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 downstream datasets that are (i) relatively large, and (ii) similar in distribution to the pretraining data. However, in many applications including personalization, it is crucial to perform well in the few-shot setting, as obtaining large amounts of labeled data may be problematic; and on images 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, 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 $32\times$ for CIFAR-100 at $\epsilon = 1$. We also find that few-shot non-private models are highly susceptible to membership inference attacks. DP provides clear mitigation against the attacks, but a small ϵ is required to effectively prevent them.

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 individual training examples from discovery. Training with 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., 2022; 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 downstream dataset, as this approach has been shown to be highly effective on non-private data (Kolesnikov et al., 2019; Shysheya et al., 2022). The subset of model parameters to fine-tune ranges from all model parameters (Kolesnikov et al., 2019) to only the final layer, with the tuning of parameter efficient adapters (Perez et al., 2018; Houlsby et al., 2019; Karimi 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 also especially important in federated learning (where a global model is learned from many distributed users) and personalized federated learning (where a model obtained via federated learning is personalized with a specific user's data) where data contributed by each user may be small and sensitive, including personal data or actions

entered on a mobile device (Differential Privacy Team, 2017; Ding et al., 2017), medical images, or (Sheller et al., 2020), personal photos (Massiceti et al., 2021).

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 overlaps significantly with 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?* We provide the first comprehensive study on the efficacy of DP few-shot image classification. Our contributions are:

- 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, dataset distribution overlap, model architecture, and the subset of learnable parameters in the model vary.
- We establish a new DP baseline for the VTAB-1k (Zhai et al., 2019) transfer learning benchmark to encourage DP researchers to test methods on more challenging datasets.
- We assess the vulnerability of DP few-shot models with a strong membership inference attack (MIA) and find the attack to perform close to the theoretical upper bound derived as a composite of DP under the *substitute* adjacency for different δ. The bound is significantly higher than indicated by naive analysis with (ϵ, δ) from the *add/remove* adjacency commonly used in DP deep learning.

2 BACKGROUND

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

Preliminaries We denote input images $\boldsymbol{x}$ and image labels $y \in \{1, \ldots, C\}$ 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 $p(y = c|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ for $c = 1, \ldots, C$ 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 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 a myriad of adaptors for both 2D convolutional and transformer networks including FiLM, Adapter (Houlsby et al., 2019), LoRA (Hu et al., 2021), VPT (Jia et al., 2022), AdaptFormer (Chen et al., 2022), 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. BatchNorm, GroupNorm, or LayerNorm — see Appendix A.5.1 for details), which can update during fine-tuning. For example, in a ResNet50, there are only $11\,648$ learnable FiLM parameters, which is fewer than 0.05% of $\boldsymbol{\theta}$.

Dataset Distribution Overlap (DDO) The overlap between the distributions of the pretraining data and the downstream dataset is a key determinant of the ease and success of transfer learning. We measure the overlap as the relative difference between the accuracy of the *All* and *Head* learnable parameter configurations for a non-private model. If two domains overlap substantially, then only

adapting the head of the network is sufficient. If the overlap is small, then the backbone must also be adapted. Table A.1 provides the DDO values for all of the datasets used in the paper.

Differential Privacy (DP) DP (Dwork et al., 2006) is the gold standard for protecting sensitive data against privacy attacks. A stochastic algorithm is differentially private if it produces similar output distributions on similar datasets. More formally, (ϵ, δ)-DP with privacy budget $\epsilon \geq 0$ (lower means more private) and additive error $\delta \in [0, 1]$ bounds how much the output distribution can diverge on adjacent datasets. We primarily use *add/remove* adjacency, where two datasets are adjacent if one can be obtained from the other by adding or removing one datapoint. We denote (ϵ, δ) the corresponding privacy parameters. When considering *substitute* adjacency, where two datasets are adjacent if one can be obtained from the other by substituting one datapoint, we use instead (ϵ_S, δ_S). (See Appendix A.2 for more details.) The additive error is typically chosen such that $\delta < 1/|\mathcal{D}|$. We refer to Dwork & Roth (2014) for a comprehensive 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 privacy loss in (ϵ, δ)-DP is a result of 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 audit DP training algorithms as they test how well the (ϵ, δ)-DP guarantee holds for trained models. While there are many types of MIA (Hu et al., 2022), 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 privacy loss on particular training examples (Carlini et al., 2022; Ye et al., 2022). In addition to black-box access to the model, 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 certainty (Carlini et al., 2022). DP implies an upper bound on TPR at a given FPR: $\mathrm{TPR} \leq \min\{e^{\epsilon_S}\mathrm{FPR} + \delta_S, 1 - \frac{1-\delta_S-\mathrm{FPR}}{\exp(\epsilon_S)}\}$ Since MIAs are defined w.r.t. substitute adjacency, this depends on (ϵ_S, δ_S) rather than (ϵ, δ).

A detailed discussion of related work on DP transfer learning in the high-shot setting using pretrained models can be found in Appendix A.3.

3 EXPERIMENTS

The experiments address the question: "Under what conditions is differentially private few-shot image classification effective?" We focus on transfer learning approaches that utilize large pretrained backbones. We vary the: (i) number of shots S; (ii) set of learnable parameters in f (*All*, *Head*, *FiLM*); (iii) downstream dataset $\mathcal{D}$ (with varying DDO); and (iv) network architecture: BiT-M-R50x1 (R-50) (Kolesnikov et al., 2019) pretrained on ImageNet-21K with 23.5M parameters, Vision Transformer VIT-Base-16 (VIT-B) (Dosovitskiy et al., 2020) pretrained on ImageNet-21K with 85.8M parameters. Source code for all experiments will be made publicly available.

Training Protocol For all experiments, we first draw $\mathcal{D}$ of the required size (usually $|\mathcal{D}| = CS$ 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 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 split. 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.5.2. We assume that any pretraining has been non-private. 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 report (ϵ, δ)-DP computed with the RDP accountant (Mironov, 2017) and set $\delta = 1/|\mathcal{D}|$. 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.5 for additional training details.

3.1 EFFECT OF SHOTS AND DP

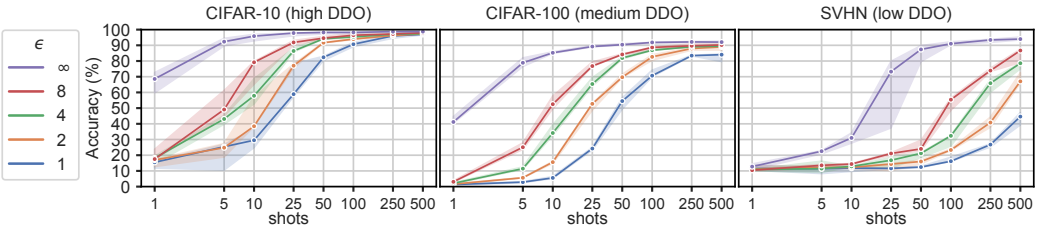

Figure 1: Classification accuracy as a function of shots and ϵ for CIFAR-10, CIFAR-100 and SVHN. DDO (low, medium, high) refers to the data distribution overlap (see Appendix A.1). Backbone is VIT-B and the best performing configuration out of *All*, *FiLM* and *Head* is used for each combination of 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.

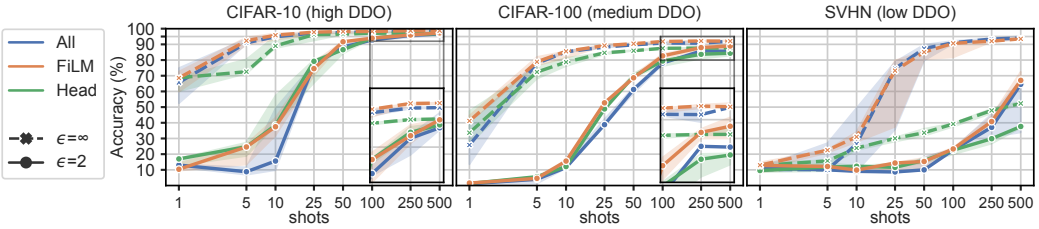

Figure 2: Classification accuracy as a function of shots and learnable parameters (*All*, *FiLM* and *Head*) on VIT-B for CIFAR-10, CIFAR-100 and SVHN for $\epsilon \in \{2, \infty\}$ with $\delta = 1/|\mathcal{D}|$. DDO (low, medium, high) refers to the data distribution overlap (see Appendix A.1). The accuracy is reported over three seeds with the line showing the median and the band the lowest and highest accuracy.

