先基于local data训练得到local model,
client再对所有public data生成logits 的加权平均,
然后server用这些logits结合global model的logits做KD

Preserving Privacy in Federated Learning with Ensemble Cross-Domain Knowledge Distillation

Xuan Gong[1], Abhishek Sharma [2], Srikrishna Karanam [2], Ziyan Wu [2], Terrence Chen [2], David Doermann [1], Arun Innanje [2]

[1] University at Buffalo, Buffalo NY
[2] United Imaging Intelligence, Cambridge MA
xuangong@buffalo.edu, abhishek.sharma@uii-ai.com, srikrishna.karanam@uii-ai.com, ziyan.wu@uii-ai.com, terrence.chen@uii-ai.com, doermann@buffalo.edu, arun.innanje@uii-ai.com

Abstract

Federated Learning (FL) is a machine learning paradigm where local nodes collaboratively train a central model while the training data remains decentralized. Existing FL methods typically share model parameters or employ co-distillation to address the issue of unbalanced data distribution. However, they suffer from communication bottlenecks. More importantly, they risk privacy leakage. In this work, we develop a privacy preserving and communication efficient method in a FL framework with one-shot offline knowledge distillation using unlabeled, cross-domain public data. We propose a quantized and noisy ensemble of local predictions from completely trained local models for stronger privacy guarantees without sacrificing accuracy. Based on extensive experiments on image classification and text classification tasks, we show that our privacy-preserving method outperforms baseline FL algorithms with superior performance in both accuracy and communication efficiency.

Introduction

The availability of large collections of data has facilitated the recent success of deep learning. However, in many cases, this wealth of data is dispersed over numerous physical locations and controlled by separate entities. Consequently, collaboration among parties, especially clinical institutions, is restricted due to the decentralized nature of the data. This is especially true for medical images where various legal, privacy, technical, and data ownership concerns often make it impractical or even impossible to gather such medical data to a centralized location.

To tackle some of these issues, federated learning (FL) (Shokri and Shmatikov 2015; Yang et al. 2019) has emerged as a practical machine learning paradigm where local models are used to collaboratively train a centralized model using data-free communication. There are several important challenges that make FL markedly different than typical distributed learning. First, privacy is a key concern. It is essential that local data remain protected. Second, communication is a critical bottleneck, so steps must be taken to minimize its detrimental effects. Third, due to the decentralized nature of the collection (leading to different settings), data across various local parties are typically heterogeneous, rendering the typical machine learning assumption of independent and identical distributions (i.i.d.) invalid.

Mainstream federated learning methods are based on the repeated sharing of parameters or gradients of local models during the training process (McMahan et al. 2017; Smith et al. 2017; Li et al. 2018; Zhao et al. 2018; Hsu, Qi, and Brown 2019; Wang et al. 2020; Karimireddy et al. 2020). Typically, such approaches involve each local model sharing its gradients with a central server after each round of local training on its local data. The central server then aggregates the local model parameters with typical data aggregation techniques (Wang et al. 2020; Li et al. 2020a; Hsu, Qi, and Brown 2020). Each local node then updates its local model with the latest global aggregation, and this process continues. These parameter-based communication methods have many known security weaknesses and are limited only to models with homogeneous architectures. While some methods have shown hope of protecting against data leakage in medical imaging (Li et al. 2019, 2020b), sharing parameters/gradients is highly susceptible to privacy leakage, and stealthy attacks. Some recent works (Zhu, Liu, and Han 2019; Geiping et al. 2020) demonstrated the ability to obtain local private data from publicly shared gradients, further highlighting the associated privacy risks in general and in medical applications in particular.

Another class of approaches in FL is to fuse local models into a single central model based on knowledge distillation (Hinton, Vinyals, and Dean 2015). Knowledge distillation eliminates the requirement for identical model architectures. While (Li and Wang 2019) distills the locally-computed knowledge on auxiliary public data to get around data privacy issues, they assume both the public and private data are sampled from the same underlying distribution. This further exposes the private data to security attacks. Recently proposed FedDF (Lin et al. 2020) relaxes the public data to be unlabeled and non-sensitive (i.e., sampled from another domain). Similarly, (Zhu, Hong, and Zhou 2021) eliminates the prerequisite of public data with a generator and aggregates knowledge in a data-free manner. However, both of them still exchange model parameters recursively, resulting in privacy vulnerabilities due to model memorization (Zhu, Liu, and Han 2019).

To address these issues, we present a new framework for federated learning (Fig. 1) with several innovations. First,

