Ensemble Attention Distillation for Privacy-Preserving Federated Learning

Xuan Gong[1,2], 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, doermann}@buffalo.edu {first.last}@uii-ai.com

Abstract

We consider the problem of Federated Learning (FL) where numerous decentralized computational nodes collaborate with each other to train a centralized machine learning model without explicitly sharing their local data samples. Such decentralized training naturally leads to issues of imbalanced or differing data distributions among the local models and challenges in fusing them into a central model. Existing FL methods deal with these issues by either sharing local parameters or fusing models via online distillation. However, such a design leads to multiple rounds of inter-node communication resulting in substantial bandwidth consumption, while also increasing the risk of data leakage and consequent privacy issues. To address these problems, we propose a new distillation-based FL framework that can preserve privacy by design, while also consuming substantially less network communication resources when compared to the current methods. Our framework engages in inter-node communication using only publicly available and approved datasets, thereby giving explicit privacy control to the user. To distill knowledge among the various local models, our framework involves a novel ensemble distillation algorithm that uses both final prediction as well as model attention. This algorithm explicitly considers the diversity among various local nodes while also seeking consensus among them. This results in a comprehensive technique to distill knowledge from various decentralized nodes. We demonstrate the various aspects and the associated benefits of our FL framework through extensive experiments that produce state-of-the-art results on both classification and segmentation tasks on natural and medical images.

1. Introduction

Modern deep learning algorithms rely on massive annotated datasets for many practical applications [63, 21, 18]. In most cases, however, this data is physically located across multiple disparate locations and regulated by different entities. This results in the challenges of centralizing the physically dispersed data, with the primary concerns being privacy and network bandwidth issues. Consequently, federated learning (FL) [46, 54, 17] has emerged as an important topic where a single centralized model is trained in a distributed, decentralized fashion using model fusion/distillation techniques.

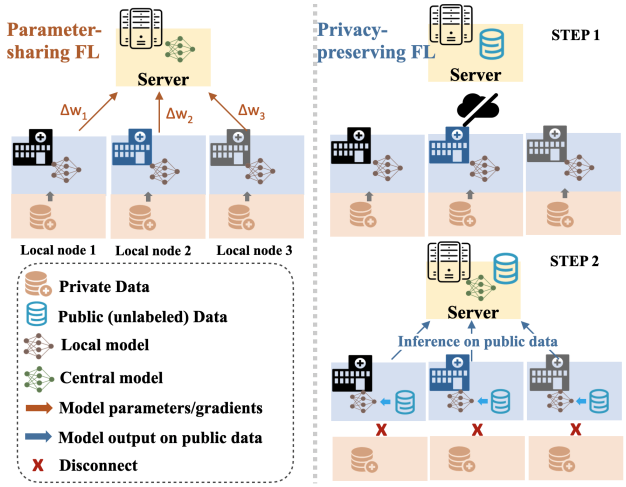

Figure 1. A schematic illustration of the proposed privacy-preserving federated learning framework compared to traditional update or parameter-sharing based federated learning frameworks. Traditional FL frameworks transfer gradients or parameter updates produced with private data from local nodes to a server, risking privacy leakage. Our framework only transfers products of unlabeled public data.

While some similarities exist, there are many more unique challenges that make FL substantially different from distributed learning. First, privacy is a critical consideration, and maintaining it is of utmost importance, particularly for applications such as healthcare [44, 3]. Second, one needs to be able to train centralized models efficiently and not get bogged down by network communication issues. Communication bandwidth may be quite pronounced depending on the task (e.g. image and video applications) and quantity of data (e.g., model parameters) being shared

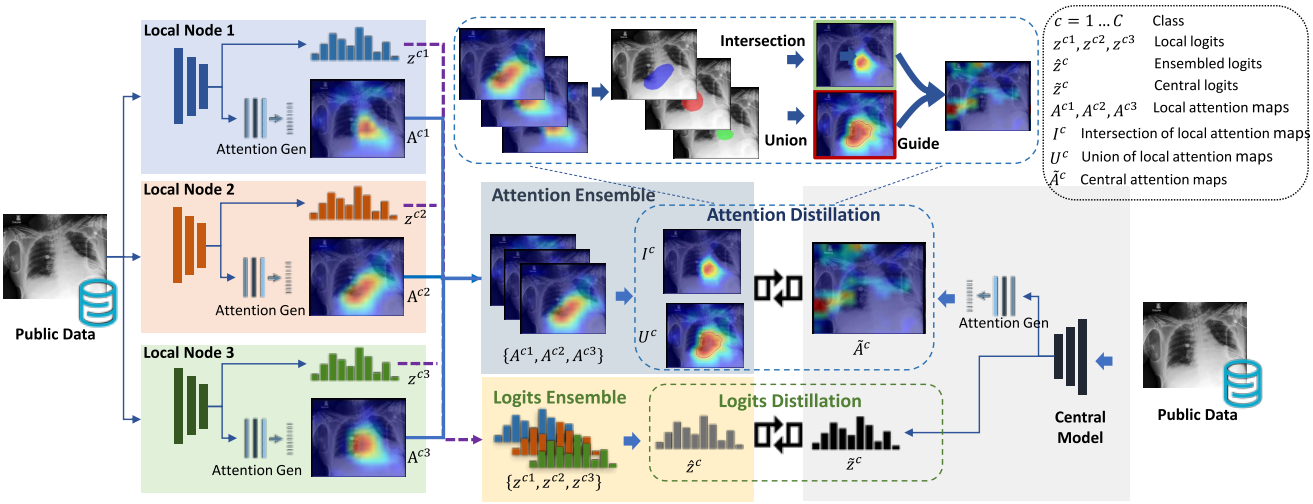

Figure 2. **Overview of the proposed FedAD framework.**

among local nodes. Finally, given that raw data typically resides locally, the data collection and preprocessing mechanisms may vary substantially among the local nodes, leading to a situation where common assumptions such as independent and identically distributed sampling do not hold.