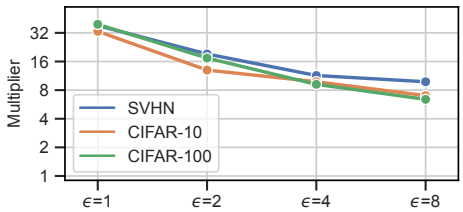

Figure 3: Multiplier of shots required to reach same accuracy as non-private with $S = 5$ for VIT-B with *FiLM* with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation (see details for more configurations in Appendix A.4.2).

We evaluate the performance of transfer learning under DP when varying S and ϵ. Results are in Figures 1 to 3, with tabular versions in Tables A.2 to A.7. In addition to the CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009), which are commonly used in DP transfer learning, we also evaluate SVHN (Netzer et al., 2011), which has a low DDO and hence requires a greater degree of adaptation of the pretrained backbone. Key observations are:

Shots Figure 1 shows that accuracy decreases as ϵ decreases. For $S \leq 10$, accuracy is poor under DP. However, if the DDO is high 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 DDO is low, 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 $S = 500$.

Learnable Parameters Referring to Figure 2, *FiLM* is at least as good 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 DDO is high, training only the *Head* is competitive with *FiLM* and *All*, but when DDO is low, *Head* falls short as it cannot adapt the backbone to a downstream dataset that has a different data distribution. See Appendix A.4.4 for heat maps showing the advantage of *FiLM* over *Head*.

Effect of DP Referring to Figure 3, we see that DP requires significantly more shots than non-private, with the multiple of shots increasing as the privacy level increases (i.e. as ϵ decreases). For all datasets,

at $\epsilon = 8$, S must be increased by approximately $8\times$ to meet $S = 5$ non-private accuracy and $32\times$ at $\epsilon = 1$ for *FiLM* and VIT-B. In effect, as the privacy level increases, the effective number of shots decreases in an exponential manner. There is evidence that these multipliers are lower for simpler forms of adaptation (e.g. *Head*) than for more complex forms (e.g. *All*), see Figures A.9 and A.10.

Backbone Referring to Figure A.11, we see that VIT-B performs comparably to or better than R-50, at the expense of having significantly more parameters (see Table A.15).

3.2 VTAB-1K

The VTAB-1k benchmark (Zhai et al., 2019) is a low to medium-shot transfer learning benchmark that consists of 19 datasets grouped into three distinct categories (natural, specialized, and structured). From each dataset, 1000 examples are drawn at random from the training split to use for the downstream dataset $\mathcal{D}$. After fine-tuning, the entire test split is used to evaluate performance.

Figure 4 shows average classification accuracy over all of the datasets in the VTAB-1k benchmark. Complete tabular results are in Tables A.8 to A.13. Key observations are: (i) DP classification

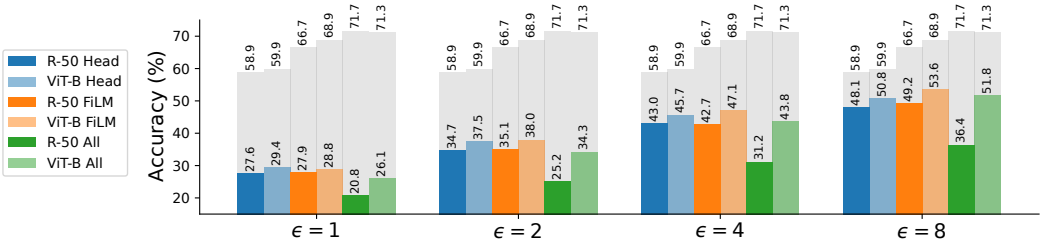

Figure 4: Average classification accuracy over all VTAB-1k datasets as a function of backbone, learnable parameters, and privacy level (ϵ) at $\delta = 10^{-3}$. Colored columns indicate results under DP, light gray indicates non-private accuracy for the corresponding configuration.

accuracy decreases significantly as ϵ is decreased and always falls short of non-private accuracy. (ii) For non-private settings, the *All* learnable parameters setting outperforms *FiLM* which outperforms *Head*. For DP settings, *All* performs worst, *FiLM* and *Head* perform similarly, though *FiLM* is better in the majority of cases. (iii) At the expense of extra parameters (85.8M vs. 23.5M), the VIT-B backbone outperforms the R-50 backbone.

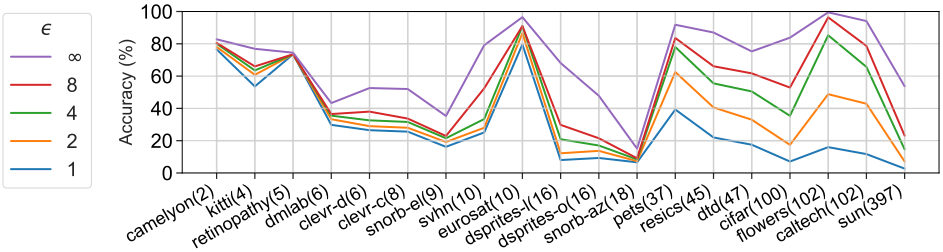

Figure 5: Classification accuracy as a function of VTAB-1k dataset and privacy level (ϵ) at $\delta = 10^{-3}$. Backbone is VIT-B and configuration is *FiLM*. The datasets are ordered increasingly by C (in parenthesis) or equivalently decreasingly by S as $|\mathcal{D}| = 1000$.

Figure 5 depicts the classification accuracy for VTAB-1k datasets ordered by the number of classes (C) in each as a function of privacy level for the VIT-B backbone in the *FiLM* configuration. Note that since the dataset $\mathcal{D}$ has a fixed size of 1000 examples, as C increases, S necessarily decreases. The key observation is that as S decreases, the degradation in accuracy is the more severe as ϵ decreases. 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. A complete set of plots for various backbones and configurations can be found in Figures A.15 and A.16.

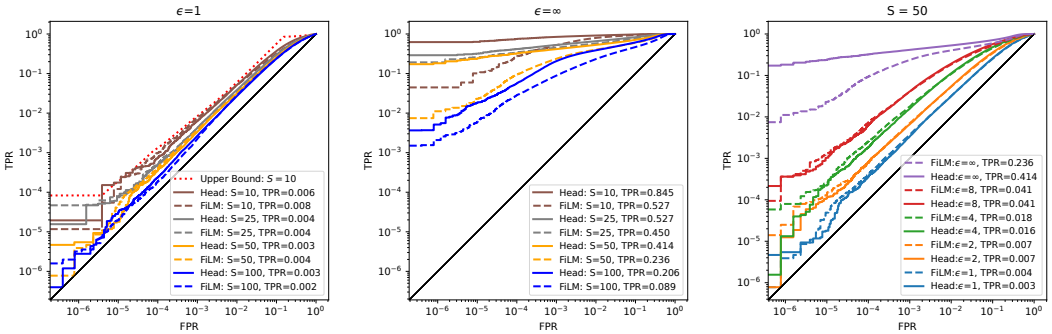

Figure 6: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for two values of ϵ (1 and ∞) where S varies and for $S = 50$ where ϵ varies. TPR values in legends are measured at FPR=0.001. Complete results in Table A.14 and Figures A.17 and A.18. The dotted red curve on the $\epsilon = 1$ plot indicates the theoretical upper bound on TPR for $S = 10$. $\delta = 1/100S$.

Figure A.14 shows the difference between the accuracy of *FiLM* and *Head* for VTAB-1k datasets as a function of ϵ. The datasets are ordered from high to low DDO (see Table A.1). At $\epsilon = 1$, *Head* has an advantage over *FiLM* on several datasets. *FiLM* shows a significant advantage when the DDO decreases and as ϵ increases.

3.3 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. 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. Refer to Appendix A.5.5 for more detail. Excerpts from attack results are shown in Figure 6. The complete set of attack ROC curves are shown in Figures A.17 and A.18, while Table A.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: (i) Non-private ($\epsilon = \infty$) models are extremely vulnerable to MIAs (see Figure 6, middle). For example, in the case of $\epsilon = \infty$, $S = 10$, and *Head* configuration, 84.5% of the examples can be successfully identified with a false positive rate of only 0.1%. (ii) Vulnerability of non-private ($\epsilon = \infty$) models decreases as S increases. Also, the *FiLM* configuration is consistently less vulnerable than *Head* (see Figure 6). We hypothesize that *FiLM* generalizes better, so training examples do not stand out as much as in the *Head* configuration. (iii) When S is fixed, vulnerability to MIAs greatly decreases with decreasing ϵ (see Figure 6, right). However, when $S = 10$ with $\epsilon = 1$, 5.1% of the examples can be successfully identified with a FPR of 1% and 0.8% of the examples with 0.1% FPR (see Table A.14). (iv) Under DP, there appears to be little or no difference between the vulnerability of the *FiLM* and *Head* configurations at the same ϵ (see Figure 6, right). (v) Under DP with small ϵ, the vulnerability to MIA decreases as S increases and TPR is close to the theoretical upper bound (see Figure 6, left). For larger ϵ there is no trend with S and the bounds are loose.

