29.3±4.3	54.1±7.4	74.3±1.2	86.1±1.7	90.5±0.7
	$\epsilon = 2$	9.8±0.3	14.0±7.9	9.8±0.2	50.2±3.5	71.2±2.3	83.3±1.9	90.4±0.6	91.6±0.1
	$\epsilon = 4$	10.0±0.1	9.8±0.3	9.8±0.3	69.2±4.1	81.9±3.1	88.2±1.0	91.5±0.3	92.0±0.3
	$\epsilon = 8$	10.0±0.0	20.1±19.7	28.0±35.9	53.0±26.2	85.9±2.0	90.1±0.8	91.8±0.5	92.8±0.8
	$\epsilon = \infty$	51.3±6.0	78.7±3.0	86.3±0.7	89.3±1.1	91.9±0.6	93.5±0.3	94.6±0.4	95.5±0.3
FiLM	$\epsilon = 1$	11.1±0.6	18.8±3.5	21.4±4.4	45.6±3.4	68.3±3.9	78.3±4.6	89.6±0.9	92.4±0.8
	$\epsilon = 2$	13.3±3.3	21.0±5.6	31.7±6.6	63.4±3.0	80.9±2.4	86.9±0.9	92.3±0.4	93.6±0.7
	$\epsilon = 4$	14.9±4.3	30.3±8.5	52.4±3.4	76.1±1.5	86.2±1.7	89.1±1.2	93.4±0.0	94.4±0.3
	$\epsilon = 8$	17.1±5.3	40.1±5.0	67.0±5.8	80.9±1.2	89.2±1.7	92.4±0.7	94.0±0.3	94.7±0.3
	$\epsilon = \infty$	48.1±3.6	79.4±6.3	86.5±2.6	91.7±0.4	94.1±0.3	94.9±0.2	95.5±0.2	95.8±0.2
Head	$\epsilon = 1$	11.7±5.0	18.7±1.6	20.5±8.3	44.7±4.7	68.7±3.3	82.0±1.6	87.3±0.2	88.9±0.8
	$\epsilon = 2$	11.9±3.8	23.7±1.5	31.1±8.4	62.2±8.8	80.1±1.2	86.2±0.6	89.3±0.4	90.0±0.7
	$\epsilon = 4$	13.1±2.6	30.0±5.2	45.5±12.1	79.4±1.8	84.7±0.7	87.7±0.5	89.8±0.2	90.8±0.4
	$\epsilon = 8$	16.8±4.4	41.5±6.5	65.5±3.1	83.2±1.7	86.2±0.7	89.1±0.3	90.1±0.3	91.2±0.1
	$\epsilon = \infty$	49.2±4.7	75.2±6.0	83.5±0.4	87.2±0.6	89.0±0.2	90.3±0.2	91.4±0.2	92.1±0.3

Table 3: Classification accuracy as a function of ϵ, S and learnable parameters for CIFAR-100. Backbone is R-50 pretrained on ImageNet-21k. Accuracy figures are percentages and the ± sign indicates the 95% confidence interval over 3 runs with different seeds.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
All	$\epsilon = 1$	1.0±0.0	1.7±0.8	1.7±1.4	7.9±1.5	23.9±0.7	48.5±2.1	50.6±0.3	62.4±5.2
	$\epsilon = 2$	1.3±0.6	1.2±0.6	6.8±0.2	23.1±0.6	46.0±2.1	59.4±2.9	61.9±6.0	65.7±5.2
	$\epsilon = 4$	1.0±0.0	1.6±1.3	15.3±1.1	43.7±3.9	57.6±2.8	64.8±0.4	63.0±0.9	68.1±0.8
	$\epsilon = 8$	1.0±0.0	5.1±7.9	31.1±2.0	55.9±2.6	59.6±3.7	67.3±0.3	67.6±0.8	74.1±2.3
	$\epsilon = \infty$	28.3±2.4	51.9±5.6	59.7±9.9	71.9±0.5	76.0±0.8	79.9±0.1	82.2±3.0	85.8±2.1
FiLM	$\epsilon = 1$	1.0±0.4	1.8±0.5	3.7±0.2	14.2±0.2	34.2±1.8	59.2±1.0	75.3±0.7	80.1±0.8
	$\epsilon = 2$	1.4±0.5	2.9±0.3	9.3±0.6	33.4±0.8	55.0±3.2	72.2±0.3	80.2±0.4	82.8±0.5
	$\epsilon = 4$	1.7±0.4	7.2±1.3	22.4±1.2	53.9±0.9	70.1±0.6	77.9±0.6	82.3±0.4	84.2±0.5
	$\epsilon = 8$	2.2±0.3	18.6±0.3	38.7±3.7	65.3±1.6	75.6±0.5	80.5±0.8	83.5±0.5	85.2±0.4
	$\epsilon = \infty$	25.8±6.1	64.8±3.2	71.8±1.1	77.6±0.1	81.4±0.4	82.9±0.4	83.8±0.8	83.4±1.2
Head	$\epsilon = 1$	1.3±0.6	2.3±0.9	4.4±1.2	13.5±1.5	32.3±0.9	52.1±0.7	66.9±0.8	71.9±0.7
	$\epsilon = 2$	1.4±0.7	3.9±0.8	8.7±1.4	31.6±1.2	51.0±1.2	63.1±1.0	71.6±0.4	75.2±0.3
	$\epsilon = 4$	1.8±0.6	7.5±1.8	22.1±1.3	46.3±0.7	61.7±1.6	68.4±0.8	74.4±0.3	77.0±0.1
	$\epsilon = 8$	2.5±0.7	17.2±1.7	38.9±1.9	58.9±0.3	67.5±0.4	72.1±0.5	76.3±0.1	78.4±0.3
	$\epsilon = \infty$	25.6±5.6	56.4±0.3	62.8±1.4	69.6±0.2	72.9±0.3	75.8±0.3	78.5±0.2	78.8±1.5

Table 4: Classification accuracy as a function of ϵ, S and learnable parameters for SVHN. Backbone is R-50 pretrained on ImageNet-21k. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

		1S	**5S**	**10S**	**25S**	**50S**	**100S**	**250S**	**500S**
	$\epsilon = 1$	7.1±2.0	12.9±6.3	8.8±1.2	9.0±1.2	8.8±2.3	20.5±3.6	25.7±0.4	32.1±9.5
	$\epsilon = 2$	7.2±2.0	9.0±1.6	9.4±1.9	8.6±0.7	12.3±6.3	23.1±4.9	35.5±2.9	46.1±5.0
All	$\epsilon = 4$	7.2±2.0	7.3±2.3	7.6±1.7	7.8±1.8	9.5±1.4	27.8±3.7	42.9±6.7	65.1±6.9
	$\epsilon = 8$	7.2±2.0	8.5±2.3	8.7±1.1	9.0±0.6	22.4±8.9	39.6±3.1	60.9±5.9	78.1±3.1
	$\epsilon = \infty$	14.4±1.5	19.6±12.8	42.2±4.1	76.1±4.5	84.1±5.5	86.8±5.0	93.1±0.5	94.6±0.2
	$\epsilon = 1$	12.6±3.6	9.5±0.6	11.4±2.8	12.3±1.4	13.9±1.6	18.6±2.1	25.1±1.1	33.8±2.5
	$\epsilon = 2$	11.3±4.0	10.8±1.3	12.0±1.6	15.0±1.8	16.3±0.5	21.7±1.4	32.9±3.5	45.3±2.9
FiLM	$\epsilon = 4$	9.3±0.4	11.4±2.0	13.5±1.3	17.0±0.8	20.7±1.8	27.6±2.1	39.8±3.4	61.1±2.6
	$\epsilon = 8$	9.9±1.0	11.9±1.6	16.6±1.4	21.2±1.8	25.6±1.9	33.3±2.1	51.4±3.9	67.3±2.0
	$\epsilon = \infty$	13.8±0.3	20.2±1.1	25.7±1.7	37.8±4.1	42.4±1.8	58.7±6.2	71.5±5.5	84.4±3.1
	$\epsilon = 1$	10.0±0.9	8.7±1.0	10.8±0.1	11.1±1.2	13.3±0.9	18.1±1.4	24.7±1.0	31.2±0.7
	$\epsilon = 2$	9.9±0.5	9.1±1.4	11.2±1.4	13.9±1.6	17.1±1.1	21.2±2.0	29.6±1.5	36.1±1.7
Head	$\epsilon = 4$	10.3±0.8	10.2±0.8	13.5±2.1	17.4±0.2	19.3±0.7	24.9±1.1	35.3±1.2	40.9±1.0
	$\epsilon = 8$	10.3±0.8	11.1±1.3	15.0±3.1	19.9±1.8	23.3±1.3	29.3±1.0	39.7±1.5	45.2±1.2
	$\epsilon = \infty$	13.8±0.5	18.5±1.6	21.2±3.4	28.7±1.4	32.8±1.6	38.0±1.4	47.1±0.5	48.5±3.2

Table 5: Classification accuracy as a function of ϵ, S and learnable parameters for CIFAR-10. Backbone is VIT-B pretrained on ImageNet-21k. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

		1S	**5S**	**10S**	**25S**	**50S**	**100S**	**250S**	**500S**
	$\epsilon = 1$	12.5±1.3	20.9±10.2	18.8±7.4	64.9±13.2	70.3±15.1	84.2±10.9	93.0±2.0	95.3±1.0
	$\epsilon = 2$	12.7±1.0	9.1±2.7	24.8±24.0	76.9±1.4	88.4±1.3	93.1±0.9	95.4±1.0	97.0±0.9
All	$\epsilon = 4$	12.6±1.4	42.7±3.2	59.2±5.2	86.8±1.5	91.8±1.0	95.3±0.7	96.9±0.7	97.6±0.5
	$\epsilon = 8$	12.7±1.7	51.7±9.6	54.9±10.7	86.2±10.5	90.9±4.9	96.6±0.8	97.3±0.3	98.1±0.3
	$\epsilon = \infty$	64.3±12.8	91.4±1.6	95.1±0.8	97.2±0.3	97.6±0.4	97.9±0.3	98.3±0.1	98.4±0.1
	$\epsilon = 1$	10.3±3.1	15.0±3.9	23.8±1.7	57.7±8.7	81.9±1.6	89.6±1.5	95.5±1.0	96.9±0.2
	$\epsilon = 2$	11.4±2.9	21.5±8.0	37.5±6.4	74.5±4.8	91.7±0.2	93.5±1.5	96.1±0.9	97.3±0.1
FiLM	$\epsilon = 4$	13.3±2.0	37.7±5.4	58.6±5.9	82.8±5.5	93.1±1.0	94.4±1.2	96.9±0.4	97.5±0.1
	$\epsilon = 8$	16.4±1.1	51.4±10.9	71.5±2.0	89.4±4.4	94.6±1.1	96.0±1.0	97.1±0.1	97.6±0.6
	$\epsilon = \infty$	67.0±7.4	92.1±2.7	95.3±2.4	97.2±1.1	97.9±0.6	98.0±0.4	98.6±0.1	98.7±0.1
	$\epsilon = 1$	14.6±2.7	19.1±5.0	30.3±6.7	56.6±2.1	81.5±1.7	90.6±1.5	95.3±0.3	96.4±0.2
	$\epsilon = 2$	15.7±2.9	23.7±4.9	44.7±12.3	74.9±9.3	86.2±2.1	93.9±0.2	96.3±0.4	96.9±0.1
Head	$\epsilon = 4$	17.3±2.9	34.9±9.8	59.7±9.3	85.3±6.5	94.4±0.6	95.1±1.4	96.7±0.3	97.0±0.4
	$\epsilon = 8$	19.5±4.1	42.2±2.8	74.7±9.3	91.7±1.6	92.9±5.5	95.4±0.7	97.0±0.2	97.1±0.3
	$\epsilon = \infty$	66.0±5.7	74.8±5.6	90.7±4.7	95.1±2.6	96.5±0.4	97.0±0.1	97.3±0.1	97.4±0.1

Table 6: Classification accuracy as a function of ϵ, S and learnable parameters for CIFAR-100. Backbone is VIT-B pretrained on ImageNet-21k. Accuracy figures are percentages and the ± sign indicates the 95% confidence interval over 3 runs with different seeds.