The current state-of-the-art FL techniques approach the aforementioned issues by repeatedly sharing local model parameters or their gradients during the training process [34, 47, 12, 24, 58, 49, 19, 30, 8]. However, for many applications, these parameter-based communication methods suffer from impractical network bandwidth overhead, are limited only to models with homogeneous architectures, and more importantly, have many known security weaknesses [4]. While some methods have taken a step towards data protection in medical imaging [26, 27], there are also counter arguments in the literature that show local private data can get exposed as a result of using publicly shared gradients in the FL technical pipeline [61].

Distillation-based techniques that aggregate locally-computed logits [16, 22, 4, 28, 62] form the basis for another line of work for building central models from multiple local models, helping eliminate the need for each local model to follow the same architecture. While some recent methods distill with public data to get around data privacy issues [22, 4], they assume both the public and private data are sampled from the same underlying distribution. This assumption invariably exposes private data to security risks and attacks. While the recently proposed FedDF [28] method provides some relaxations (e.g., public data can be unlabeled and domain robust, i.e., sampled from another domain), it still exchanges model parameters recursively, resulting in privacy vulnerabilities due to model memorization [61, 4]. Despite the known bottleneck of communication in FL, all the above methods jointly (online) optimize the central and local models by synchronizing local inferred predictions. This approach requires a high degree of synchronization and communication bandwidth. It is clear that these co-distillation methods require recursive communication primarily because these methods ensemble with dark knowledge, such as averaging to soften labels, leaving structural knowledge unexplored. In other words, in addition to distilling the final *what* (i.e., logits), using more feature information depicting the *why* (e.g., attention maps) should lead to improved performance and efficiency, which is largely ignored by the current state-of-the-art.

In this work, we present a new distillation-based federated learning framework to address the aforementioned issues (Figure 1). First, our framework presents stronger privacy guarantees of local data by only using model outputs of unlabeled public data during distillation *without any exchange of local model parameters or gradients*. This, by design, eliminates the vulnerabilities identified by prior work. Second, in our framework, local models are fully trained and then distilled to the central server, in contrast to prior works [28, 4, 22] that synchronously update local models through online distillation.

Our key insight is that such well-trained local models, as opposed to incremental snapshots of "half-baked" models [28, 4, 22], provide more structural knowledge about their expertise. This design choice immediately enables top-down class-specific attention maps that capture the fully-trained local model's reasoning process (note that for methods that do online distillation, this would not be possible since their attention maps would be incomplete due to incremental training). We ensemble local knowledge with both predicted logits and these attention maps, capturing each local model's final output as well as the underlying reasoning process. We also use these attention maps to capture the knowledge diversity across models and reach a consensus to effectively coordinate local expertise in the FL

paradigm. This in-depth ensemble strategy enables our federated distillation to be completed in an offline fashion in a single round (which we call one-shot), helping keep local model training independent and asynchronous. This results in a FL framework that is substantially more efficient and flexible when compared to prior art.

To demonstrate efficacy, we conduct extensive experiments on CIFAR10/100 and large-scale chest x-ray datasets. We also show our framework is flexible to be used in other tasks by conducting preliminary proof-of-concept experiments on the segmentation tasks, where we demonstrate the state-of-the-art privacy/performance trade-offs compared to prior methods.

To summarize, our key contributions are below:

- We propose a one-shot federated learning framework with one-way distillation to explicitly preserve the privacy of local data by only distilling model outputs on unlabeled and domain robust public data.
- Our framework addresses the communication inefficiencies of prior work by communicating high-level logits and model-agnostic attention maps.
- We introduce a seminal distillation algorithm that aggregates structural knowledge with explicit balance between both local model diversity as well as consensus to deal with the inherent heterogeneity of decentralized federated learning.
- We show that the proposed framework can be extended to other applications such as semantic segmentation with evaluations on Cityscapes and BraTS dataset.

2. Related Work

Our proposed FL framework is related to areas such as knowledge transfer and the type of FL algorithms and how they deal with issues such as privacy. Here, we briefly review methods that are relevant to these topics.

Knowledge Transfer. Recent methods can be categorized based on their ensemble strategy and the kind of information distilled to the student. Following the work of Hinton et al. [11], there has been much progress in model ensemble, with a particular focus on the student-teacher learning paradigm [43]. These techniques aggregate the knowledge of multiple teachers before distilling it to the student model. This has led to a variety of aggregation schemes, with gate learning being quite popular in the supervised setting [43, 1, 53, 48, 59]. In semi-supervised and self-supervised scenarios, techniques based on relative sample similarity have been proposed [56, 52]. Going beyond the use of soft label distillation [11], recent approaches have explored the transfer of structural knowledge such as intermediate feature representations [41], Gram matrices [55], maximum mean discrepancy [14], or mutual information [40]. Finally, there have also been some attempts at combining both structure as well as label ensemble strategies, with FEED [39] and knowledge flow [29] being notable examples. In contrast to these techniques, our feature-level ensemble method is label-free, model agnostic, and can also be used in heterogeneous knowledge distillation scenarios.

Federated Learning. Most existing FL methods are either parameter-based or distillation-based. 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, e.g., by averaging [34]. This result 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 number of extensions to FedAVG [34, 49, 25, 13] have been proposed with new aggregation schemes such as momentum [12] or local weighting [25, 13], or new local training strategies such as the use of proximal term [24] or control variations[19]. However, such parameter/gradient sharing can certainly be a straightforward way of information exchange, it is highly susceptible to privacy leakage and stealth attacks, as also demonstrated elsewhere [61, 4].

