ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094822

FINE-GRAINED PRIVATE KNOWLEDGE DISTILLATION

Yuntong Li[1], *Shaowei Wang*[1,⋆], *Yingying Wang*[1], *Jin Li*[1], *Yuqiu Qian*[2], *Bangzhou Xin*[3], *Wei Yang*[3]

[1]Institute of Artificial Intelligence and Blockchain, Guangzhou University
[2]Interactive Entertainment Group, Tencent Games
[3]Institute of Advanced Research, University of Science and Technology of China

ABSTRACT

Knowledge distillation has emerged as a scalable and effective way for privacy-preserving machine learning. One remaining drawback is that it consumes privacy in a client-level manner. In order to attain fine-grained privacy accountant and improve utility, this work proposes a model-free *reverse k-NN labeling* method towards record-level private knowledge distillation, where each private record is employed for labeling at most k queries. Theoretically, we provide bounds of labeling error rate under the centralized/local model of differential privacy. Experimentally, we demonstrate that it achieves new state-of-the-art accuracy in MNIST/SVHN/CIFAR-10 dataset with one order of magnitude lower of privacy loss.

Index Terms— Differential Privacy, Federated Learning, Knowledge Distillation

1. INTRODUCTION

Federated learning benefits from data across multiple individuals or organizations. However, data privacy has been a critical issue during collaboration, especially under increasingly rigid privacy laws, such as Data Security Law of the PRC and General Data Protection Regulation in the Europe Union. In contrast to transmitting raw data among clients, the seminal work of federated learning [1] proposes to share gradients. Subsequent works [2, 3] further impose rigorous protections (e.g., differential privacy [4]) on the gradients.

Since iteratively transmitting gradients is inefficient, researchers [5, 6] begin to employ the paradigm of knowledge distillation [7]. Federated clients are asked to label public-available data with locally-trained models (see the top of Figure 1), meanwhile preserving the privacy of clients' local records. Because labels have much lower dimensionality than gradients, federated knowledge distillation has become a communication & privacy efficient and thus prevalent way to federated deep learning [8, 9]. One line of these studies inject Laplace/Gaussian random noise for preserving centralized differential privacy on aggregated labels from teacher models [6, 10]; another line of studies sanitize the pseudo label predicted by teacher model locally [11, 8] to enable local differential privacy.

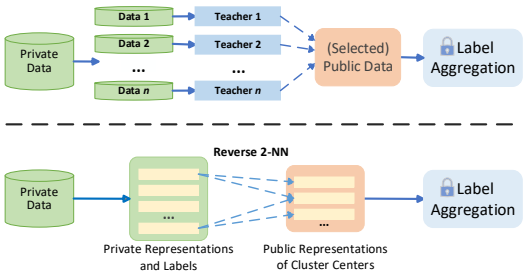


Fig. 1. Comparison of the current paradigm of federated knowledge distillation (top) and our record-level private knowledge distillation with reverse k-NN (bottom).

Despite many advantages over sharing gradients, the current knowledge distillation paradigm is still suffering from a critical drawback on privacy accountant. Instead of accounting privacy loss at the record level as the gradient-sharing paradigm, the current knowledge distillation paradigm summarizes local records to a privacy-sensitive local model. As a result, it inevitably reckons in each private record's contribution to the answers (i.e. client-level privacy).

In contrast, we propose the *reverse k-NN labeling* method[1] to limit the single record's maximum impact on the answers to multiple queries (i.e. record-level privacy). Besides, instead of relying on a locally-trained model that needs hundreds of training records, we utilize the unsupervised learning to generate public representation space where the private records are connected with public queries. Accordingly, our method is naturally immune to data *Non-I.I.D.* settings. As demonstrated in Figure 1, every private record is associated with k-nearest neighboring query samples, hence the privacy loss scales with only k instead of the number of total queries. The contributions of this paper are as follow:

(a) We initialize the study of federated knowledge distillation with record-level privacy preservation, and propose the model-free reverse k-NN query labeling method. (b) We for-

⋆ Corresponding author: wangsw@gzhu.edu.cn. This work is supported by NSFC (No.62102108), NSF of Guangdong Province (No.2022A1515010061), CSSC Systems Engineering Research Institute (No. 193-A11-107-01-33), NSFC for Joint Fund Project (No. U1936218).

