DIFFERENTIALLY PRIVATE LEARNING NEEDS BETTER FEATURES (OR MUCH MORE DATA)

Anonymous authors
Paper under double-blind review

ABSTRACT

We demonstrate that differentially private machine learning has not yet reached its "AlexNet moment" on many canonical vision tasks: linear models trained on handcrafted features significantly outperform end-to-end deep neural networks for moderate privacy budgets. To exceed the performance of handcrafted features, we show that private learning requires either much more private data, or access to features learned on public data from a similar domain. Our work introduces simple yet strong baselines for differentially private learning that can inform the evaluation of future progress in this area.

1 INTRODUCTION

Machine learning (ML) models have been successfully applied to the analysis of sensitive user data such as medical images (Lundervold & Lundervold, 2019), text messages (Chen et al., 2019) or social media posts (Wu et al., 2016). Training these ML models under the framework of *differential privacy* (DP) (Dwork et al., 2006b; Chaudhuri et al., 2011; Shokri & Shmatikov, 2015; Abadi et al., 2016) can protect deployed classifiers against unintentional leakage of private training data (Shokri et al., 2017; Song et al., 2017; Carlini et al., 2019).

Yet, training deep neural networks with strong DP guarantees comes at a significant cost in utility (Abadi et al., 2016; Yu et al., 2020; Bagdasaryan et al., 2019; Feldman, 2020). In fact, on many ML benchmarks the reported accuracy of private deep learning still falls short of "shallow" (non-private) techniques. For example, on CIFAR-10, Papernot et al. (2020b) train a neural network to 66.2% accuracy for a large DP budget of $\varepsilon = 7.53$, the highest accuracy we are aware of for this privacy budget. Yet, without privacy, higher accuracy is achievable with linear models and unlearned "handcrafted" features, e.g., (Coates & Ng, 2012; Oyallon & Mallat, 2015). This leads to the central question of our work:

Can differentially private learning benefit from handcrafted features?

We answer this question affirmatively by introducing simple and strong handcrafted baselines for differentially private learning, that significantly improve the privacy-utility guarantees on canonical vision benchmarks.

Our contributions. We leverage the *Scattering Network* (ScatterNet) of Oyallon & Mallat (2015)—an *unlearned* SIFT-like feature extractor (Lowe, 1999)—to train linear models that improve upon the privacy-utility guarantees of deep learning on MNIST, Fashion-MNIST and CIFAR-10 (see Table 1). For example, on CIFAR-10 we exceed the accuracy reported by Papernot et al. (2020b) while simultaneously improving the provable DP-guarantee by $130\times$. On MNIST, we match the privacy-utility guarantees obtained with PATE (Papernot et al., 2018) *without requiring access to any public data*. We find that privately training deeper neural networks on handcrafted features also significantly improves over end-to-end deep learning, but fails to exceed the simpler linear models. Our results show that private deep learning remains outperformed by handcrafted priors on many tasks, and thus has yet to reach its "*AlexNet moment*" (Krizhevsky et al., 2012).

Table 1: Test accuracy of linear models with handcrafted ScatterNet features compared to prior results with end-to-end CNNs for various DP budgets ($\varepsilon, \delta = 10^{-5}$). Lower ε values provide stronger privacy. The CNNs with maximal accuracy for each privacy budget are underlined. We select the best ScatterNet model for each DP budget $\varepsilon \leq 3$ with a hyper-parameter search, and show the mean and standard deviation in accuracy for five runs.

Data	ε-DP	Source	CNN Accuracy (%)	ScatterNet Accuracy (%)
MNIST	1.2	Feldman (2020)	96.6	$\mathbf{98.1 \pm 0.1}$
	2.0	Abadi et al. (2016)	95.0	$\mathbf{98.5 \pm 0.0}$
	2.32	Bu et al. (2019)	96.6	$\mathbf{98.6 \pm 0.0}$
	2.5	Chen & Lee (2020)	90.0	$\mathbf{98.7 \pm 0.0}$
	2.93	Papernot et al. (2020a)	98.1	$\mathbf{98.7 \pm 0.0}$
	3.2	Nasr et al. (2020)	96.1	–
	6.78	Yu et al. (2019b)	93.2	–
Fashion-MNIST	2.7	Papernot et al. (2020a)	86.1	$\mathbf{89.5 \pm 0.0}$
	3.0	Chen & Lee (2020)	82.3	$\mathbf{89.7 \pm 0.0}$
CIFAR-10	3.0	Nasr et al. (2020)	55.0	$\mathbf{67.0 \pm 0.1}$
	6.78	Yu et al. (2019b)	44.3	–
	7.53	Papernot et al. (2020a)	66.2	–
	8.0	Chen & Lee (2020)	53.0	–