4 Discussion

Our work shows that there is still much to be done in order to realize effective transfer learning under DP constraints for few-shot, low DDO datasets. 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 training process (Golatkar et al., 2022) in order to improve utility.

REFERENCES

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

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

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.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, 2018.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. *arXiv preprint arXiv:2201.04845*, 2022.

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 arXiv:1612.03801*, 2016.

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.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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.

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

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

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

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. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

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

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.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8376–8386, 2022.

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

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems, NeurIPS 2021*, volume 34, pp. 11631–11642, 2021.

Rob Hall, Alessandro Rinaldo, and Larry A. Wasserman. Differential privacy for functions and functional data. *J. Mach. Learn. Res.*, 14(1):703–727, 2013. URL https://jmlr.csail.mit.edu/papers/v14/hall13a.html.

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.

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

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

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, 2022.

Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Investigating membership inference attacks under data dependencies. *arXiv preprint arXiv:2010.12112 [cs.CR]*, 2020. doi: 10.48550/ARXIV.2010.12112.

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

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

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

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017. doi: 10.1109/TIT.2017.2685505.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *2011 ACM SIGMOD International Conference on Management of Data, SIGMOD 2011*, pp. 193–204. ACM, 2011. doi: 10.1145/1989323.1989345.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

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

Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2560–2569. PMLR, 2020. URL `http://proceedings.mlr.press/v108/koskela20b.html`.

Antti Koskela, Joonas Jälkö, Lukas Prediger, and Antti Honkela. Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using FFT. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3358–3366. PMLR, 2021. URL `http://proceedings.mlr.press/v130/koskela21a.html`.

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. *CoRR*, 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.

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, 2022. URL https://openreview.net/forum?id=bVuP3ltATMz.

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.

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

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

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

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.

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 CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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

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 arXiv:2206.08671 [stat.ML]*, 2022. doi: 10.48550/arXiv.2206.08671.

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.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining, 2022. URL https://arxiv.org/abs/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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

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.

Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.

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

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. *CoRR*, 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.

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 arXiv:1910.04867*, 2019.

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

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4782–4817. PMLR, 2022. URL https://proceedings.mlr.press/v151/zhu22c.html.

A APPENDIX

A.1 DATASET DISTRIBUTION OVERLAP

Table A.1 shows the data distribution overlap between the ImageNet-21K pretraining data and each of the datasets used in our experiments.

Table A.1: Amount of distribution overlap between each of the 19 VTAB-1k datasets (plus CIFAR-10) and the ImageNet-21K pretraining data. 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 more the data distribution overlap between the pretraining data and the downstream VTAB-1k dataset. In the Distribution Overlap column, we map the score into three buckets: a score of 0-5 is High, 5-10 is Medium, and greater than 10 is Low. 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**	**DISTRIBUTION OVERLAP**
CALTECH101 (FEI-FEI ET AL., 2006)	0.4	HIGH
CIFAR10 (KRIZHEVSKY, 2009)	1.0	HIGH
CIFAR100 (KRIZHEVSKY, 2009)	7.8	MEDIUM
FLOWERS102 (NILSBACK & ZISSERMAN, 2008)	0.2	HIGH
PETS (PARKHI ET AL., 2012)	1.1	HIGH
SUN397 (XIAO ET AL., 2010)	8.8	MEDIUM
SVHN (NETZER ET AL., 2011)	52.9	LOW
DTD (CIMPOI ET AL., 2014)	1.3	HIGH
EUROSAT (HELBER ET AL., 2019)	1.7	HIGH
RESICS45 (CHENG ET AL., 2017)	6.7	MEDIUM
PATCH CAMELYON (VEELING ET AL., 2018)	3.8	HIGH
RETINOPATHY (KAGGLE & EYEPACS, 2015)	0.2	HIGH
CLEVR-COUNT (JOHNSON ET AL., 2017)	26.2	LOW
CLEVR-DIST (JOHNSON ET AL., 2017)	38.7	LOW
DSPRITES-LOC (MATTHEY ET AL., 2017)	71.4	LOW
DSPRITES-ORI (MATTHEY ET AL., 2017)	37.7	LOW
SMALLNORB-AZI (LECUN ET AL., 2004)	33.3	LOW
SMALLNORB-ELEV (LECUN ET AL., 2004)	28.2	LOW
DMLAB (BEATTIE ET AL., 2016)	21.9	LOW
KITTI-DIST (GEIGER ET AL., 2013)	13.6	LOW

A.2 Differential Privacy and Membership Inference Attacks

We provide privacy protection for trained models using DP. In this section, we give the mathematical definition of DP and discuss its relationship to membership inference attacks.

Definition A.1. (Dwork et al., 2006) A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is $(\epsilon, \delta, \sim)$-DP if for all adjacent datasets $\mathcal{D} \sim \mathcal{D}'$ and all $S \subseteq \mathcal{Y}$,

$$\Pr(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\epsilon}\Pr(\mathcal{M}(\mathcal{D}') \in S) + \delta \, .$$

Our definition differs from the usual one by making explicit its dependence on the adjacency relation $\sim$. The most common choice in the centralized setting is the add/remove relation $\sim_{AR}$, where adjacent datasets are obtained by adding or removing one example and are of different size. This setting is sometimes called *unbounded* DP (Kifer & Machanavajjhala, 2011).[1] Another widely used choice is the substitute relation $\sim_S$, where adjacent data sets are obtained by substituting one example and are of the same size. This setting is called *bounded* DP.

The values of ϵ and δ for these two settings are not directly comparable. If we use $(\epsilon_{AR}, \delta_{AR})$ to denote parameters under $\sim_{AR}$ (these correspond to (ϵ, δ) in the main text) and (ϵ_S, δ_S) under $\sim_S$, we have the following general reduction, which follows immediately from group privacy:

Theorem A.2. *Any $(\epsilon_{AR}, \delta_{AR}, \sim_{AR})$-DP mechanism is $(\epsilon_S, \delta_S, \sim_S)$-DP with $\epsilon_S = 2\epsilon_{AR}$ and $\delta_S = (1 + e^{\epsilon_{AR}})\delta_{AR}$.*

Most works in DP deep learning use $\sim_{AR}$, and privacy accountants implemented in DP training libraries report $(\epsilon_{AR}, \delta_{AR})$. This is because privacy accounting with $\sim_S$ is much more difficult (Zhu et al., 2022).

A.2.1 Differential Privacy Bounds for Hypothesis Testing

Given the output y of a DP mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ on either of two adjacent datasets $\mathcal{D} \sim \mathcal{D}'$, an attack against the $(\epsilon, \delta, \sim)$-DP of $\mathcal{M}$ can be formulated as a hypothesis test where the null hypothesis is that the input to $\mathcal{M}$ was $\mathcal{D}$ and the alternative hypothesis is that it was $\mathcal{D}'$. A false positive occurs when the null hypothesis holds but the the attack rejects it, while a false negative occurs when the null hypothesis does not hold but the attack does not reject it. Kairouz et al. (2017) characterize DP in terms of the FPR and FNR of hypothesis tests (this strengthens an earlier result of Hall et al. (2013) that shows that the conditions are necessary, but not that they are sufficient). Given that $\mathrm{TPR} = 1 - \mathrm{FNR}$, we restate this result as follows:

Theorem A.3 (Kairouz et al. (2017))**.** *A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is $(\epsilon, \delta, \sim)$-DP if and only if for all adjacent $\mathcal{D} \sim \mathcal{D}'$,*

$$\mathrm{TPR} \leq \min\{e^{\epsilon}\mathrm{FPR} + \delta, 1 - e^{-\epsilon}(1 - \delta - \mathrm{FPR})\} \, . \tag{A.1}$$

The set of TPR and FPR that can be achieved under $(\epsilon, \delta, \sim)$-DP thus define a *privacy region* $\mathcal{R}(\epsilon, \delta)$, as illustrated in Figure A.7.

A.2.2 Formalising Membership Inference as Hypothesis Testing

Yeom et al. (2020) (and similarly Carlini et al. (2022)) formalize membership inference using the following probabilistic experiment parameterized by n, the size of the training dataset, where an adversary interacts with a challenger:

1. The challenger samples a dataset $\mathcal{D}$ of n examples and trains a model f on it;
2. The challenger chooses $b \in \{0, 1\}$ uniformly at random;
3. If $b = 0$, the challenger draws an example $(\boldsymbol{x}, y)$ from $\mathcal{D}$, otherwise it draws $(\boldsymbol{x}, y)$ from the underlying training data distribution;

[1] Unlike e.g. Balle et al. (2018), we use $\sim_{AR}$ to denote the add/remove relation to distinguish this from separate add and remove relations used by Zhu et al. (2022).