[1]Source Code: https://github.com/liyuntong9/rknn

mulate the reverse k-NN labeling as Bucketized Sparse Vector Summation problem, and provide concrete mechanisms & theoretical guarantees for the problem under centralized/local differential privacy. (c) Through extensive experiments in centralized/local privacy setting, our method achieves a significant accuracy boost with one magnitude lower of privacy consumption when compared to existing approaches.

2. PROPOSED APPROACH

2.1. Background

Federated knowledge distillation. Every data record (x, y) is sampled from a Cartesian domain $\mathcal{X} \times \mathcal{Y}$, where the sample x might be a tabular vector, an image, and etc. The class label y can be a binary value (i.e. $\mathcal{Y} = \{0, 1\}$) or categorical value (e.g., $\mathcal{Y} = \{0, 1, .., 9\}$). The label y is often represented as vector form for the convenience of calculation. Assume that the client i possesses m_i records, let $D^i = [(x_1^i, y_1^i), ..., (x_{m_i}^i, y_{m_i}^i)]$ denote these records, let $D_{priv} = \bigcup_{i=1}^n D^i$ denote the union of all local datasets, and let $D_{pub} = [(x_1, ?), ..., (x_{m_{pub}}, ?)]$ denote the public unlabeled dataset possessed by the federated server. The primary goal of federated knowledge distillation is then labeling D_{pub} with the knowledge from D_{priv}.

Centralized differential privacy. Let $\mathcal{Z}$ denote the output domain. A randomized mechanism K satisfies (ϵ, δ)-differential privacy [4] (DP) if for any neighboring datasets D, D' and any outputs $\mathbf{z} \subseteq \mathcal{Z}$ it holds that $\mathbb{P}[K(D) \in \mathbf{z}] \leq \exp(\epsilon) \cdot \mathbb{P}[K(D') \in \mathbf{z}] + \delta$.

Local differential privacy. Assume that each client holds a dataset D with only one record. A randomized mechanism K satisfies local ϵ-differential privacy [12] (local DP) if for any data pair $D, D' \in \mathcal{X} \times \mathcal{Y}$ and any output $z \in \mathcal{Z}$ it holds that $\mathbb{P}[K(D) = z] \leq \exp(\epsilon) \cdot \mathbb{P}[K(D') = z]$.

2.2. Reverse k-NN Labeling

In this subsection, we provide the detailed description of reverse k-NN labeling in federated learning. The whole algorithm is illustrated in Algorithm 1.

Learning to represent: Since raw pixels are unstable w.r.t. semantic labels, we measure sample distance by their latent representations. One can use pre-trained representation models or train an unsupervised representation model from scratch with public-available D_{pub}.

Selecting queries: Labeling all samples in D_{pub} is privately expensive. Following current approaches [6, 10], we select representative samples from D_{pub}. At the first iteration, we cluster D_{pub} into s groups in the representation space, and treat cluster centers $Q = [q_1, q_2, ..., q_s]$ as query samples. For later iterations, samples are selected adaptively w.r.t. uncertainty of the current model M_S.

Local labeling: Given queries $Q = [q_1, q_2, ..., q_s]$ and public representation model, every local record (x_j^i, y_j^i) is connected to k-nearest query samples in the representation space. Let $N_j^i \subseteq [1 : s]$ denote the set of query indices that are k-nearest neighbors of x_j^i. Then the labeling answer from client i is $A^i = [a_1^i, a_2^i, ..., a_s^i]$, where $a_l^i = \sum_{j=1}^{m_i} [\![l \in N_j^i]\!] y_j^i \in \mathbb{R}^{|\mathcal{Y}|}$ for each $l \in [1 : s]$.

Algorithm 1 Fine-grained Private Knowledge Distillation

Input: n clients, private datasets $\mathcal{D}^1, ..., \mathcal{D}^n$, unlabeled public dataset $\mathcal{D}_{pub}$.
Parameter: number of iterations T, number of nearest neighbors k, number of query samples s, privacy budget ϵ.
Output: student model M_S that satisfies ϵ-differential privacy

