SEMI-SUPERVISED FEDERATED LEARNING FOR KEYWORD SPOTTING

Enmao Diao, Eric W. Tramel, Jie Ding, Tao Zhang

Amazon, Alexa AI

ABSTRACT

Keyword Spotting (KWS) is a critical aspect of audio-based applications on mobile devices and virtual assistants. Recent developments in Federated Learning (FL) have significantly expanded the ability to train machine learning models by utilizing the computational and private data resources of numerous distributed devices. However, existing FL methods typically require that devices possess accurate ground-truth labels, which can be both expensive and impractical when dealing with local audio data. In this study, we first demonstrate the effectiveness of Semi-Supervised Federated Learning (SSL) and FL for KWS. We then extend our investigation to Semi-Supervised Federated Learning (SSFL) for KWS, where devices possess completely unlabeled data, while the server has access to a small amount of labeled data. We perform numerical analyses using state-of-the-art SSL, FL, and SSFL techniques to demonstrate that the performance of KWS models can be significantly improved by leveraging the abundant unlabeled heterogeneous data available on devices.

***Index Terms*—** Keyword Spotting, Semi-Supervised Learning, Federated Learning, Semi-Supervised Federated Learning

1. INTRODUCTION

Keyword spotting (KWS) focuses on identifying pre-specified keywords in audio signals derived from human speech. Recognizing speech commands is crucial for audio-based interactions on mobile devices and virtual assistants [1]. This work investigates the potential of Federated Learning (FL) [2, 3] for training a supervised global model on distributed edge devices. In several application domains, including KWS, it may not be feasible for clients to provide precise annotations for unlabeled on-device data. These observations have prompted us to develop an FL framework that utilizes the unlabeled audio streams on distributed clients' devices.

In this work, we demonstrate the efficacy of Semi-Supervised Learning (SSL) and Federated Learning (FL) for Keyword Spotting (KWS) and propose a novel SSL and FL integration for Semi-Supervised Federated Learning (SSFL) in KWS. Additionally, we show that the alternate training method used in SSFL can mitigate the Non-IID problem in FL and transfer pre-trained KWS models with unlabeled on-device data. Our research focuses on label-skewed data heterogeneity, where devices may have different keyword distributions. We conduct extensive experiments and ablation studies to assess the effectiveness of data augmentation techniques for KWS to address the SSL and SSFL challenges. Figure 1 demonstrates that leveraging unlabeled on-device data in a distributed manner can significantly enhance the performance of KWS models trained with labeled data.

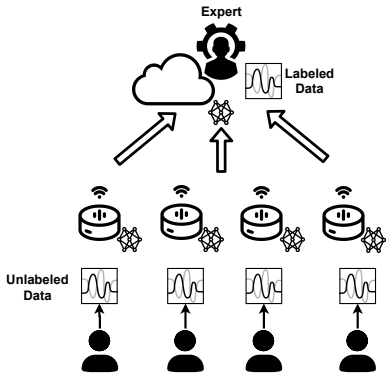

Fig. 1. An illustration of Semi-Supervised Federated Learning (SSFL) for Keyword Spotting (KWS).

2. RELATED WORKS

Keyword spotting aims to detect predetermined specific words from audio streams [1]. Usually, audio streams are processed locally on users' devices to save computation and communication expenses. Recently, data augmentation and deep neural networks have proved to be effective for KWS [4, 5]. Semi-Supervised Learning (SSL) refers to the process of training a model using partially labeled data, especially when the amount of labeled data is significantly less than that of unlabeled data. Recently, SSL methods based on image augmentation techniques have been developed for computer vision tasks to achieve state-of-the-art performance [6, 7]. Federated Learning (FL) aims to expedite and scale the distributed training of models [2]. Several studies have been proposed to reduce the computation and communication costs in FL [2, 3]. It has been demonstrated that for speech recognition, FL is advantageous for KWS to exploit local data and computation resources [8]. The integration of Semi-Supervised Learning and Federated Learning is commonly referred to as Semi-Supervised Federated Learning (SSFL) [7].