		$\mathbf{1}S$	$\mathbf{5}S$	$\mathbf{10}S$	$\mathbf{25}S$	$\mathbf{50}S$	$\mathbf{100}S$	$\mathbf{250}S$	$\mathbf{500}S$
All	$\epsilon = 1$	1.1±0.3	1.0±0.3	3.4±2.1	18.7±3.3	41.0±2.0	62.7±2.0	80.2±4.9	85.7±2.9
	$\epsilon = 2$	1.1±0.1	3.2±1.9	11.9±1.4	39.9±2.5	60.8±2.2	78.0±0.8	87.3±0.1	89.5±0.6
	$\epsilon = 4$	0.9±0.2	10.5±2.1	24.3±2.2	56.4±3.5	68.5±15.8	82.9±5.3	89.0±1.3	90.2±1.2
	$\epsilon = 8$	1.3±0.4	17.3±7.7	18.8±2.9	60.8±18.8	84.2±0.3	86.9±2.4	90.3±0.3	91.2±0.2
	$\epsilon = \infty$	26.2±14.5	78.1±1.0	85.3±0.6	88.4±0.7	89.6±0.3	90.9±0.4	92.1±0.2	93.0±0.0
FiLM	$\epsilon = 1$	1.3±0.1	2.1±1.2	5.1±1.3	22.4±0.4	53.5±4.1	71.6±2.5	84.9±0.2	89.4±0.3
	$\epsilon = 2$	1.6±0.2	4.5±1.5	15.5±0.7	51.4±2.7	69.1±1.3	82.4±2.0	89.1±0.6	90.3±0.5
	$\epsilon = 4$	1.6±0.1	11.2±1.9	35.3±4.0	66.2±3.2	82.0±0.8	87.0±0.4	90.6±0.4	92.0±0.2
	$\epsilon = 8$	2.7±0.7	25.6±2.4	53.3±4.6	77.6±1.8	83.6±3.5	88.1±1.2	91.6±0.1	92.3±0.5
	$\epsilon = \infty$	42.1±2.5	79.1±3.1	84.2±2.8	89.4±0.5	90.6±0.5	91.6±0.4	91.9±1.8	90.9±1.3
Head	$\epsilon = 1$	1.3±0.4	2.8±0.3	5.7±0.7	24.2±0.9	49.0±3.2	70.9±0.3	82.1±0.5	85.1±0.3
	$\epsilon = 2$	1.4±0.2	5.7±0.1	12.7±2.4	48.7±1.1	70.2±1.3	78.5±2.4	85.1±0.2	87.0±0.3
	$\epsilon = 4$	2.2±0.4	11.7±0.9	29.5±5.0	65.7±3.9	76.0±5.0	83.7±0.6	86.9±0.4	88.1±0.3
	$\epsilon = 8$	3.2±0.5	21.7±2.0	51.9±4.6	74.5±4.2	82.0±0.3	85.4±0.5	87.2±1.3	88.5±0.9
	$\epsilon = \infty$	35.8±12.2	72.2±4.5	78.7±3.0	84.3±0.8	86.1±0.3	87.4±0.4	88.0±0.8	88.4±0.4

Table 7: Classification accuracy as a function of ϵ, S and learnable parameters for SVHN. Backbone is VIT-B pretrained on ImageNet-21k. Accuracy figures are percentages and the ± sign indicates the 95% confidence interval over 3 runs with different seeds.

		$\mathbf{1}S$	$\mathbf{5}S$	$\mathbf{10}S$	$\mathbf{25}S$	$\mathbf{50}S$	$\mathbf{100}S$	$\mathbf{250}S$	$\mathbf{500}S$
All	$\epsilon = 1$	11.9±1.9	9.5±2.1	9.7±2.4	9.2±1.5	10.3±1.4	14.1±4.6	22.4±2.8	33.5±14.9
	$\epsilon = 2$	11.7±1.7	10.1±0.3	9.9±1.9	8.6±0.5	12.6±5.5	22.8±0.4	37.9±8.4	55.2±20.9
	$\epsilon = 4$	10.7±0.5	10.9±2.1	10.8±1.7	9.7±2.0	15.6±6.2	28.6±5.3	45.8±17.7	66.1±22.0
	$\epsilon = 8$	10.5±0.5	10.5±0.6	9.1±0.5	14.3±5.1	25.5±4.0	36.6±18.3	64.6±31.4	84.4±5.3
	$\epsilon = \infty$	10.5±1.2	15.6±11.2	28.4±22.9	63.0±26.9	86.1±5.0	91.2±1.0	93.0±1.1	94.2±1.0
FiLM	$\epsilon = 1$	11.7±3.0	10.7±2.4	11.1±1.4	10.0±1.2	11.4±2.4	17.0±1.7	26.4±1.1	43.7±4.5
	$\epsilon = 2$	11.6±2.8	12.2±0.6	10.3±0.9	13.1±2.8	14.4±4.2	23.5±0.6	41.0±3.6	68.6±4.0
	$\epsilon = 4$	10.3±2.2	11.1±2.2	12.5±2.2	15.8±4.0	20.5±1.9	30.3±4.4	64.8±4.6	77.5±2.3
	$\epsilon = 8$	9.1±1.6	13.1±1.4	14.4±2.1	20.9±0.9	23.4±2.1	53.4±4.8	74.3±1.2	83.7±0.5
	$\epsilon = \infty$	12.6±2.6	22.1±1.2	31.0±3.6	65.9±21.8	83.7±3.1	87.7±5.9	92.2±0.7	93.6±0.7
Head	$\epsilon = 1$	9.1±0.5	10.0±2.9	10.9±1.5	11.5±0.3	12.5±0.4	15.5±0.9	21.5±1.6	31.1±1.5
	$\epsilon = 2$	9.2±0.7	11.7±2.9	12.1±0.7	12.4±1.6	15.8±1.4	22.0±2.1	28.8±2.1	37.4±0.6
	$\epsilon = 4$	9.7±0.5	12.2±4.1	12.9±0.1	15.5±2.2	19.5±0.5	24.8±2.1	35.5±1.2	42.9±1.0
	$\epsilon = 8$	9.3±0.8	10.7±0.8	14.5±0.2	18.4±0.5	23.2±2.1	29.6±2.3	40.5±0.9	46.3±1.1
	$\epsilon = \infty$	12.7±2.0	14.6±2.8	23.5±0.9	29.6±1.6	33.8±2.6	38.8±1.3	47.6±1.7	52.6±1.1

A.2.2 Additional Shot Multiplier Variations

We compute the multiplier for a configuration and dataset at ϵ as follows: Using the median accuracy obtained through the experiments depicted in Tables 2 to 7 ($S = 1, 5, 10, 25, 50, 100, 250, 500$) we linearly interpolate the median accuracy in the complete $S = [1, 500]$ grid. We determine the minimum S required to reach at least the same accuracy as for non-private at $S \in \{5, 10\}$ using the $S = [1, 500]$ grid. The multiplier is then the minimum S required for DP divided by the S for non-private.

The Figs. 10 and 11 display the same analysis as Fig. 2 for all backbones (VIT-B, R-50) and non-private shots of $S \in \{5, 10\}$. The Figs. 12 and 13 display the same analysis grouped by datasets.

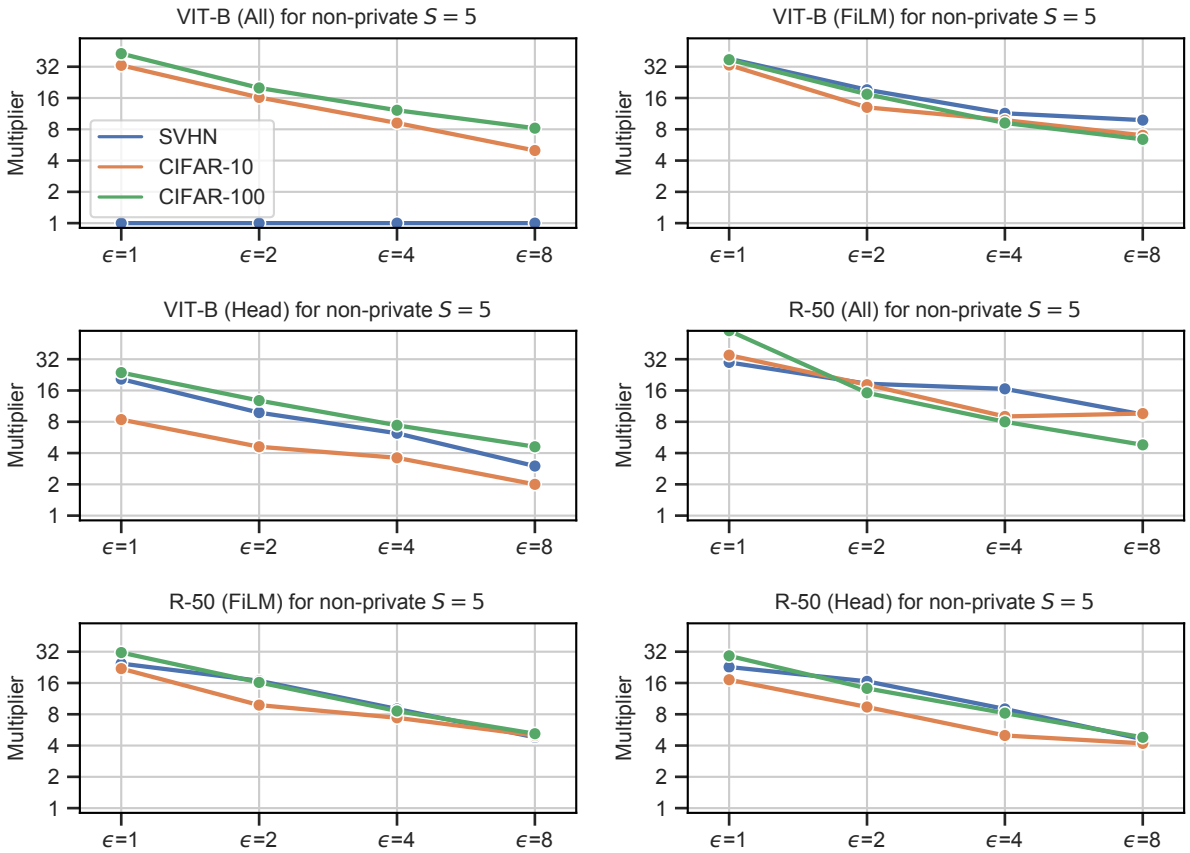

Figure 10: Multiplier of shots required to reach same accuracy as non-private with $S = 5$ for VIT-B and R-50 on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Appendix A.2.1. The multiplier is 1 for all ϵ for ViT-B with All parameters on SVHN at $S = 5$ (top left plot) because non-DP achieves a random accuracy and achieving random accuracy requires $S = 1$ for all configurations in the experiment.

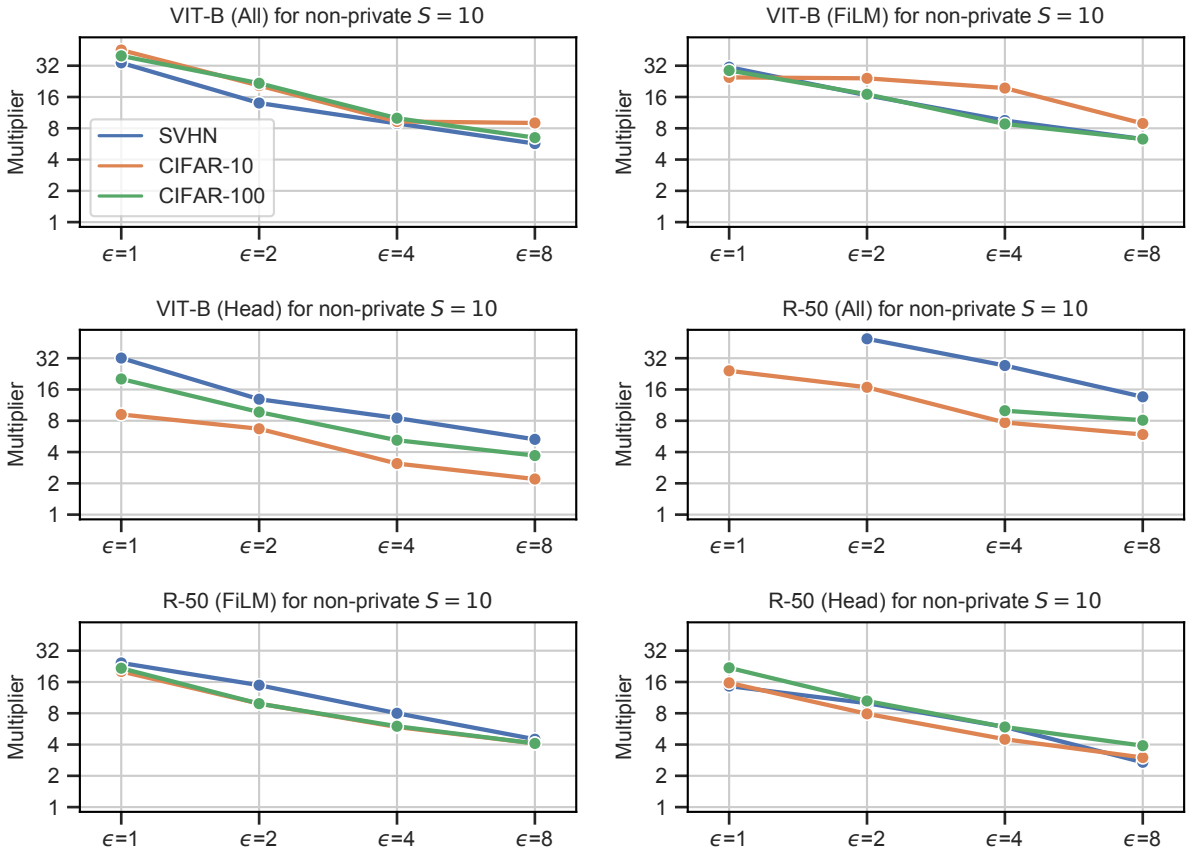

Figure 11: Multiplier of shots required to reach same accuracy as non-private with $S = 10$ for VIT-B and R-50 on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Appendix A.2.1.