for t=1,2,...,T **do**
  select s query samples $\mathcal{D}_{query}$ from $\mathcal{D}_{pub}$
  // Client side
  for i=1,2,...,n **do**
    connect each local record $(x_j^i, y_j^i) \in \mathcal{D}^i$ to k-nearest queries in $\mathcal{D}_{query}$
    find k-nearest neighbors $N_j^i \subseteq [1 : s]$ of x_j^i
    represent the labeling answer on each query $l \in [1 : s]$ as $a_l^i = \sum [\![l \in N_j^i]\!] y_j^i \in \mathbb{R}^{|\mathcal{Y}|}$
    send all labeling answers $A^i = [a_1^i, a_2^i, ..., a_s^i]$
  end for
  // Sever side
  aggregate the label counts $a_l = \sum_{i=1}^{n} a_l^i \in \mathbb{R}^{|\mathcal{Y}|}$ from all clients for each $l \in [1 : s]$
  ensemble $A = [a_1, a_2, ..., a_s]$ and add noise to A by ϵ-differential privacy
  derive labels $\{\hat{y}_l\}_{l=1}^s$ from noisy label counts A
  train student model M_S on $\mathcal{D}_{query}$ with labels $\{\hat{y}_l\}_{l=1}^s$
end for
return M_S

Label aggregation: Given the labeling answers $A^1, ..., A^n$ from n clients, we summarize them as $A = [a_1, a_2, ..., a_s]$, where $a_l = \sum_{i=1}^n a_l^i$. Then we add noise to A for privacy preservation, and the final hard labeling results is $(q_1, \hat{y}_1), ..., (q_s, \hat{y}_s)$, where $\hat{y}_l = \arg\max_{c=1}^{|\mathcal{Y}|} a_l(c)$ for each $l \in [1 : s]$. After assigning every sample in D_{pub} with the label of its cluster center, a student model is built upon labeled $\mathcal{D}_{pub}$ or $\mathcal{D}_{query}$ with cross-entropy loss. The proposed method is highly efficient. The computational/communication cost of the client i is linear to the number of local samples $|D^i|$ and queries s (i.e., $O(|D^i| \cdot s \cdot T + s \cdot |\mathcal{Y}| \cdot T)$ and $O(s \cdot |\mathcal{Y}| \cdot T)$).

k-NN vs. reverse k-NN: When the k-NN classifier works as a local model for labeling (e.g., in [13, 14]), each query is associated with k private records. However, from one private record's perspective, it might be associated with all queries. The influence of one private record is limited to k queries in the reverse k-NN labeling.

2.3. Centralized Private Mechanisms

In this subsection, we reformulate the reverse k-NN labeling as Bucketized Sparse Vector Summation (BSVS). Then we present centralized DP mechanisms and provide corresponding labeling error bounds.

Definition 1 (Bucketized Sparse Vector Summation)**.** *In the BSVS problem, each datum corresponds to a set $T_j \subseteq T$ of k buckets and a sparse vector $y_j \in \{0,1\}^{|\mathcal{Y}|}$ and $|y_j| = r$. The goal is to determine, for a given bucket $t \in T$, the vector sum of t, which is $a_t := \sum_{j=1}^{\sum_{i=[1:n]} m_i} y_j [\![t \in T_j]\!]$. An approximate oracle $\tilde{a}$ is said to be (η, β)-accurate at bucket t if we have $|a_t - \tilde{a}_t|_{+\infty} < \eta$ with probability $1-\beta$.*

In the above reformulation, the number of all buckets is equal to the number of query samples: $|T| = s$. Note that in conventional multi-class classification, we have $r \equiv 1$.

When centralized $(\epsilon, 0)$-DP is imposed on the BSVS problem, we employ the classical Laplace mechanism for privacy preservation. Apparently, the maximum possible changing magnitude of $A^i = [a_1^i, a_2^i, ..., a_s^i]$ (i.e. the sensitivity) is $2k \cdot r$ in our record-level labeling method, while in client-level methods the sensitivity is $2s \cdot r$ where $s \gg k$ in most scenarios. Therefore, we inject $Laplace(\frac{2k \cdot r}{\epsilon})$ to every element of A. The corresponding accuracy guarantee is presented in Proposition 1.