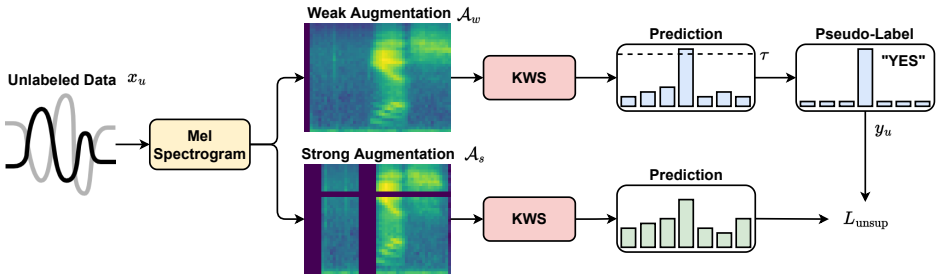

Fig. 2. An illustration of leveraging unlabeled audio streams in Semi-Supervised Learning (SSL).

3. METHOD

Semi-Supervised Federated Learning In this research, we investigate Semi-Supervised Federated Learning (SSFL) in a scenario where the data available on the device-side is entirely unlabeled, and only the server possesses some labeled data. Previous studies [7] have shown that if the model parameters trained by the server and devices are directly aggregated, the performance may significantly deteriorate when devices train multiple local epochs [2]. Therefore, we apply the alternate training technique [7] that employs unlabeled and labeled data iteratively to update models.

We consider a scenario where there are M devices. At the t-th round of Federated Learning (FL), M_t active devices upload their trained model θ_m^t to the server for aggregation, resulting in the aggregated model $\theta^t = \frac{1}{M_t}\sum_{m=1}^{M_t}\theta_m^t$. The server has access to a small labeled dataset $\mathcal{L} = x_{l,i}, y_{l,i}{}^{N_\mathcal{L}} i = 1$, which can be utilized to train a server model with the supervised loss Lsup. Here, $\mathcal{A}_w$ denotes a function of weak data augmentation [6], $f\theta$ is a parameterized Keyword Spotting (KWS) model, and ℓ is the cross-entropy function. The supervised loss is expressed as:

$$L_{\text{sup}} = \ell(f_\theta(\mathcal{A}_w(xl)), y_l). \tag{1}$$

We use an alternate training technique to balance the training of labeled and unlabeled data. In the subsequent iteration, active devices receive the fine-tuned model from the server and generate pseudo-labels for training the unlabeled data $x_{u,i}{}^{N_\mathcal{U}} i = 1$ using the unsupervised loss Lunsup. Here, $\mathcal{A}s$ denotes a function of strong data augmentation [6], and τ is a threshold value. The unsupervised loss is expressed as:

$$\begin{aligned} L\text{unsup} &= \ell(f_\theta(\mathcal{A}s(xu)), y_u), \\ y_u &= f_\theta(\mathcal{A}_w(xu)), \max(y_u) \geq \tau. \end{aligned} \tag{2}$$

Data augmentation Data augmentation is a crucial component of state-of-the-art SSL methods. Prior works on KWS have employed various data augmentation techniques, such as random shifting and resampling of audio streams, addition of background noises, and masking of frequency and time components [4, 5]. To better leverage SSL methods, we treat mel-spectrograms as images and employ image-based data augmentation techniques such as RandAugment [9]. Mixup data augmentation [7, 10], which uses a random convex combination to mix labeled and unlabeled data, is another type of SSL method. We conduct extensive ablation studies and demonstrate that appropriate utilization of data augmentation strategies can leverage the abundance of unlabeled data to enhance the performance of KWS models.

Transfer from pretrained models Classical SSL and SSFL methods assume a small labeled dataset and a large unlabeled dataset, i.e., $N_\mathcal{L} \ll N_\mathcal{U}$. When a large labeled dataset is available, SSFL can adapt pre-trained KWS models to new data domains. We propose fine-tuning pre-trained models $\tilde{\theta}$ with a small labeled dataset on the server and a large unlabeled dataset on devices. The transferred KWS model parameters can be obtained as $\hat{\theta} = \underset{\tilde{\theta}}{\text{argmin}}, \ell(f_{\tilde{\theta}}(\mathcal{A}_w(x_l)), y_l) + \ell(f_{\tilde{\theta}}(\mathcal{A}_s(x_u)), y_u)$, where $\mathcal{A}_w$ and $\mathcal{A}_s$ are functions of weak and strong data augmentation, respectively, and ℓ is the cross-entropy loss function.