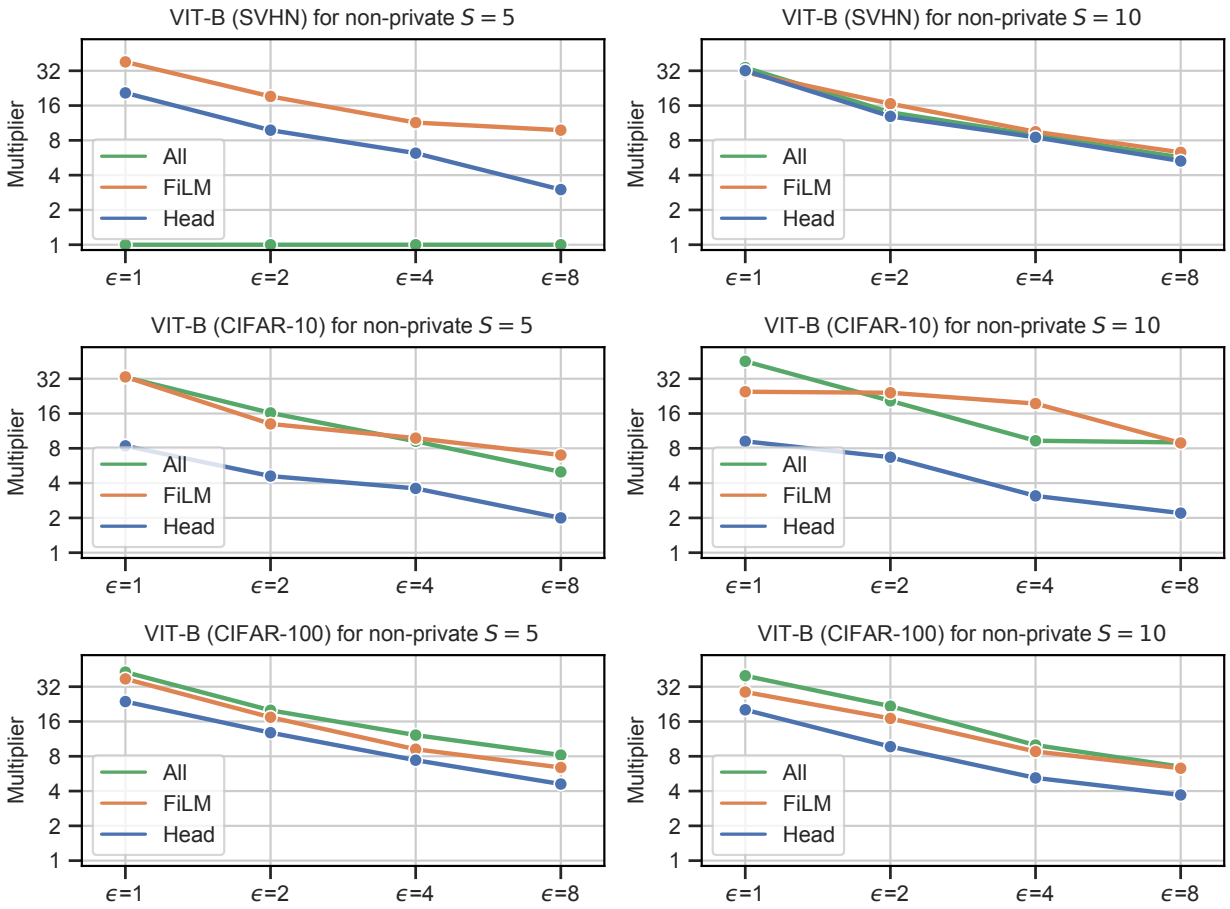

Figure 12: Multiplier of shots required to reach same accuracy as non-private with $S \in \{5, 10\}$ for VIT-B on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Appendix A.2.1. The multiplier is 1 for all ϵ for ViT-B with All parameters on SVHN at $S = 5$ (top left plot) because non-DP achieves a random accuracy and achieving random accuracy requires $S = 1$ for all configurations in the experiment.

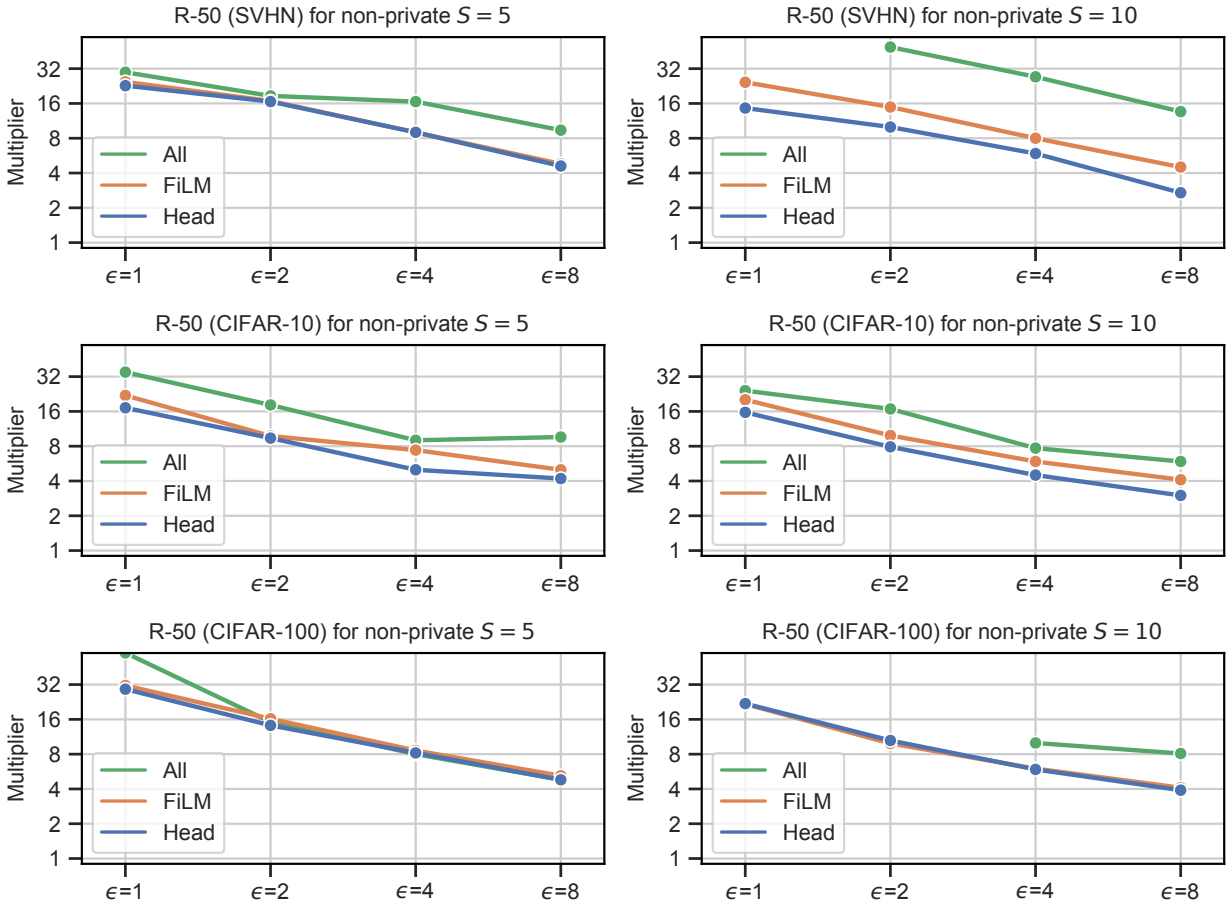

Figure 13: Multiplier of shots required to reach same accuracy as non-private with $S \in \{5, 10\}$ for R-50 on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Appendix A.2.1.

A.2.3 Additional versions of Fig. 4

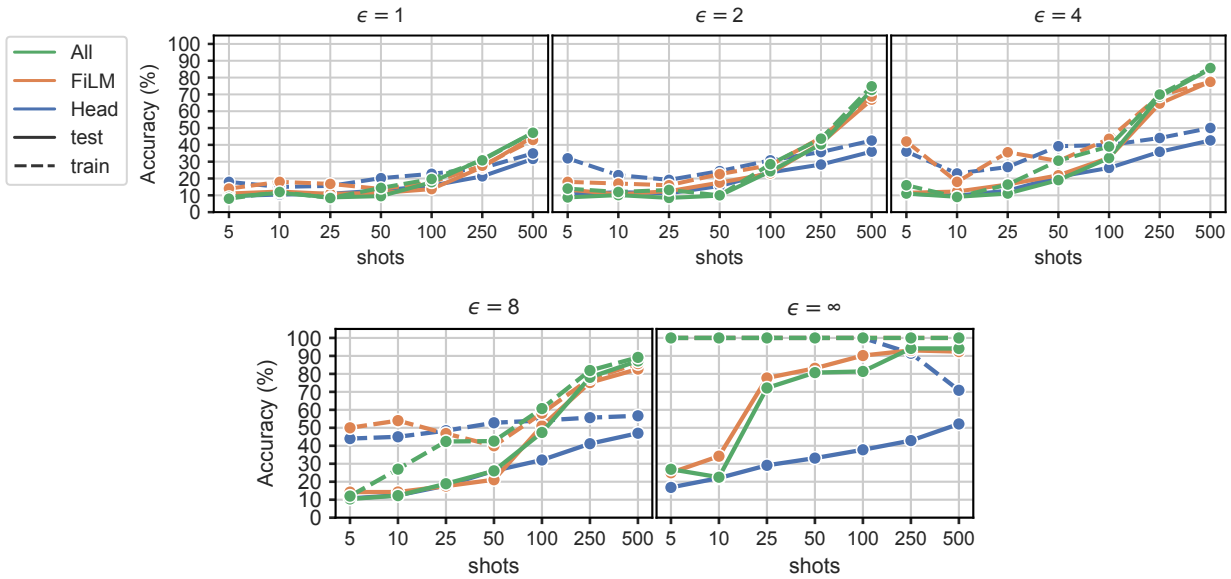

Figure 14: Test and train classification accuracy as a function of shots and learnable parameters (*All*, *FiLM* and *Head*) on VIT-B for SVHN for different ϵ with $\delta = 1/|\mathcal{D}|$. The accuracy is reported for the median runs of Table 7.

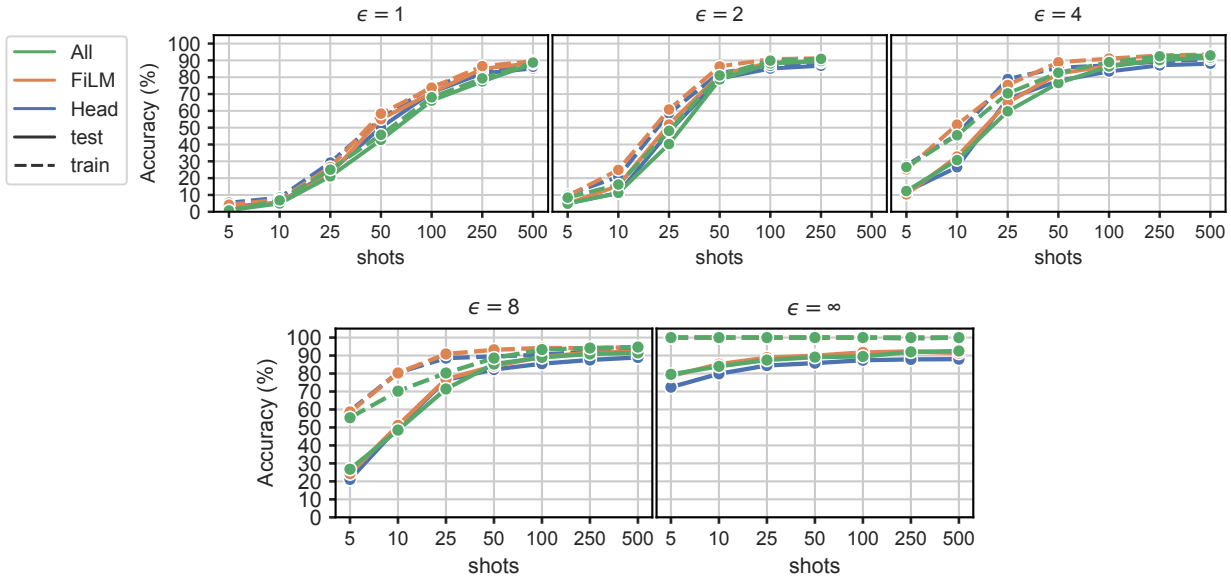

Figure 15: Test and train classification accuracy as a function of shots and learnable parameters (*All*, *FiLM* and *Head*) on VIT-B for CIFAR-100 for different ϵ with $\delta = 1/|\mathcal{D}|$. The accuracy is reported for the median runs of Table 6.

