Scalable Differential Privacy with Sparse Network Finetuning

Zelun Luo Daniel J. Wu Ehsan Adeli Li Fei-Fei
Stanford University

Abstract

We propose a novel method for privacy-preserving training of deep neural networks leveraging public, out-domain data. While differential privacy (DP) has emerged as a mechanism to protect sensitive data in training datasets, its application to complex visual recognition tasks remains challenging. Traditional DP methods, such as Differentially-Private Stochastic Gradient Descent (DP-SGD), perform well only on simple datasets and shallow networks, while recent transfer learning-based DP methods often make unrealistic assumptions about the availability and distribution of public data. In this work, we argue that minimizing the number of trainable parameters is the key to improving the privacy-performance tradeoff of DP on complex visual recognition tasks. Inspired by this argument, we also propose a novel transfer learning paradigm that finetunes a very sparse subnetwork with DP. We conduct extensive experiments and ablation studies on two visual recognition tasks: CIFAR-100 → CIFAR-10 (standard DP setting) and the CD-FSL challenge (few-shot, multiple levels of domain shifts) and demonstrate competitive experimental performance.

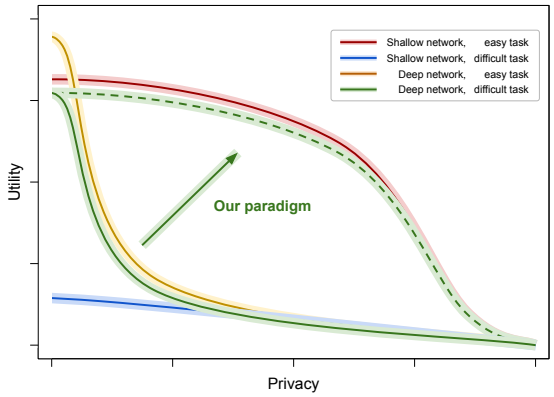

Figure 1. We propose a training framework that improves the privacy-utility trade-off for deep neural networks on complex visual recognition tasks. By introducing a novel transfer learning paradigm, our model trained with differential privacy is able to achieve a performance comparable to its non-private counterpart.

1. Introduction

As computer vision becomes increasingly ubiquitous, the robustness and privacy of vision models are a growing concern. In fact, there are ample examples of privacy attacks on standard deep learning models successfully revealing the contents of training data [40, 16, 35] – one attack was able to reconstruct credit card and social security numbers [7]. This is particularly concerning in the field of computer vision, where many applications, e.g., medical imaging, work with sensitive and legally-protected data. The onus of protecting private data falls upon machine learning practitioners, and, indeed, this responsibility may be encoded into law by regulations such as the EU's General Data Privacy Regulation (GDPR) and the California Consumer Privacy Act [11].

Although many notions of privacy have been proposed, notably including k-anonymity [43], and its extension l-diversity [31], differential privacy [13] has emerged as the gold-standard for the field. Differential privacy is a formalization of the notion of data privacy, providing strict upper bounds on the information about a data record, which may be obtained from resulting models [13, 9]. This is an attractive guarantee – many applications of computer vision involve sensitive datasets, which carry a strong obligation to protect user data in which the exposure of even singular data records is problematic [42, 3]. Differentially private models also benefit from several auxiliary guarantees, such as robustness to adversarial examples [27], and compliance under data privacy laws [11]. Although differential privacy is not a concept native to machine learning research, arising instead from research into database privacy [13], differentially private machine learning has been a burgeoning field of research [1].

In practice, there are many obstacles to building powerful differentially private machine learning systems. There is an inherent tradeoff between model utility and privacy [18] – larger networks, in particular, suffer from far greater disruption during training compared to their non-private

counterparts, which results in significant penalties to utility. This is due to the implementation of differentially private machine learning – namely, differential privacy requires bounding the influence of each example on the mini-batch gradient. Given a data sample x_i, the DP-SGD [1] algorithm clips the per-sample gradient $\mathbf{g}(x_i)$ in ℓ_2 norm, i.e., the gradient vector $\mathbf{g}$ is replaced by $\mathbf{g}/\max(1, \frac{\|\mathbf{g}\|_2}{C})$ for a clipping threshold C. It is evident that the norm is proportional to the number of training parameters, and thus will be large in deep neural networks, which leads to dramatically greater gradient clipping in larger networks. On the other hand, deep convolutional neural networks have enjoyed great success in large-scale image and video recognition tasks, and it has been shown that the depth of neural networks is crucial for the expressive power of deep learning [23].

In this work, we propose a novel solution to improve the privacy-utility tradeoff in deep neural networks with differential privacy (Figure 1). Our key idea is to leverage additional, public, datasets to instill strong representations in large models, which are then adapted to private datasets at a minimal privacy cost. To further minimize the negative effect of differential privacy, we minimize the number of trainable parameters to only those necessary for effective transfer learning. Not all neurons are created equal – in particular, we identify normalization parameters [8, 51, 5] as carrying domain-specific information, and find that the domain gap between public and private datasets can be significantly minimized by only finetuning these parameters. Besides, we draw insights from model pruning [15] and propose a novel approach to selecting and finetuning a very small subset of parameters in convolutional layers.