Proposition 1. *There is an $(\frac{2k \cdot r \cdot \log(|\mathcal{Y}|/\beta)}{\epsilon}, \beta)$-accurate centralized ϵ-DP algorithm for the BSVS problem.*

For the t-th query/bucket, we define the non-private count gap between the true label $y^* \in [1 : |\mathcal{Y}|]$ and false labels as $Gap_t = a_t(y^*) - \max_{c \in [1:|\mathcal{Y}|] \text{ and } c \neq y^*} a_t(c)$. Then we have the following conclusion on the private labeling accuracy w.r.t. the accuracy of the BSVS problem:

Remark 1. *If $Gap_t \geq 2\eta$ and the private algorithm is (η, β)-accurate, then with probability $1-\beta$, the estimated hard labeling result is accurate (equals to the true label y^*).*

2.4. Locally Private Mechanisms

Considering the most stringent case of imposing local DP on every client who holds only one record (i.e., $m_i \equiv 1$), every client i now sanitize the labeling answer $A^i = [a_1^i, a_2^i, ..., a_s^i]$ independently. We employ the randomized response mechanism [12] for local differential privacy, which randomly flips every binary value in A^i with probability $\frac{1}{e^{\epsilon/(2kr)}+1}$. We show randomized response is $(O(\sqrt{\frac{nk^2r^2 \log(|\mathcal{Y}|/\beta)}{\epsilon^2}}), \beta)$-accurate (in Theorem 1).

Theorem 1. *The local ϵ-DP randomized response mechanism is an $(\frac{e^{\epsilon/(2kr)}+1}{e^{\epsilon/(2kr)}-1}\sqrt{3n \log(|\mathcal{Y}|/\beta)/(e^{\epsilon/(2kr)}+1)}, \beta)$-accurate algorithm for the BSVS problem when $\epsilon = O(1)$.*

Proof. For a binary value b flipped with probability $\frac{1}{e^{\epsilon/(2kr)}+1}$, the unbiased estimation given the observation b' is $\tilde{b} = \frac{b' - 1/(e^{\epsilon/(2kr)}+1)}{(e^{\epsilon/(2kr)}-1)/(e^{\epsilon/(2kr)}+1)}$. The total count of observed ones is a summation of n Bernoulli variables with a success rate of either $\frac{1}{e^{\epsilon/(2kr)}+1}$ or $\frac{e^{\epsilon/(2kr)}}{e^{\epsilon/(2kr)}+1}$. Let u denote the estimation bias of one element in $\tilde{a}_t$, and we have $\mathbb{P}[|u| > \eta \cdot \frac{e^{\epsilon/(2kr)}+1}{e^{\epsilon/(2kr)}-1}] \leq \exp(\frac{-\eta^2(e^{\epsilon/(2kr)}+1)}{3n})$. Therefore, with probability of $1-\beta$, we have $|a_t - \tilde{a}_t|_{+\infty} \leq \frac{e^{\epsilon/(2kr)}+1}{e^{\epsilon/(2kr)}-1}\sqrt{\frac{3n \log(|\mathcal{Y}|/\beta)}{e^{\epsilon/(2kr)}+1}}$. □

Additionally, we can adopt an optimal sparse vector summation oracle (Collision Mechanism) in the high privacy regime [15] for the BSVS problem and achieve an error rate of $\tilde{\Theta}(\frac{\sqrt{k \cdot r}}{\epsilon})$.

3. EXPERIMENTS

3.1. Datasets, Networks and Performance Metric

To validate the proposed private knowledge distillation algorithm, we conduct extensive experiments on real-world datasets: **MNIST**[2] that contains $70,000$ gray-scale images of size 28×28 and 10 categories; **SVHN**[3] that contains $630,420$ digit images of size 32×32 and 10 categories; **CIFAR-10**[16] that contains $60,000$ images of size 32×32 and 10 categories.

Following common settings in the literature, for the MNIST/SVHN dataset, the public data D_{pub} are $5,000/26,000$ samples from the test dataset, the remaining $5,000/1,000$ test samples are used for evaluating the performance of the student classifier, and the training dataset (together with the extended data in SVHN) is used as the private data D_{priv}. For the CIFAR-10 dataset, the public data D_{pub} are $30,000$ samples from the training set, the $1,000$ samples from the test dataset are used for evaluation, and other $29,000$ samples are used as the private data D_{priv}.

For the MNIST dataset, the architecture of the student classifier is from [17], and the DTI [18] is employed for general-purpose representation & clustering on D_{pub} (denoted as **[general]**). For the SVHN dataset, the architecture of the student classifier is Mixmatch [19], and the histogram of oriented gradients (HOG) [20] and k-means++[21] are employed for general-purpose representation & clustering on the D_{pub}. For the CIFAR-10 dataset, the network architecture is DenseNet121 [22], and the SimCLR [23] and k-means++ ate used for representation learning & clustering on the D_{pub}.