A.2.4 Comparison of Backbones for Effect of Shots and ϵ

Fig. 16 compares the backbones (VIT-B, R-50) using their best performing configuration. The VIT-B backbone achieves comparable or better performance.

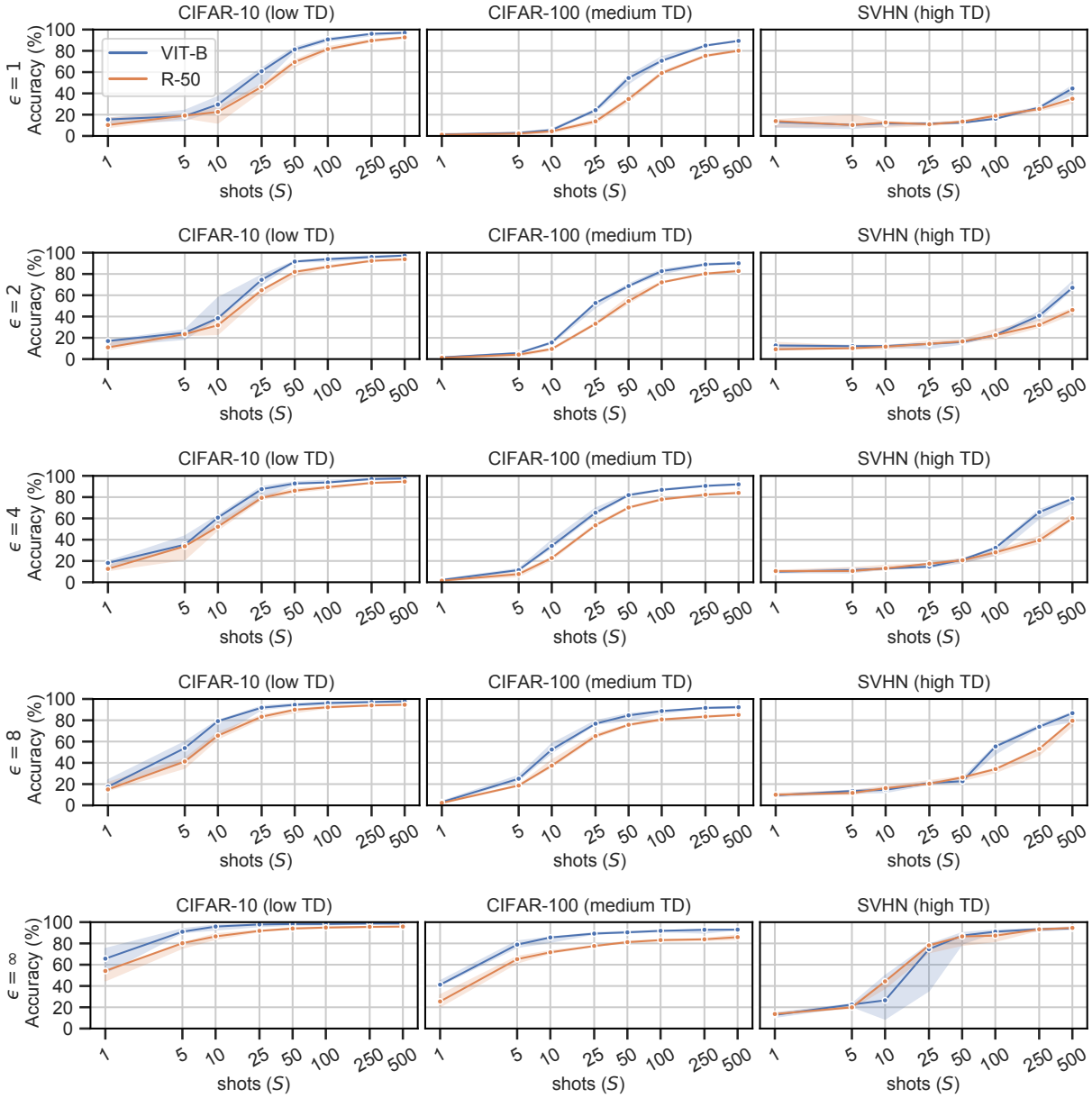

Figure 16: Classification accuracy for different ϵ as a function of S and backbone (VIT-B, R-50) for CIFAR-10, CIFAR-100 and SVHN. TD (low, medium, high) refers to the transfer difficulty and is computed as in Appendix A.1. The best performing configuration out of *All*, *FiLM* and *Head* for each combination of ϵ, S and backbone is used. The accuracy is reported over three seeds with the line showing the median and the band reporting the lowest and highest accuracy.

A.2.5 Advantage of *FiLM* as a Function of Shots

Figs. 17 and 18 show the difference between the mean classification accuracy of *FiLM* and *Head*. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better.

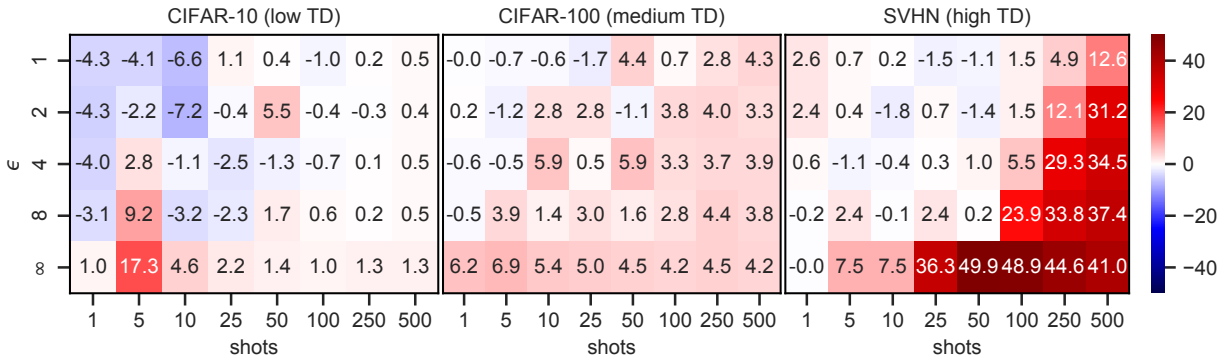

Figure 17: Heat map showing the accuracy advantage of *FiLM* over *Head* for CIFAR-10, CIFAR-100 and SVHN as a function of ϵ. Backbone is VIT-B. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from low to high TD.

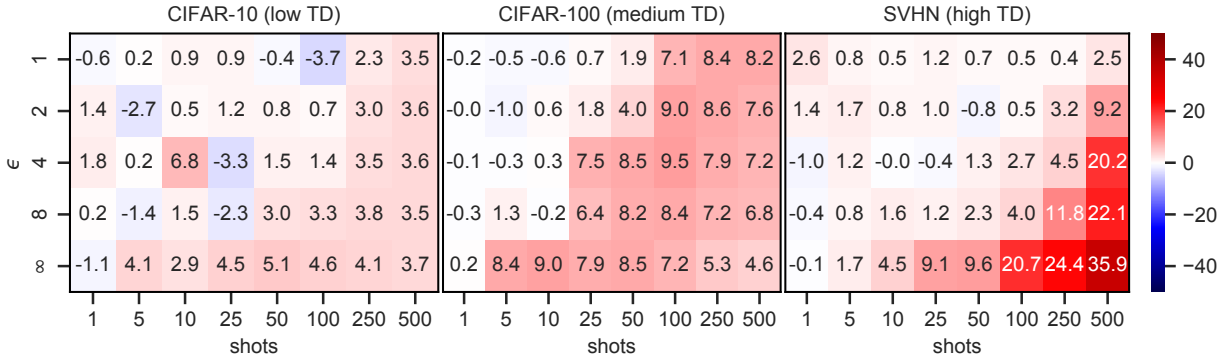

Figure 18: Heat map showing the accuracy advantage of *FiLM* over *Head* for CIFAR-10, CIFAR-100 and SVHN as a function of ϵ. Backbone is R-50. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from low to high TD.

A.2.6 Additional VTAB-1k Results

Tables 8 to 13 depict tabular results for different backbones (R-50, ViT-B), different learnable parameter sets (*Head*, *FiLM*, *All*), and various privacy levels ($\epsilon = 1, 2, 4, 8, \infty$), all at $\delta = 10^{-3}$.

Table 8: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is R-50 pretrained on ImageNet-21k. Learnable parameters are Head. Accuracy figures are percentages and the ± sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
Caltech101 (Fei-Fei et al., 2006)	102	11.8±6.9	30.0±4.7	57.1±3.6	69.3±1.4	87.9±0.2
CIFAR100 (Krizhevsky, 2009)	100	4.2±1.2	10.6±1.0	20.8±1.6	34.7±2.2	61.5±0.6
Flowers102 (Nilsback & Zisserman, 2008)	102	11.3±2.7	33±6.4	73.1±0.7	89.8±2.1	98.4±0.1
Pets (Parkhi et al., 2012)	37	28.6±5.4	50±2.4	65.6±1.5	73.4±1.0	84.4±0.3
Sun397 (Xiao et al., 2010)	397	4.7±0.2	8.4±0.2	13.4±1.2	21.5±0.7	46.2±0.4
SVHN (Netzer et al., 2011)	10	23.0±1.0	26.8±2.5	30.6±1.9	34.9±2.0	41.9±2.2
DTD (Cimpoi et al., 2014)	47	19.6±4.0	36.2±1.3	51.2±1.2	61.3±0.9	72.0±0.4
EuroSAT (Helber et al., 2019)	10	77.2±1.9	85.3±1.4	88.4±1	91.1±0.0	94.3±0.2
Resics45 (Cheng et al., 2017)	45	19.3±3.5	33.4±2.8	48.6±2.5	60.9±0.8	78.5±0.2
Patch Camelyon (Veeling et al., 2018)	2	77.6±2.3	79.2±1.0	80.8±1.6	80.6±0.3	81.2±0.2
Retinopathy (Kaggle & EyePacs, 2015)	5	73.2±0.6	73.7±0.3	73.4±0.6	74.0±0.4	75.2±0.1
CLEVR-count (Johnson et al., 2017)	8	27.5±1.5	30.1±1.9	33.6±1.5	36.8±2.1	51.2±1.2
CLEVR-dist (Johnson et al., 2017)	6	26.4±2.2	28.7±0.7	29.8±0.9	30.9±1.6	36.2±0.9
dSprites-loc (Matthey et al., 2017)	16	6.6±0.1	6.8±0.7	7.6±0.5	7.5±1.2	18.9±6.5
dSprites-ori (Matthey et al., 2017)	16	9.3±0.9	10.8±0.5	13.2±0.7	16.2±0.2	45.9±2.5
SmallNORB-azi (LeCun et al., 2004)	18	6.1±0.5	7.5±0.5	8.1±0.4	8.7±0.7	11.7±0.1
SmallNORB-elev (LeCun et al., 2004)	9	17±3.3	19.8±1.1	22.5±0.9	24.5±0.7	31±0.5
DMLab (Beattie et al., 2016)	6	26.9±0.8	28.8±0.6	31±0.2	32.4±0.1	34.6±3.7
KITTI-dist (Geiger et al., 2013)	4	54.5±2.5	59.6±3.5	66.9±1.6	65.4±1.3	69.2±0.7
All		27.6	34.7	43.0	48.1	58.9
Natural		14.7	27.9	44.5	55.0	70.4
Specialized		61.8	67.9	72.8	76.7	82.3
Structured		21.8	24.0	26.6	27.8	37.3