On the other hand, distillation-based methods exchange model outputs (on local private data) [60, 45] rather than the parameters, leading to growing concerns on the privacy of local data. While some methods address this issue by distilling on public data [16, 22, 4], they often select public data based on some prior knowledge of private data, leading to similar security vulnerabilities as above. While the recently proposed FedDF [28] attempts to address this issue, it is quite inefficient (e.g., high network bandwidth) due to the iterative exchange of models over hundreds of rounds, which in turn also leads to more susceptibility to stealth attacks and hence privacy concerns. The framework most similar to ours is PATE [38], where all local models can be trained independently without inter-institutional communication, and the central model is trained with hard pseudo labels voted by local models. In contrast, our method preserves privacy by exchanging the soft prediction of domain robust public data, and further exploiting in-depth feature level information for high efficient communication.

Privacy Concerns for FL. Parameter-based FL methods have shown to be highly susceptible to privacy leakage [61, 4]. Furthermore, as noted above, some distillation-based FL methods are also at privacy risk when recursively model exchanges are involved [28]). Utilizing unlabeled public data during distillation, while also restricting the local nodes' access to the server model, has been shown to be more resilient to attacks and guarantee privacy [38, 9]. Therefore, our framework explicitly protects local private

data by using one-shot communication, and public data from different domain or distributions.

3. Approach

We first begin with a formal introduction to the problem before discussing the design of our proposed framework, called Ensemble Attention Distillation federated learning (FedAD).

3.1. Problem Definition

Let there be K local/private nodes with each node playing host to labeled (private) dataset $\mathcal{D}_k = \{(x_k^i, y_k^i)|i = 1, \ldots, |\mathcal{D}_k|\}$, where $y_k \in \mathcal{C}_k$, $\mathcal{C}_k$ is the set of existing classes in dataset $\mathcal{D}_k$, and $\mathcal{C}_k \subset \{1, \ldots, C\}$ (C is the overall number of classes across all local nodes). The shared public dataset $\mathcal{D}_0 = \{x_0^i|i = 1, \ldots, |\mathcal{D}_0|\}$ can be accessed by any local node and can unlabeled or labeled.

As illustrated in Figure 2, in the first stage of FedAD, the model at each local node k is trained using the corresponding private data $\mathcal{D}_k$ (θ_k represents these local model parameters after local training). Since our proposed FedAD is agnostic to the type of network architecture, each local neural network can be customized to have its own architecture, helping adapt the model to its local data distribution.

In the second stage, we disconnect local datasets from the network to minimize security risks and protect privacy. Instead, the central public dataset $\mathcal{D}_0$ is used at each local node as part of a one-way knowledge distillation framework from the local models to the central server. The set of local models and the central model form a teacher-student knowledge transfer setup where the teacher is a local-model ensemble (one local model at each local node).

3.2. Logits Ensemble Distillation

During each distillation step t, we sample (randomly) a subset $\mathcal{K}_t$ of models from all local nodes. This subset comprises a fraction γ of all local models. The standard distillation pipeline uses the Kullback-Leibler divergence to ensemble all teachers' soft labels:

$$\mathcal{L} = \sum_c p(y = c) log \frac{p(y = c)}{q(y = c)}, \tag{1}$$

where p is the sample's probability being class c for the teacher model and q is the corresponding value for the student model. Given a sample x_0 from the public dataset, let $z^{ck} = f(x_0, \theta_k, c)$ be its logits (on class c from local model at node k) and $\widetilde{z}^c = f(x_0, \theta_s, c)$, where $c \in \{1, \ldots, C\}$ be the corresponding central model output. Given these notations, the standard ensemble strategy $\widehat{z}^c = \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} z^{ck}$ uses the average operation on logits from all teachers' with a softmax activation as:

$$p(y = c) = \frac{exp(\widehat{z}^c/\tau)}{\sum_c exp(\widehat{z}^c/\tau)}, q(y = c) = \frac{exp(\widetilde{z}^c/\tau)}{\sum_c exp(\widetilde{z}^c/\tau)} \tag{2}$$

It has been shown in prior work [11] that minimizing Eq. 1 with a high temperature parameter τ is the same as to minimizing the ℓ_2 error between student and teacher logits, hence establishing a relationship between matching logits and the cross entropy objective. Without any loss of generality, let the activation be $q^c = \sigma(\widetilde{z}^c)$ and $p^c = \sigma(\widehat{z}^c)$.

Since the FL setting involves extreme heterogeneity, the standard distillation process discussed above is not directly suitable. This is because of the inability to tackle a general scenario where all local models do not share the same set of target classes. To explicitly consider this in our proposed framework, during our distillation procedure, we introduce a new variable, called the importance weight ω, for each local model so as to capture the corresponding local data distribution used in the initial training step:

$$\widehat{z}^c = \sum_{k \in \mathcal{K}_t} \omega_k^c z^{ck}, \omega_k^c = \frac{N_k^c}{\sum_{k \in \mathcal{K}_t} N_k^c}, \tag{3}$$

where $N_k^c = \sum_{i=1}^{|\mathcal{D}_k|} (y_k^i = c)$ is the number of c-class samples used in training the local model at node k.

Following the observation above, we set $\tau \to \infty$ and rewrite the logits learning objective as:

$$\mathcal{L}_{\text{logits}}(\widetilde{\mathbf{z}}, \widehat{\mathbf{z}}) = \frac{1}{C} \|\widetilde{\mathbf{z}} - \widehat{\mathbf{z}}\|, \tag{4}$$

where $\widetilde{\mathbf{z}} = [\widetilde{z}^1, \ldots, \widetilde{z}^C]$ and $\widehat{\mathbf{z}} = [\widehat{z}^1, \ldots, \widehat{z}^C]$. Note that the aforementioned methods can be easily applied pixel-wise, making it readily applicable for segmentation tasks.

Multi-label Classification: We next show our proposed method can be easily extended to multi-class classification (to go with single class above). Here, we denote the private data as $\mathcal{D}_k = \{(x_k^i, y_k^i)|i = 1, \ldots, |\mathcal{D}_k|\}$ with $y_k^i \in \{-1, 0, 1\}^C$ where -1, 0, and 1 represent unknown, negative and positive for class c. Finally, the adaptation requires a few more changes. Instead of softmax, we use the sigmoid $p^c = \sigma(\widehat{z}^c)$ and $q^c = \sigma(\widetilde{z}^c)$ as the activation function. Next, we modify Eq. 3 by defining $N_k^c = \sum_{i=1}^{|\mathcal{D}_k|} (y_k^i(c) = 1)$.