In summary, our key contributions are as follows:

1. We develop a method for effectively scaling differential privacy to large neural networks, by leveraging out-of-domain transfer learning and sparse network finetuning.
2. We enable differential private training of models on extremely small (few-shot) private datasets at a reasonable privacy cost.
3. We achieve state-of-the-art performance with a smaller privacy budget on CIFAR-10, a prototypical benchmark for differentially private machine learning.

2. Related Work

The challenge of scaling up machine learning models with limited access to data is not a new problem; in this section, we review recent discoveries in the fields of non-private few-shot learning and domain adaptation and draw parallels to our work.

2.1. Differential privacy

Differential privacy was first proposed in the context of securing statistical databases [13, 2] against user queries, but quickly gained traction in the machine learning community in the context of adversarial privacy attacks; formally, a model acts as a database to which an adversary may submit an arbitrary number of queries.

Research in differentially private machine learning models tracks a relaxed variant of differential privacy, known as Renyi differential privacy (RDP) [32]. Deep learning models attain RDP guarantees via two alterations to the training process: the clipping of per-sample gradients, and the addition of Gaussian noise to gradients, collectively known as DP-SGD [14, 1, 20]. Since the amount of noise added is a function of the number of parameters in the model, the large model architectures favored in non-private settings often do poorly under the regime of differential privacy. Furthermore, privacy guarantees decay as the number of training iterations increases, so previous work in differential privacy has focused primarily on lightweight architectures which may be trained rapidly, on relatively simple datasets [33, 44].

2.2. Differential privacy with additional data

One promising avenue of research has been the application of representation learning and transfer learning to differentially private training. A good initial representation, garnered via pre-training on a public dataset, can often offer reasonable accuracy on the target task with minimal training time, and thus minimizing exposure to noise and to private data.

Generally, the private data is assumed to be small, and the task is to produce a model that achieves reasonable accuracy on a large public dataset in the same domain as the private data, minimizing the privacy leakage of the training data by leveraging a learned representation. PATE trained an ensemble of teacher models, which was then used to train a student model, where ensemble voting and selective response to student queries provided a suitable degree of obfuscation [39]. Similarly, Private-KNN trained a feature extractor on private data and classified queries via the k-nearest neighbors in the feature space, where a random subsampling of private data was used to augment the privacy guarantee [55].

2.3. Few-shot learning

In settings where our private dataset is small, finetuning models on this private data incurs a large privacy cost, as individual data samples may be seen many times over the course of finetuning the model up to a reasonable accuracy. This is quite similar to the setting considered in few-shot learning, where the goal is to quickly achieve competitive

and generalizable performance with exposure to only a few examples of each class.

Much of the research in few-shot learning has focused on meta-learning [53, 30, 54, 41]. Although these meta-learning methods can be successfully applied under the setting of differential privacy, the construction and use of a meta-learner incurs an additional privacy cost [28]. However, recent work [45] suggests that the performance of these meta-learning algorithms has yet to outperform simple finetuning on top of a pre-trained embedding, and that a good representation space produces a strong few-shot learner.

2.4. Transfer learning and domain adaptation

To produce a good representation, we naturally wish to leverage large public datasets; transfer learning is a well-known method in which external datasets are used to pre-train models, leading to faster convergence and greater utility resulting from a strong initial representation space [4, 36]. These large datasets are generally quite different in domain from our private data, particularly when our private dataset is small or specialized. To tackle this, we turn to the sizable corpus of work in domain adaptation. There are three main tracks of research in this area [50]: methods incentivizing the learning of domain-agnostic features via direct optimization [17], adversarial approaches [47, 25], and data-reconstruction approaches [6]. Most of these methods involve whole-model retraining and sizable data from the target domain, making them impractical under the regime of differential privacy.

There is promise in retraining only some part of the model; in particular, the tuning of normalization layers has been shown to improve model robustness [5], and indeed, improve domain adaptation [8, 49]. This method is encouraging because it is lightweight – normalization layers contain a minuscule fraction of the total parameters of a model – but also maintains representational power, as the bulk of network parameters are isolated from the noise introduced by DP-SGD.

3. Groundwork

Definition 1 (Differential Privacy). Suppose a model $M : \mathcal{D} \to \mathcal{R}$ is trained on two datasets, $D, D' \in \mathcal{D}$, which differ only by a single data record. Then, for any subset of outputs $R \in \mathcal{R}$, the model is said to satisfy (ϵ, δ)-differential privacy [13] if

$$\mathbf{Pr}[M(D) \in R] \leq e^{\epsilon} \cdot \mathbf{Pr}[M(D') \in R] + \delta.$$

In other words, ϵ bounds the privacy loss on any individual sample, and δ is the probability that this bound does not hold.