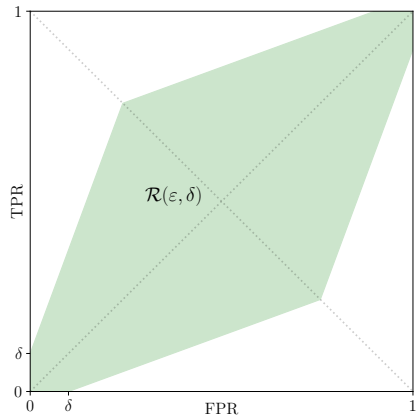

Figure A.7: Privacy region of a $\mathcal{R}(\epsilon, \delta)$-DP mechanism. The region is symmetric w.r.t. the TPR = FPR diagonal to account for complementary tests.

4. The adversary produces a guess $\hat{b}$ for b, given f, $(\boldsymbol{x}, y)$, n, the training data distribution and the training algorithm.

A membership inference attack defined in this way is a hypothesis test for the $\sim_S$ adjacency relation, because it asks the adversary to distinguish between two possible points in a training set of fixed size.

A.2.3 DP Bounds on Membership Inference

Following the recommendation of Carlini et al. (2022), we primarily measure the performance of a MIA as its true positive rate (TPR) at a fixed small false positive rate (FPR). To give a more complete picture, we also provide full ROC curves in Appendix A.4.6. For $\epsilon \neq \infty$, these curves must be contained between the corresponding privacy region $\mathcal{R}(\epsilon, \delta)$.

Taking into account that a MIA is a hypothesis test for the $\sim_S$ adjacency relation, we can combine Theorems A.2 and A.3 to obtain the following result.

Proposition A.4. *Let $\mathcal{M}$ be an $(\epsilon_{AR}, \delta_{AR}, \sim_{AR})$-DP training algorithm. The FPR and TPR of a membership inference attack distinguishing members of the training dataset of models trained with $\mathcal{M}$ from non-members satisfy*

$$\text{TPR} \leq \min\{e^{2\epsilon_{AR}}\text{FPR} + (1 + e^{\epsilon_{AR}})\delta_{AR},\ 1 - e^{-2\epsilon_{AR}}(1 - (1 + e^{\epsilon_{AR}})\delta_{AR} - \text{FPR})\}\ . \quad \text{(A.2)}$$

The bound in Equation (A.2) applies to all (ϵ, δ)-DP algorithms. Most algorithms satisfy (ϵ, δ)-DP for multiple values of (ϵ, δ), which can be parameterized as $(\epsilon(\delta), \delta)$. For DP-SGD, we can evaluate the $(\epsilon(\delta), \delta)$ curve using a privacy accountant such as a Rényi DP (RDP, Mironov, 2017) or a numerical DP accountant (Koskela et al., 2020; 2021; Gopi et al., 2021). The accounting results with different δ will give slightly different TPR bounds that are all valid. To get the maximally tight bounds, we can take the minimum of the bounds over all $(\epsilon(\delta), \delta)$ to obtain the final composite bound:

$$\text{TPR} \leq \min_{(\epsilon_{AR},\, \delta_{AR}) \text{ s.t. } \mathcal{M} \text{ is } (\epsilon_{AR},\, \delta_{AR})\text{-DP}} \min\{e^{2\epsilon_{AR}}\text{FPR}+(1+e^{\epsilon_{AR}})\delta_{AR},\ 1-e^{-2\epsilon_{AR}}(1-(1+e^{\epsilon_{AR}})\delta_{AR}-\text{FPR})\}. \quad \text{(A.3)}$$

Considering the bound of equation A.2, we observe that in order to have a meaningful bound, we need to have $\epsilon_S \leq 2$ or equivalently $\epsilon_{AR} \leq 1$ with small corresponding δ, although $\epsilon_S \leq 1$ (equivalently $\epsilon_{AR} \leq 0.5$) would be preferable. These are smaller than commonly used values $\epsilon_{AR} > 1$. Conflation of results from the two relations, i.e. using the natural bound of Equation (A.1) but carelessly substituting ϵ_{AR} for ϵ_S, may misrepresent the residual risk and DP protection when using the standard definition of MIA (Carlini et al., 2022; Yeom et al., 2020).

The connection between DP and MIA via Theorem A.3 can be used to derive bounds for other metrics. For instance, Humphries et al. (2020) provide a tight bound for the MI advantage (defined by Yeom et al. (2020) as TPR - FPR) of attacks against general DP mechanisms.

A.2.4 COMPARING THE DIFFERENT BOUNDS

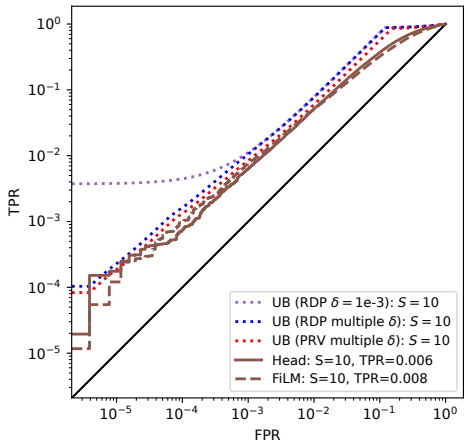

Figure A.8: Comparison of the bounds given at single δ in Equation (A.2) with the composite bound in Equation (A.3), using both the RDP accountant and the PRV accountant (Gopi et al., 2021), for a model trained with RDP target bound $\epsilon = 1.0, \delta = 10^{-3}$.

In Figure A.8 we compare the bounds given at single δ in Equation (A.2) with the composite bound in Equation (A.3), using both the RDP accountant and the PRV accountant (Gopi et al., 2021). We find that using the composite bound is very important to obtain bounds that are tight over a broad FPR range, while moving from the RDP accountant to an accurate numerical accountant such as the PRV accountant further makes the bounds slightly tighter.

A.3 RELATED WORK

Section 1 describes various works where DP transfer learning using models pretrained on large public datasets achieves 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.

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 aims at resolving this issue by evaluating DP models on datasets that have a wide range of DDO.

A.4 ADDITIONAL RESULTS

A.4.1 ADDITIONAL EFFECT OF SHOTS PER CLASS AND DP RESULTS

Tables A.2 to A.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 A.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 $\pm$ 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 A.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 $\pm$ 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	53.0±6.7	49.9±8.5
	$\epsilon = 2$	1.3±0.6	1.2±0.6	6.8±0.2	23.1±0.6	46.0±2.1	59.4±2.9	63.2±0.6	59.1±6.3
	$\epsilon = 4$	1.0±0.0	1.6±1.3	15.3±1.1	43.7±3.9	57.6±2.8	64.8±0.4	65.9±1.9	66.6±0.6
	$\epsilon = 8$	1.0±0.0	5.1±7.9	31.1±2.0	55.9±2.6	59.6±3.7	67.3±0.3	70.5±1.4	68.2±3.4
	$\epsilon = \infty$	28.3±2.4	51.9±5.6	59.7±9.9	71.9±0.5	76.0±0.8	79.9±0.1	83.6±0.4	82.5±1.4
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	72.4±0.3	72.2±0.4
	$\epsilon = 2$	1.4±0.5	2.9±0.3	9.3±0.6	33.4±0.8	55.0±3.2	72.2±0.3	79.0±0.2	78.7±0.7
	$\epsilon = 4$	1.7±0.4	7.2±1.3	22.4±1.2	53.9±0.9	70.1±0.6	77.9±0.6	81.4±0.3	81.5±0.2
	$\epsilon = 8$	2.2±0.3	18.6±0.3	38.7±3.7	65.3±1.6	75.6±0.5	80.5±0.8	83.0±0.2	83.3±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	84.9±0.3	84.4±0.1
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	63.5±0.9	64.1±0.8
	$\epsilon = 2$	1.4±0.7	3.9±0.8	8.7±1.4	31.6±1.2	51.0±1.2	63.1±1.0	70.1±0.3	70.2±0.1
	$\epsilon = 4$	1.8±0.6	7.5±1.8	22.1±1.3	46.3±0.7	61.7±1.6	68.4±0.8	73.2±0.5	72.8±0.6
	$\epsilon = 8$	2.5±0.7	17.2±1.7	38.9±1.9	58.9±0.3	67.5±0.4	72.1±0.5	75.4±0.3	75.4±0.1
	$\epsilon = \infty$	25.6±5.6	56.4±0.3	62.8±1.4	69.6±0.2	72.9±0.3	75.8±0.3	78.3±0.3	78.3±0.3