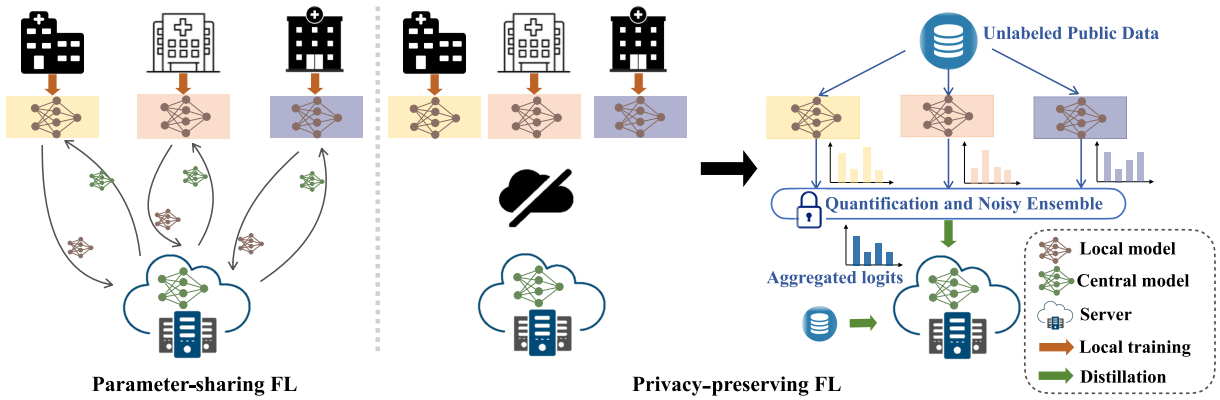


Figure 1: Traditional methods transfer private parameters or gradients from local nodes to a server, risking privacy leakage. Our framework trains local models independently, and only transfers products of the unlabeled public data. We further perturb the local predictions with a quantized and noisy ensemble for a stronger privacy guarantee.

unlike existing FL methods for either general or medical applications, our framework only shares the outputs of public data with one-shot (single round) distillation. The public data is unlabeled and decoupled from the private data. This, by design, eliminates the security vulnerabilities identified in prior works. Second, in contrast to the existing distillation-based FL work (Li and Wang 2019; Lin et al. 2020; Sui et al. 2020; Zhu, Hong, and Zhou 2021) that exclusively train local models incrementally and update them synchronously through online distillation, we keep the local training asynchronous and independent, and then aggregate the local predictions on unlabeled cross-domain public data. This offline strategy largely limits the server's exposure to local models' knowledge, reducing the consumption of communication bandwidth and reducing the risk of local information leakage. Furthermore, we deploy quantized and noisy aggregation on the locally-computed logits for stronger privacy guarantees. We experiment with CIFAR10/100 and large-scale chest x-ray datasets, showing very competitive classification results in accuracy, bandwidth, and privacy guarantees. Extensive experiments on text classification tasks also demonstrate our method outperforms prior works with higher accuracy, lower bandwidth as well as stronger privacy guarantee.

Our contributions can be summarized as:

- We propose a one-shot federated learning framework with one-way knowledge distillation (FedKD) on unlabeled, cross-domain, non-sensitive public data, explicitly addressing the communication bottleneck and preserving the privacy of local proprietary data without sacrificing accuracy.
- We introduce a seminal quantized and noisy ensemble before distillation, so that the privacy cost is meaningfully decreased with stronger security guarantees.
- We demonstrated the flexibility and efficiency of the proposed framework with extensive evaluations showing superior performance on accuracy, bandwidth and privacy-preserving capability compared to prior arts, on both image classification and text classification tasks.

Related Work

Knowledge Ensemble

With the success of knowledge transfer (Hinton, Vinyals, and Dean 2015), recent advancements on ensemble networks are dominated by the student-teacher learning paradigm (Shazeer et al. 2017; Zhou et al. 2021; Song et al. 2021). Ensemble learning aggregates the knowledge of multiple teachers before it distills the knowledge into the student network. Supervised ensemble learning is dominated by gate learning to design the weight for aggregation (Shazeer et al. 2017; Asif, Tang, and Harrer 2019; Xiang, Ding, and Han 2020). In semi-supervised and self-supervised scenarios, (Wu et al. 2019) and (You et al. 2017) exploit the relative similarity between samples for aggregation weights. Furthermore, co-distillation extends one-way transfer to bidirectional collaborative learning (Song and Chai 2018; Zhu, Gong et al. 2018; Dvornik, Schmid, and Mairal 2019; Guo et al. 2020).

Federated Learning.

In parameter-based FL methods, each local model shares its parameters/gradients with the central server after every round of local training on its local data, following which the central server aggregates them by average (McMahan et al. 2017). The result of this aggregation step is then shared by the central server with the local nodes, which in turn update their corresponding local model and proceed with the next training round. This process is then repeated until the stopping criterion is met. A variety of extensions of FedAVG (McMahan et al. 2017; Wang et al. 2020; Li et al. 2020a; Hsu, Qi, and Brown 2020) employ improved aggregation schemes, such as adding momentum (Hsu, Qi, and Brown 2019), and local weighting (Li et al. 2020a; Hsu, Qi, and Brown 2020). Another set of approaches improve local training by incorporating proximal term (Li et al. 2018) or control variations (Karimireddy et al. 2020) to restrict local training. However, such sharing of model parameters or gradients can be thought of as a naive way of information exchange, it is highly susceptible to privacy leakage and stealth

attacks, as also demonstrated elsewhere (Zhu, Liu, and Han 2019; Geiping et al. 2020).

Federated distillation methods exchange model outputs rather than model parameters. Given that some methods produce central models by distilling knowledge from private data (Zhou et al. 2020; Shin et al. 2020) in the same spirit as those above, there is a growing concern on local data privacy. In contrast, some works (Jeong et al. 2018; Li and Wang 2019) distill with the output of public data. Although model agnostic, these methods select public data based on the prior knowledge of private data. Recently proposed methods FedDF (Lin et al. 2020) and FedGEN (Zhu, Hong, and Zhou 2021) relax the prerequisites of distillation data, but they are still far from privacy-preserving or communication efficient due to the iterative exchange of models over hundreds of rounds. Besides, the above mentioned approaches exclusively require many rounds of back-and-forth communication, leading to bandwidth bottlenecks and other inefficiencies.