Definition 2 (Rényi Differential Privacy). Rényi Differential Privacy (RDP) is a generalization of (ϵ, δ)-Differential Privacy that uses Rényi divergence as a distance metric. The Rényi divergence of order α between two distributions P and Q is defined as:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim P} \left[\left(\frac{P(x)}{Q(x)} \right)^{\alpha - 1} \right].$$

A model satisfies (α, ϵ)-RDP if

$$D_\alpha(M(D)\|M(D')) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim M(D)} \left[\left(\frac{\mathbf{Pr}[M(D) = x]}{\mathbf{Pr}[M(D') = x]} \right)^{\alpha - 1} \right] \leq \epsilon.$$

It can be shown [32] that pure $(\epsilon, 0)$-differential privacy is equivalent to (∞, ϵ)-RDP, and, further, that if a model M satisfies (α, ϵ)-RDP, then M also satisfies $\left(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta\right)$-differential privacy for any $\delta \in (0, 1)$.

RDP in machine learning. The implementation of RDP-certifiable training methods in machine learning depends on two components: per-sample gradients are clipped at some fixed L2 norm threshold C, and Gaussian noise of magnitude $\sigma^2 C^2$ is added to the gradient updates for a cleverly chosen noise scale parameter σ. This training procedure, discussed in greater detail in Section 4.1, is critical to our method's inspiration.

Definition 3 (Moments Accountant). The moments accountant is a method to measure the privacy cost ϵ incurred in the training of a model according to the DP-SGD algorithm summarized above, and discussed in detail by Abadi et al [1]. Suppose a model M is trained for T steps with a batch size L on a dataset of size N. Then, there exists two constants c_1, c_2 for which for any $\epsilon < \frac{c_1 L^2 T}{N^2}$, the model is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose the noise scale σ to be

$$\sigma \geq c_2 \frac{L\sqrt{T \log(1/\delta)}}{N\epsilon}.$$

Note that as the number of training steps T increases, or as the size of the dataset N decreases, the privacy cost ϵ increases, making small, challenging datasets particularly difficult in the differentially private setting.

4. Our Approach

Our method aims to improve the privacy-utility trade-off of deep learning models trained on private datasets. In Section 4.1, we explain why it is difficult to apply differential privacy on large-scale visual recognition tasks. Then,