4. EXPERIMENTS

4.1. Experimental Setup

Speech Commands datasets We conduct experiments with the public benchmark datasets Speech Commands V1 and V2 datasets [11]. The datasets comprise twelve class labels, including ten words ('yes', 'no', 'up', 'down', 'left', 'right', 'on', 'off', 'stop', and 'go'), and two classes ('silence' and 'unknown'). The 'silence' class is randomly sampled from the background noise, while the 'unknown' class comprises the remaining keywords with equal size from the datasets. As our backbone model, we employ Temporal Convolution ResNet18 (TC-ResNet18) [12]. Our experiments involve 100 devices, and we maintain a constant ratio of active devices per communication round $C = 0.1$ in all experiments [2]. We equally allocate data examples for the IID data partition. For a balanced Non-IID data partition, we adopt the method outlined in earlier work [3,7] and set $K = 2$ to simulate label-skewed data heterogeneity. We sample data from a Dirichlet distribution $\text{Dir}(\alpha)$ [7] for an unbalanced Non-IID data partition. We conduct four random experiments with different seeds and report the standard deviation in parentheses for tables and by error bars in figures.

Wakeword dataset In this study, we conduct an evaluation of SSL and SSFL techniques on an internal audio dataset that has been de-identified. The dataset comprises approximately 14K hours of training data and 1.5K hours of test data, and is used for the purpose of wakeword detection. Wakeword

detection is a task that involves using an on-device model to identify a specified keyword and thereby initiate device activation. Specifically, our experiment focuses on a binary classification task with two output classes: wakeword detected and wakeword absent. To this end, we represent audio streams as mel-spectrograms and employ a Convolutional Neural Network (CNN) with multiple convolution and max-pooling layers. To evaluate the performance of the resulting models, we use the Relative False Reject Rate at a False Accept Rate (Relative FRR@FAR) metric. This metric is defined as the ratio of the FRR of the test model to that of the baseline model, given a fixed FAR of the baseline model. Consequently, a Relative FRR@FAR value less than one is preferred. We conduct four random experiments with different seeds, and the standard deviation is within 0.01.

4.2. Experimental Results

Semi-Supervised Federated Learning In this study, we compare our findings with two baseline methods, namely, 'Fully Supervised' and 'Partially Supervised' in the centralized setting. 'Fully Supervised' refers to the approach of training the model with all data being labeled, while 'Partially Supervised' denotes training the model with partially labeled data. To evaluate the results, we use Accuracy in Table 1 and Relative FRR@FAR in Table 2. Our analysis reveals that SSL and SSFL techniques outperform the 'Partially Supervised' case and perform comparably to the 'Fully Supervised' case. Specifically, with 250 and 2500 labeled data from the Speech Command datasets, we achieve approximately 40% and 10% improvement over the 'Partially Supervised' case in Accuracy, respectively. With 2500 labeled wakeword streams, we improve the performance of the pretrained models and the 'Partially Supervised' case by around 0.2 and 0.1 in Relative FRR@FAR, respectively. Overall, our results demonstrate the effectiveness of training KWS models from scratch or transferring from pretrained models with abundant unlabeled on-device data to improve performance.

In this study, we present the learning curves of SSL and SSFL as depicted in Figure 3. In addition to Test Accuracy, we provide visual representations of the outcomes of 'Fully Supervised' and 'Partially Supervised' techniques, the accuracy of pseudo-labels before (Label Accuracy) and after (Threshold Accuracy) employing a confidence threshold, and the proportion of unlabeled data above the confidence threshold (Label Ratio). The findings demonstrate an increase in the Accuracy of generated labels subsequent to the filtration of pseudo-labels employing a confidence threshold. Additionally, the labeling ratio increases, indicating a reduction in the prediction's entropy.