Privacy Issues

As noted above, parameter-based FL works have been shown to be highly susceptible to privacy leakage (Zhu, Liu, and Han 2019; Geiping et al. 2020). Distillation-based FL works with recursive model exchanges involved (Lin et al. 2020; Zhu, Hong, and Zhou 2021) also post privacy risk. Utilizing unlabeled public data during distillation has proven to be effective in protecting private local data from attackers (Hamm, Cao, and Belkin 2016). PATE (Papernot et al. 2017) also suggests that restricting the student network's access to the teacher's network and training with non-overlapping public datasets can further guarantee privacy protection. Unlike PATE, which uses topmost local votes to train the central model, we quantize and add noise on logits for aggregation and distillation. This retains more local expertise information and therefore improves the utility of the target model without sacrificing the protection of private data.

Method

In a federated learning setting with K local nodes, each local note hosts a private, labeled dataset $\mathcal{D}^k = \{(\boldsymbol{x}_i^k, y_i^k)|i = 1, \ldots, |\mathcal{D}^k|\}$. A shared, unlabeled public dataset $\mathcal{D}^0 = \{\boldsymbol{x}_i^0|i = 1, \ldots, |\mathcal{D}^0|\}$ is accessible by the central server and all local nodes. In the first stage of FedKD, the model at each local node k is initialized with model parameters θ^k by training with its own local private data $\mathcal{D}^k$. Note that FedKD is agnostic to the type of neural network architecture, and hence each local node can have its own specialized architecture suited for the particular distribution of its local data.

In the second stage, the local, private datasets are first disconnected from local training servers to minimize the risk of any data leakage and to protect privacy. The public dataset $\mathcal{D}^0$ that is hosted on the server and deployed at each local node is then used for one-way knowledge distillation from the local nodes to the server. Local models θ^k, together with the central model θ^s on the server, constitutes a student-teacher knowledge transfer configuration. The teacher here is an ensemble of multiple local models, one at each local node. The following sections introduce our privacy-preserving ensemble and distillation schemes for various tasks.

Privacy-Preserving Ensemble

The private dataset is denoted $\mathcal{D}^k = \{(\boldsymbol{x}_i^k, y_i^k)|i = 1, \ldots, |\mathcal{D}_k|\}$ ($k \in \mathcal{K}$), where $\mathcal{K} = \{1, ..., K\}$, $y^k \in \mathcal{C}^k$, $\mathcal{C}^k$ is the set of existing classes in the dataset $\mathcal{D}^k$, and $\mathcal{C}^k \subset \{1, \ldots, C\}$ (C is the number of classes across all local nodes). Let $z_i^{ck} = f(\boldsymbol{x}_i^0, \theta_k, c)$ be the logits of a public data sample $\boldsymbol{x}_i^0$ corresponding to class $c \in \mathcal{C}^k$, produced by the model at local node k, where $c \in \{1, \ldots, C\}$. We omit i in the following descriptions for simplicity. The conventional aggregation $\widehat{z}^c = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} z^{ck}$ takes an average of all teachers' logits. However, under the FL setting with a high degree of heterogeneity, a conventional ensemble algorithm is not appropriate primarily due to its inability to cope with the more general scenarios when local nodes are not sharing the exact same set of target classes. To take this into consideration, we introduce an importance weight ω for each local node to reflect the distribution of the local private data:

$$\omega_k^c = \frac{N_k^c}{\sum_{k \in \mathcal{K}} N_k^c}, \tag{1}$$

where for single-label classification, $N_k^c = \sum_{i=1}^{|\mathcal{D}_k|} (y_i^k = c)$ denotes the number of samples of class c used in training the model at local node k.

Inspired by PATE (Papernot et al. 2017), we perturb the locally computed logits with a quantized and noisy ensemble for a stronger privacy guarantee: +DP

$$\widehat{z}^c = \sum_{k \in \mathcal{K}} \omega_k^c \cdot Q(z^{ck}; S) + Lap(\frac{1}{\gamma}), \tag{2}$$

where $Q(\cdot, S)$ is the quantization function with S as quantization scale, and $Lap(\frac{1}{\gamma})$ is the Laplacian distribution with location 0 and scale $\frac{1}{\gamma}$. γ is a privacy parameter to trade off between privacy-preserving capability and accuracy. A smaller γ (i.e., higher noise level) results in stronger privacy guarantee and relatively lower accuracy.

To achieve better communication efficiency, we apply uniform quantization to floating point logits so they occupy fewer bits:

$$Q(z^{ck}; S) = q_s, \text{ if } z^{ck} \in (q_{s-1}, q_s], \tag{3}$$

We determine the quantization intervals $(q_{s-1}, q_s]$ with $q_s = \frac{2(s-1)z^{\max}}{S-1} - z^{\max} (s = 1, \ldots, S)$, where $z^{\max} = \max_{i,c,k} |z_i^{ck}|$ is the maximum absolute value of all the logits across public data samples $i = 1, \ldots, |\mathcal{D}_0|$ and classes $c = 1, \ldots, C$. Thus, Equation (3) becomes:

$$Q(z^{ck}; S) = \lceil \frac{S \cdot z^{ck}}{2z^{\max}} \rceil \cdot \frac{2z^{\max}}{S}, \tag{4}$$

where a smaller S sacrifices more logits precision, while maintaining a higher level of privacy.