3.3. Attention Ensemble Distillation

Logits distillation discussed above essentially captures the divergence between the output vectors of teacher and student models. However, comparing only the output vectors in this fashion only ensures the outputs can match and does not necessarily mean the underlying structural knowledge, or the model's reasoning can be transferred. We posit that such knowledge, e.g., intermediate feature representations of models, can result in more accurate distillation of

knowledge, more so in cases such as FL with its high degree of heterogeneity in local data sources. While this can seem intuitive, there are practical challenges in implementing this method. Specifically, transferring mostly bulky matrices such as intermediate feature representations is quite resource intensive, e.g., from the perspective of network bandwidth burden, and furthermore, relies on restrictive requirements such as identical network architecture among the student and teacher models.

On the other hand, representations such as class-specific top-down attention maps generated with methods such as Grad-CAM [42] have been widely used in providing location cues for weakly supervised semantic segmentation, and has been shown to be effective in providing more precise and efficient guidance to the learning process [23]. Furthermore, recent studies such as [31] extend Grad-CAM to generic embedding models, making it applicable to any tasks that use or need a fully-connected feature extractor. We argue that such top-down model interpretations can transfer knowledge in a more efficient (instead of full feature tensors) and effective (instead of just the output vectors) way, without sacrificing communication efficiency or risking privacy leakage.

Sampling by importance. Let the attention map from local node k for class c be $\boldsymbol{A}^{ck} \in \mathbb{R}^{HW}$, where H, W are the size of attention maps, and the complete set of attention maps from all local nodes be $\boldsymbol{\mathcal{A}} = \{\boldsymbol{A}^{ck} | k \in \mathcal{K}_t, c = 1, \ldots, C\}$. Due to the high degree of heterogeneity among local nodes, we weight attention maps from each local node differently based on the sample distribution in its corresponding private dataset. Similar to FedVC [13] that resamples local nodes based on the probability proportional to local data size, we sample the attention maps based on the class-specific importance weight ω_k^c in Eq. 3. That is, we independently decide on whether to use $\boldsymbol{A}^{ck}$ with the probability $\hat{\omega}_k^c$ in each batch.

To ensure at least $\widehat{K}$ samples are selected each time, we set $\hat{\omega}_k^c = \max(1, \frac{\omega_k^c}{\mathrm{Top}_{\widehat{K}}(\omega_k^c)})$, where $\mathrm{Top}_{\widehat{K}}(\omega_k^c)$ denotes the $\widehat{K}$-th maximum value among $\{\omega_k^c | k \in \mathcal{K}_t\}$. In our experiments, we set $\widehat{K} = 2$ by default. At each sample step t, we obtain a set of selected local indexes $\widehat{\mathcal{K}}_t^c$ for class c as:

$$\widehat{\mathcal{K}}_t^c = \mathrm{sample}(\{\hat{\omega}_k^c | k \in \mathcal{K}_t\}), \tag{5}$$

where $\widehat{\mathcal{K}}_t^c \subset \mathcal{K}_t$ and $|\widehat{\mathcal{K}}_t^c| \geqslant \widehat{K}$. Thus, the selected set of attention maps with respect to class c is $\boldsymbol{\mathcal{A}}_t^c = \{\boldsymbol{A}^{ck} | k \in \widehat{\mathcal{K}}_t^c\}$. To simplify notions, from this point on, we refer to $\boldsymbol{\mathcal{A}}_t^c$ as $\boldsymbol{\mathcal{A}}^c = \{\boldsymbol{A}^{ck} | k \in \widehat{\mathcal{K}}^c\}$.

Attention bound constraint. While there have been a few recent efforts in enforcing constraints directly on gradient-based attention maps [7, 50, 31], they cannot directly be applied to our scenario with an ensemble of attention maps produced by highly heterogeneous models.

Algorithm 1 FedAD on K local nodes with C classes.

Input: Labeled private data $\{\mathcal{D}_k\}$, unlabeled public data $\mathcal{D}_0$, central model θ_s, local models $\{\theta_k\}$, T distillation steps, batch-size S, sample fraction γ .
Local Training: Train each local model θ_k with $\mathcal{D}_k$
for each distillation step $t = 1, ..., T$ **do**
  $\mathcal{K}_t \leftarrow$ random subset (γ fraction) from K locals
  $\mathbf{x}_0 \leftarrow$ a batch of public data from $\mathcal{D}_0$ with size S
  for each local $k \in \mathcal{K}_t$ **do**
    $\mathbf{z}^k, \mathbf{A}^k \leftarrow f(\mathbf{x}_0; \theta_k)$
  end for
  $\widehat{\mathbf{z}} \leftarrow$ ensemble $\{\mathbf{z}^k | k \in \mathcal{K}_t\}$ ▷ Eq. 3
  for each class $c = 1, ..., C$ **do**
    $\hat{\mathcal{K}}_t^c \leftarrow$ sample a subset from $\mathcal{K}_t$ ▷ Eq. 5
    $\boldsymbol{I}^c, \boldsymbol{U}^c \leftarrow$ ensemble $\boldsymbol{\mathcal{A}}^c = \{\boldsymbol{A}^{ck} | k \in \widehat{\mathcal{K}}_t^c\}$ ▷ Eq. 6
  end for
  $\widetilde{\mathbf{z}}, \widetilde{\mathbf{A}} \leftarrow f(\mathbf{x}_0, \theta_s)$
  Update: $\theta_s \leftarrow \theta_s - \frac{1}{S} \nabla_{\theta_s} \mathcal{L}(\widetilde{\mathbf{z}}, \widehat{\mathbf{z}}, \widetilde{\mathbf{A}}, \mathbf{I}, \mathbf{U})$ ▷ Eq. 10
end for