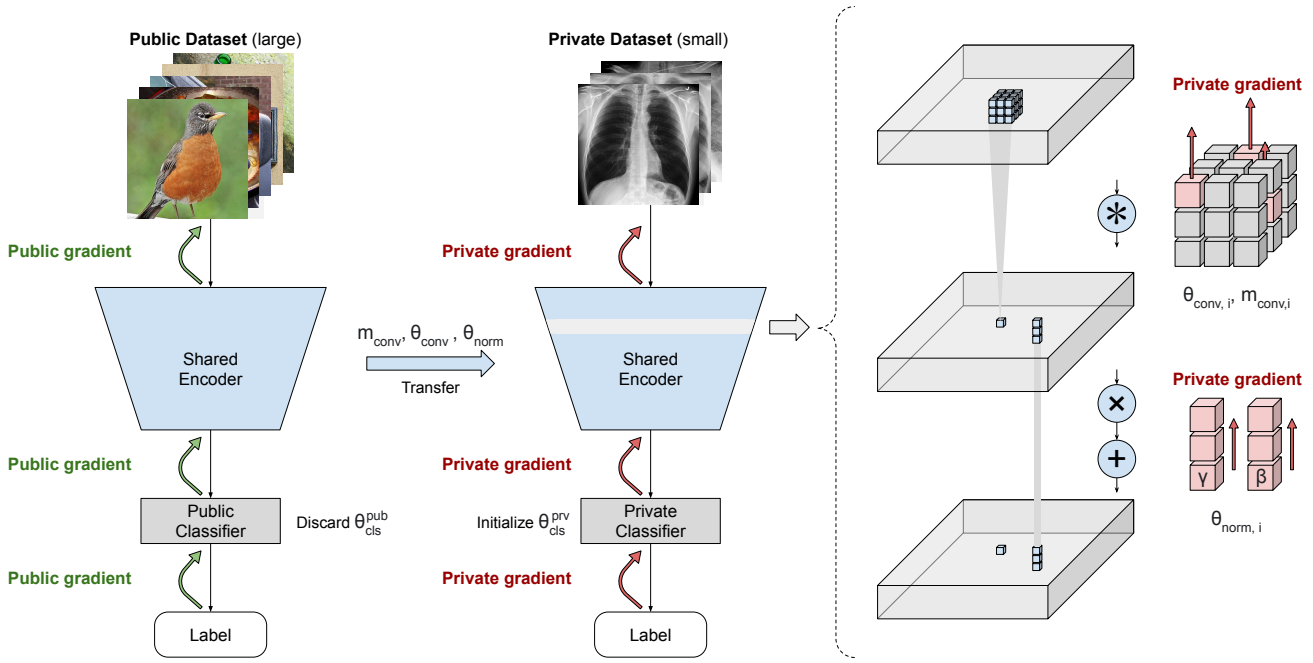

Figure 2. An overview of our method. A model is first pre-trained on public data (left). Then, convolution parameters θ_{conv} and normalization parameters θ_{norm} are transferred to private data (middle). We also construct a sparse mask m_{conv} representing the important, high-magnitude convolution parameters. During private data finetuning, only classifier parameters θ^{prv}_{cls}, normalization parameters θ_{norm} and unmasked convolution parameters $m_{conv} \bigodot \theta_{conv}$, in red, receive gradient updates (right).

in Section 4.2, we discuss a naive solution – vanilla transfer learning – that inspired our approach. Finally, in Section 4.3, we introduce our solution by first explaining our overall framework, and then the technical components that facilitate differentially private training.

Notation. We introduce a transfer learning algorithm which transfers information from a large, public dataset $\mathcal{D}_s = \{(X_s, y_s)\}$ to a small, private dataset $\mathcal{D}_t = \{(X_t, y_t)\}$. We first pre-train the model $f^{pub}_\theta(x)$ with parameters $\theta = [\theta_1, \theta_2]$ on the public dataset with an optimizer O_1, then fine-tune a subset of the model parameters θ_1 with an differentially-private optimizer O_2 to produce a private model $f^{prv}_\theta(x)$.

4.1. Challenges of deep private models

In the realm of non-private learning, it is well known that more complex tasks require deeper, larger, and more expressive models. However, larger models are not favored in the differential privacy literature, and state-of-the-art results have largely been with small models on simple datasets [33]. The lackluster performance of large models is because as the number of trainable parameters increases, so too does the overall L2 norm of the per-sample gradient. This means that for a fixed clipping magnitude C, the gradient on each weight must be clipped more aggressively, often leading to the added Gaussian noise overwhelming the gradient on each weight.

To demonstrate this phenomenon, suppose we naively train a ResNet-18 ($\approx$ 11 million parameters) on CIFAR-10 under DP-SGD with a generously low noise scale of $\sigma = 1$, which only allows for 13 epochs before we breach $\epsilon > 2.0$. Even so, since the number of parameters is so large, the magnitude of the clipped gradients on each weight is significantly smaller than the magnitude of the noise we add (Figure 3), which leads to poor performance compared to shallow networks. In short, an increase in the number of trainable parameters dramatically increases the amount of training disruption caused by DP-SGD. Thus, in order to make differentially private models competitive on large datasets, our goal is to minimize the number of trainable parameters during differentially-private training, while maintaining the expressive power of deep networks.

4.2. Transfer learning

Transfer learning from a public dataset to a private dataset is a natural solution in the context of learning with differential privacy [37]. Previous studies have shown that transfer learning can reduce the generalization error as well as increase the convergence speed, effectively reducing the spent privacy budget. In our parlance, transfer learning considers a model $f^{pub}_\theta(x)$ composed of encoder parameters θ_{enc} and classifier parameters θ^{pub}_{cls}. First, $f_\theta(x)$ is trained

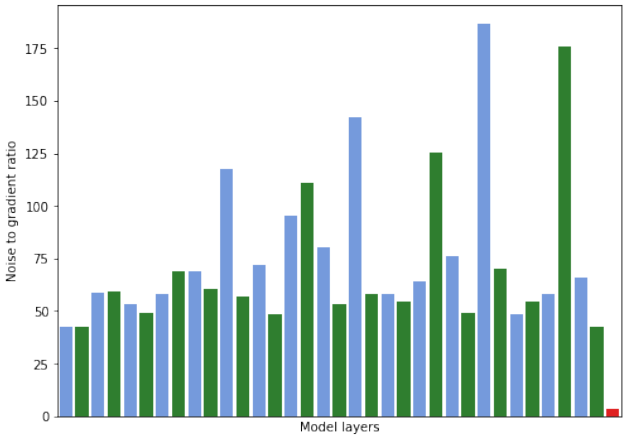

Figure 3. The noise to gradient ratio in each layer. While training ResNet-18 on CIFAR10 with differential privacy, we see that the magnitude of Gaussian noise added greatly overwhelms that of the clipped gradients in convolution layers (blue) and normalization layers (green), and slightly overwhelms that of the classification layer (red).