Table A.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.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
ALL	$\epsilon = 1$	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
	$\epsilon = 4$	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
FiLM	$\epsilon = 1$	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
	$\epsilon = 4$	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
HEAD	$\epsilon = 1$	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
	$\epsilon = 4$	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 A.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.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
ALL	$\epsilon = 1$	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
	$\epsilon = 4$	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
FiLM	$\epsilon = 1$	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
	$\epsilon = 4$	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
HEAD	$\epsilon = 1$	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
	$\epsilon = 4$	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 A.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 $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds. Note that for three entries not all runs finished before submission deadline. The confidence intervals for these entries will be added for the rebuttal.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
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	62.2±24.4	63.7±25.9
	$\epsilon = 2$	1.1±0.1	3.2±1.9	11.9±1.4	39.9±2.5	60.8±2.2	78.0±0.8	71.3	75.9±19.1
	$\epsilon = 4$	0.9±0.2	10.5±2.1	24.3±2.2	56.4±3.5	68.5±15.8	82.9±5.3	84.2±9.6	84.3±7.9
	$\epsilon = 8$	1.3±0.4	17.3±7.7	18.8±2.9	60.8±18.8	84.2±0.3	86.9±2.4	88.3±3.4	87.1±5.2
	$\epsilon = \infty$	26.2±14.5	78.1±1.0	85.3±0.6	88.4±0.7	89.6±0.3	90.9±0.4	88.9±3.9	86.4
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	83.5±0.2	82.7±3.0
	$\epsilon = 2$	1.6±0.2	4.5±1.5	15.5±0.7	51.4±2.7	69.1±1.3	82.4±2.0	87.8±0.9	88.8±1.9
	$\epsilon = 4$	1.6±0.1	11.2±1.9	35.3±4.0	66.2±3.2	82.0±0.8	87.0±0.4	88.9±1.4	89.7±0.9
	$\epsilon = 8$	2.7±0.7	25.6±2.4	53.3±4.6	77.6±1.8	83.6±3.5	88.1±1.2	89.9±0.9	90.3±0.6
	$\epsilon = \infty$	42.1±2.5	79.1±3.1	84.2±2.8	89.4±0.5	90.6±0.5	91.6±0.4	92.1±0.3	92.2±0.4
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	79.7±1.1	80.4±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	83.0±2.2	83.9±1.1
	$\epsilon = 4$	2.2±0.4	11.7±0.9	29.5±5.0	65.7±3.9	76.0±5.0	83.7±0.6	85.3±1.7	86.1±0.5
	$\epsilon = 8$	3.2±0.5	21.7±2.0	51.9±4.6	74.5±4.2	82.0±0.3	85.4±0.5	86.6±1.1	87.4±0.2
	$\epsilon = \infty$	35.8±12.2	72.2±4.5	78.7±3.0	84.3±0.8	86.1±0.3	87.4±0.4	87.6±0.1	88.0±0.8

Table A.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 $\pm$ sign indicates the 95% confidence interval over 3 runs with different seeds.

		$1S$	$5S$	$10S$	$25S$	$50S$	$100S$	$250S$	$500S$
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.4.2 ADDITIONAL VARIATIONS OF FIGURE 3

We compute the multiplier for a configuration and dataset at ϵ as follows: Using the median accuracy obtained through the experiments depicted in Tables A.2 to A.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 Figures A.9 and A.10 display the same analysis as Figure 3 for all backbones (VIT-B, R-50) and non-private shots of $S \in \{5, 10\}$.

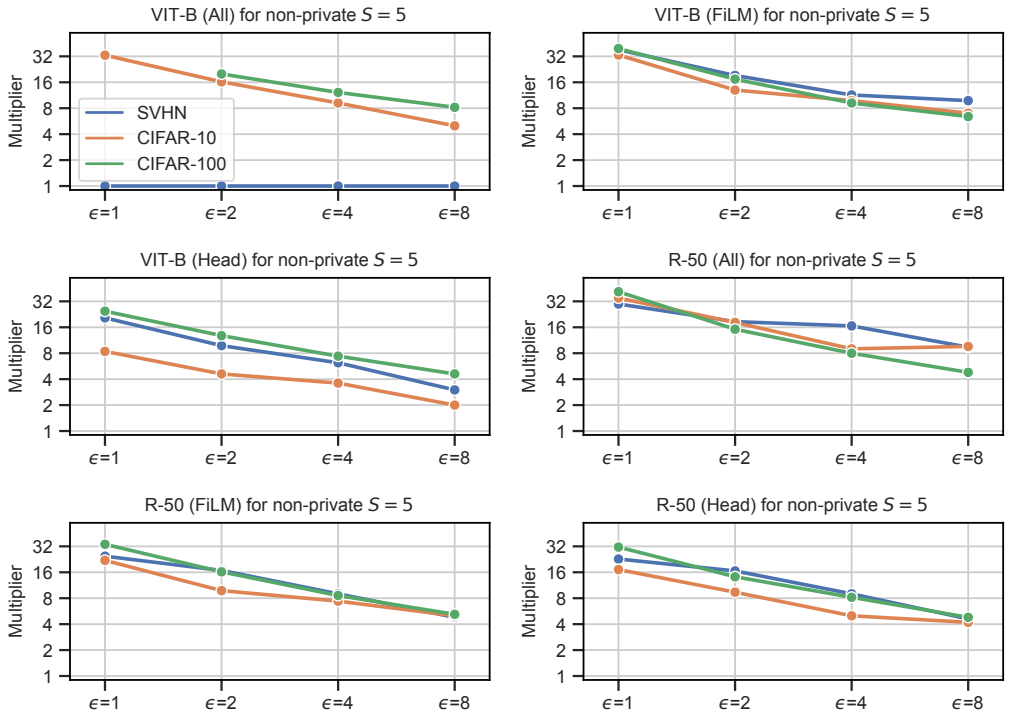

Figure A.9: Multiplier of shots required to reach same accuracy as non-private with $S = 5$ for VIT-B and R-50 on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Section 3.1.

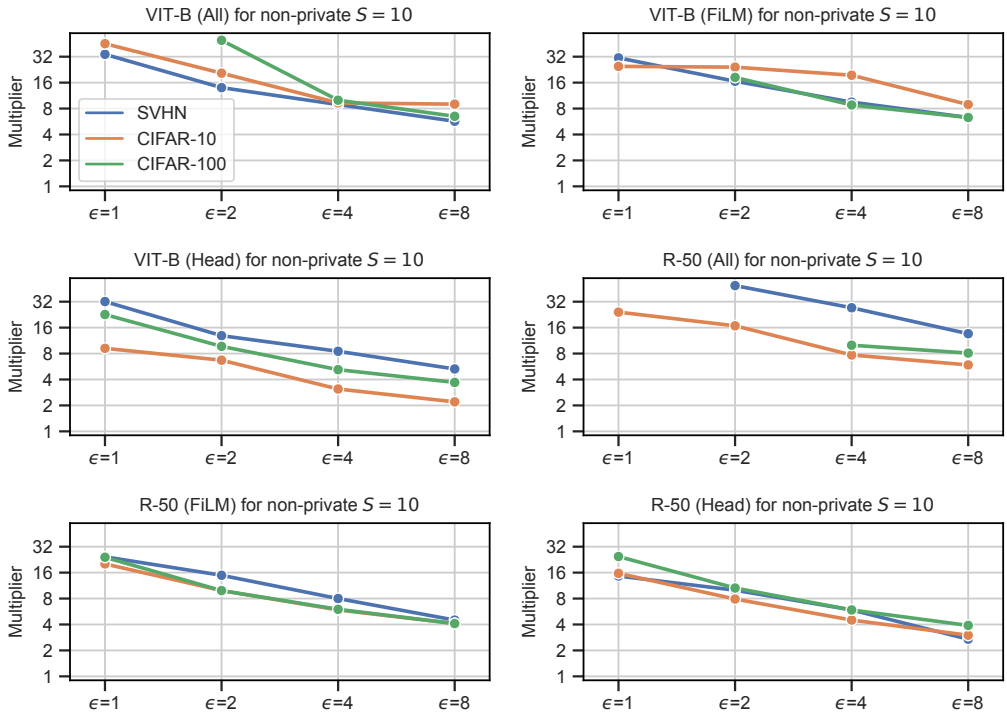

Figure A.10: Multiplier of shots required to reach same accuracy as non-private with $S = 10$ for VIT-B and R-50 on CIFAR-10, CIFAR-100 and SVHN with $\delta = 1/|\mathcal{D}|$. The data is obtained using linear interpolation of the median results of the experiments of Section 3.1.

A.4.3 Comparison of Backbones for Effect of Shots and DP

Figure A.11 compares the backbones (VIT-B, R-50) using their best performing configuration. The VIT-B backbone achieves comparable or better performance.