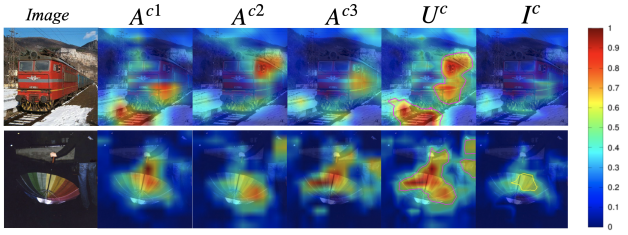

Figure 3. Illustration of Eq. 6. Suppose $k \in \widehat{\mathcal{K}}^c = \{1, 2, 3\}$ with respect to class c, $\boldsymbol{A}^{ck}$ indicate local attention maps, $\boldsymbol{U}^c$ and $\boldsymbol{I}^c$ indicate union attention maps and intersection attention maps respectively.

Specifically, the training objective based on the ℓ_1 or ℓ_2 functions in these methods seek to enforce exactly the same activation strength in each location of the attention map, which in our case may likely introduce and amplify noise from local models. To achieve consensus while also maintaining the diversity that is inherent among the local nodes, we design a novel attention bound constraint based on the intersection and union of local attention maps with respect to the same input data and class output. Given the set of local attention maps $\boldsymbol{\mathcal{A}}^c = \{\boldsymbol{A}^{ck} | k \in \widehat{\mathcal{K}}^c\}$, we denote $\boldsymbol{I}^c, \boldsymbol{U}^c \in \mathbb{R}^{HW}$ as the intersection and union among all the local attention maps $\boldsymbol{\mathcal{A}}^c$ with respect to class c, respectively. Let h,w be the pixel index, we have

$$I_{hw}^c = \min_{k \in \widehat{\mathcal{K}}^c} A_{hw}^{ck}, U_{hw}^c = \max_{k \in \widehat{\mathcal{K}}^c} A_{hw}^{ck}, \tag{6}$$

where $\boldsymbol{I}^c$ denotes a consensus on the high-response region among all the local attention maps, that has a high probability to comprise the object of interest. While $\boldsymbol{U}^c$ considers all of the high-response regions among the local attention maps, it also preserves diversity of "expertise" among the local models by means of the union operation. Figure 3 shows an example of such attention maps. Given the at-

tention map $\boldsymbol{A}$, let $T(\cdot)$ be a soft-masking operation (with sigmoid) to have all values lie between 0 and 1 [23]:

$$T(\boldsymbol{A}) = \frac{1}{1 + exp(-\rho(\boldsymbol{A} - b))}. \quad (7)$$

For simplicity, we denote $\widetilde{\boldsymbol{A}}^c$ as the attention map generated by central model with respect to class c. Considering the attention intersection $\boldsymbol{I}^c$ as the consensus achieved by all locals, we enforce the high-response region in $\widetilde{\boldsymbol{A}}^c$ to explicitly include that of $\boldsymbol{I}^c$ using our proposed attention intersection loss, defined as:

$$\mathcal{L}_{\text{inter}}(\widetilde{\mathbf{A}}, \mathbf{I}) = -\frac{1}{C}\sum_c \frac{\sum_{hw} I^c_{hw} \cdot T(\widetilde{A}^c_{hw}; \rho_1, b_1)}{\sum_{hw} I^c_{hw}}, \quad (8)$$

where $\widetilde{\mathbf{A}} = [\widetilde{\boldsymbol{A}}^1, \ldots, \widetilde{\boldsymbol{A}}^C]$, $\mathbf{I} = [\boldsymbol{I}^1, \ldots, \boldsymbol{I}^C]$, $\rho_1 = 10$, and $b_1 = 0.6$. With the attention union $\boldsymbol{U}^c$ including all the high-response regions among all locals, we enforce the high-response region in $\widetilde{\boldsymbol{A}}^c$ to be explicitly inside that of $\boldsymbol{U}^c$ using our attention union loss $\mathcal{L}_{\text{union}}$. The intuition here we seek each high-response pixel in the $\widetilde{\boldsymbol{A}}^c$ to have support from at least one local model to take model diversity into account. The $\mathcal{L}_{\text{union}}$ loss is defined as:

$$\mathcal{L}_{\text{union}}(\widetilde{\mathbf{A}}, \mathbf{U}) = -\frac{1}{C}\sum_c \frac{\sum_{hw} \widetilde{A}^c_{hw} \cdot T(U^c_{hw}; \rho_2, b_2)}{\sum_{hw} \widetilde{A}^c_{hw}}, \quad (9)$$

where $\mathbf{U} = [\boldsymbol{U}^1, \ldots, \boldsymbol{U}^C]$, $\rho_2 = 10$, and $b_2 = 0.3$.

Our method optimizes the learning of the central model so that $\widetilde{\boldsymbol{A}}^c$ is encouraged to activate in the pixels which are activated in $\boldsymbol{I}^c$, and penalized for activating in the pixels not activated in $\boldsymbol{U}^c$. The attention bound constraint combines the two constraints balancing the local consensus and diversity. This is a relaxed constraint as it is capable of tolerating up to $|\widehat{\mathcal{K}}^c_t| - 1$ incorrect/biased attention maps, and maintaining high robustness to outliers, which is particularly important to tackle with heterogeneity in the FL setting. The overall loss for optimization is

$$\mathcal{L} = \mathcal{L}_{\text{logits}}(\widetilde{\mathbf{z}}, \widehat{\mathbf{z}}) + \mathcal{L}_{\text{inter}}(\widetilde{\mathbf{A}}, \mathbf{I}) + \mathcal{L}_{\text{union}}(\widetilde{\mathbf{A}}, \mathbf{U}). \quad (10)$$

Segmentation. Crucially, our framework by no means is limited to just classification problems. The distillation algorithm of our proposed FedAD framework can be used to aggregate knowledge for other kinds of tasks as well, e.g., segmentation. In this case, we aggregate predicted masks $\boldsymbol{z}^{ck}$ in Eq. 5 and N^c_k is counted pixel-wisely. Thus the logits in Eq. 4 can be written as $\widetilde{\mathbf{z}} = [\widetilde{\boldsymbol{z}}^1, \ldots, \widetilde{\boldsymbol{z}}^C]$ and $\widehat{\mathbf{z}} = [\widehat{\boldsymbol{z}}^1, \ldots, \widehat{\boldsymbol{z}}^C]$. The other modification is we obtain the class-specific attention map of local model through activation: $\boldsymbol{A}^{ck} = \sigma(\boldsymbol{z}^{ck})$, where σ can be softmax or sigmoid. More details are in supplementary materials.