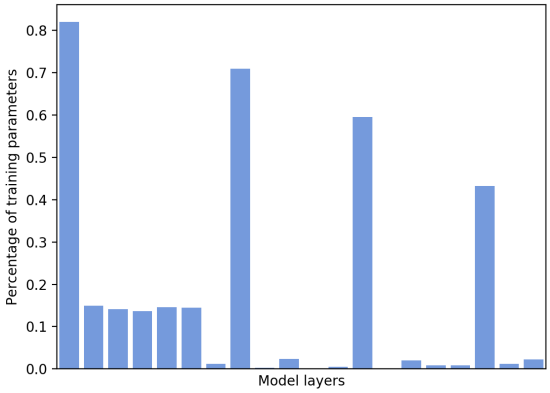

Figure 4. The percentage of training parameters in each layer of ResNet-18 during finetuning. Our method chooses to finetune the first few layers and all three 1×1 convolutional layers heavily and while keeping most of the parameters in deeper layers frozen. This coincides with the findings in domain adaptation that the earlier layers are critical for handling domain shift.

on $\mathcal{D}_s$, and then the transfer model $f_\theta^{prv}(x) = \{\theta_{enc}, \theta_{cls}^{prv}\}$, composed of the source model's encoder parameters and a fresh set of classifier parameters, is finetuned on $\mathcal{D}_t$. We consider both classifier finetuning, in which only θ_{cls}^{prv} is trained on $\mathcal{D}_t$, and whole-network finetuning, as benchmarks for our method. Unfortunately, these vanilla transfer learning methods are generally uncompelling; the former suffers from a low model utility, while the latter suffers from high privacy cost (Section 5.7).

4.3. Network Sparsity

The mission is clear – we want to pick some subset of model parameters to train under differential privacy. Choosing larger subsets increases the adaptability of the model, but also increases the amount of disruption in training; we think that finding the balance between these two factors is a field of research unto itself.

At first glance, the minimal set of parameters we must tune are the classifier parameters θ_{cls}^{prv}, i.e., vanilla transfer learning. As demonstrated in Table 1, however, this approach fails to bridge the domain gap, leading to low accuracy.

In this work, we pilot two simplistic ways of selecting a subset of trainable parameters inspired by other fields of research. First, we finetune only the normalization and classification parameters under differential privacy. This is inspired by work in domain adaptation that found retraining only normalization layers could capture the vast majority of the performance of whole-network finetuning [8, 51, 45], and indeed, that even simply recomputing the running data statistics of normalization layers can effectively close the gap [29]. We note that although prior work largely considers adaption of batch normalization layers, batch normalization is incompatible with the computation of per-sample gradients in DP-SGD, and so we tune group normalization as a close analog. The parameters in normalization layers generally account for a minuscule proportion ($\ll 1\%$) of the total parameter count in these models, which makes this methodology particularly lightweight; the amount of additional training disruption as compared to vanilla classifier finetuning is negligible.

Second, we finetune a subset of convolution parameters, *in addition* to tuning normalization and classification parameters. The convolution parameters make up the bulk of total model parameters in image recognition models, and tuning all convolution parameters leads to high training disruption, as demonstrated in Figure 3. Not all parameters are created equal; our method selects a small subset of convolution parameters to finetune.

4.4. Selecting transfer parameters

Formally, we consider a vision model $f_\theta(x)$ composed of convolution parameters θ_{conv}, normalization parameters θ_{norm}, and classifier parameters θ_{cls}. As a generalization of vanilla transfer learning, we first train a source model

$$f_\theta^{pub}(x) = \{\theta_{conv}, \theta_{norm}, \theta_{cls}^{pub}\}$$

on $\mathcal{D}_s$, and then construct a transfer model

$$f_\theta^{prv}(x) = \{\theta_{conv}, \theta_{norm}, \theta_{cls}^{prv}\}.$$

Then, we choose some subset $\tau \subset \bigcup\{\theta_{conv}, \theta_{norm}, \theta_{cls}^{prv}\}$ to finetune on $\mathcal{D}_t$, while freezing the rest of the parameters.

For those parameters not in τ, we still compute their gradients during the backward pass, but we simply do not update those model parameters, thereby not increasing the norm of the effective gradient update.

Procedure 1: Normalization layer transfer. First, we take $\tau = \theta_{norm} \bigcup \theta_{cls}^{prv}$. When tuning normalization layers, we first pre-train models with group normalization on our public dataset, and then finetune the normalization layers on our private dataset, both tuning normalization parameters, as well as recomputing running normalization statistics. This approach errs on the side of minimal training disruption, at the cost of model expressiveness.

Procedure 2: Convolution parameter transfer. We refine our first approach by choosing a small set of convolution parameters $\hat{\theta}_{conv} \subset \theta_{conv}$, to tune *in addition* to tuning normalization and classification parameters. We pick

$$\hat{\theta}_{conv} \coloneqq \{\theta : \theta \in \theta_{conv}, \|\theta\| > t\}$$

for some magnitude threshold $t > 0$. We select t dynamically so that we isolate out some fixed proportion of parameters p, i.e. $|\hat{\theta}_{conv}| = p|\theta_{conv}|$, and in doing so, we create a mask m_{conv} over θ_{conv} so that $\hat{\theta}_{conv} = m_{conv} \bigodot \theta_{conv}$.