Algorithm 1: Federated Knowledge Distillation (FedKD)

Input: Labeled private datasets $\{\mathcal{D}^k|k \in \mathcal{K}\}$ ($\mathcal{K} = \{1, \ldots, K\}$), unlabeled public data $\mathcal{D}^0$, central model θ^s, local models $\{\theta^k|k \in \mathcal{K}\}$, T distillation steps, batchsize B, quantization scale S, privacy hyperparameter γ.
Local Training:
Train each local model θ^k with private data $\mathcal{D}^k$.
Logits Ensemble:
for each sample $\boldsymbol{x}_i^0$ in $\mathcal{D}^0$ **do**
  for each local $k \in \mathcal{K}$ **do**
    $\boldsymbol{z}_i^k \leftarrow f(\boldsymbol{x}_i^0, \theta^k)$
  end for
  $\widehat{\boldsymbol{z}}_i \leftarrow$ aggregate $\{\boldsymbol{z}_i^k; S, \gamma | k \in \mathcal{K}\}$ ▷ Eq. 2
end for
Distillation:
for each distillation step $t = 1, ..., T$ **do**
  $\boldsymbol{x}^0 \leftarrow$ a batch of public data from $\mathcal{D}^0$ with size B
  $\widetilde{\boldsymbol{z}} \leftarrow f(\boldsymbol{x}^0, \theta^s)$
  Update the central model: $\theta^s \leftarrow \theta^s - \frac{1}{B}\nabla_{\theta^s}\mathcal{L}$ ▷ Eq. 7
end for

1，根据本地数据集训练本地模型
2，量化加权平均logits+DP噪声（基于所有public data）
*还是需要一个server来aggregate
3，KD（基于batch 的public data），更新全局模型

During ensemble, we protect the private data at each local node by: (1) transferring only the final prediction inferred with the non-proprietary public data $\mathcal{D}_0$; and (2) perturbing the local outputs with quantization and random noise .

One-shot Distillation

Conventional knowledge distillation aggregates all teachers' soft labels subject to the Kullback-Leibler divergence:

$$\mathcal{L} = \sum_c p^c log \frac{p^c}{q^c}, \tag{5}$$

where p^c and q^c denote the probabilities of a sample of class c for the teacher and student models, respectively. The aggregated logits $\widehat{z}^c$ can be viewed as teacher knowledge, and the output logits of the central model $\widetilde{z}^c = f(\boldsymbol{x}_0, \theta_s, c)$ can be viewed as student knowledge. Without loss of generality, we denote the activation by $p^c = \sigma(\widehat{z}^c)$ and $q^c = \sigma(\widetilde{z}^c)$. For single-label classification, we obtain the probabilities using softmax activation:

$$p^c = \sigma(\widehat{z}^c) = \frac{e^{\widehat{z}^c/\tau}}{\sum_c e^{\widehat{z}^c/\tau}}, \; q^c = \sigma(\widetilde{z}^c) = \frac{e^{\widetilde{z}^c/\tau}}{\sum_c e^{\widetilde{z}^c/\tau}}, \tag{6}$$

where τ is a temperature parameter. Hinton et al. (Hinton, Vinyals, and Dean 2015) showed that minimizing Eq. 5 with high τ is equivalent to minimizing the ℓ_2 error between the teacher and student logits, thereby relating cross-entropy minimization to matching logits.

Based on the observations above by Hinton et al. (Hinton, Vinyals, and Dean 2015), we consider the case of $\tau \to \infty$ so the loss can be written as:

$$\mathcal{L} = \|\widetilde{\boldsymbol{z}} - \widehat{\boldsymbol{z}}\|, \tag{7}$$

where $\widetilde{\boldsymbol{z}} = [\widetilde{z}^1, ...\widetilde{z}^C]$, and $\widehat{\boldsymbol{z}} = [\widehat{z}^1, ...\widehat{z}^C]$.

Note that we use one-shot offline distillation where the local nodes predict with each public data sample only once, and the predicted logits are used to train the central model iteratively. This distillation strategy (1) provides a higher privacy guarantee by executing fewer queries to the local model (limiting the access to local knowledge); and (2) eliminates the iterative and repetitive communication requirement of synchronous updates, improving communication efficiency and flexibility. The overall process is described in Algorithm 1.

Cross-domain Analysis

We argue that with cross-domain public data our framework can distill knowledge from multiple locals with generalizability. In this section we present a performance bound for the aggregated central model, which is built upon prior arts from domain adaptation (Ben-David et al. 2010).

Let the input space be $\mathcal{X}$, $\mathcal{D}^S$ and $\mathcal{D}^T$ be source and target domain respectively, We denote the ground-truth labeling function as g and the hypothesis function as f, we get the error as $\epsilon_{\mathcal{D}^S}(h, g) = \mathbb{E}_{x\sim\mathcal{D}^S}[|h(x) - g(x)|]$. We denote the risk of h on $\mathcal{D}^S$ and $\mathcal{D}^T$ as $\epsilon_{\mathcal{D}^S}$ and $\epsilon_{\mathcal{D}^T}$. (Ben-David et al. 2010) introduces $\mathcal{H}$-divergence to evaluate the distance between two domain distributions $\mathcal{U}$, $\mathcal{U}'$ on the a hypothesis space $\mathcal{H}$. $\mathcal{H}$-divergence is defined as $d_{\mathcal{H}}(\mathcal{U}, \mathcal{U}') = 2 \sup_{A\in\mathcal{A}_{\mathcal{H}}} |\Pr_{\mathcal{D}}(A) - \Pr_{\mathcal{D}'}(A)|$, where $\mathcal{A}_{\mathcal{H}}$ denotes a collection of subsets of $\mathcal{X}$ which support the hypothesis in $\mathcal{H}$. The symmetric different space is defined as $\mathcal{H}\Delta\mathcal{H} = \{h(x) \bigoplus h'(x) | h, h' \in \mathcal{H}\}$ ($\bigoplus$ represents the XOR operation). For the generalizability between two domains, we have the following theorem (Blitzer et al. 2007):