We explain the benefits of handcrafted features by analyzing the convergence rate of *non-private* gradient descent. First, we observe that with low enough learning rates, training converges similarly with or without privacy (both for models with and without handcrafted features). Second, we show that handcrafted features significantly boost the convergence rate of *non-private* learning at low learning rates. As a result, when training with privacy, handcrafted features lead to more accurate models for a fixed privacy budget.

Informed by this analysis, we show that the utility of private deep learning can be increased by collecting more private data, which allows for training for more steps at a fixed privacy budget. On CIFAR-10, we use an additional $500{,}000$ labeled Tiny Images from Carmon et al. (2019) to privately train deep neural networks that outperform our handcrafted features baselines. Yet, the high sample-complexity of private deep learning could be detrimental for tasks that cannot leverage "internet-scale" data collection (e.g., most medical applications).

We further consider private learning with access to *public* data from a similar domain. In this setting, handcrafted features can be replaced by features learned from public data via *transfer learning* (Razavian et al., 2014). We show that for transfer learning, strong privacy comes at only a minor cost in accuracy. For example, given public *unlabeled* ImageNet data, we train a CIFAR-10 model to 92.7% accuracy for a DP budget of $\varepsilon = 2$.

Our work demonstrates that higher quality features—whether handcrafted or transferred from public data—are of paramount importance for improving the performance of private classifiers in low (private) data regimes.

2 Strong Shallow Baselines for Differentially Private Learning

We consider the standard *central* model of differential privacy (DP): a trusted party trains an ML model f on a private dataset $D \in \mathcal{D}$, and publicly releases the model. The learning algorithm A satisfies (ε, δ)-differential privacy (Dwork et al., 2006a), if for any datasets D, D' that differ in one record, and any set of models S:

$$\Pr[A(D) \in S] \leq e^{\varepsilon} \Pr[A(D') \in S] + \delta.$$

DP bounds an adversary's ability to infer information about any individual training point from the model. Cryptography can split the trust in a central party across users (Jayaraman et al., 2018; Bonawitz et al., 2017).

Prior work has trained private deep neural networks "end-to-end" (e.g., from image pixels), with large losses in utility (Shokri & Shmatikov, 2015; Abadi et al., 2016; Papernot et al., 2020b). In contrast, we study the

Table 2: Effect of feature normalization on the test accuracy of *non-private* ScatterNet models after 20 epochs. We also report the maximal test accuracy upon convergence (mean and standard deviation over five runs).

	Normalization (Test acc. after 20 epochs)			
Dataset	None	Group Normalization	Data Normalization	Maximal Accuracy
MNIST	95.9 ± 0.0	$\mathbf{99.1 \pm 0.0}$	$\mathbf{99.1 \pm 0.0}$	99.3 ± 0.0
Fashion-MNIST	82.6 ± 0.1	$\mathbf{90.9 \pm 0.1}$	$\mathbf{91.0 \pm 0.2}$	91.5 ± 0.0
CIFAR-10	58.0 ± 0.1	67.8 ± 0.2	$\mathbf{70.7 \pm 0.1}$	71.1 ± 0.0

benefits of handcrafted features that encode *priors* on the learning task's public domain (e.g., edge detectors for images). Although end-to-end neural networks outperform such features in the non-private setting, our thesis is that handcrafted features result in an easier learning task that is more amenable to privacy. We focus on computer vision, a canonical domain for private deep learning (Abadi et al., 2016; Yu et al., 2019b; Papernot et al., 2020b; Nasr et al., 2020)), with a rich literature on handcrafted features (Lowe, 1999; Dalal & Triggs, 2005; Bruna & Mallat, 2013). Our approach can be extended to handcrafted features in other domains, e.g., text or speech.