Table 9: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is R-50 pretrained on ImageNet-21k. Learnable backbone parameters are FiLM. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\boldsymbol{\epsilon = 1}$	$\boldsymbol{\epsilon = 2}$	$\boldsymbol{\epsilon = 4}$	$\boldsymbol{\epsilon = 8}$	$\boldsymbol{\epsilon = \infty}$
Caltech101 (Fei-Fei et al., 2006)	102	11.3±1.3	35.8±4.1	55.7±0.5	72.2±1.9	88.8±0.5
CIFAR100 (Krizhevsky, 2009)	100	3.4±0.8	10.2±0.5	23.2±0.8	38.5±1.3	71.7±1.3
Flowers102 (Nilsback & Zisserman, 2008)	102	10.4±0.6	34.2±5.4	70.4±1.0	89.3±0.3	98.7±0.1
Pets (Parkhi et al., 2012)	37	28.4±1.4	48.8±3.0	64.1±1.1	75±1.1	88±0.4
Sun397 (Xiao et al., 2010)	397	4.2±0.5	8.0±0.1	14.1±0.8	21.7±0.8	46.8±0.7
SVHN (Netzer et al., 2011)	10	23.3±1.4	28.1±0.9	32.9±0.7	38.6±3.2	56.6±2.4
DTD (Cimpoi et al., 2014)	47	20.8±2.7	36.7±1.5	50.3±4.5	61.4±1.3	72.4±0.4
EuroSAT (Helber et al., 2019)	10	79.2±0.8	85.1±1.5	88.8±2.2	92.1±0.7	95±0.1
Resics45 (Cheng et al., 2017)	45	21.1±1.7	35.2±0.6	49.5±1.6	61.3±0.8	81.9±0.1
Patch Camelyon (Veeling et al., 2018)	2	76.8±0.8	77.3±2.7	79.1±1.2	79.4±0.4	81.3±0.1
Retinopathy (Kaggle & EyePacs, 2015)	5	73.4±0.3	73.5±0.1	73.9±0.5	74.4±0.2	74.0±3.2
CLEVR-count (Johnson et al., 2017)	8	29.1±1.5	31.0±0.4	34.6±1.3	38±1.4	73±1.3
CLEVR-dist (Johnson et al., 2017)	6	26.7±1.3	28.7±0.7	30.5±0.6	31.8±0.6	49.3±1.6
dSprites-loc (Matthey et al., 2017)	16	6.6±0.3	6.4±0.3	6.7±0.5	8.5±1.4	64.0±8.7
dSprites-ori (Matthey et al., 2017)	16	9.1±2.0	11.2±1.7	12.3±1.0	16.7±0.8	56.8±3.8
SmallNORB-azi (LeCun et al., 2004)	18	6.4±0.4	7.3±0.8	8.2±0.8	9.4±0.5	14.6±0.2
SmallNORB-elev (LeCun et al., 2004)	9	17.6±1.0	20.8±0.3	22.7±1.1	25.6±0.3	32.0±4.0
DMLab (Beattie et al., 2016)	6	25.8±0.7	28.7±0.7	30.5±0.9	32.1±0.8	41.8±0.4
KITTI-dist (Geiger et al., 2013)	4	56.3±1.6	60.7±3.9	63.4±2.5	68.5±1.6	80.4±0.5
All		27.9	35.1	42.7	49.2	66.7
Natural		14.6	28.8	44.4	56.7	74.7
Specialized		62.6	67.8	72.8	76.8	83.1
Structured		22.2	24.3	26.2	28.8	51.5

Table 10: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is R-50 pretrained on ImageNet-21k. All parameters are learnable. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\boldsymbol{\epsilon = 1}$	$\boldsymbol{\epsilon = 2}$	$\boldsymbol{\epsilon = 4}$	$\boldsymbol{\epsilon = 8}$	$\boldsymbol{\epsilon = \infty}$
Caltech101 (Fei-Fei et al., 2006)	102	8.2±8.5	17.6±7.0	26.0±3.9	33.0±3.9	86.8±2.0
CIFAR100 (Krizhevsky, 2009)	100	1.0±0.1	2.3±1.5	6.9±2.9	13.3±2.7	59.3±7.0
Flowers102 (Nilsback & Zisserman, 2008)	102	6.2±2.8	6.9±7.1	33.7±11.6	69.7±7.3	95.8±1.8
Pets (Parkhi et al., 2012)	37	12.8±0.5	22.7±2.2	32.4±4.1	24.7±9.0	83.0±0.2
Sun397 (Xiao et al., 2010)	397	3.3±0.4	3.0±0.5	2.7±0.4	3.4±1.2	38.3±0.8
SVHN (Netzer et al., 2011)	10	19.2±0.8	23.6±5.5	26.8±7.0	37.2±2.3	88.5±2.5
DTD (Cimpoi et al., 2014)	47	13.7±3.6	21.6±2.3	27.7±3.8	34.0±1.9	72.4±0.1
EuroSAT (Helber et al., 2019)	10	49.8±9.5	69.2±5.5	72.7±1.2	82.4±3.4	96.0±1.0
Resics45 (Cheng et al., 2017)	45	11.9±1.4	13.8±6.4	24.3±1.0	26.8±6.8	84.1±1.1
Patch Camelyon (Veeling et al., 2018)	2	65.9±15.7	70.5±20.1	80.5±1.2	79.5±2.9	85.0±0.8
Retinopathy (Kaggle & EyePacs, 2015)	5	73.6±0.0	73.6±0.0	73.6±0.0	73.6±0.0	76.0±1.3
CLEVR-count (Johnson et al., 2017)	8	18.1±4.6	26.3±2.0	36.2±2.7	41.4±7.0	93.2±0.2
CLEVR-dist (Johnson et al., 2017)	6	23.8±1.4	22.7±2.2	25.2±1.7	36.9±3.3	62.1±1.7
dSprites-loc (Matthey et al., 2017)	16	6.2±0.1	6.4±0.3	6.3±0.1	6.2±0.1	89.1±3.7
dSprites-ori (Matthey et al., 2017)	16	7.5±0.0	6.6±1.8	7.2±0.1	8.8±2.9	61.0±5.2
SmallNORB-azi (LeCun et al., 2004)	18	5.4±0.2	5.7±0.1	5.7±0.3	6.2±0.8	21.9±3.3
SmallNORB-elev (LeCun et al., 2004)	9	12.3±1.2	13.8±2.4	21.2±2.7	22.7±5.5	39.5±6.3
DMLab (Beattie et al., 2016)	6	22.5±0.3	24.9±3.1	28.5±1.4	30.3±5.2	48.4±0.8
KITTI-dist (Geiger et al., 2013)	4	34.4±6.7	46.9±1.8	55.2±0.5	60.6±2.5	81.1±0.2
All		20.8	25.2	31.2	36.4	71.7
Natural		9.2	13.9	21.1	28.7	73.3
Specialized		50.3	56.8	61.1	65.6	85.3
Structured		16.3	19.2	23.2	26.6	62.0

Table 11: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is VIT-B pretrained on ImageNet-21k. Learnable backbone parameters are Head. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\boldsymbol{\epsilon = 1}$	$\boldsymbol{\epsilon = 2}$	$\boldsymbol{\epsilon = 4}$	$\boldsymbol{\epsilon = 8}$	$\boldsymbol{\epsilon = \infty}$
Caltech101 (Fei-Fei et al., 2006)	102	20.8±2.1	39.7±5.7	65.6±1.1	79.9±0.3	93.3±0.3
CIFAR100 (Krizhevsky, 2009)	100	7.0±1.4	15.9±2.7	33.3±1.5	49.9±2.3	77.6±2.4
Flowers102 (Nilsback & Zisserman, 2008)	102	13.7±3.0	47.2±1.5	85.4±1.8	93.5±2.6	99.3±0.3
Pets (Parkhi et al., 2012)	37	38.3±2.8	65.6±0.2	76.0±3.9	81.1±2.4	90.7±0.1
Sun397 (Xiao et al., 2010)	397	3.5±0.5	6.9±0.8	13.2±1.5	24.0±0.3	51.0±3.4
SVHN (Netzer et al., 2011)	10	23.3±1.4	27.2±1.5	31.6±1.2	35.3±0.3	43.1±0.4
DTD (Cimpoi et al., 2014)	47	20.4±2.6	37.0±3.1	49.9±4.0	61.6±3.2	75.7±0.3
EuroSAT (Helber et al., 2019)	10	81.3±1.3	87.0±1.0	89.9±0.9	91.6±1.1	94.6±0.4
Resics45 (Cheng et al., 2017)	45	23.2±2.8	41.4±2.1	58.0±2.7	67.9±2.1	82.5±0.5
Patch Camelyon (Veeling et al., 2018)	2	79.8±2.9	78.5±2.1	81.6±1.8	82.8±0.4	83.8±0.7
Retinopathy (Kaggle & EyePacs, 2015)	5	73.3±0.6	72.6±1.3	73.6±0.6	74.0±0.2	73.8±2.3
CLEVR-count (Johnson et al., 2017)	8	25.5±0.9	27.7±1.3	30.8±0.4	33.3±0.5	42.5±0.5
CLEVR-dist (Johnson et al., 2017)	6	26.1±0.7	27.5±0.5	30.1±0.3	31.5±0.5	35.1±0.3
dSprites-loc (Matthey et al., 2017)	16	6.9±0.5	7.8±0.7	8.7±0.1	9.4±0.6	19.1±2.7
dSprites-ori (Matthey et al., 2017)	16	11.2±0.9	13.3±1.2	15.6±0.9	18.9±1.5	31.2±0.6
SmallNORB-azi (LeCun et al., 2004)	18	6.9±0.6	7.8±0.5	8.1±1.3	9.0±0.8	12.2±0.1
SmallNORB-elev (LeCun et al., 2004)	9	16.9±1.4	19.3±0.4	20.4±1.1	22.7±1.8	27.5±0.4
DMLab (Beattie et al., 2016)	6	29.2±1.7	33.0±1.6	35.0±1.0	37.3±1	40.2±0.6
KITTI-dist (Geiger et al., 2013)	4	51.3±8.4	57.1±5.6	61.2±0.6	61.4±3.0	65.7±3.2
All		29.4	37.5	45.7	50.8	59.9
Natural		18.1	34.2	50.7	60.8	75.8
Specialized		64.4	69.9	75.8	79.1	83.7
Structured		21.7	24.2	26.2	27.9	34.2

Table 12: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is VIT-B pretrained on ImageNet-21k. Learnable backbone parameters are FiLM. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\boldsymbol{\epsilon = 1}$	$\boldsymbol{\epsilon = 2}$	$\boldsymbol{\epsilon = 4}$	$\boldsymbol{\epsilon = 8}$	$\boldsymbol{\epsilon = \infty}$
Caltech101 (Fei-Fei et al., 2006)	102	11.7±6.0	42.9±4.9	65.7±3.0	78.7±2.6	94.1±0.8
CIFAR100 (Krizhevsky, 2009)	100	7.1±0.2	17.4±1.2	35.4±2.0	52.9±2.0	83.8±0.6
Flowers102 (Nilsback & Zisserman, 2008)	102	16.0±2.8	48.8±5.2	85.3±1.7	96.4±0.7	99.5±0.0
Pets (Parkhi et al., 2012)	37	39.3±2.2	62.5±3.6	78.0±0.9	83.6±2.4	91.8±0.4
Sun397 (Xiao et al., 2010)	397	2.7±0.4	7.1±1.2	14.6±1.0	23.1±0.5	53.7±2.0
SVHN (Netzer et al., 2011)	10	25.1±1.5	28.0±1.1	33.4±0.7	52.4±7.4	79.1±2.6
DTD (Cimpoi et al., 2014)	47	17.5±2.0	33.0±3.5	50.5±1.4	61.7±1.1	75.3±3.9
EuroSAT (Helber et al., 2019)	10	79.6±1.8	86.6±2.2	90.9±0.2	91.0±0.5	96.5±0.2
Resics45 (Cheng et al., 2017)	45	22.0±3.1	40.6±2.3	55.5±3.9	66.1±1.7	87.0±0.5
Patch Camelyon (Veeling et al., 2018)	2	76.6±2.7	78.1±2.2	80.1±1.6	80.6±0.4	82.8±1.0
Retinopathy (Kaggle & EyePacs, 2015)	5	73.5±0.1	73.4±0.4	73.5±0.5	73.5±0.9	74.5±0.6
CLEVR-count (Johnson et al., 2017)	8	25.6±1.5	28.0±1.3	31.6±0.5	33.7±0.6	52.0±3.3
CLEVR-dist (Johnson et al., 2017)	6	26.5±0.9	29.0±0.7	32.6±1.3	38.0±3.2	52.6±6.4
dSprites-loc (Matthey et al., 2017)	16	8.0±1.6	12.2±1.1	20.9±6.1	29.8±5.4	68.1±10
dSprites-ori (Matthey et al., 2017)	16	9.3±0.7	13.7±1.5	17.0±1.4	21.5±1.1	47.8±3.9
SmallNORB-azi (LeCun et al., 2004)	18	6.6±0.3	7.3±0.5	8.4±0.3	9.1±0.4	15.1±1.5
SmallNORB-elev (LeCun et al., 2004)	9	16.2±1.6	19.2±1.3	21.5±1.2	22.9±1.4	35.3±4.6
DMLab (Beattie et al., 2016)	6	29.8±1.7	33.3±0.5	35.5±0.8	36.5±0.9	43.3±3.6
KITTI-dist (Geiger et al., 2013)	4	53.6±2.6	60.8±2.1	63.5±0.8	66±2.9	76.9±4.4
All		28.8	38.0	47.1	53.6	68.9
Natural		17.0	34.3	51.9	64.1	82.4
Specialized		62.9	69.7	75.0	77.8	85.2
Structured		21.9	25.4	28.9	32.2	48.6