Theorem 1. Generalization bounds. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d, $\mathcal{U}^S$ and $\mathcal{U}^T$ be unlabeled samples of size N each, drawn from $\mathcal{D}^S$ and $\mathcal{D}^T$ respectively. For any $h \in \mathcal{H}$ and $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$ (over the choice of the samples):*

$$\begin{aligned}\epsilon_{\mathcal{D}^T}(h) \leq & \epsilon_{\mathcal{D}^S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^S, \mathcal{U}^T) \\ & + 4\sqrt{\frac{2d\log(2N) + \log(\frac{2}{\delta})}{N}} + \lambda,\end{aligned} \tag{8}$$

where $\lambda = \epsilon_{\mathcal{D}^S}(h^*) + \epsilon_{\mathcal{D}^T}(h^*)$ and h^* is the ideal joint hypothesis minimizing the combined error: $h^* = \text{argmin}_{h\in\mathcal{H}} \epsilon_{\mathcal{D}^S}(h^*) + \epsilon_{\mathcal{D}^T}(h^*)$.

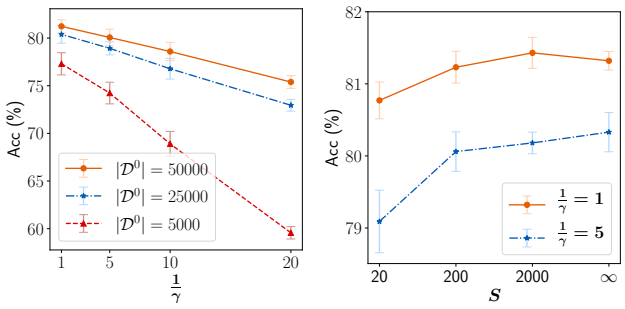
In our case, $\mathcal{D}^S$ is the domain of private data distributed across K local nodes: $\mathcal{D}^S = \{\mathcal{D}^k|k \in \mathcal{K}\}$, and $\mathcal{D}^T = \mathcal{D}^0$ is the domain of public data. We assume $|\mathcal{D}^0| = N$, $\sum_{k\in\mathcal{K}} |\mathcal{D}^k| = N$. Given the local model $h_{\mathcal{D}^k}$ trained on data $\mathcal{D}^k$, we learn central model $h_{\mathcal{D}^0}$ from public data $\mathcal{D}^0$ through weighted aggregation: $h_{\mathcal{D}^0} = \sum_{k\in\mathcal{K}} \omega_k(h_{\mathcal{D}^k} + n_k(\gamma))$, where $\sum_{k\in\mathcal{K}} \omega_k = 1$, and $n_k(\gamma)$ is the introduced noise parameterized by γ to strengthen the privacy. We have

Method	CIFAR-10			CIFAR-100		
	Accuracy(%) ↑		**Bandwidth**	**Accuracy**(%) ↑		**Bandwidth**
	$\alpha = 1$	$\alpha = 0.1$	(GB)↓	$\alpha = 1$	$\alpha = 0.1$	(GB)↓
FedAvg (McMahan et al. 2017)	78.57 ±0.22	68.37±0.50	58	42.54±0.51	36.72±1.50	63
FedProx (Li et al. 2018)	76.32 ±1.95	68.65±0.77	58	42.94±1.23	35.74±1.00	63
FedAvgM (Hsu, Qi, and Brown 2019)	77.79±1.22	68.63±0.79	58	42.83±0.36	36.29±1.98	63
FedDF (Lin et al. 2020)	80.69±0.43	71.36±1.07	58	**47.43±0.45**	39.33±0.03	63
FedGEN (Zhu, Hong, and Zhou 2021)	80.31±0.97	68.13±1.37	58	45.97±0.23	35.97±0.31	63
FedMD (Li and Wang 2019)	80.37±0.37	**69.23±1.31**	6.24	45.83±0.58	38.86±0.78	160
Standalone	61.11±24.90	28.99±27.24	-	27.49±14.76	16.31±15.75	-
FedKD	**80.98±0.11**	65.46±3.45	**0.078**	45.55±0.38	**40.61±2.54**	**2**

Table 1: Comparisons on the CIFAR-10 and CIFAR-100 datasets with ResNet-8 when K=20. Our FedKD uses S=200, γ=1 for knowledge ensemble, while the competing methods use the setting in FedDF (Lin et al. 2020) with 100 rounds and a sampled fraction as 1 at each communication round. *Standalone*: mean/std performance of all local models. Both logits and parameters are of type float64 for bandwidth calculation.

aggregation scheme	baseline	Eq. 1	Eq. 1	Eq. 1
logits distillation	τ=∞	τ=3	τ=∞	τ=∞
# local prediction	$\lvert\mathcal{D}^0\rvert$	$\lvert\mathcal{D}^0\rvert$	$\lvert\mathcal{D}^0\rvert$	50×$\lvert\mathcal{D}^0\rvert$
Accuracy(%)↑	79.92	80.01	80.98	**81.89**
Bandwidth (GB) ↓	0.078			3.91

Table 2: Ablation study on CIFAR-10 with ResNet-8, K=20, α=1, S=200, γ=1. With the commonly used distillation scheme (temperature $\tau = 3$) as baseline, we show the comparison on different ensemble and distillation schemes. $|\mathcal{D}^0|$ indicates the number of samples in the public dataset $\mathcal{D}^0$, and 50×$|\mathcal{D}^0|$ indicates local model predicts 50 times on each sample of $\mathcal{D}^0$ with different augmentation seeds.