2.1 Scattering Networks

We use the Scattering Network (ScatterNet) of Oyallon & Mallat (2015), a feature extractor that encodes natural image priors (e.g., invariance to small rotations and translations) using a cascade of wavelet transforms (Bruna & Mallat, 2013). As this cascade of transforms is data independent, we can obtain a differentially private classifier by privately fine-tuning a (linear) model on top of locally extracted features. In Appendix A, we discuss other candidate "non-deep" approaches that we believe to be less suitable for differentially private learning.

We use the default parameters in (Oyallon & Mallat, 2015), a ScatterNet $S(\boldsymbol{x})$ of depth two with wavelets rotated along eight angles. For images of size $H \times W$, this network extracts features of dimension $(K, H/4, W/4)$, with $K = 81$ for grayscale images, and $K = 243$ for RGB images. More details on ScatterNets are in Appendix C.1.

2.2 Differentially Private ScatterNet Classifiers

To train private classifiers, we use the DP-SGD algorithm[1] of Abadi et al. (2016) (see Appendix B). DP-SGD works as follows: (1) batches of expected size B are sampled at random;[2] (2) gradients are *clipped* to norm C; (3) Gaussian noise of variance $\sigma^2 C^2$ is added to the mean gradient. DP-SGD guarantees privacy for *gradients*, and is thus oblivious to preprocessing applied independently to each data sample, such as the ScatterNet transform.

When training a supervised classifier on top of ScatterNet features with gradient descent, we find that *normalizing* the features is crucial to obtain strong performance. We consider two approaches:

- *Group Normalization* (Wu & He, 2018): the channels of $S(\boldsymbol{x})$ are split into G groups, and each is normalized to zero mean and unit variance. Data points are normalized independently so this step incurs no privacy cost.
- *Data Normalization*: the channels of $S(\boldsymbol{x})$ are normalized by their mean and variance across the training data. This step incurs a privacy cost as the per-channel means and variances need to be privately estimated.

Table 2 shows that normalization significantly accelerates convergence of *non-private* linear models trained on ScatterNet features, for MNIST, Fashion-MNIST and CIFAR-10. For CIFAR-10, Data Normalization performs

[1] Yu et al. (2019a) show that DP-SGD outperforms other algorithms for private convex optimization, e.g., logistic regression with output or objective perturbation (Chaudhuri et al., 2011; Bassily et al., 2014; Kifer et al., 2012). In Appendix D.3, we show that DP-SGD also outperforms *Privacy Amplification by Iteration* (Feldman et al., 2018) in our setting.

[2] Existing DP-SGD implementations (tensorflow/privacy, 2019; pytorch/opacus, 2020) and many prior works (e.g., (Abadi et al., 2016; Papernot et al., 2020b)) heuristically split the data into random batches of size *exactly* B. We use the same heuristic and show in Appendix D.4 that using the correct batch sampling does not affect our results.

significantly better than Group Normalization, so the small privacy cost of estimating channel statistics is warranted. While the maximal test accuracy of these models falls short of state-of-the-art CNNs, it exceeds all previously reported results for differentially private neural networks (even for large privacy budgets).

3 EVALUATING PRIVATE SCATTERNET CLASSIFIERS

We compare differentially private ScatterNet classifiers and deep learning models on MNIST (LeCun et al., 2010), Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009). Many prior works have reported improvements over the DP-SGD procedure of Abadi et al. (2016) for these datasets. As we will show, ScatterNet classifiers outperform all prior approaches *while making no algorithmic changes to DP-SGD*. ScatterNet classifiers can thus serve as a strong canonical baseline for evaluating proposed improvements over DP-SGD in the future.

3.1 EXPERIMENTAL SETUP

Most prior works find the best model for a given DP budget using a hyper-parameter search. As the private training data is re-used many times, this overestimates the privacy guarantees. Private hyper-parameter search is possible at a small cost in the DP budget (Liu & Talwar, 2019), but we argue that fully accounting for this privacy leakage is hard as even our choices of architectures, optimizers, hyper-parameter *ranges*, etc. are informed by prior analysis of the same data. As in prior work, we thus do not account for this privacy leakage, and instead compare ScatterNet models and end-to-end CNNs with similar hyper-parameter searches. Moreover, we find that ScatterNet models are very robust to hyper-parameter changes and achieve near-optimal utility with random hyper-parameters (see Table 3). To evaluate ScatterNet models, we apply the following hyper-parameter search:

- We begin by fixing a *privacy schedule*. We target a moderate differential privacy budget of $(\varepsilon = 3, \delta = 10^{-5})$ and compute the noise scale σ of DP-SGD so that the privacy budget is consumed after T epochs. We try different values of T, with larger values resulting in training for more steps but with higher noise.
- We fix the gradient clipping threshold for DP-SGD to $C = 0.1$ for all our experiments. Thakkar et al. (2019) suggest to vary this threshold adaptively, but we did not observe better performance by doing so.
- We try various batch sizes B and base learning rates η, with linear learning rate scaling (Goyal et al., 2017).[3]
- We try both Group Normalization (Wu & He, 2018) with different choices for the number of groups, and private Data Normalization with different choices of privacy budgets (see Appendix B for details).