Alternate training for heterogeneous on-device data We compare parallel and alternate training for Federated Learning (FL) using Speech Commands V1 and V2 datasets, as presented in Tables 1 and 2. 'Parallel' training denotes the traditional combination of Semi-Supervised Learning (SSL) and FL, while 'Alternate' training involves the server and devices training in an alternating manner [7]. Our results demonstrate that alternate training significantly improves the performance of FL in Non-IID data distributions. Specifically, in the most heterogeneous case (Non-IID, $K = 2$), alternate training outperforms parallel training by 30% and 20% in Accuracy for Speech Commands V1 and V2, respectively. Furthermore, when transferring from pretrained models with highly imbalanced on-device data, we observe that alternate training outperforms parallel training by 0.04 in Relative False Rejection Rate (FRR) at a given False Acceptance Rate (FAR). The findings suggest that a small amount of labeled IID data at the server can effectively address Non-IID data partition in the FL setting.

Data augmentation Our investigation focuses on four data augmentation strategies, namely 'BasicAugment,' 'SpecAugment,' 'RandAugment,' 'RandAugment (Selected),' and 'MixAugment.' 'BasicAugment' involves time shift, resampling, and random background noise, as proposed by Rybakov et al. [5]. 'SpecAugment' utilizes masking on frequency and time components of mel-spectrograms, as proposed by Park et al. [4]. 'RandAugment' employs image-based augmentation methods proposed by Cubuk et al. [9], while 'RandAugment (Selected)' uses selected image-based augmentation strategies. 'MixAugment' utilizes a random

Table 1. Comparison of SSL and SSFL with baselines for the public Speech Commands datasets using Accuracy (↑).

Dataset			Speech Commands V1		Speech Commands V2	
Proportion of Supervised Data			≈1%	≈10%	≈1%	≈ 10%
Fully Supervised			99.7(0.0)		99.6(0.0)	
Partially Supervised			62.4(0.7)	90.3(0.2)	60.2(0.9)	89.7(0.1)
Semi-Supervised			95.7(0.2)	97.8(0.0)	96.5(0.1)	98.1(0.1)
Non-IID, $K = 2$	Parallel	FL	68.2(1.7)		77.9(2.4)	
	Alternate	FL	98.3(0.1)		97.6(0.0)	
		SSFL	88.6(0.3)	94.8(0.1)	86.9(0.4)	93.0(0.1)
Non-IID, $\mathrm{Dir}(0.1)$	Parallel	FL	97.6(0.2)		97.1(0.2)	
	Alternate	FL	98.3(0.2)		98.1(0.0)	
		SSFL	90.1(0.0)	95.6(0.1)	89.1(0.2)	94.2(0.2)
Non-IID, $\mathrm{Dir}(0.3)$	Parallel	FL	99.0(0.1)		98.7(0.1)	
	Alternate	FL	99.0(0.1)		98.8(0.0)	
		SSFL	93.3(0.4)	96.2(0.0)	93.1(0.1)	95.4(0.1)
IID	Parallel	FL	99.6(0.0)		99.3(0.0)	
	Alternate	FL	99.5(0.0)		99.3(0.0)	
		SSFL	95.3(0.2)	97.8(0.0)	95.9(0.1)	97.7(0.0)

Table 2. Comparison of SSL and SSFL with baselines for wakeword dataset using Relative FRR@FAR (↓) with respect to pretrained models.

Proportion of of Supervised Data			≈ 3%
Fully Supervised			0.80
Partially Supervised			0.90
Semi-Supervised			0.81
Imbalanced	Parallel	FL	0.85
	Alternate	FL	0.81
		SSFL	0.85