Table 13: Classification accuracy as a function of ϵ for each of the datasets in the VTAB-1k benchmark. Backbone is VIT-B pretrained on ImageNet-21k. All parameters are learnable. Accuracy figures are percentages and the $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

dataset	**classes**	$\boldsymbol{\epsilon = 1}$	$\boldsymbol{\epsilon = 2}$	$\boldsymbol{\epsilon = 4}$	$\boldsymbol{\epsilon = 8}$	$\boldsymbol{\epsilon = \infty}$
Caltech101 (Fei-Fei et al., 2006)	102	16.1±5.2	34.9±2.1	55.3±1.0	69.9±2.7	93.7±0.4
CIFAR100 (Krizhevsky, 2009)	100	7.1±0.4	14.3±0.7	24.2±1.5	36.2±4.0	84.2±0.3
Flowers102 (Nilsback & Zisserman, 2008)	102	10.6±2.9	33±4.9	77.3±6.9	96±1.2	99.5±0.0
Pets (Parkhi et al., 2012)	37	26.7±6.0	56.9±7.0	76.0±3.7	84.2±0.7	91.7±0.2
Sun397 (Xiao et al., 2010)	397	2.4±2.1	5.7±1.5	7.7±0.4	11.6±3.1	55.9±0.2
SVHN (Netzer et al., 2011)	10	22.9±1.5	28.8±0.7	34±5.6	44.3±9.0	91.6±0.8
DTD (Cimpoi et al., 2014)	47	17.3±1.1	29.3±2.4	41.1±1.2	51.7±5.0	76.7±0.5
EuroSAT (Helber et al., 2019)	10	74.3±1.4	78.9±2.2	86±1.4	91.6±1.6	96.3±0.5
Resics45 (Cheng et al., 2017)	45	16±2.4	28±1.6	45.7±3.3	60.8±2.1	88.4±0.4
Patch Camelyon (Veeling et al., 2018)	2	74.1±1.5	76.6±1.4	78.9±2.1	76.2±5.3	87.1±0.7
Retinopathy (Kaggle & EyePacs, 2015)	5	73.4±0.5	73.1±0.5	73.6±0.1	73.6±0.1	74.0±1.3
CLEVR-count (Johnson et al., 2017)	8	21.5±5.6	28.8±1.5	33.6±2.4	38.2±0.7	57.6±8.7
CLEVR-dist (Johnson et al., 2017)	6	27.0±1.8	36.4±3.5	42.2±3.2	45.8±1.3	57.2±2.5
dSprites-loc (Matthey et al., 2017)	16	6.4±0.5	7.9±3.4	22.7±2.6	37.6±5.0	66.8±5.2
dSprites-ori (Matthey et al., 2017)	16	7.9±2.1	11.1±3.7	13.5±6.5	19.9±2.9	50.1±1.1
SmallNORB-azi (LeCun et al., 2004)	18	5.9±0.7	7.9±0.8	8.5±0.4	11.4±2.3	18.3±0.7
SmallNORB-elev (LeCun et al., 2004)	9	14.5±1.0	17.0±3.8	18.7±4.4	26.7±0.1	38.3±2.9
DMLab (Beattie et al., 2016)	6	29.2±1.3	32.7±1.4	35.4±2.0	39.3±1.1	51.5±1.9
KITTI-dist (Geiger et al., 2013)	4	41.7±5.8	51.2±3.4	57.9±8.6	68.9±0.3	76.0±0.7
All		26.1	34.3	43.8	51.8	71.3
Natural		14.7	29.0	45.1	56.3	84.8
Specialized		59.4	64.2	71.0	73.6	86.4
Structured		19.3	24.1	29.1	36.0	52.0

Figs. 19 and 20 depict the complete set of VTAB-1k accuracy results as a function of dataset, privacy level (ϵ), backbone, and learnable parameters. The datasets are ordered increasingly by transfer difficulty (TD). Although classifiers for the Retinopathy dataset appear to perform equally well independently of ϵ, a closer inspection reveals that this dataset is unbalanced and learned classifiers predict the most common class in all settings.

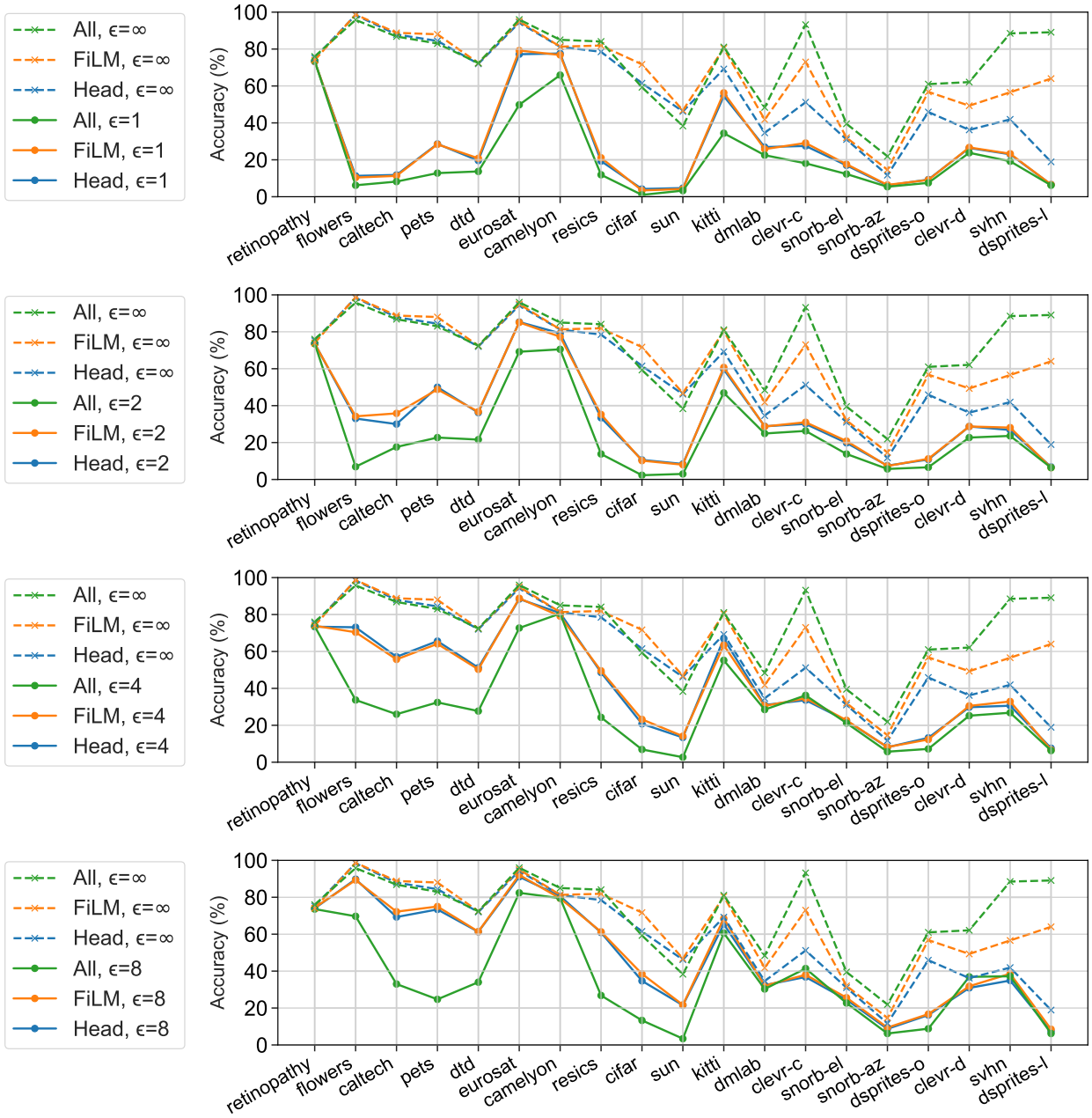

Figure 19: Classification accuracy for VTAB-1k datasets as a function of privacy level (ϵ) and learnable parameters. Backbone is R-50. Dashed lines in all plots indicate non-private accuracy as a reference. The datasets are ordered increasingly by transfer difficulty (TD).

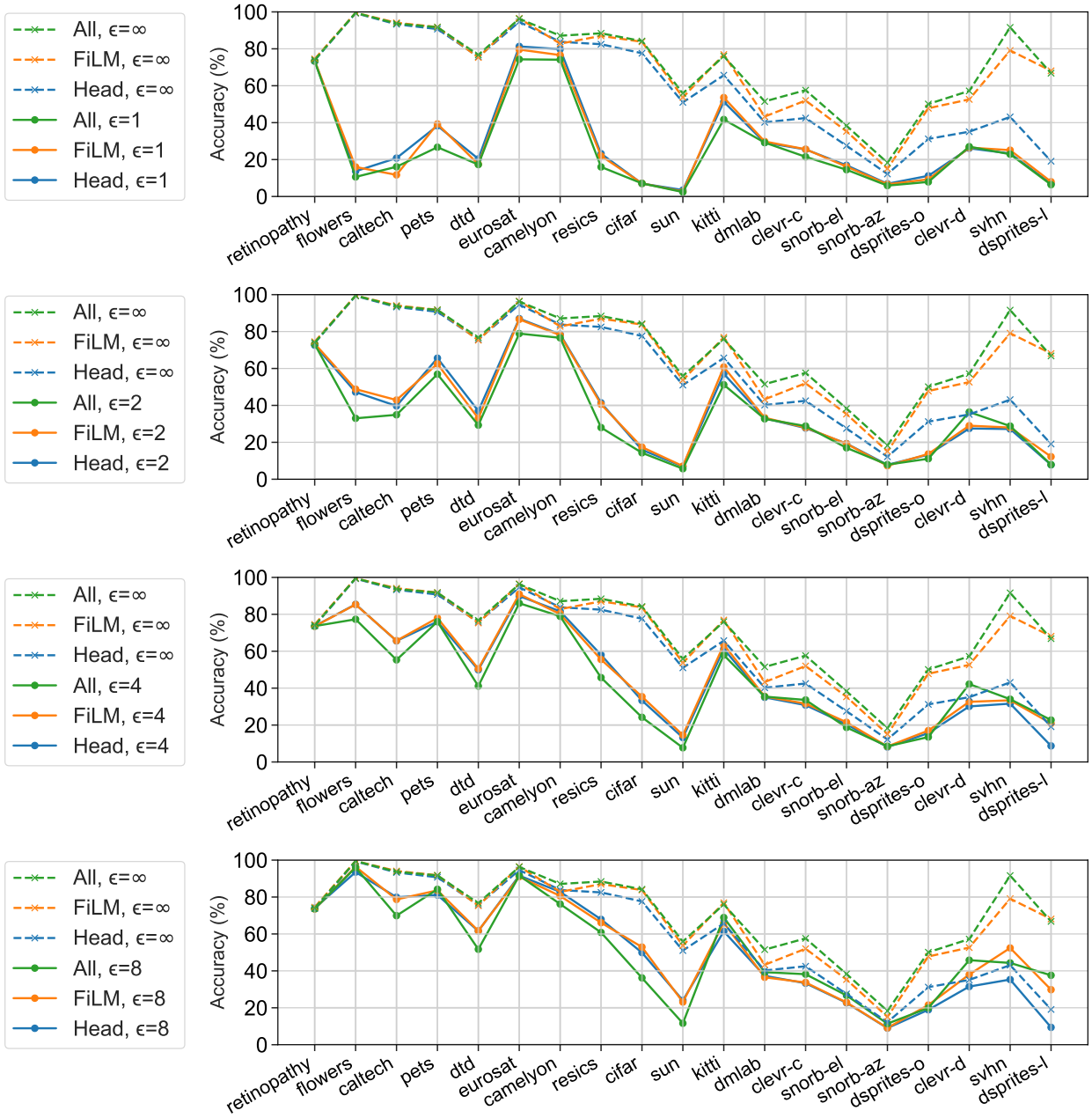

Figure 20: Classification accuracy for VTAB-1k datasets as a function of privacy level (ϵ) and learnable parameters. Backbone is ViT-B. Dashed lines in all plots indicate non-private accuracy as a reference. The datasets are ordered increasingly by transfer difficulty (TD).