(a) Varying $\frac{1}{\gamma}$ and $|\mathcal{D}^0|$. (b) Varying S and $\frac{1}{\gamma}$.

Figure 2: Ablation study on the CIFAR-10 dataset (K=20, α=1) with varying public data size $|\mathcal{D}^0|$, noise $\frac{1}{\gamma}$, and quantization scale S.

the following weighted noisy generalization bound:

$$\epsilon_{\mathcal{D}^0}(h_{\mathcal{D}^0}) \leq \epsilon_{\mathcal{D}^S}\left(\sum_{k\in\mathcal{K}}\omega_k(h_{\mathcal{D}^k}+n_k(\gamma))\right)+\lambda_\omega + \sum_{k\in\mathcal{K}}\omega_k\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^k,\mathcal{U}^0)\right) + 4\sqrt{\frac{2d\log(2N)+\log(\frac{2}{\delta})}{N}}. \quad (9)$$

Extension to Other Tasks

While Eq. 6 corresponds to the single-label classification scenario, our method is also extensible to multi-label classification. In this case, the private data notation from above is changed to $\mathcal{D}^k = \{(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k)|i = 1, \ldots, |\mathcal{D}^k|\}$ with $\boldsymbol{y}_i^k \in \{-1, 0, 1\}^c$ where -1, 0, and 1 indicate unknown, negative, and positive for class $c \in 1, ..., C$, respectively. We have made two other modifications: first, a sigmoid is used as the activation instead of softmax so $p^c = \sigma(\widehat{z}^c)$ and $q^c = \sigma(\widetilde{z}^c)$; second, in Eq. 1, we define $N_k^c = \sum_{i=1}^{|\mathcal{D}^k|}(\boldsymbol{y}_i^k(c) = 1)$ as the number of samples labeled as class c for training the model of local node k.

Experiments

We conduct experiments on natural image classification (single-label), medical image classification (multi-label), and extensive experiments on text classification. We construct local training sets using heterogeneous data splits with a Dirichlet distribution as in prior works (Hsu, Qi, and Brown 2019). The value of α controls the degree of non-IID-ness. An α of positive infinity indicates identical local data distributions, and a smaller α indicates higher non-IID-ness.

CIFAR10/100 Classification

For natural image classification task we use CIFAR-10/100 (Krizhevsky, Hinton et al. 2009) as datasets. To keep consistency with the prior arts, we use the same experimental settings as in FedDF (Lin et al. 2020): CIFAR-100 as unlabeled public data when CIFAR-10 as private data, and downsampled version of ImageNet (32 × 32) (Deng et al. 2009) as unlabeled public data corresponding to CIFAR-100 as private data. For each experiment, we sample over three different random seeds as private data split for local train-

Method	Private data $\mathcal{D}^k$	Test	$K=3$	$K=5$
FedKD (Single-domain)	CXR14	CXR14	75.02	74.80
	Xpert	Xpert	82.41	82.35
			$K=2$	$K=6$
FedKD (Cross-domain)	CXR14+Xpert	CXR14	79.03	76.13
	CXR14+Xpert	Xpert	79.77	80.91

Table 3: Multi-label classification experiments on chest-x-ray images with single/cross domain private data. We report the test mAUC (%) on NIH CXR14 and CheXpert over 12 and 8 classes respectively.

ing. We report the average accuracy metrics on CIFAR-10 and CIFAR-100 test set corresponding to its private data respectively.

Implementation Details. Following (Lin et al. 2020; Gong et al. 2021), we use ResNet-8 as backbone. We train each local model individually with SGD and Cosine Annealing (Loshchilov and Hutter 2016), decreasing the learning rate from 0.0025 to 0.001 in 500 epochs with a batch size of 16. For distillation, we use the Adam optimizer, a constant learning rate of 1e-3, and a batch size of 512. We use 200 and 10 epochs for CIFAR-10 and CIFAR-100 respectively. The weight decay is 3e-4 and 0 for local training and distillation, respectively.

Results. The comparison in Table 1 shows that our method achieves a significantly stronger privacy guarantee as well as a far better communication efficiency compared to prior arts, without sacrificing accuracy. On CIFAR-10 ($\alpha = 1$) and CIFAR-100 ($\alpha = 0.1$), our method demonstrates better accuracy with significantly lower communication cost than the prior arts. On CIFAR-10 ($\alpha = 0.1$) and CIFAR-100 ($\alpha = 1$), our method achieves the best performance-bandwidth trade-off compared with the prior arts. More importantly, our method does not share any locally trained model parameters and further adds noise perturbation on the transferred product of non-sensitive public data, demonstrating stronger privacy guarantee than the prior arts.

Ablation Studies. We perform ablation studies to validate the efficacy of our ensemble and distillation strategy and show the results in Table 2. The extensive experiments in Table 2 show the distillation accuracy can be improved by a large margin with more access to local information (e.g., local models predicted on dynamically augmented public data multiple times). For an accuracy-privacy trade-off, we restrict that local model to only predict each public sample once in our method. Besides, we do ablation study with different temperatures τ for logits distillation (Hinton, Vinyals, and Dean 2015).

In Figure 2 we study the impact of quantization/noise on the accuracy for different sized public datasets. The left figure suggests that increased noise degrades the ensemble distillation performance, but a (unlabeled) larger public dataset can substantially improves the robustness to noise perturbation. We observe from the right figure that the distillation results are insensitive to data precision, which is also observed in prior work (Shazeer et al. 2017). Thus we use $S = 200$ and $\gamma = 1$ as default setting in the following experiments.