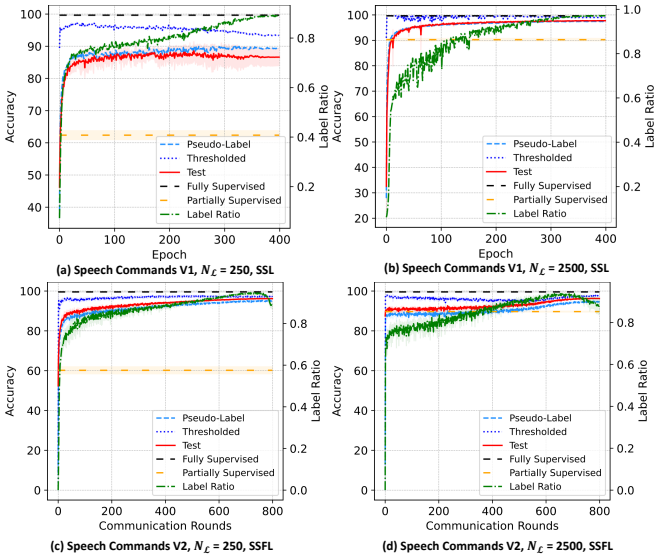

Fig. 3. Learning curves of SSL and SSFL for TC-ResNet18 with Speech Commands datasets and $N_{\mathcal{L}} = \{250, 2500\}$.

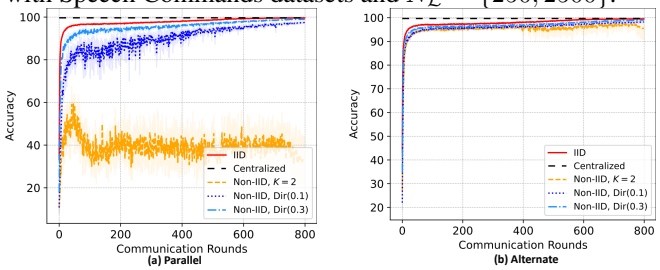

Fig. 4. Comparison between 'Parallel' and 'Alternate' training for heterogeneous on-device data.

convex combination to mix labeled and unlabeled data, as proposed by Berthelot et al. [10]. We conduct ablation studies for strong data augmentation and loss functions, as presented in Table 3. Specifically, we use 'BasicAugment' for weak data augmentation $\mathcal{A}_w$ and experiment with various combinations of strong data augmentation methods $\mathcal{A}_s$. Our findings demonstrate that using strong data augmentation can significantly enhance the performance of SSL methods. Moreover, we observe that image-based augmentation methods do not outperform 'SpecAugment,' and 'MixAugment' can be combined with 'SpecAugment' to further improve performance.

Table 3. Ablation studies of strong data augmentation methods used in SSL for KWS models with $N_{\mathcal{L}} = 250$.

Strong Augmentation $\mathcal{A}_s$	Speech Commands V1	Speech Commands V2
BasicAug	86.1(0.6)	85.2(0.6)
BasicAug, RandAug	90.4(0.2)	89.1(0.2)
BasicAug, RandAug (Selected)	89.8(0.8)	89.9(0.3)
BasicAug, SpecAug	95.2(0.3)	95.8(0.2)
BasicAug, SpecAug, RandAug (Selected)	94.4(0.3)	94.0(0.1)
BasicAug, SpecAug, MixAug	**95.7(0.2)**	**96.5(0.1)**

5. CONCLUSION

This work studie the efficacy of SSFL methods for KWS and different strong data augmentation techniques. Our experiments suggest that a small amount of labeled data on the server can enable training from scratch or transfer from pretrained models, leveraging heterogeneous unlabeled on-device data.

6. REFERENCES

[1] Iván López-Espejo, Zheng-Hua Tan, John Hansen, and Jesper Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, 2021.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*. PMLR, 2017, pp. 1273–1282.

[3] Enmao Diao, Jie Ding, and Vahid Tarokh, "HeteroFL: Computation and communication efficient federated learning for heterogeneous clients," in *ICLR*, 2021.

[4] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[5] Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirko Visontai, and Stella Laurenzo, "Streaming keyword spotting on mobile devices," *arXiv preprint arXiv:2005.06720*, 2020.

[6] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[7] Enmao Diao, Jie Ding, and Vahid Tarokh, "SemiFL: Communication efficient semi-supervised federated learning with unlabeled clients," *NeurIPS*, 2022.

[8] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, "Federated learning for keyword spotting," in *ICASSP*. IEEE, 2019, pp. 6341–6345.

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[10] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[11] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[12] Seungwoo Choi, Seokjun Seo, Beomjun Shin, Hyeongmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, and Sungjoo Ha, "Temporal convolution for real-time keyword spotting on mobile devices," *arXiv preprint arXiv:1904.03814*, 2019.