We perform a grid-search over all parameters as detailed in Appendix C.5. We compare our ScatterNet classifiers to the CNN models of Papernot et al. (2020b) (see Appendix C.2), which achieve the highest reported accuracy for our targeted privacy budget for all three datasets. We also perform a grid-search for these models, which reproduces the results of Papernot et al. (2020b). We use the ScatterNet implementation from `Kymatio` (Andreux et al., 2020), and the DP-SGD implementation in (pytorch/opacus, 2020) (formerly called `pytorch-dp`).

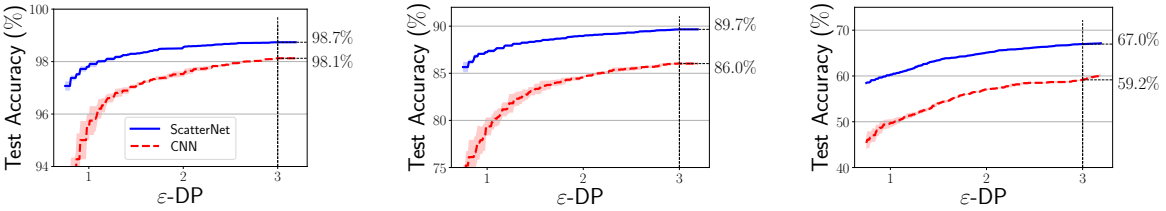
3.2 RESULTS

To measure a classifier's accuracy for a range of privacy budgets, we compute the test accuracy as well as the DP budget ε after each training epoch (with the last epoch corresponding to $\varepsilon = 3$). For various DP budgets $(\varepsilon, \delta = 10^{-5})$ used in prior work, Table 1 shows the maximal test accuracy achieved by a ScatterNet model in our hyper-parameter search, averaged over five runs. Figure 1 further compares the full privacy-accuracy

[3] Our decision to try various batch sizes is inspired by Abadi et al. (2016) who found that this parameter has a large effect on the performance of DP-SGD. Yet, in Appendix D.1 we show empirically, and argue formally that with a *fixed privacy schedule* and with linear learning rate scaling (Goyal et al., 2017), DP-SGD performs similarly for a range of batch sizes.

Table 3: Variability across hyper-parameters. For each model, we report the minimum, maximum, median and median absolute deviation (MAD) in test accuracy (in $\%$) achieved for a DP budget of $(\varepsilon = 3, \delta = 10^{-5})$. The maximum accuracy below may exceed those in Table 1 and Figure 1, which are averages of five runs.

	MNIST				Fashion-MNIST				CIFAR-10			
Model	Min	Max	Median	MAD	Min	Max	Median	MAD	Min	Max	Median	MAD
ScatterNet	**96.8**	**98.8**	**98.4**	**0.2**	**85.3**	**89.8**	**88.7**	**0.5**	**59.5**	**67.0**	**65.4**	**0.9**
CNN	86.1	98.2	97.4	0.5	20.2	86.2	83.6	1.8	39.4	59.2	52.5	5.4



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 1: Highest test accuracy achieved for each DP budget $(\varepsilon, \delta = 10^{-5})$ for ScatterNet classifiers and the end-to-end CNNs of Papernot et al. (2020b). We plot the mean and standard deviation across five runs.