The overall process is explained in Algorithm 1.

Accuracy(%)	**Shared** Param.	*CIFAR-10* $\alpha = 1$	*CIFAR-10* $\alpha = 0.1$	*CIFAR-100* $\alpha = 1$	*CIFAR-100* $\alpha = 0.1$
FedAvg [34]	Y	78.57 ±0.22	68.37 ±0.50	42.54 ±0.51	36.72 ±1.50
FedProx [24]	Y	76.32 ±1.95	68.65 ±0.77	42.94 ±1.23	35.74 ±1.00
FedAvgM [12]	Y	77.79 ±1.22	68.63 ±0.79	42.83 ±0.36	36.29 ±1.98
FedDF [28]	Y	80.69 ±0.43	71.36 ±1.07	47.43 ±0.45	39.33 ±0.03
FedMD [22]	N	80.37 ±0.37	69.23 ±1.31	45.83 ±0.58	39.86 ±0.78
Standalone	N	61.11 ±24.90	28.99 ±27.24	27.49 ±14.76	16.31 ±15.75
FedAD	N	**82.48** ±0.21	**73.11** ±1.25	**50.34** ±0.33	**48.43** ±1.01

Table 1. Results on CIFAR-10 and CIFAR-100 with ResNet-8 when γ=0.4 and K=20, comparing our FedAD with several existing parameter-based [34, 24, 12, 28] and distillation-based [22] FL methods. *Standalone*: mean/std accuracy of all local models.

		FedDF [28] $\gamma = 0.4$	FedAD (Ours) $\gamma = 0.4$	$\gamma = 0.8$	$\gamma = 1$
Accuracy (%) ↑	$\alpha = 1$	80.69	82.48	83.16	**83.68**
	$\alpha = 0.1$	71.36	73.11	73.29	**73.40**
Bandwidth (GB) ↓		29.4	**10.1**	22.1	27.6

Table 2. FL communication efficiency on CIFAR-10. In FedDF, both logits and parameters are of type float64, whereas attention maps are of type float16.

aggregation scheme	[28, 13]	Eq. 3	Eq. 3	Eq. 3
logits distillation	$\tau = \infty$	$\tau = 3$	$\tau = \infty$	$\tau = \infty$
attention distillation	✗	✗	✗	✓
Accuracy(%)↑	65.73	67.81	72.67	**73.11**

Table 3. Ablation study on CIFAR-10 with ResNet-8, γ=0.4, K=20, and α=0.1. We compare our aggregation scheme (Eq. 3) with the existing aggregation schemes [28, 13], and different temperatures for logits distillation, we only list the result with typical value τ=3 [11].

4. Experiments

We conduct a number of experiments on natural images/ medical images for both classification/segmentation tasks. Additionally, we experiment on text classification tasks and provide the results in the supplementary.

4.1. Classification

In constructing local training sets, we use heterogeneous data splits using a Dirichlet distribution as in prior works [12]. The value of α controls the degree of non-IID-ness: a plus infinite α indicates identical local data distribution, and a smaller α indicates higher non-IID-ness. While distilling knowledge with task-relevant data from the same domain is ideal, FedAD is compatible with using public data from different domains. Here, we thus consider a more general heterogeneous setting where the public data come from different domains. Note we save augmentation seeds locally and transmit the predictions of each seed during distillation

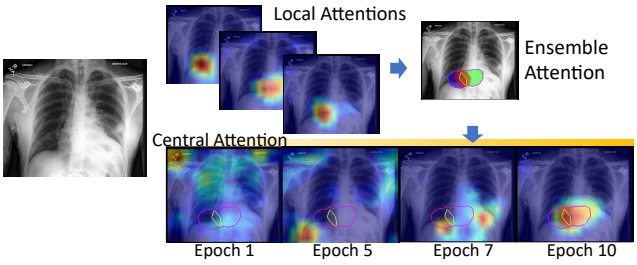

Figure 4. One example showing how ensemble attention can effectively guide the central model focus on the right region.

Methods	Distill	Test	mAUC(%)↑	Distill	Test	mAUC(%)↑
FedMD[22]	C^3	C	71.23	X^3	X	77.66
	C^3+X^3	C	70.97	C^3+X^3	X	77.02
FedAD	C^3 *alone*	C	75.08 ± 4.44	X^3 *alone*	X	73.67 ± 7.31
	C^3	C	75.17	X^3	X	82.62
	C^3+X^3	C	75.55	C^3+X^3	X	**82.65**
Centralized	C^c	C	79.73	X^c	X	82.58
	C^c+X^c	C	**80.47**	C^c+X^c	X	80.91

Table 4. Results on CXR14 and CheXpert when private data is inside/cross domain. We split each dataset with $K_d = 3$, $\alpha = 1$,$\gamma = 1$. 'C' refers to CXR14, 'X' refers to CheXpert, and $\{C, X\}^n$ denotes n local nodes with data from the dataset, where an additional *alone* denotes mean/std of all local models. $\{C, X\}^c$ denotes centralized training all local data.

for classification tasks.