Chest X-Ray Image Classification

Although mainstream FL methods experiment exclusively with private data from the same dataset (domain), this is typically not realistic in practical applications. For example, data acquired at different hospitals may come from different sources. We thus consider a more general heterogeneous setting where the private data at different local nodes and the unlabeled public data all come from different domains.

Here we implement multi-label classification on chest-x-ray images, using the NIH CXR14 (Wang et al. 2017) and CheXpert (Irvin et al. 2019) datasets to represent different domains for private data. We ignore ambiguous categories (Effusion, Pleural Effusion, Pleural Other and Support Device), remaining a total of 14 annotation classes, of which NIH CXR14 has annotations for 12 classes and CheXpert for 8 classes, with 6 overlapping classes. So there are totally 86,524 images come from NIH CXR14 and 64,346 images come from CheXpert dataset. For each dataset, we randomly sample 90% for training and the rest 10% for validation. We use 26,684 images from the RSNA Pneumonia Detection Challenge (RSNA and Kaggle 2018) without using their labels as public data.

Implementation Details. We use ResNet-34 as the backbone. For local training, we use a batch size of 32, same data augmentation strategies as in prior work (Ye et al. 2020). We train each local model individually with SGD and Cosine Annealing , decreasing the learning rate from 1e-4 to 1e-6 in 50 epochs. For distillation, we use SGD and a constant learning rate of 1e-3 and 50 epochs. For samples with multiple classes labeled as positive, we choose the most infrequent one (the class with least positive samples) as its label for the Dirichlet data split. In the setting with cross-domain private data (two datasets as private data), each dataset is distributed to $K_d = K/2$ local nodes when there is a total of K local nodes.

Results with Unlabeled Public Data. In Table 3, we first study the hyper-parameters K with $\alpha = 1$, $S = 200$, $\gamma = 1$ and local data from a single dataset (domain). It shows larger numbers of locals K negatively affects the distillation performance. Table 3 also shows cross-domain, cross-site evaluations using both datasets as private data, with a total of K local nodes ($K_d = K/2$ for each dataset, and each node hosts data from only one of the datasets). We can see that the introduction of additional cross-domain local nodes will help to improve the performance of the source domain: CXR14 ($K = 3$) as private datasets achieves 75.02%

	Homogeneous						Heterogeneous					
	CM	ED	CS	AT	PE	**mAUC**	CM	ED	CS	AT	PE	**mAUC**
Standalone	78.57 ±2.27	85.82 ±1.95	88.16 ±2.12	79.87 ±4.22	84.60 ±1.58	83.67 ±1.24	69.12 ±5.15	82.63 ±3.48	83.26 ±2.74	70.71 ±0.63	80.32 ±3.49	77.21 ±1.29
Public-only	67.34	79.76	79.24	76.38	80.37	82.43	45.28	78.03	77.36	66.98	75.43	68.60
Centralized	**82.88**	**87.04**	**91.53**	80.90	**87.02**	**85.88**	75.38	82.28	86.37	**75.36**	**85.93**	**81.07**
FedKD	81.81	86.12	91.15	**83.34**	86.59	85.81	**75.62**	**82.83**	**87.95**	74.61	83.48	80.90

Table 4: Comparisons of AUCs (%) on the homogeneous/heterogeneous positive data distribution with $K = 5$ and labeled public data. *Standalone*: averaged AUCs of all local models. *Public-only*: training with only labeled public data. *Centralized*: central training with all public and private data. *CM*: Cardiomegaly, *ED*: Edema, *CS*: Consolidation, *AT*: Atelectasis, *PE*: Pleural Effusion.

		FedAvg	FedDF	FedKD	*Standalone*	*Centralized*
AG News	Accuracy (%) ↑	91.98	92.57	**92.58**	86.30±5.21	93.11
	Bandwidth(MB) ↓	10217	10235	**36.6**	-	-
SST2	Accuracy (%) ↑	87.13	88.51	**91.50**	74.80±5.05	90.07
	Bandwidth(MB) ↓	10217	10221	**10.3**	-	-
Privacy (NO shared Param.)		✗	✗	✓	-	-

Table 5: Comparisons on AG News and SST2 datasets with K=10 under the same experiment setting. *Standalone*: mean ± std of local models trained with individual private data. *Centralized*: centralized training all local private data.

on CXR14 test set while CXR14+Xpert ($K = 6$) as private datasets achieves 76.13%. Note that the model trained with this cross-domain setting is capable of classifying all 14 classes, whereas training with a single domain can only classify 12 and 8 classes, respectively.

Ablation Studies on Heterogeneity with Labeled Public Data. In this experiment, we use labeled public data $\mathcal{D}^0 = \{(\boldsymbol{x}_i^0, \boldsymbol{y}_i^0)|i = 1, \ldots, N_0\}$ which is accessible by all local nodes and included in local training along with local private data. Since medical image datasets are usually characterized by a high degree of imbalance (e.g., far more negative samples than positive samples with abnormalities), we study the heterogeneity of the positive distribution, with each local node having an equal number of private samples. We set the number of local nodes to $K = 5$ and the data size to $N_k = 6000$, $N_0 = 1000$ and use the official validation set for testing. Table 4 shows results with homogeneous and heterogeneous distributions (w.r.t. positive samples). Notably, under both homogeneous and heterogeneous settings, our method achieves results comparable to centralized training on all public and local data. This can be viewed as an upper bound.