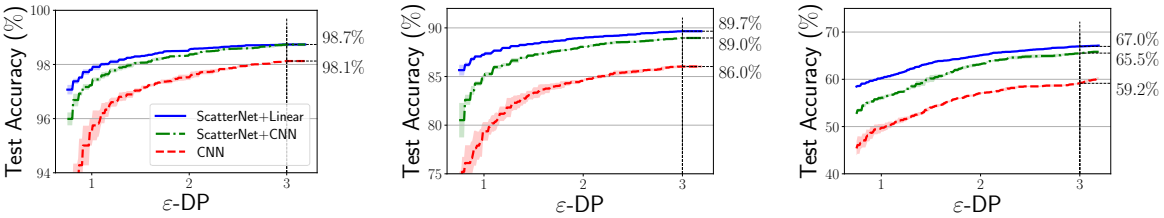
curves of our ScatterNets and of the CNNs of Papernot et al. (2020b). Linear models with handcrafted features significantly outperform prior results with end-to-end CNNs, for all privacy budgets $\varepsilon \leq 3$ we consider. Even when prior work reports results for larger budgets, they do not exceed the accuracy of our baseline.

In particular, for CIFAR-10, we match the best CNN accuracy in (Papernot et al., 2020b)—namely 66.2% for a budget of $\varepsilon = 7.53$—with a much smaller budget of $\varepsilon = 2.6$. This is an improvement in the DP-guarantee of $e^{4.9} \approx 134$. On MNIST, we significantly improve upon CNN models, and match the results of PATE (Papernot et al., 2018), namely 98.5% accuracy at $\varepsilon = 1.97$, in a more restricted setting (PATE uses $5{,}000$ *public* unlabeled MNIST digits). In Appendix C.5, we provide the hyper-parameters that result in the highest test accuracy for our target DP budget of $(\varepsilon = 3, \delta = 10^{-5})$. We did not consider larger privacy budgets for ScatterNet classifiers, as the accuracy we achieve at $\varepsilon = 3$ is close to the accuracy of non-private ScatterNet models (see Table 2).

As noted above, our models (as well as those of most prior work) are the result of a hyper-parameter search. While we do not account for the privacy cost of this search, Table 3 shows that an additional advantage of ScatterNet classifiers is an increased robustness to hyper-parameter changes. In particular, for CIFAR-10 we find that the *worst* configuration for ScatterNet classifiers outperforms the *best* configuration for end-to-end CNNs.

Training CNNs on Handcrafted Features. Since *linear* models trained on handcrafted features outperform the privacy-utility guarantees of deep models trained end-to-end, a natural question is whether training deeper models on these features achieves even better results. We repeat the above experiment with a similar CNN model trained on ScatterNet features (see Appendix C.2). The privacy-accuracy curves for these models are in Figure 2. Although handcrafted features also improve the utility of private deep models, the CNNs remain outperformed by linear models for all privacy budgets.

Note that on each dataset we consider, end-to-end CNNs can outperform ScatterNet models when trained *without privacy*. Thus, end-to-end CNNs trained with DP-SGD must eventually surpass ScatterNet models for large enough privacy budgets. But this currently requires settling for weak provable privacy guarantees. On CIFAR-10 for example, ScatterNet classifiers still outperform end-to-end CNNs for $\varepsilon = 7.53$ (Papernot et al., 2020b). While the analysis of DP-SGD might not be tight, Jagielski et al. (2020) suggest that the true ε guarantee of DP-SGD is at most one order of magnitude smaller than the current analysis suggests. Thus, surpassing handcrafted features for small privacy budgets on CIFAR-10 may require improvements beyond a tighter analysis of DP-SGD.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 2: Highest test accuracy achieved for each DP budget $(\varepsilon, \delta = 10^{-5})$ for linear ScatterNet classifiers, CNNs on top of ScatterNet features, and end-to-end CNNs. Shows mean and standard deviation across five runs.

4 How Do Handcrafted Features Help?

In this section, we analyze *why* private models with handcrafted features outperform end-to-end CNNs. We first consider the *dimensionality* of our models, but show that this does not explain the utility gap. Rather, we find that the higher accuracy of ScatterNet classifiers is due to their faster convergence rate *when trained without noise*.