This procedure is inspired by work done in model pruning. Pruning is an active area of research focused on decreasing the memory footprint of models at a minimal penalty to model utility. Pruning has been fairly successful; the Lottery Ticket Hypothesis paper found that by staged pruning of convolution parameters with low magnitude, models could be reduced by one to two orders of magnitude with a little-to-no decrease in model utility [15]. While the original methodology in the non-private setting simply trains the network on the target dataset, we note that pruning based on the target private dataset incurs an additional privacy cost [19]. Instead, we identify important parameters (i.e., those with large magnitudes) based on single-stage model training on our public dataset, under the hypothesis that sub-networks that are valuable on the public dataset will continue to be useful on the private data. Then, rather than pruning out unimportant parameters, we instead freeze them during training on the private data, masking out their gradient updates with m_{conv}, so that they may still provide some model utility without exacerbating training disruption. This procedure can be summarized as follows:

1. Train the source model $f_\theta(x)$ on our public dataset from scratch.
2. Select $p\%$ of convolution parameters with the highest magnitude, creating a mask m_{conv}.
3. Identify the set of transfer parameters τ, composed of our selected convolution parameters, as well as normalization and classifier parameters.
4. Finetune only the set of transfer parameters τ with DP-SGD on the private dataset, to produce a differentially private transfer model $f'_\theta(x)$.

We emphasize that under the paradigm of DP-SGD, the privacy ϵ is independent of the choice of sparsity parameter p; however, higher sparsity reduces the magnitude of noise and severity of clipping done during training, leading to faster convergence and thus smaller ϵ.

5. Experiments

We apply our method to two vision tasks: first, we tackle image classification on CIFAR-10 [26], using CIFAR-100 as the public data. Then, we consider the context of very small, out-of-domain, private datasets, and attempt the Cross-Domain Few-Shot Learning (CD-FSL) challenge [22], with a subset of ImageNet [12] as our public data.

5.1. Datasets

CIFAR-10 and CIFAR-100. CIFAR [26] is a well-known subset of the 80 million tiny images dataset [46], which is commonly used as an image classification benchmark. There are two disjoint variants: CIFAR-10, which contains 10 object classes, and CIFAR-100, containing 100 object classes. These datasets each contain 60,000 32x32 RGB images split evenly among their object classes. This is the standard sandbox for research in differential privacy, and indeed, shallow networks trained with differential privacy are able to achieve high accuracies with little difficulty.

CD-FSL. The CD-FSL challenge [22] contains a single source domain, miniImageNet [48], which is a subset of ImageNet [12] containing 60,000 84 × 84 RGB images evenly split among 100 classes. MiniImageNet is commonly used for in-memory fast prototyping. The challenge also contains four target image classification datasets, with progressively increasing domain differences from miniImageNet: CropDiseases [34], EuroSAT [24], ISIC2018 [10], and ChestX [52], which contains images of plant disease, satellite photography, skin lesions, and X-ray scans, respectively.

We use this challenge to highlight the two main foci of our method: our approach enables differentially private models to learn across domain gaps, and with minimal exposure to the private data. This is a close analog to real-world use cases, where small private datasets, e.g., the chest X-ray scans of a single clinic, are likely to lack large public in-domain datasets, and for which the size of the private dataset necessitates a few-shot approach.

5.2. Metrics

Since all training tasks are image classification challenges, we use the top-1 accuracy as our metric for

model utility. For the CD-FSL challenge, we follow the challenge's proposed assessment regime [21], and evaluate 5-way top-1 accuracy under 5, 20, and 50-shot constraints. We assess only on the CropDiseases, EuroSAT, and ISIC2018 datasets, and assess trained models on each dataset with 60 randomly selected few-shot 5-way classification trials.

Our primary metric for the privacy expenditure of a model is the tolerance parameter ϵ. For models trained under the regime of differential privacy, we report model accuracy at $\epsilon = \{0.5, 1, 1.5\}$.

5.3. Baselines and prior state-of-the-art

We compare our method against models trained with DP-SGD from scratch [1], transfer models first trained on public data, and with state-of-the-art DP methods that utilize additional data [55]. We consider two variants of transfer models: models in which only the classification parameters are finetuned, and whole-model finetuning. To the best of our knowledge, Private-KNN is currently the state-of-the-art for differentially private learning on CIFAR-10, and so we use this method as our benchmark. Finally, we consider models trained without differential privacy to be our upper bound.