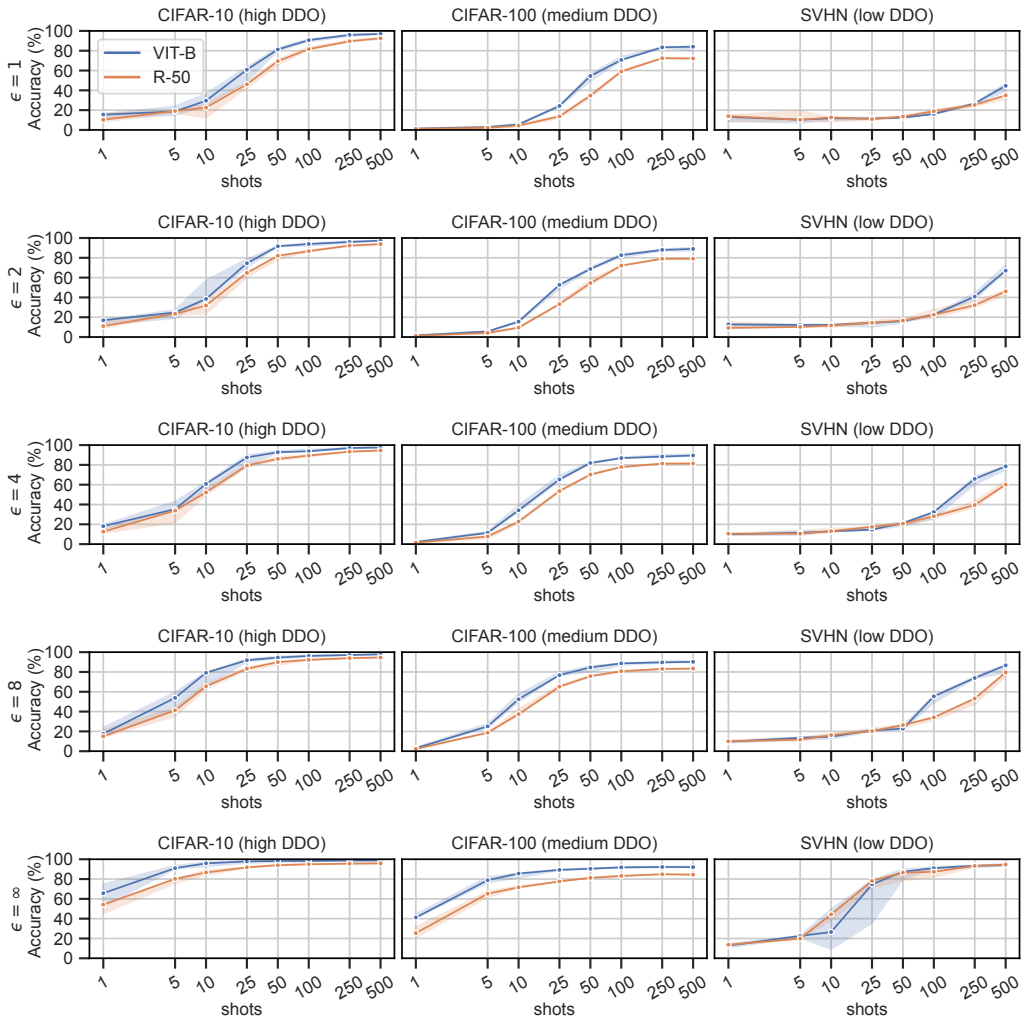

Figure A.11: Classification accuracy for different ϵ as a function of S and backbone (VIT-B, R-50) for CIFAR-10, CIFAR-100 and SVHN. DDO (low, medium, high) refers to the data distribution overlap and is computed as in Appendix A.1. The best performing configuration out of *All*, *FiLM* and *Head* for 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.4.4 ADVANTAGE OF *FiLM* AS A FUNCTION OF SHOTS

Figures A.12 to A.14 show the difference between the mean classification accuracy of *FiLM* and *Head*. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better.

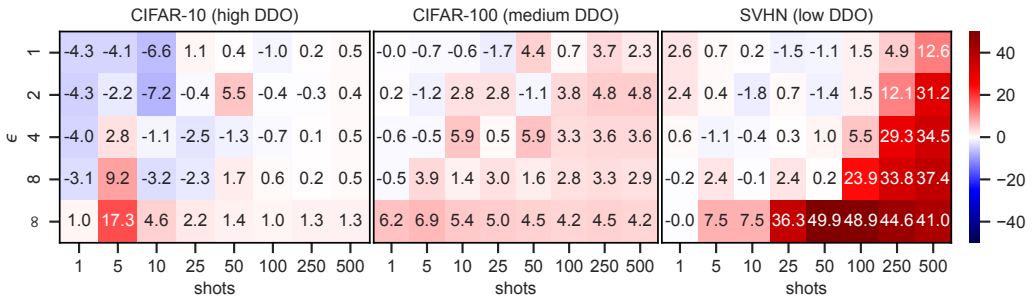

Figure A.12: Heat map showing the accuracy advantage of *FiLM* over *Head* for CIFAR-10, CIFAR-100 and SVHN as a function of ϵ. Backbone is VIT-B. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from highest to lowest DDO (data distribution overlap).

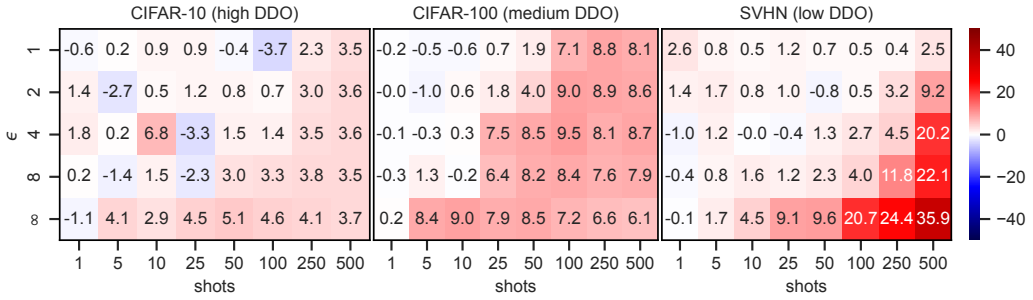

Figure A.13: Heat map showing the accuracy advantage of *FiLM* over *Head* for CIFAR-10, CIFAR-100 and SVHN as a function of ϵ. Backbone is R-50. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from highest to lowest DDO (data distribution overlap).

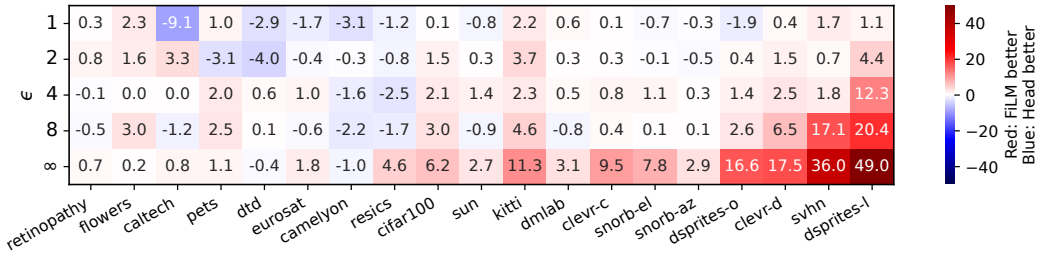

Figure A.14: Heat map showing the accuracy difference between *FiLM* and *Head* for the VTAB-1k datasets as a function of ϵ. Backbone is VIT-B. Darker red indicates *FiLM* is better. Darker blue indicates *Head* is better. Datasets ordered from highest to lowest DDO. $\delta = 10^{-3}$.