Smaller models are not easier to train privately. The utility of private learning typically degrades as the model's dimensionality increases (Chaudhuri et al., 2011; Bassily et al., 2014). This is also the case with DP-SGD which adds Gaussian noise, of scale proportional to the gradients, to *each* model parameter. We thus expect smaller models to be easier to train privately. Yet, as we see from Table 4, for MNIST and Fashion-MNIST the linear ScatterNet model has *more* parameters than the CNNs. For CIFAR-10, the CNN we used is larger, so we repeat the experiment from Section 3 with a CNN of comparable size to the linear ScatterNet classifier (see Appendix D.5). This has a minor effect on the performance of the CNN, both when trained end-to-end or on ScatterNet features. *Thus, the dimensionality of ScatterNet classifiers fails to explain their better performance.*

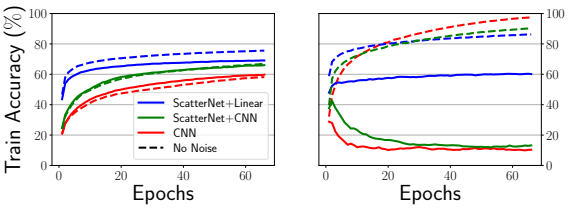
Models with handcrafted features converge faster *without privacy*. DP-SGD typically requires a smaller learning rate than noiseless (clipped) SGD, so that the added noise gets averaged out over small steps. We indeed find that the optimal learning rate when training with DP-SGD is an order of magnitude lower than the optimal learning rate for training without noise addition (with gradients clipped to the same norm in both cases).

To understand the impact of gradient noise on the learning process, we conduct the following experiment: we select a low learning rate that is near-optimal for training models with gradient noise, and a high learning rate that is near-optimal for training without noise. For both learning rates, we train CIFAR-10 models both with and without noise (with gradient clipping in all cases). Figure 3 shows that with a high learning rate, all classifiers converge rapidly when trained without noise, but gradient noise vastly degrades performance. With a low learning rate however, training converges similarly whether we add noise or not. What distinguishes the ScatterNet models is the faster convergence rate of *noiseless* SGD. The experimental setup and similar results on MNIST and Fashion-MNIST are in Appendix C.6. Thus, we find that handcrafted features are beneficial for private learning because they result in a simpler learning task where training converges rapidly even with small update steps. Our analysis suggests two avenues towards obtaining higher accuracy with private deep learning:

- *Faster convergence:* Figure 3 suggests that faster convergence of *non-private* training could translate to better private learning. DP-SGD with adaptive updates (e.g., Adam (Kingma & Ba, 2015)) indeed sometimes leads to small improvements (Papernot et al., 2020b; Chen & Lee, 2020; Zhou et al., 2020). Investigating private variants of *second-order optimization methods* is an interesting direction for future work.
- *More training steps (a.k.a more data):* For a fixed DP-budget ε and noise scale σ, increasing the training set size N allows for running more steps of DP-SGD (McMahan et al., 2018). In Section 5.1, we confirm that collecting more data indeed significantly improves the utility of private deep learning.

Table 4: Number of trainable parameters of our models. For CIFAR-10, we consider two different CNN architectures (see Appendix C.2), the smaller of which has approximately as many parameters as the linear ScatterNet model.

	MNIST / F-MNIST	CIFAR-10
ScatterNet+Linear	40K	155K
ScatterNet+CNN	33K	267K / 118K
CNN	26K	551K / 168K



Low LR ($\eta = 0.25$) High LR ($\eta = 4.0$)

Figure 3: Convergence of DP-SGD with and without noise on CIFAR-10, for ScatterNet classifiers and end-to-end CNNs. (Left): low learning rate. (Right): high learning rate.

5 Towards Better Private Deep Learning

We have shown that on standard vision tasks, private learning strongly benefits from *handcrafted features*. Further improving our private baselines seems hard, as they come close to the maximal accuracy of linear ScatterNet models (see Table 2). We thus turn to other avenues for obtaining stronger privacy-utility guarantees. We focus on CIFAR-10, and discuss two paths towards better private models: (1) access to a larger *private* training set, and (2) access to a *public* image dataset from a different distribution (some works also consider access to public unlabeled data from the same distribution as the private data (Papernot et al., 2017; 2018; Zhu et al., 2020)).

5.1 Improving Privacy by Collecting More Data

We first analyze the benefits of additional *private labeled data* on the utility of private models. Since the privacy budget consumed by DP-SGD scales inversely with the size of the training data N, collecting more data allows either to train for more steps, or to lower the amount of noise added per step—for a fixed DP budget ε.