4.1.1 CIFAR-10/100 Classfication

We first consider CIFAR-10/100 [20]. For a fair comparison, we follow the same setup as in FedDF [28], using CIFAR-10 and CIFAR-100 as private datasets, and CIFAR-100 and a downsampled version of ImageNet (32×32) as the corresponding public datasets for CIFAR-10 and CIFAR-100 respectively. We use test datasets on the central server and report the average accuracy over three different random seeds. Implementation details are provided in the supplementary material.

The comparison in Table 1 shows that our method outperforms current state-of-the-art methods by a large margin, particularly for the higher non-IID-ness scenario. Next, we compare the accuracy and communication efficiency at different sample fractions $\gamma = \{0.4, 0.8, 1.0\}$ with the current state-of-the-art method [28] in Table 2, where our method achieves higher accuracy on CIFAR-10 while consuming significantly lower communication bandwidth. We observed that the distillation result is not sensitive to data precision, which has also been shown in prior work [43], and hence used float16 for attention maps. In Table 3, we show the results of ablation studies to validate the efficacy of our proposed ensemble and distillation strategy.

4.1.2 Chest X-Ray Images Classification

With privacy being a particularly important topic for real-world medical applications, we believe our proposed FL framework will be a good fit to facilitate privacy-preserving learning across various hospital sites. To this end, we evaluate our method on cross-domain, cross-site learning with private data, which is relatively under-explored in the contemporary FL methods.

We use NIH chestX-ray14 (NIH CXR14) [51] and CheXpert [15] as two domains where private data come from. To ensemble their labels, we disregard ambiguous categories such as Effusion, Pleural Effusion, and Pleural Other, and other samples labeled "Support Device", leaving 86,524 images in NIH CXR14 and 64,346 images in CheXpert for training across 14 classes. Of these, NIH CXR14 has annotations for 12 classes and CheXpert for 8 classes, with 6 of these classes overlapping across both datasets. For public data, we use 26,684 x-ray images in the RSNA Pneumonia Detection Challenge public data [1] without using their labels. For K local nodes, each private dataset is distributed to $K_d = K/2$ local nodes. Implementation details and ablation studies are in supplementary material.

Figure 4 illustrates the effectiveness of bound attention constraint during ensemble distillation. In Table 4, we compare FedAD to FedMD[22] on multi-label chest-x-ray image classification under the same settings, where one can note our FedAD outperforms FedMD by a significant margin on both test datasets. While distilling and testing on the data from the same, cross domain with FedAD obtains the best result for CheXpert dataset, distilling from one domain gives comparable performance. Note that our proposed FedAD uses additional unlabeled public data during distillation when compared to the centralized training setting. Under this cross-domain setting, the trained model is capable of classifying all 14 classes, whereas training with a single domain can only classify 12 and 8 classes, respectively due to the annotation limitations as noted above.

4.2. Extension to Segmentation

We evaluate our method on segmentation on the Cityscapes dataset and 3D Brain Tumor Segmentation (BraTS) dataset. Note each local model predicts one time on the public data, thus the communication cost is independent of the distillation iterations.

4.2.1 Cityscapes Segmentation

For natural images, we use Cityscapes dataset [5] for semantic segmentation of urban street scenes with 50 cities. In constructing local training sets, we split 2975 training images into three subsets based on the countries the images

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

Method				mIoU (%) ↑	Pixel Accuracy(%)↑
Standalone(mean ± std)				64.11 ± 10.58	93.74 ± 1.79
Centralized				75.65	95.87
PATE [38]				64.48	93.99
	τ=1	τ=∞	*w/ att.*		
FedAD	✓		✗	58.56	92.91
		✓	✗	71.86	94.94
		✓	✓	**72.97**	**95.07**

Table 5. Segmentation results on Cityscapes. We calculate IoU between the ground truth mask and the predicted segmentation mask for each class, and report mean IoU of all classes.

Standalone(mean ± std)	73.38 ± 3.44	*Centralized*	82.13
	Li *et al.* [26]	FedMD [22]	FedAD
Average Dice (%) ↑	84.33	75.71	77.85
Bandwidth (GB)↓	64.37	2154.84	**13.36**
Privacy (NO Shared Param.)	✗	✓	✓
Flexibility	online	online	on/offline
One-shot	✗	✗	✓

Table 6. Segmentation results on BraTS. We compare average Dice score over three types of tumor regions, communication bandwidth, privacy guarantee, and other attributes such as flexibility.

are collected from, e.g., Germany, France and Switzerland. We use 1,525 unlabeled test images as public data. Note the cities where the public images are collected from are different from with either of that of the private data. We test 19 classes on the validation dataset of cityscapes which contains 500 images.

We adopt PSPNet [57] with ResNet101 [10], with the backbone pretrained on ImageNet [6]. We use mini-batch SGD with the learning rate initialized as 0.01 and decreased to 0 in 100,000 iterations, and the same data augmentation as [32]. In Table 5, we compare with the other offline decentralized learning method PATE [38], which is also privacy-guaranteed (no shared weights/gradients). With the same locally trained models, our FedAD demonstrates superior performance on both mIoU and pixel accuracy. For ablation study, Table 5 also compares with the commonly used [32, 33] $\tau = 1$ case, showing the efficacy of our distillation method for knowledge aggregation.

4.2.2 Brain Tumor Segmentation

For medical images, we use the BraTS 2018 dataset [35, 2] that contains multi-parametric preoperative 3D MRI scans of 285 subjects with brain tumors. Each subject was scanned under the T1-weighted, T1-weighted with contrast enhancement, T2-weighted, and T2 fluid-attenuated inversion recovery (T2-FLAIR) modalities. Following the experimental protocol in prior work [26], we use 242 subjects for the training set and 43 subjects for held-out test set. The training set is stratified into three subsets according to the institution the data is originated ("2013", "CBICA", "TCIA") and assigned each to federated local client. We use half of the unlabeled validation set of the BraTS 2020 dataset [35, 2] as the public data comprising 62 subjects independent of either private dataset.