A.4.5 ADDITIONAL VTAB-1K RESULTS

Tables A.8 to A.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 A.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 $\pm$ 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 A.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	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
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 A.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	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
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 A.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	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
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 A.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	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
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 A.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	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
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

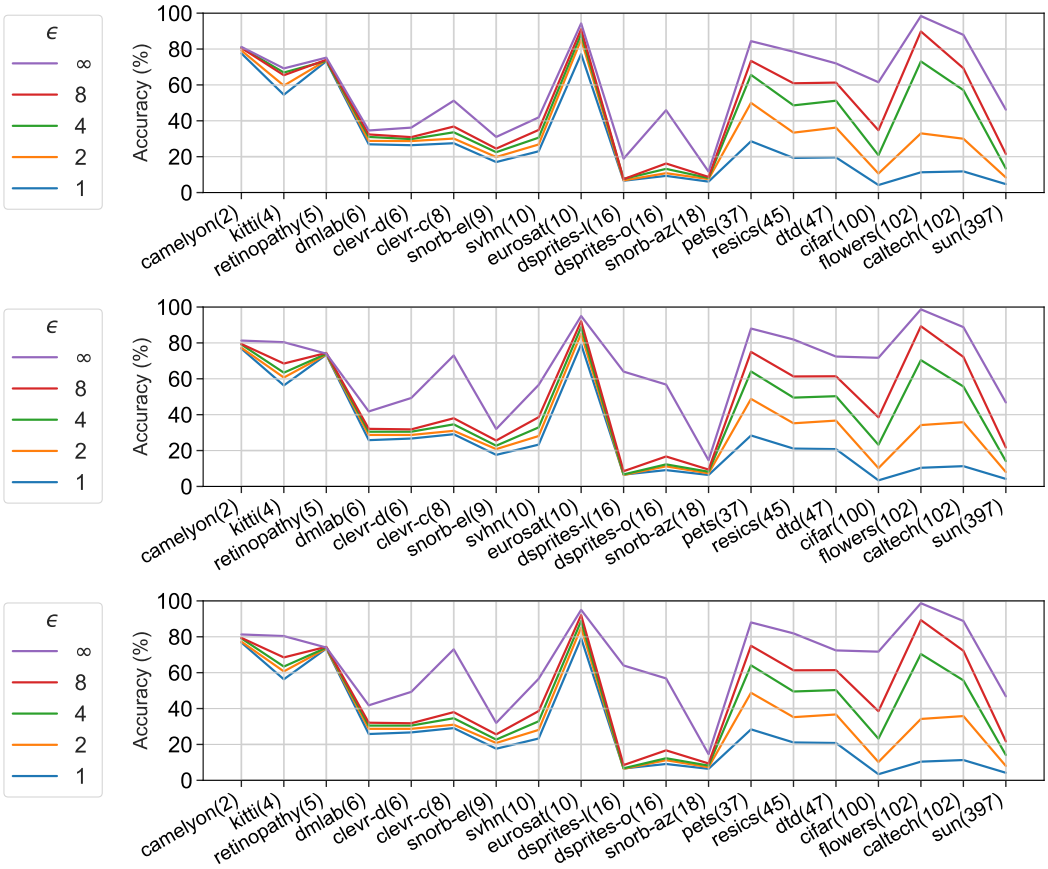

Figure A.15: Classification accuracy for VTAB-1k datasets as a function of privacy level (ϵ). Backbone is R-50. Top is *Head*, middle is *FiLM*, and bottom is *All*. The datasets are ordered increasingly by C (in parenthesis) or equivalently decreasingly by S as $|\mathcal{D}| = 1000$.

Figures A.15 and A.16 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 C or or equivalently decreasingly by S as $|\mathcal{D}| = 1000$ and $S = |\mathcal{D}|/C$.

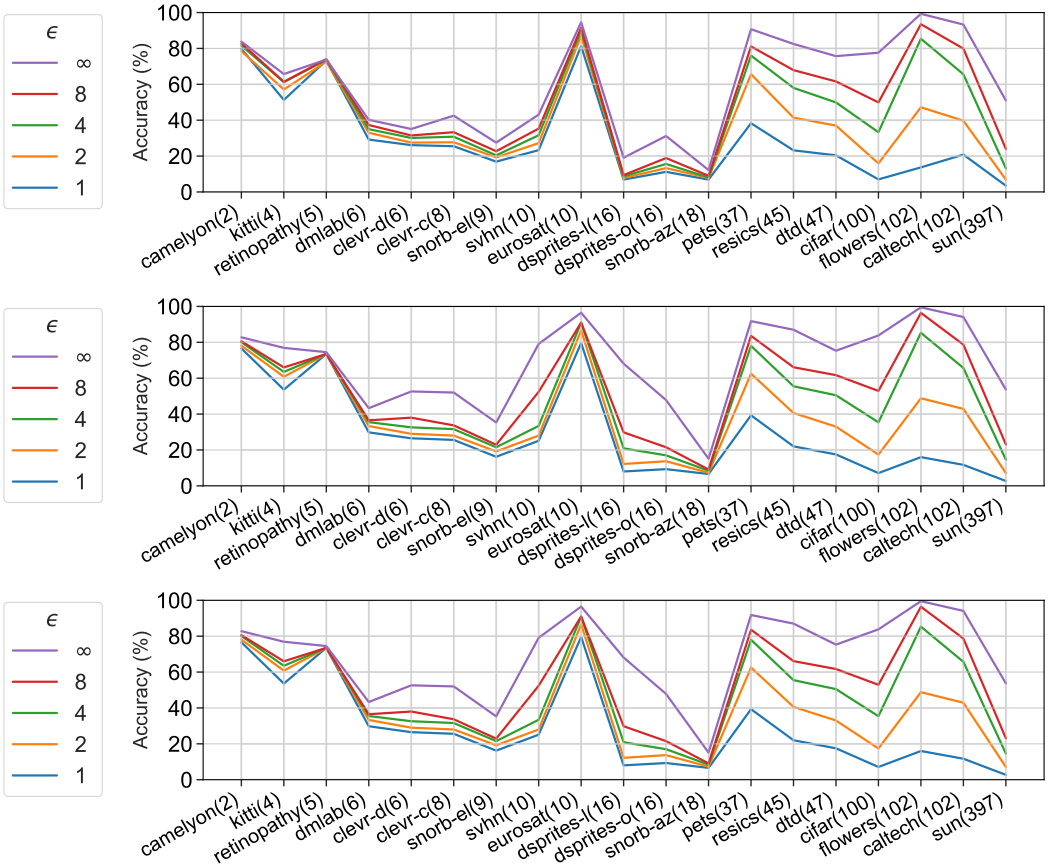

Figure A.16: Classification accuracy for VTAB-1k datasets as a function of privacy level (ϵ). Backbone is VIT-B. Top is *Head*, middle is *FiLM*, and bottom is *All*. The datasets are ordered increasingly by C (in parenthesis) or equivalently decreasingly by S as $|\mathcal{D}| = 1000$.

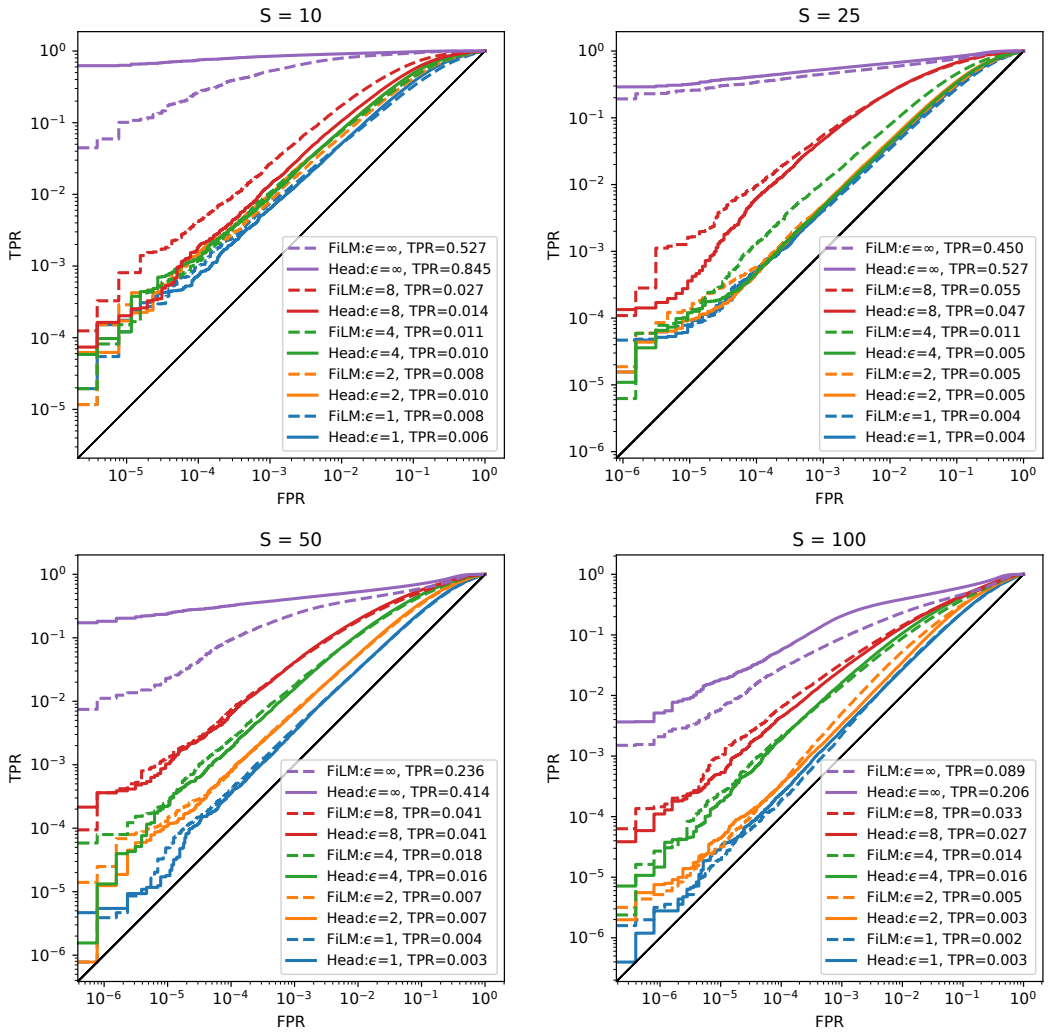

Figure A.17: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for various privacy levels (ϵ) and backbone configurations *Head* and *FiLM* at a fixed S. TPR values in legends are measured at FPR=0.001.