Fig. 21 depicts the final training and test accuracy as a function of ϵ and learnable parameters for all 19 VTAB-1k datasets with the ViT-B backbone. Although classifiers for the Retinopathy dataset appear to perform equally well independently of ϵ, a closer inspection reveals that this dataset is unbalanced and learned classifiers predict the most common class in all settings.

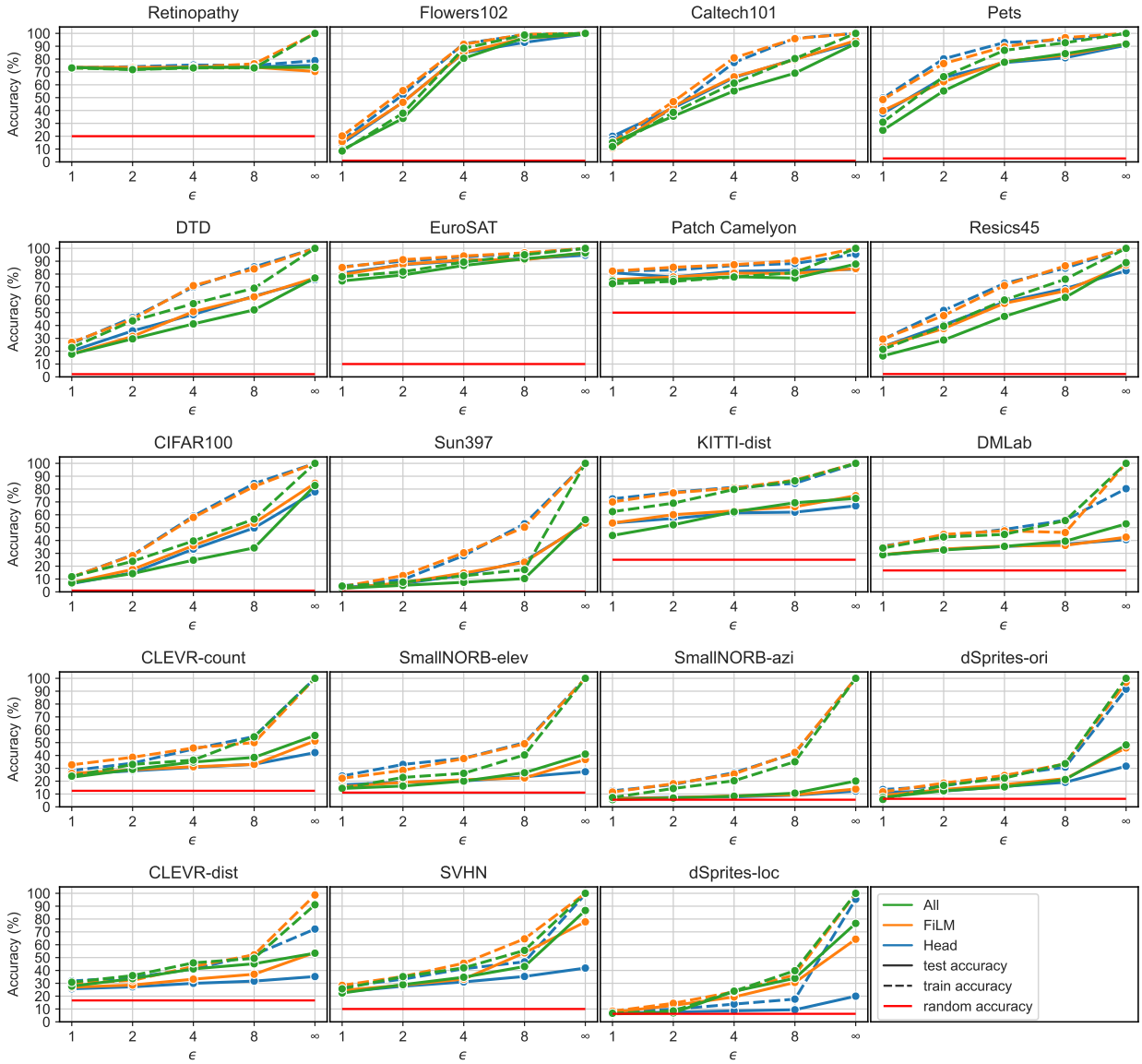

Figure 21: Test and train classification accuracy as a function of ϵ and learnable parameters (*All*, *FiLM* and *Head*) on VIT-B for all VTAB datasets with $\delta = 1/|\mathcal{D}|$. The accuracy is reported for the median run of Tables 11 to 13. The datasets are in order of increasing transfer difficulty (TD) from left-to-right and top-to-bottom.

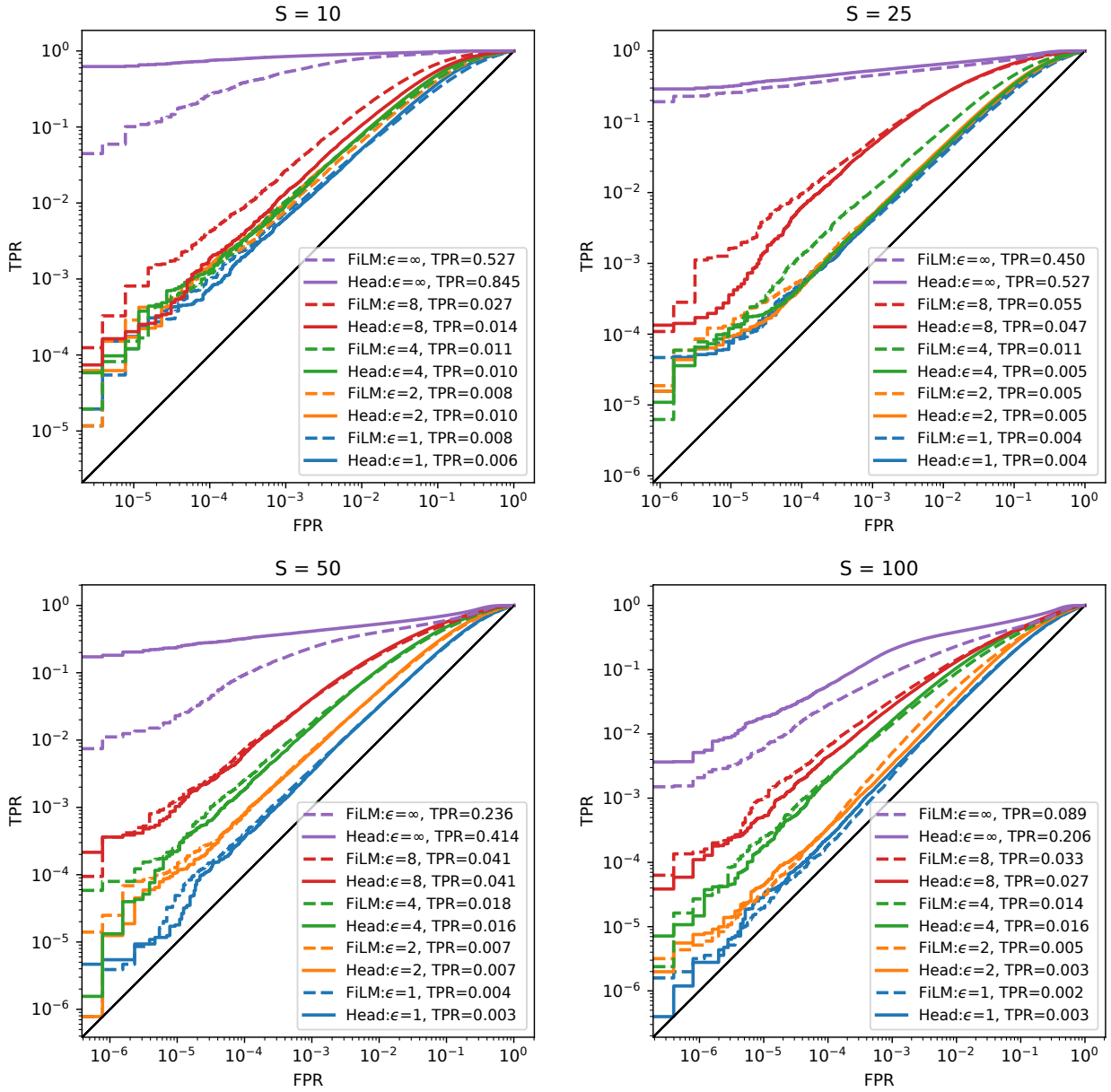

Figure 22: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for various privacy levels (ϵ) and backbone configurations *Head* and *FiLM* at a fixed S. TPR values in legends are measured at FPR=0.001.

A.2.7 Additional Membership Inference Attack Results

Fig. 22 depicts the complete set of ROC curves for LiRA on CIFAR-100 with the R-50 backbone for various privacy levels (ϵ) and learnable parameters *Head* and *FiLM* at a fixed S.

Fig. 23 depicts the complete set of ROC curves for LiRA on CIFAR-100 with the R-50 backbone for various shots S at fixed privacy levels (ϵ) and learnable parameters *Head* and *FiLM*.

Table 14 presents the True Positive Rates (TPR) at various False Positive Rates (FPR), Area Under Receiver Operating Curve (AUC), and Attack Advantage (Attack Adv) (Yeom et al., 2018) for various privacy levels (ϵ) and shots per class (S) corresponding to the plots in Figs. 22 and 23.

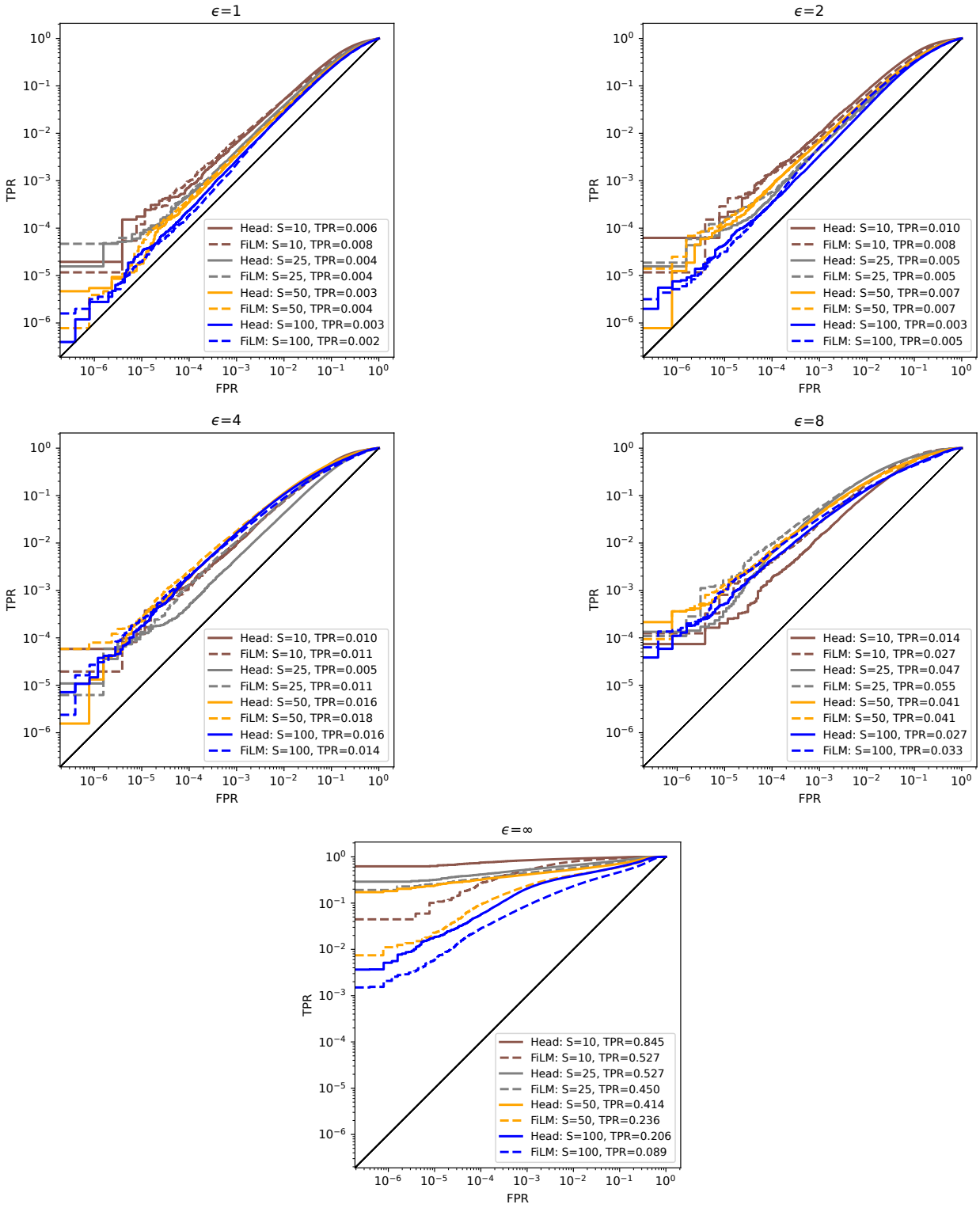

Figure 23: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for various S at fixed privacy levels (ϵ) and backbone configurations *Head* and *FiLM*. TPR values in legends are measured at FPR=0.001.