While the aforementioned distillation method is designed for general tasks which use cross entropy objective, we optimize with dice loss [36] for brain tumor segmentation. We thus modify Eq. 4 with soft dice to constrain the predicted mask $\widehat{\boldsymbol{z}}^c$ to match the aggregated masks $\widetilde{\boldsymbol{z}}^c$. Let $\boldsymbol{p}^c = \sigma(\widehat{\boldsymbol{z}}^c)$ be the predicted probabilities of 3D voxel being class c, $\boldsymbol{q}^c = \sigma(\widetilde{\boldsymbol{z}}^c)$ be the 3D soft pseudo labels, we have:

$$\mathcal{L}_{\text{logits}} = -\frac{1}{C}\sum_c \frac{2 \cdot \sum \boldsymbol{p}^c \cdot \boldsymbol{q}^c}{\sum (\boldsymbol{p}^c)^2 + \sum (\boldsymbol{q}^c)^2 + \epsilon}, \tag{11}$$

where the summation on $\boldsymbol{p}^c$, $\boldsymbol{q}^c$ are voxel-wise, and ϵ is a small constant to avoid numerical instability. Note in Eq. 8 and Eq. 9 the summation in numerator and denominator are modified to be voxel-wise as well for this 3D case.

We use the same network backbone as [26]. The training strategy is the same as [37]. We train each local model individually with 20,000 iterations, with local-to-central distillation taking 5,000 iterations. The weight decay is 5e-4 and 0 for local training and distillation, respectively. Table 6 compares the segmentation performance and other utilities with parameter-based [26] and distillation-based [22] federated methods. We can note that our method achieves the state-of-the-art privacy/performance trade-offs compared with other existing methods.

5. Conclusions

In this work, we propose a one-shot federated learning framework, called FedAD, that can in principle preserve local data privacy using only unlabeled and domain robust public data and be efficient in the utilization of available network bandwidth resources when compared to competing prior art. Another key challenge in federated learning is its inherent heterogeneity, which can manifest itself in various ways: different differing domain distributions, different local model architectures, or simply the diversity in knowledge across all local models. To comprehensively address these issues, our framework also includes a novel knowledge distillation algorithm that is based on ensemble distillation of prediction logits as well as structural knowledge by means of model attention. Extensive experiments on classification and segmentation tasks with both natural image and medical image datasets demonstrated the efficacy of FedAD while also not risking privacy leakage. Furthermore, with experiments and analyses using cross-domain and heterogeneous data distributions, we also demonstrated FedAD's applicability in real-world cross-institutional learning with medical imaging data.

References

[1] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*, 2019. 3

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 8

[3] Christopher M Bishop and Markus Svensén. Bayesian hierarchical mixtures of experts. *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 2012. 1

[4] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279*, 2019. 2, 3

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 8

[7] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019. 5

[8] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2423–2432, 2021. 2

[9] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *Proceedings of the International Conference on Machine Learning*, pages 555–563, 2016. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Deep Learning Workshop*, 2015. 3, 4, 6

[12] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 2, 3, 6

[13] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. *European Conference on Computer Vision*, pages 76–92, 2020. 3, 5, 6

[14] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 3

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 7

[16] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018. 2, 3

[17] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021. 1

[18] Srikrishna Karanam, Ren Li, Fan Yang, Wei Hu, Terrence Chen, and Ziyan Wu. Towards contactless patient positioning. *IEEE Transactions on Medical Imaging*, 2020. 1

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *Proceedings of the International Conference on Machine Learning*, 2020. 2, 3

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[21] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2019. 1

[22] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019. 2, 3, 6, 7, 8

[23] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 5, 6

[24] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *arXiv preprint arXiv:1812.06127*, 2018. 2, 3, 6

[25] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *Proceedings of the International Conference on Learning Representations*, 2020. 3

[26] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng,

Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141, 2019. 2, 8

[27] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020. 2

[28] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *34th Conference on Neural Information Processing Systems*, 2020. 2, 3, 6, 7

[29] Iou-Jen Liu, Jian Peng, and Alexander G Schwing. Knowledge flow: Improve upon your teachers. In *Proceedings of the International Conference on Learning Representations*, 2019. 3

[30] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2

[31] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020. 5

[32] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 8

[33] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 8

[34] 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*, pages 1273–1282. PMLR, 2017. 2, 3, 6

[35] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014. 8

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 8

[37] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320, 2018. 8

[38] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the International Conference on Learning Representations*, 2017. 3, 8

[39] SeongUk Park and Nojun Kwak. Feed: Feature-level ensemble for knowledge distillation. In *Proceedings of the AAAI conference on Artificial Intelligence*, 2020. 3

[40] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2020. 3

[41] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations*, 2015. 3

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2017. 5

[43] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*, 2017. 3, 7

[44] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104, 2018. 1

[45] MyungJae Shin, Chihoon Hwang, Joongheon Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Xor mixup: Privacy-preserving data augmentation for one-shot federated learning. In *Proceedings of the International Conference on Machine Learning*, 2020. 3

[46] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321, 2015. 1

[47] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017. 2

[48] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Robust knowledge transfer via hybrid forward on the teacher-student model. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 2558–2566, 2021. 3

[49] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *Proceedings of the International Conference on Learning Representations*, 2020. 2, 3

[50] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N Metaxas. Sharpen focus: Learning with attention separability and consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 512–521, 2019. 5

[51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 3462–3471, 2017. 7

[52] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1187–1196, 2019. 3

[53] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263, 2020. 3

[54] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 1

[55] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 3

[56] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 3

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 8

[58] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 2

[59] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In *International Conference on Learning Representations*, 2021. 3

[60] Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020. 3

[61] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, pages 14774–14784, 2019. 2, 3

[62] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the International Conference on Machine Learning*, 2021. 2

[63] Yiming Zuo, Weichao Qiu, Lingxi Xie, Fangwei Zhong, Yizhou Wang, and Alan L Yuille. Craves: Controlling robotic arm with a vision-based economic system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4214–4223, 2019. 1