A.4.6 Additional Membership Inference Attack Results

Figure A.17 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.

Figure A.18 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 A.14 presents the True Positive Rates (TPR) at various False Positive Rates (FPR), along with their upper bounds (UB), 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 Figures A.17 and A.18.

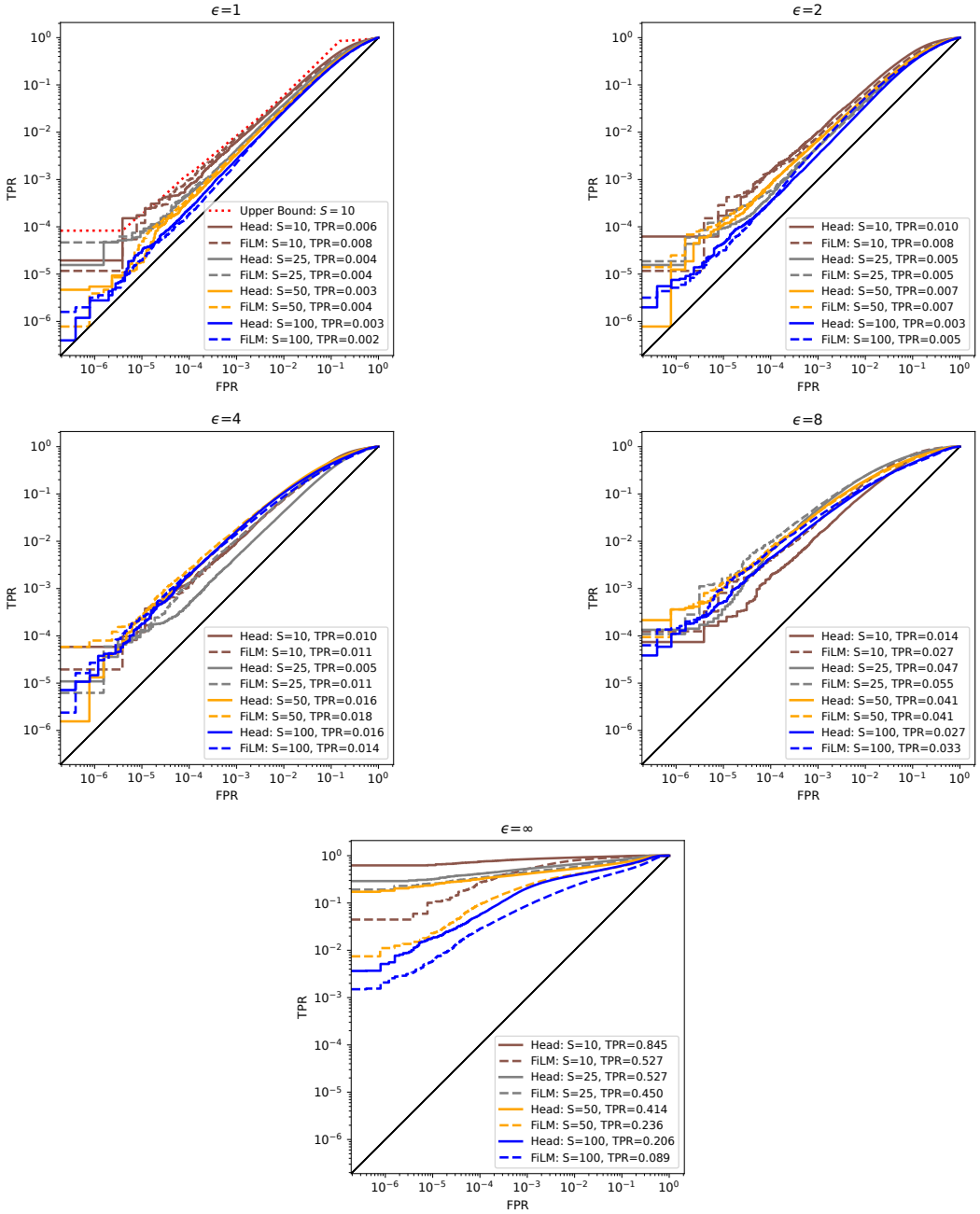

Figure A.18: ROC curves for LiRA (Carlini et al., 2022) on CIFAR-100 with R-50 backbone for various S at fixed privacy levels (ϵ) and backbone configurations *Head* and *FiLM*. TPR values in legends are measured at FPR=0.001. The dotted red curve on the $\epsilon = 1$ plot indicates the theoretical upper bound on TPR for $S = 10$ and $\epsilon = 1$.

Table A.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 Figures A.17 and A.18. Dataset ($\mathcal{D}$) is CIFAR-100. Backbone is R-50 pretrained on ImageNet-21k. The UB columns list the theoretical upper bound on TPR. See Appendix A.2 for details.

		TPR (%) @ 0.1% FPR			TPR (%) @ 1% FPR			TPR (%) @ 10% FPR			AUC		Attack Adv	
ϵ	S	Head	FiLM	UB	Head	FiLM	UB	Head	FiLM	UB	Head	FiLM	Head	FiLM
1	10	0.62	0.77	0.84	5.13	5.11	5.97	37.12	32.20	56.66	0.765	0.731	0.419	0.354
	25	0.43	0.40	0.71	3.85	3.45	5.87	30.14	27.37	57.52	0.721	0.702	0.343	0.312
	50	0.34	0.37	0.65	3.16	3.20	5.91	25.92	25.18	58.44	0.689	0.682	0.291	0.279
	100	0.27	0.22	0.63	2.70	2.79	6.0	23.77	23.82	59.67	0.670	0.668	0.260	0.252
2	10	1.02	0.79	3.99	7.90	6.47	33.8	48.20	41.37	97.3	0.819	0.788	0.509	0.454
	25	0.49	0.47	3.75	4.55	4.13	34.99	35.41	31.39	97.42	0.754	0.730	0.397	0.356
	50	0.66	0.71	3.71	5.42	5.36	35.84	35.87	34.30	97.48	0.751	0.740	0.380	0.362
	100	0.34	0.51	3.75	3.54	5.31	36.85	30.65	32.27	97.56	0.717	0.719	0.325	0.319
4	10	0.96	1.05	99.92	7.91	7.56	99.92	48.84	43.26	99.93	0.821	0.797	0.516	0.465
	25	0.47	1.07	99.92	4.24	7.86	99.92	33.77	45.66	99.93	0.744	0.805	0.382	0.472
	50	1.58	1.78	99.93	11.24	10.83	99.93	48.08	45.50	99.94	0.809	0.794	0.466	0.438
	100	1.63	1.42	99.93	10.63	8.89	99.93	43.09	38.18	99.94	0.778	0.749	0.402	0.356
8	10	1.36	2.68	100.0	10.57	17.41	100.0	55.23	68.03	100.0	0.847	0.895	0.558	0.643
	25	4.69	5.49	100.0	24.28	24.15	100.0	67.23	65.19	100.0	0.888	0.880	0.612	0.590
	50	4.12	4.05	100.0	19.25	18.36	100.0	56.89	52.75	100.0	0.845	0.825	0.520	0.480
	100	2.69	3.35	100.0	13.37	14.25	100.0	45.15	43.12	100.0	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.5 TRAINING AND EVALUATION DETAILS

A.5.1 FiLM LAYER IMPLEMENTATION

Table A.15 details the locations and count of the parameters that are updateable for the *FiLM* configuration in each of the backbones used in the experiments.

Table A.15: 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.5.2 HYPERPARAMETER TUNING

For all 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 A.16. 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 A.16: 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.5.3 EFFECT OF SHOTS PER CLASS AND DP EXPERIMENTS

For each evaluated configuration, we draw $|\mathcal{D}| = CS$ examples from the dataset training split, tune hyperparameters as described in Appendix A.5.2, 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.5.4 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.5.2, 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 A.17.

Table A.17: 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.5.5 MEMBERSHIP INFERENCE ATTACKS EXPERIMENTS

Our implementation is based on code from the TensorFlow Privacy library (Google, 2019). 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 Section 3.1 that yielded the highest accuracy for the particular configuration. Approximate training times for all 257 models in each configuration are listed on Table A.18. The value of ϵ did not alter the training times to a significant degree.

Table A.18: 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