5.4. Comparison with state-of-the-art

Our method is architecture-agnostic, so we simply chose a prominent image recognition architecture – ResNet-18 – for all of our experiments. As mentioned above, because batch normalization layers are incompatible with the computation of per-sample gradients necessary for DP-SGD, we modify our ResNet to use group normalization instead. Our results are given in Table 1. Our method outperforms previous methods in both privacy budget usage and accuracy – we bridge the gap between vanilla DP-SGD and non-private training.

5.5. Few-shot learning

To show that our method performs well under a low-data regime, we compare our method against our baselines in the context of 5, 20, and 50-shot learning (Table 2), on the CD-FSL challenge. The results show that our differentially private model achieves compelling results under a low-data regime, and outperforms state-of-the-art non-private meta-learning methods for few-shot learning.

5.6. Domain gaps

We also investigate the effect of domain gaps on our methodology, using the CD-FSL challenge. The three datasets we consider: CropDisease, EuroSAT, and ISIC2018, have progressively larger domain gaps from the source domain, MiniImageNet. The CropDisease dataset is simply a different classification task than MiniImageNet, but still contains RGB images of natural objects with perspective. EuroSAT is composed of natural RGB images, but lacks perspective, and finally, ISIC2018 contains medical RGB images without perspective. The results of our methodology across these domain gaps is summarized in Table 2.

Table 1. CIFAR10 results. Our number for "Ours (public)" was obtained by finetuning a ResNet-18 without differential privacy on CIFAR10. Both Private-KNN and the original DP-SGD paper used CIFAR-100 as an additional dataset for model pre-training. The model used in the original DP-SGD paper was a 5-layer CNN, far smaller than our ResNet-18.

Method	ϵ	Accuracy
DP-SGD [1]	2.00	0.6700
Private-KNN [55]	2.92	0.7080
DP-SGD [1]	4.00	0.7000
Ours (private)	0.50	**0.7328**
Ours (private)	1.00	**0.7664**
Ours (private)	1.50	**0.8157**
Ours (public)	∞	0.9410

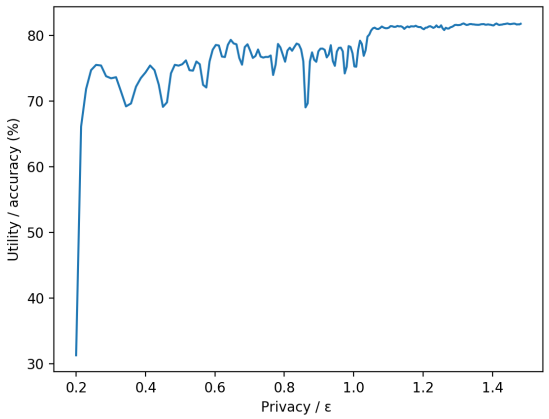

Figure 5. ϵ-accuracy tradeoff of our method on CIFAR10. With transfer learning and sparse finetuning, our model quickly converges to a reasonable performance at very low privacy budget ($\epsilon \approx 0.25$).

5.7. Ablation

We deconstruct the components of our proposed method, and demonstrate the performance gain from each element on CIFAR-10 (Table 3). Note that although classifier-only finetuning converges quickly and at a low privacy budget, the overall performance of the model is lackluster. We ascribe this to the limited ability of the final classification layer to effectively undertake domain adaptation. Indeed,

Table 2. Results across domain gaps. MatchingNet, RelationNet, and ProtoNet are non-private meta-learning methods [21].

	Plant Disease (small gap)			EuroSAT (medium gap)			ISIC2018 (large gap)		
Method	5-shot	20-shot	50-shot	5-shot	20-shot	50-shot	5-shot	20-shot	50-shot
MatchingNet	0.6639	0.7638	0.5853	0.6445	0.7710	0.5444	0.3674	0.4572	0.5458
RelationNet	0.6899	0.8045	0.8508	0.6131	0.7443	0.7491	0.3941	0.4177	0.4932
ProtoNet	0.7972	0.8815	0.9081	0.7329	0.8227	0.8048	0.3957	0.4950	0.5199
Ours	0.8715	0.9349	0.9687	0.7933	0.8728	0.9008	0.4648	0.5979	0.6377

Table 3. An ablation study of our method on CIFAR10. We finetuned only the classification layer, the classification and normalization layers, and the entire method – finetuning classification, normalization, and a subset of convolution parameters. All three methods achieve decent accuracy with a small privacy budget ($\epsilon = 0.5$), but the flexibility and expressive power granted by the finetuning of additional parameters allows our method to achieve higher accuracies compared to vanilla transfer learning.

Method	$\epsilon = 0.50$	$\epsilon = 1.00$	$\epsilon = 1.50$
Cls only	0.7068	0.7128	0.7120
Cls+conv	0.7370	0.7278	0.7621
All	0.7328	0.7664	0.8157

as we enable the finetuning of both the normalization layers and a subset of convolution parameters, we see that model utility significantly increases with little cost to privacy.

5.8. Implementation details

We implement our method using the Opacus [1] library. Here, we describe the implementation details.