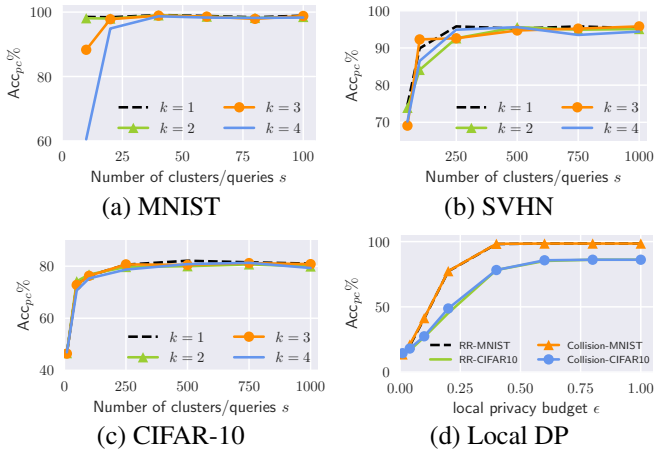
Two accuracy indications are employed for measuring the performances. One is the accuracy of the private label answering (Acc_{pl}), the other is the test accuracy of the privately learned classifier (Acc_{pc}). As we use unsupervised clustering for query selection at iteration 1, here the Acc_{pl} is the number of public samples receiving correct labels divided by $|D_{pub}|$.

[2] http://yann.lecun.com/exdb/mnist

[3] http://ufldl.stanford.edu/housenumbers

Dataset	Methods	#Queries	ϵ	Test Acc.	Label Acc.	Non-priv Acc.
MNIST	LNMAX [6]	1000	8.03	98.1%		99.2%
	GNMAX [6]	286	1.97	98.5%		
	Private k-NN [13]	735	0.47	98.8%		
	Noisy SGD [24]		1.0	81.2%		91.1%
	Ours **[general]**	40	**0.1**	**99.1%**	98.5%	99.2%
	Ours **[general]**	40	**0.04**	**98.6%**	97.7%	
	Ours **[end2end]**	10	**0.01**	**98.5%**	97.5%	98.7%
	Ours **[end2end]**	10	**0.004**	**98.2%**	97.3%	
SVHN	LNMAX [6]	1000	8.19	90.1%		92.8%
	GNMAX [6]	3098	4.96	91.6%		
	Private k-NN [13]	2939	0.49	91.6%		
	Noisy SGD [24]		4.0	76.0%		84.4%
	Ours **[general]**	500	**0.1**	**95.6%**		96.7%
	Ours **[general]**	500	**0.04**	**95.3%**		
CIFAR-10	GNMAX [6]	286		$< 50\%$		80.5%
	Private k-NN [13]	3877	2.92	70.8%		
	Finetuning Noisy SGD [3]		1.0	76.6%		
	Layer-1 Noisy SGD [24]		4.0	73.7%		77.7%
	Layer-2 Noisy SGD [24]		4.0	78.5%		80.9%
	Ours **[general]**	500	**1.0**	**82.1%**	77.1%	82.3%
	Ours **[general]**	500	**0.29**	**79.4%**	73.9%	
	Ours **[end2end]**	10	**0.01**	**86.1%**	85.9%	86.2%
	Ours **[end2end]**	10	**0.005**	**86.0%**	85.7%	

Table 1. Test accuracy & privacy consumption comparison of centralized differentially private methods.



(a) MNIST (b) SVHN (c) CIFAR-10 (d) Local DP

Fig. 2. Results of varying number of clusters/nearest neighbors on MNIST/SVHN/CIFAR-10 dataset (a, b, c) and results of local differential privacy on MNIST/CIFAR-10 dataset (d).

3.2. Varying Number of Clusters and Nearest Neighbors

We here explore the choice of s and k of our method in Figures 2. The purity of clusters (w.r.t. class labels) upper bounds the Acc_{pl}. Increasing the number of clusters s can roughly increase purity, but reduce the number of local records associated with one query. Besides there is no noticeable difference between choosing the number of nearest neighbors k at 1, 2, 3 or 4. Theoretically, as k gets larger, the label count of each query grows with k, but the count gap grows sublinearly with k and the standard deviation of the privacy noise grows with k. The k here in $[1, 2, 3, 4]$ are all small, thus the sublinearity is negligible and the noises hardly overwhelm count gaps.