To obtain a larger dataset comparable to CIFAR-10, we use 500K pseudo-labeled Tiny Images[4] (Torralba et al., 2008) collected by Carmon et al. (2019).[5] We then train private models on subsets of size $10{,}000 \leq N \leq 550{,}000$ from this dataset. Figure 4 reports the highest test accuracy achieved for a privacy budget of $(\varepsilon = 3, \delta = 1/2N)$ (see Appendix C.7 for the experimental setup). We find that larger private datasets enable end-to-end CNNs to outperform ScatterNet features. As we show in Appendix C.7, larger datasets allow DP-SGD to be run for more steps at a fixed privacy budget and noise level (as also observed in (McMahan et al., 2018))—thereby overcoming the slow convergence rate we uncovered in Section 4. While the increased sample complexity of private deep learning might be viable for "internet-scale" applications (e.g., language modeling across mobile devices), it is detrimental for sensitive applications with more stringent data collection requirements, such as in healthcare.

5.2 Transfer Learning: Better Features from Public Data

Transfer learning is a natural candidate for privacy-preserving computer vision, as features learned on public image data often significantly outperform handcrafted features (Razavian et al., 2014). We first consider transfer learning from CIFAR-100 to CIFAR-10, where the labeled CIFAR-100 data is assumed public. We extract features from the penultimate layer of a ResNeXt (Xie et al., 2017) model trained on CIFAR-100. A non-private

[4]The Tiny Images dataset has been withdrawn after the discovery of offensive class labels (Prabhu & Birhane, 2020). The subset used by Carmon et al. (2019) is filtered to match the CIFAR-10 labels, and is thus unlikely to contain offensive content.

[5]The privacy guarantees obtained with this dataset could be slightly overestimated, as the pseudo-labels of Carmon et al. (2019) are obtained using a model pre-trained on CIFAR-10, thus introducing dependencies between private data points.

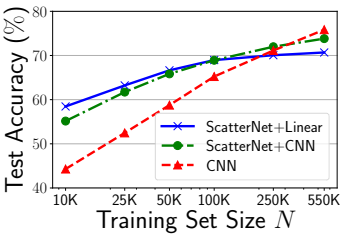


Figure 4: CIFAR-10 test accuracy for a training set of size N and a DP budget of $(\varepsilon = 3, \delta = 1/2N)$. For $N > 50$K, we augment CIFAR-10 with pseudo-labeled Tiny Images collected by Carmon et al. (2019).

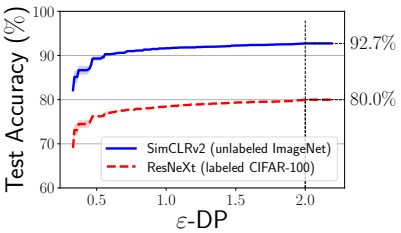


Figure 5: Privacy-utility tradeoffs for transfer learning on CIFAR-10. We fine-tune linear models on features from a ResNeXt model trained on CIFAR-100, and from a SimCLR model trained on unlabeled ImageNet.

linear model trained on these features achieves 84% accuracy on CIFAR-10. When training linear models with DP-SGD, we get the privacy-utility curve in Figure 5 (see Appendix C.8 for details). We reach an accuracy of 80.0% at a budget of $(\varepsilon = 2, \delta = 10^{-5})$, a significant improvement over prior work for the same setting and privacy budget, e.g., 67% accuracy in (Abadi et al., 2016) and 72% accuracy in (Papernot et al., 2020a).

We further consider access to a public dataset of *unlabeled* images. We extract features from the penultimate layer of a SimCLR model (Chen et al., 2020a) trained on unlabeled ImageNet. A non-private linear model trained on these features achieves 95% accuracy on CIFAR-10 (using *labeled* ImageNet data marginally improves non-private transfer learning to CIFAR-10 (Chen et al., 2020a)). With the same setup as for CIFAR-100 (see Appendix C.8), we train a linear model to 92.7% accuracy for a DP budget of $(\varepsilon = 2, \delta = 10^{-5})$ (see Figure 5).

6 Conclusion and Open Problems

We have demonstrated that differentially private learning benefits from "handcrafted" features that encode priors on the learning task's domain. In particular, we have shown that private linear ScatterNet classifiers outperform end-to-end CNNs on MNIST, Fashion-MNIST and CIFAR-10. We have further found that handcrafted features can be surpassed when given access to more data, either a larger private training set, or a public dataset from a related domain. In addition to introducing strong baselines for evaluating future improvements to private deep learning and DP-SGD, our work suggests a number of open problems and directions for future work:

Improving DP by accelerating convergence: Our analysis in Section 4 shows that a limiting factor of private deep learning is the slow convergence rate of deep models. While the existing literature on second-order optimization for deep learning has mainly focused on improving the overall *wall-clock time* of training, it suffices for DP to reduce the number of private *training steps*—possibly at an increase in computational cost.