Text Classification Tasks

We evaluate our framework on two text classification datasets: AG News (Zhang, Zhao, and LeCun 2015) and SST2 (Socher et al. 2013). Following FedDF (Lin et al. 2020), we use pre-trained DistilBERT (Sanh et al. 2019) as the transformer language model. Local training and distillation takes 100 and 20 epochs, respectively, and the training strategy is the same as FedDF. From Table 5, we can note that our method gives the best performance on both datasets. On bot AG News and SST2 dataset, our proposed framework achieves superior accuracy and substantially lower communication bandwidth compared to the prior arts. More importantly, our method does not share parameters/gradients of local models during communication, which offers much stronger privacy guarantee compared to the prior arts.

Conclusions

In this work, we propose a novel distillation-based federated learning framework, namely FedKD, which can preserve local data privacy by learning with only unlabeled and domain robust public data. To comprehensively address the communication bottleneck, we employ a one-shot and one-way (offline) knowledge distillation process with an efficient ensemble scheme. Experiments on both image classification and text classification tasks demonstrate the efficacy of FedKD with better privacy guarantee compared to prior arts. Given the increasing importance of privacy, we believe our proposed FL method will be a practical solution to facilitate privacy-preserving decentralized learning across multiple sites in real-world scenarios, especially for medical applications where leveraging valuable local data at different hospitals without exposing proprietary data to privacy risks is essential.

Acknowledgements

We thank the reviewers for their constructive comments and thank Liangchen Song and Barry M. Yao for the discussion and assistance.

References

Asif, U.; Tang, J.; and Harrer, S. 2019. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2007. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dvornik, N.; Schmid, C.; and Mairal, J. 2019. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 3723–3731.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting Gradients–How easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*.

Gong, X.; Sharma, A.; Karanam, S.; Wu, Z.; Chen, T.; Doermann, D.; and Innanje, A. 2021. Ensemble Attention Distillation for Privacy-Preserving Federated Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15076–15086.

Guo, Q.; Wang, X.; Wu, Y.; Yu, Z.; Liang, D.; Hu, X.; and Luo, P. 2020. Online Knowledge Distillation via Collaborative Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11020–11029.

Hamm, J.; Cao, Y.; and Belkin, M. 2016. Learning privately from multiparty data. In *Proceedings of the International Conference on Machine Learning*, 555–563.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *NeurIPS Deep Learning Workshop*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2020. Federated Visual Classification with Real-World Data Distribution. *European Conference on Computer Vision*, 76–92.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, 590–597.

Jeong, E.; Oh, S.; Kim, H.; Park, J.; Bennis, M.; and Kim, S.-L. 2018. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for on-device federated learning. In *Proceedings of the International Conference on Machine Learning*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Citeseer*.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2018. Federated optimization in heterogeneous networks. In *arXiv preprint arXiv:1812.06127*.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2020a. Fair resource allocation in federated learning. *Proceedings of the International Conference on Learning Representations*.

Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M. J.; et al. 2019. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 133–141.

Li, X.; Gu, Y.; Dvornek, N.; Staib, L. H.; Ventola, P.; and Duncan, J. S. 2020b. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*, 65: 101765.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble Distillation for Robust Model Fusion in Federated Learning. *34th Conference on Neural Information Processing Systems*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the International Conference on Learning Representations*.

RSNA; and Kaggle. 2018. Radiological Society of North America (RSNA) pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. Accessed: 2018-11-30.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*.

Shin, M.; Hwang, C.; Kim, J.; Park, J.; Bennis, M.; and Kim, S.-L. 2020. XOR Mixup: Privacy-Preserving Data Augmentation for One-Shot Federated Learning. In *Proceedings of the International Conference on Machine Learning*.

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1310–1321.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 4424–4434.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Song, G.; and Chai, W. 2018. Collaborative learning for deep neural networks. In *Advances in Neural Information Processing Systems*, 1832–1841.

Song, L.; Wu, J.; Yang, M.; Zhang, Q.; Li, Y.; and Yuan, J. 2021. Robust Knowledge Transfer via Hybrid Forward on the Teacher-Student Model. In *Proceedings of the AAAI conference on Artificial Intelligence*, 2558–2566.

Sui, D.; Chen, Y.; Zhao, J.; Jia, Y.; Xie, Y.; and Sun, W. 2020. Feded: Federated learning via ensemble distillation for medical relation extraction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2118–2128.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020. Federated learning with matched averaging. In *Proceedings of the International Conference on Learning Representations*.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3462–3471.

Wu, A.; Zheng, W.-S.; Guo, X.; and Lai, J.-H. 2019. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1187–1196.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, 247–263.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–19.

Ye, W.; Yao, J.; Xue, H.; and Li, Y. 2020. Weakly Supervised Lesion Localization With Probabilistic-CAM Pooling. *arXiv preprint arXiv:2005.14480*.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing systems*, 649–657.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhou, H.; Song, L.; Chen, J.; Zhou, Y.; Wang, G.; Yuan, J.; and Zhang, Q. 2021. Rethinking Soft Labels for Knowledge Distillation: A Bias-Variance Tradeoff Perspective. In *International Conference on Learning Representations*.

Zhou, Y.; Pu, G.; Ma, X.; Li, X.; and Wu, D. 2020. Distilled One-Shot Federated Learning. *arXiv preprint arXiv:2009.07999*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 14774–14784.

Zhu, X.; Gong, S.; et al. 2018. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, 7517–7527.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-Free Knowledge Distillation for Heterogeneous Federated Learning. In *Proceedings of the International Conference on Machine Learning*.