3.3. Comparison with Existing Approaches and Local DP

The competitive approaches include SOTA private knowledge distillation methods by adding Laplace noise (LNMAX) or Gaussian noise (GNMAX) in [5, 6], private k-NN [13] and the noisy SGD methods in [3]. Given the representation model trained on the D_{pub} and the labeled samples in D_{priv}, one may simply train a prediction head with noisy SGD [24] as the classifier. We denote this straight-forward approach as Layer-1/Layer-2 noisy SGD (with one/two prediction layer(s)).

In Table 1, we compare our method with reported results of existing approaches under the same settings. The hyper-parameter of our method is set to $T = 1$ and $k = 1$. When employing general-purpose unsupervised representation learning and clustering, our method achieves better accuracy with an order magnitude smaller privacy compared to LNMAX, GNMAX, noisy SGD and private k-NN. Specifically, if we employ end-to-end unsupervised clustering [18, 25] (denoted as **[end2end]**), we are able to achieve average accuracy of 86.1% for CIFAR-10, and 99.1% for MNIST with centralized privacy budget $\epsilon = 0.01$. Note that when equipped with more relaxed DP and tighter privacy accountant [26], our experiment results will be better.

For the local DP, we present results in Figure 2 on MNIST/CIFAR-10 dataset. Since noises due to local DP easily dominate Gap_t, we fix hyper-parameters at $s = |\mathcal{Y}| = 10$ and $k = 1$, and employ end-to-end unsupervised clustering on MNIST with DTI [18] and on CIFAR-10 with SCAN [25]. It is observed that the Collision mechanism [15] achieves test accuracy of 98.5% for MNIST and 78.2% for CIFAR-10 with privacy budget $\epsilon = 0.4$. To the best of our knowledge, it is the first time locally private deep learning provides meaningful privacy/accuracy trade-offs.

4. CONCLUSION

This work proposes a fine-grained private knowledge distillation method (i.e., reverse k-NN labeling). Theoretically, we provide concrete differentially private mechanisms that are guaranteed for labeling accuracy. Experimentally, our solution improves significantly upon existing methods.

5. REFERENCES

[1] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[2] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, 2020.

[3] Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei, "Scalable differential privacy with sparse network finetuning," in *CVPR*, 2021.

[4] Cynthia Dwork, "Differential privacy," in *ICALP*. Springer, 2006.

[5] 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.

[6] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson, "Scalable private learning with pate," *ICLR*, 2018.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] Lingjuan Lyu and Chi-Hua Chen, "Differentially private knowledge distillation for mobile analytics," in *SIGIR*, 2020.

[9] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou, "Data-free knowledge distillation for heterogeneous federated learning," *ICML*, 2021.

[10] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip, "Private model compression via knowledge distillation," in *AAAI*, 2019.

[11] Lichao Sun and Lingjuan Lyu, "Federated model distillation with noise-free differential privacy," *arXiv preprint arXiv:2009.05537*, 2020.

[12] John C Duchi, Michael I Jordan, and Martin J Wainwright, "Local privacy and statistical minimax rates," *FOCS*, 2013.

[13] Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang, "Private-knn: Practical differential privacy for computer vision," in *CVPR*, 2020.

[14] Yuqing Zhu, Xiang Yu, Yi-Hsuan Tsai, Francesco Pittaluga, Masoud Faraki, Yu-Xiang Wang, et al., "Voting-based approaches for differentially private federated learning," *arXiv preprint arXiv:2010.04851*, 2020.

[15] Shaowei Wang, Jin Li, Yuqiu Qian, Jiachun Du, Wenqing Lin, and Wei Yang, "Hiding numerical vectors in local private and shuffled messages," *IJCAI*, 2021.

[16] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," *Technical Report*, 2009.

[17] Sanghyeon An, Minjun Lee, Sanglee Park, Heerin Yang, and Jungmin So, "An ensemble of simple convolutional neural network models for mnist digit recognition," *arXiv preprint arXiv:2008.10400*, 2020.

[18] Tom Monnier, Thibault Groueix, and Mathieu Aubry, "Deep transformation-invariant clustering," in *NeurIPS*, 2020.

[19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] 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)*. Ieee, 2005, vol. 1, pp. 886–893.

[21] David Arthur and Sergei Vassilvitskii, "k-means++: The advantages of careful seeding," Tech. Rep., Stanford, 2006.

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," *ICML*, 2020.

[24] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *CCS*, 2016.

[25] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, "Scan: Learning to classify images without labels," in *ECCV*, 2020.

[26] Ilya Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017, pp. 263–275.