Federated learning: While we have focused on a standard centralized setting for DP, our techniques can be extended to decentralized training schemes such as Federated Learning (McMahan et al., 2017; Bonawitz et al., 2017; Kairouz et al., 2019). DP has been considered for Federated Learning (Geyer et al., 2017; McMahan et al., 2018), but has also been found to significantly degrade performance in some settings (Yu et al., 2020).

Handcrafted features for ImageNet and non-vision domains: To our knowledge, there have not yet been any attempts to train ImageNet models with DP-SGD, partly due to the cost of computing per-sample gradients. While linear classifiers are unlikely to be competitive on ImageNet, handcrafted features could help private learning by accelerating the convergence of CNNs, as we have shown in Figure 2. Notably, Oyallon et al. (2018) match the (non-private) accuracy of AlexNet (Krizhevsky et al., 2012) on ImageNet with a small six-layer CNN trained on ScatterNet features. Another interesting direction is to extend our results to domains beyond vision, e.g., with handcrafted features for text (Manning & Schutze, 1999) or speech (Andén & Mallat, 2014).

REFERENCES

tensorflow/privacy, 2019. URL https://github.com/tensorflow/privacy.

pytorch/opacus, 2020. URL https://github.com/pytorch/opacus.

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*, pp. 308–318, 2016.

Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, Joan Bruna, Vincent Lostanlen, Matthew J. Hirn, Edouard Oyallon, Sixin Zhang, Carmine Cella, and Michael Eickenberg. Kymatio: Scattering transforms in Python. *Journal of Machine Learning Research*, 21(60):1–6, 2020.

Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. In *International Conference on Learning Representations (ICLR)*, 2020.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pp. 15479–15488, 2019.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1175–1191. ACM, 2017.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with Gaussian differential privacy. *arXiv preprint arXiv:1911.11607*, 2019.

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*, pp. 267–284, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11192–11203, 2019.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chen Chen and Jaewoo Lee. Stochastic adaptive line search for differentially private optimization. *arXiv preprint arXiv:2008.07978*, 2020.

Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2287–2295, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pp. 561–580. Springer, 2012.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. IEEE, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126, 2015.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. IEEE, 2018.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? *arXiv preprint arXiv:2006.07709*, 2020.

Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Distributed learning without distress: Privacy-preserving empirical risk minimization. In *Advances in Neural Information Processing Systems*, pp. 6343–6354, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1, 2012.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs*, 2010.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 298–309, 2019.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. 1999.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled Gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Milad Nasr, Reza Shokri, and Amir houmansadr. Improving deep learning with differential privacy using gradient encoding and denoising. *arXiv preprint arXiv:2007.11524*, 2020.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2865–2873, 2015.

Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew Blaschko, and Eugene Belilovsky. Scattering networks for hybrid representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2208–2221, 2018.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR)*, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations (ICLR)*, 2018.

Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy, 2020a. URL `https://openreview.net/forum?id=rJg851rYwH`.

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. In *Theory and Practice of Differential Privacy*, 2020b.

Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pp. 512–519. IEEE, 2014.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Rubinstein, Peter Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Ludwig Schmidt, Jonathan Ragan-Kelley, and Benjamin Recht. Neural kernels without tangents. In *International Conference on Machine Learning (ICML)*, 2020.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

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.

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 587–601, 2017.

Om Thakkar, Galen Andrew, and H Brendan McMahan. Differentially private learning with adaptive clipping. *arXiv preprint arXiv:1905.03871*, 2019.

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.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. *Proceedings of Machine Learning Research*, 89:1226–1235, 16–18 Apr 2019.

Shaomei Wu, Hermes Pique, and Jeffrey Wieland. Using artificial intelligence to help blind people 'see' Facebook. `https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/`, 2016.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Da Yu, Huishuai Zhang, Wei Chen, Tie-Yan Liu, and Jian Yin. Gradient perturbation is underrated for differentially private convex optimization. *arXiv preprint arXiv:1911.11363*, 2019a.

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 332–349. IEEE, 2019b.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Yingxue Zhou, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Arindam Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

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*, pp. 11854–11862, 2020.