Privacy. We hold δ constant at 10^{-5}. We use a clipping threshold of $C = 1$, and a noise multiplier of $\sigma = 15$.

Pre-training. For our experiments on CIFAR, ResNet-18 is first pre-trained on CIFAR-100, while for CD-FSL, ResNet-18 is pre-trained on MiniImageNet. These models are pre-trained for 200 epochs, with a batch size of 128. We optimize the models with SGD with momentum, with an initial learning rate of 0.1, a momentum coefficient of 0.9, and weight decay of 10^{-4}. We use cosine learning rate decay over the full 200 epochs.

Finetuning. During private finetuning, we do not perform any random data augmentation, as DP-SGD is already a strong regularizer, and augmented data incurs the same privacy cost with less utility than simply training on unaugmented images. For finetuning on CIFAR, we take the ResNet-18 which was pre-trained on CIFAR-100, and finetune it on CIFAR-10; similarly, for CD-FSL, we take the MiniImageNet-pre-trained ResNet, and finetune it on each of our three target datasets. Finetuning was done with DP-SGD for 200 epochs, with a batch size of 5000. Large batch sizes have been shown to assist differentially private machine learning [38], and we achieve this batch size via virtual gradient updates. We optimize the model with SGD with momentum, with an initial learning rate of 0.8 on the classification parameters and of 0.01 on the normalization and convolution parameters, a momentum coefficient of 0.9, and no weight decay. We use linear learning rate warmup to our initial learning rate during the first epoch, and cosine learning rate decay over the full 200 epochs.

Pruning. We select the magnitude threshold t so that only $p = 1\%$ of the convolution parameters receive gradient updates from the private data. We note that we cannot tune over the parameter p without violating the privacy constraint, so we selected this setting with heuristics from previous work [15].

6. Conclusion

We propose a simple yet effective method to scale up differential privacy to large neural networks at reasonable privacy budgets. Our key insight was to minimize the number of trainable parameters during private dataset finetuning, by leveraging additional public data. By pre-training large models on public data, we obtain a strong representation at no privacy cost. Next, finetuning a small subset of parameters on private data, we maintain the expressiveness of large models while introducing minimal training disruption during the process of domain adaptation.

We note that our two proposed approaches – normalization transfer and convolution parameter transfer – albeit outperforming previous methods, are naive stabs-in-the-dark; an exploration into the precise parameter subset to choose in order to optimize over the fundamental privacy-accuracy tradeoff in differential privacy is likely to be a fruitful area of research.

[1] https://opacus.ai/

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Nabil R Adam and John C Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

[3] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

[4] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[5] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. *arXiv preprint arXiv:2010.03630*, 2020.

[6] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016.

[7] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284, 2019.

[8] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[9] Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–93. IEEE, 2013.

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[11] Rachel Cummings and Deven Desai. The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[15] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[16] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[18] Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Privacy and utility tradeoff in approximate differential privacy. *arXiv preprint arXiv:1810.00877*, 2018.

[19] Lovedeep Gondara, Ke Wang, and Ricardo Silva Carvalho. The differentially private lottery ticket mechanism. *arXiv preprint arXiv:2002.11613*, 2020.

[20] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and Alex Kai Qin. A survey on differentially private machine learning. *IEEE Comput. Intell. Mag.*, 15(2):49–64, 2020.

[21] Yunhui Guo, Noel C. Codella, Leonid Karlinsky, James V. Codella, John R. Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning, 2020.

[22] Yunhui Guo, Noel CF Codella, Leonid Karlinsky, John R Smith, Tajana Rosing, and Rogerio Feris. A new benchmark for evaluation of cross-domain few-shot learning. *arXiv preprint arXiv:1912.07200*, 2019.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] 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.

[25] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[27] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[28] Jeffrey Li, Mikhail Khodak, Sebastian Caldas, and Ameet Talwalkar. Differentially private meta-learning. *arXiv preprint arXiv:1909.05830*, 2019.

[29] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[30] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[31] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[32] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[33] Fatemehsadat Mirshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*, 2020.

[34] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

[35] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753. IEEE, 2019.

[36] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[37] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[38] Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. 2019.

[39] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[40] 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)*, pages 3–18. IEEE, 2017.

[41] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2019.

[42] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. *arXiv preprint arXiv:2010.06667*, 2020.

[43] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[44] Harry Chandra Tanuwidjaja, Rakyong Choi, and Kwangjo Kim. A survey on deep learning techniques for privacy-preserving. In *International Conference on Machine Learning for Cyber Security*, pages 29–46. Springer, 2019.

[45] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

[46] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[47] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[49] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[50] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[51] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 1953–1963, 2019.

[52] X Wang, Y Peng, L Lu, Z Lu, M Bagheri, and RM Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, 2017.

[53] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[54] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018.

[55] Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. Private-knn: Practical differential privacy for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11854–11862, 2020.