Table 14: True Postive Rates (TPR) at various False Positive Rates (FPR), Area Under Receiver Operating Curve (AUC), and Attack Advantage (Yeom et al., 2018) for various privacy levels (ϵ) and shots per class (S) corresponding to the plots in Figs. 22 and 23. Dataset ($\mathcal{D}$) is CIFAR-100. Backbone is R-50 pretrained on ImageNet-21k.

		TPR (%) @ 0.1% FPR		**TPR (%) @ 1% FPR**		**TPR (%) @ 10% FPR**		**AUC**		**Attack Adv**	
ϵ	S	**Head**	**FiLM**	**Head**	**FiLM**	**Head**	**FiLM**	**Head**	**FiLM**	**Head**	**FiLM**
1	10	0.62	0.77	5.13	5.11	37.12	32.20	0.765	0.731	0.419	0.354
	25	0.43	0.40	3.85	3.45	30.14	27.37	0.721	0.702	0.343	0.312
	50	0.34	0.37	3.16	3.20	25.92	25.18	0.689	0.682	0.291	0.279
	100	0.27	0.22	2.70	2.79	23.77	23.82	0.670	0.668	0.260	0.252
2	10	1.02	0.79	7.90	6.47	48.20	41.37	0.819	0.788	0.509	0.454
	25	0.49	0.47	4.55	4.13	35.41	31.39	0.754	0.730	0.397	0.356
	50	0.66	0.71	5.42	5.36	35.87	34.30	0.751	0.740	0.380	0.362
	100	0.34	0.51	3.54	5.31	30.65	32.27	0.717	0.719	0.325	0.319
4	10	0.96	1.05	7.91	7.56	48.84	43.26	0.821	0.797	0.516	0.465
	25	0.47	1.07	4.24	7.86	33.77	45.66	0.744	0.805	0.382	0.472
	50	1.58	1.78	11.24	10.83	48.08	45.50	0.809	0.794	0.466	0.438
	100	1.63	1.42	10.63	8.89	43.09	38.18	0.778	0.749	0.402	0.356
8	10	1.36	2.68	10.57	17.41	55.23	68.03	0.847	0.895	0.558	0.643
	25	4.69	5.49	24.28	24.15	67.23	65.19	0.888	0.880	0.612	0.590
	50	4.12	4.05	19.25	18.36	56.89	52.75	0.845	0.825	0.520	0.480
	100	2.69	3.35	13.37	14.25	45.15	43.12	0.787	0.771	0.413	0.375
∞	10	84.46	52.68	91.44	79.49	97.51	93.75	0.993	0.982	0.914	0.851
	25	52.70	45.04	65.80	57.70	82.95	76.38	0.952	0.930	0.730	0.664
	50	41.43	23.63	53.02	40.25	71.10	60.06	0.911	0.862	0.612	0.506
	100	20.58	8.88	38.70	23.24	61.30	46.85	0.870	0.790	0.524	0.381

A.2.8 Additional Federated Learning Results

Table 15 shows the non-private performance on the FLAIR dataset, while Table 16 shows the same performance under DP guaranties with $\epsilon = 2$.

Table 15: 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 16: 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

CIFAR-100 and Federated EMNIST Additionally, we perform experiments on CIFAR-100 and Federated EMNIST, which are commonly used to benchmark FL methods. We opt for these datasets as they have different degree of TD: CIFAR-100 has medium TD, while Federated EMNIST had high TD. 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). Federated EMNIST Caldas et al. (2018) is 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. high TD) 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. For CIFAR-100 and Federated EMNIST, we report standard test classification accuracy. All training details and hyperparameters are in Appendix A.3.6.

Fig. 24 shows the performance of different model configurations on CIFAR-100 and Federated EMNIST with and without DP. Table 17 illustrates private with $\epsilon = 8$ and non-private performance on CIFAR-100 and Federated EMNIST. These tables present a tabular version of the results in Fig. 9.

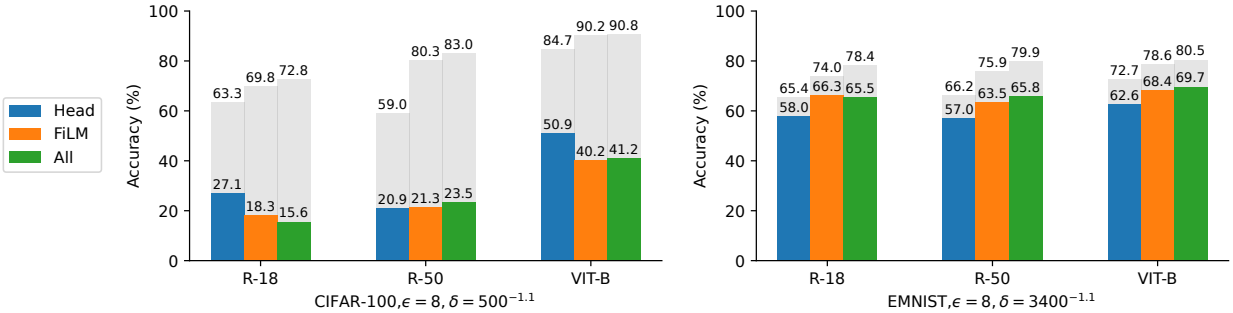

Figure 24: Private ($\epsilon = 2$, colored) and non-private ($\epsilon = \infty$, gray) FL performance on CIFAR-100 (left) and Federated EMNIST (right) as a function of backbone and learnable parameters. We report accuracy on test clients. R-18 backbone is pretrained on ImageNet-1k, VIT-B and R-50 are pretrained on ImageNet-21k.

Table 17: 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**

A.3 Training and Evaluation Details

A.3.1 FiLM Layer Implementation

Table 18 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 18: 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 Hyperparameter Tuning

For all centralized experiments, we first draw $\mathcal{D}$ of the required size ($|\mathcal{D}| = CS$, or $|\mathcal{D}| = 1000$ in the case of VTAB-1k) from the entire training split of the current dataset under evaluation. For the purposes of hyperparameter tuning, we then split $\mathcal{D}$ into 70% train and 30% validation. We then perform 20 iterations of hyperparameter tuning using the tree-structured parzen estimator (Bergstra et al., 2011) strategy as implemented in Optuna (Akiba et al., 2019) to derive a set of hyperparameters that yield the highest accuracy on the validation split. This set of parameters are subsequently used to train a final model on all of $\mathcal{D}$. We the evaluate the final, tuned model on the entire test split of the current dataset. Details on the set of hyperparameters that are tuned and their ranges can be found in Table 19. For DP training, we compute the required noise multiplier depending on the target (ϵ, δ)-DP guarantee. The hyperparameter ranges are purposely broad and have been empirically derived. We fine-tune models for at most 200 epochs to limit the amount of compute necessary.

Table 19: Hyperparameter ranges used for the Bayesian optimization.

	lower bound	**upper bound**
EPOCHS	1	200
LEARNING RATE	1E-7	1E-2
BATCH SIZE	10	$\lvert\mathcal{D}\rvert$
CLIPPING NORM	0.2	10
NOISE MULTIPLIER	BASED ON TARGET ϵ	

A.3.3 Effect of Shots per Class and ϵ Experiments

For each evaluated configuration, we draw $|\mathcal{D}| = CS$ examples from the dataset training split, tune hyperparameters as described in Appendix A.3.2, and then test on the entire test split of the dataset. We use the DP-Adam optimizer as implemented in Opacus (Yousefpour et al., 2021) for all private experiments. For non-private experiments, we used the Adam (Kingma & Ba, 2015) optimizer for the *Head* and *FiLM* parameter configurations and the SGD optimizer for the *All* configuration. No data augmentation was used and images were scaled to 224×224 pixels.

All of the effect of S and ϵ experiments were carried out on 1 (for *Head* and *FiLM*) and up to 3 (for *All*) NVIDIA V100 GPUs with 32GB of memory. The runtime for executing the whole experiment depends on the the size of the few-shot training set and the number of parameters resulting from the choice of the backbone

and the number of learnable parameters (*All* > *FiLM* > *Head*). For CIFAR-10 and SVHN the runtime for one configuration ranges from less than 5 GPU minutes ($S = 1 +$ *Head*) to 60 GPU hours ($S = 500 +$ *All*). For CIFAR-100, the range is from 15 GPU minutes ($S = 1 +$ *Head*) to over 700 GPU hours ($S = 500 +$ *All*).

A.3.4 VTAB-1k Experiments

For each evaluated configuration of each of the 19 datasets in the VTAB-1k benchmark, we draw $|\mathcal{D}| = 1000$ examples from the dataset training split, tune hyperparameters as described in Appendix A.3.2, and then test on the entire test split of the dataset. We use the DP-Adam optimizer as implemented in Opacus (Yousefpour et al., 2021) for all private experiments. For non-private experiments, we used the Adam (Kingma & Ba, 2015) optimizer for the *Head* and *FiLM* parameter configurations and the SGD optimizer for the *All* configuration.

No data augmentation was used. For the R-50 backbone, images were scaled to 384×384 pixels unless the image size was 32×32 pixels or less, in which case the images were scaled to 224×224 pixels. For the VIT-B backbone, images were scaled to 224×224 pixels.

All of the VTAB-1k transfer learning experiments were carried out on a single NVIDIA A100 GPU with 80GB of memory. Processing times for each configuration of each dataset will vary with the selected hyperparameters and the size of the test split, but approximate times are listed in Table 20.

Table 20: Approximate time to tune, train, and test a single configuration of parameters on a single VTAB-1k dataset for various backbones and parameter configurations. Units are wall clock GPU hours.

	Parameter Configuration		
Backbone	**None**	**FiLM**	**All**
R-50	0.6	0.9	2.7
VIT-B	1.3	2.4	6.5

A.3.5 Membership Inference Attacks Experiments

For each setting of S and ϵ, we first sample $2|\mathcal{D}|$ examples (recall $|\mathcal{D}| = CS = 100S$) from the CIFAR-100 training set, and then train 257 different models (1 target model plus 256 shadow models) where each sample for the training set is randomly selected with 50% probability from the $2|\mathcal{D}|$ examples. This ensures that approximately half of the models are trained on each example and half are not so that we can create distributions over the losses for each example being in and out of the training set as described in the LiRA algorithm (Carlini et al., 2022). We use each of the trained models in turn as the target model and then accumulate the attack predictions over all 257 targets to produce the ROC curve for the attack. Due to the extreme computation demand in training a large number of shadow models for each setting of S and ϵ, we restrict the attacks to the R-50 backbone and the *Head* and *FiLM* parameter configurations.

Our implementation is based on code from the TensorFlow Privacy library (Google, 2019b). All of the VTAB-1k transfer learning experiments were carried out on a single NVIDIA A100 GPU with 80GB of memory. When training the 257 models for each attack configuration, we do not perform hyperparameter tuning, instead we used the hyperparameter set from the CIFAR-100 experiments in Table 3 that yielded the highest accuracy for the particular configuration. Approximate training times for all 257 models in each configuration are listed on Table 21. The value of ϵ did not alter the training times to a significant degree.

A.3.6 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

Table 21: Approximate time to train 257 models for a single configuration of parameters for a LiRA attack on the CIFAR-100 dataset for various parameter and shot configurations. Units are wall clock GPU hours.

	Shot (S)			
Parameter Configuration	**10**	**25**	**50**	**100**
Head	6	12	16	46
FiLM	8	25	49	96

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.

Processing times for each configuration on FLAIR are given in Table 22.

Table 22: Approximate time to train and test a single configuration of parameters on FLAIR dataset for various backbones and parameter configurations. Units are wall clock GPU hours.

	Parameter Configuration		
Backbone	**Head**	**FiLM**	**All**
R-18	18	30	-
R-50	30	43	60
VIT-B	40	60	75

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 $\in \{0.1, 0.5, 0.8\}$