arXiv:2305.05110v1 [cs.LG] 9 May 2023

半监督 用于关键词检出的监督式联邦学习

Enmao Diao、*Eric W. Tramel*、*Jie Ding*、*Tao Zhang*

亚马逊，Alexa AI

摘要

关键词检出（KWS）是移动设备和虚拟助手中音频应用的关键组成部分。联邦学习（FL）的最新进展显著提升了利用众多分布式设备的计算资源和私有数据资源训练机器学习模型的能力。然而，现有的联邦学习方法通常要求设备拥有准确的真实标签，而在处理本地音频数据时，获取此类标签的成本可能很高，也不切实际。本研究首先展示了半监督学习（SSL）和联邦学习在关键词检出中的有效性。随后，我们进一步研究了用于关键词检出的半监督联邦学习（SSFL）：设备拥有完全无标签的数据，而服务器则可以访问少量有标签数据。我们采用最先进的半监督学习、联邦学习和半监督联邦学习技术进行数值分析，结果表明，利用设备上丰富的无标签异构数据，可以显著提升关键词唤醒模型的性能。

索引词— 关键词检出、半监督学习、联邦学习、半监督联邦学习

1. 引言

关键词检测（KWS）致力于识别从人类语音中提取的音频信号里预先指定的关键词。在移动设备和虚拟助手中，识别语音命令对于基于音频的交互至关重要 [1]。本研究探讨了利用联邦学习（FL） [2, 3] 在分布式边缘设备上训练有监督全局模型的潜力。在包括关键词检测在内的多个应用领域中，客户端可能无法为无标签的设备端数据提供精确标注。基于这些观察，我们开发了一种联邦学习框架，利用分布式客户端设备上的无标签音频流。

在本研究中，我们展示了半监督学习（SSL）和联邦学习（FL）在关键词检测（KWS）中的有效性，并提出了一种新颖的 SSL 与 FL 集成方法，用于关键词检测中的半监督联邦学习（SSFL）。此外，我们表明，SSFL 中采用的交替训练方法能够缓解 FL 中的非独立同分布问题，并利用端侧无标签数据迁移预训练关键词检测模型。我们的研究聚焦于标签偏斜的数据异质性，即不同设备可能具有不同的关键词分布。我们开展了大量实验和消融研究，以评估数据增强技术在关键词检测中的有效性，从而应对 SSL 和 SSFL 所面临的挑战。图 1 表明，以分布式方式利用设备端数据中的无标签数据，可以显著提升使用有标签数据训练的关键词唤醒模型的性能。

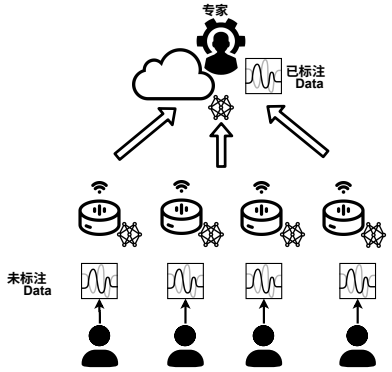

图 1. 用于关键词检出的半监督联邦学习（SSFL）示意图。

2. 相关工作

关键词检出旨在从音频流中检测预定特定词语 [1]。通常，为了节省计算开销和通信开销，音频流会在用户设备上进行本地处理。近年来，数据增强和深度神经网络已被证明对关键词检测有效[4, 5]。半监督学习（SSL）是指使用部分有标签的数据训练模型的过程，尤其是在有标签数据量显著少于无标签数据量时。近年来，基于图像增强技术的半监督学习方法已应用于计算机视觉任务，并取得了最先进的性能[6, 7]。联邦学习（FL）旨在加速并扩展模型的分布式训练 [2]。已有多项研究提出降低联邦学习中的计算开销和通信开销 [2, 3]。研究表明，在语音识别领域，联邦学习有利于关键词检出充分利用本地数据和计算资源 [8]。半监督学习与联邦学习的结合通常称为半监督联邦学习（SSFL） [7]。

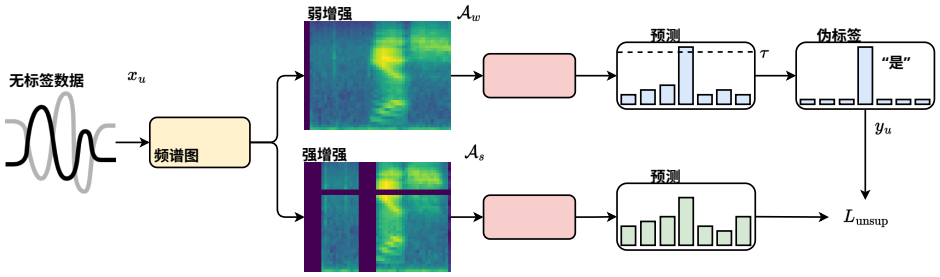

图 2。利用无标签音频流进行半监督学习（SSL）的示意图。

3. 方法

半监督联邦学习 本研究探讨一种半监督联邦学习（SSFL）场景：设备端可用的数据全部为无标签数据，只有服务器拥有一些有标签数据。先前的研究 [7] 表明，如果直接聚合由服务器和设备训练得到的模型参数，那么当设备进行多个本地训练轮次时，模型性能可能会显著下降 [2]。因此，我们采用交替训练技术 [7]，迭代使用无标签数据和有标签数据来更新模型。

我们考虑一个包含 M 台设备的场景。在联邦学习（FL）的第 t 轮中，M_t 台活跃设备将其训练好的模型 θ_m^t 上传至服务器进行聚合，从而得到聚合模型 $\theta^t = \frac{1}{M_t}\sum_{m=1}^{M_t}\theta_m^t$。服务器可以访问一个小型有标签数据集 $\mathcal{L} = \{x_{l,i}, y_{l,i}\}_{i=1}^{N_\mathcal{L}}$，并利用它通过监督损失 L sup 训练服务器模型。这里，$\mathcal{A}_w$ 表示弱数据增强函数 [6]，f_θ 是一个参数化的关键词检测（KWS）模型，而 ℓ是交叉熵函数。监督损失表示为：

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

我们采用交替训练技术，以平衡有标签数据和无标签数据的训练。在后续迭代中，活跃设备从服务器接收微调后的模型，并生成伪标签，以无监督损失 Lunsup 使用无标签数据 $\{x_{u,i}\}_{i=1}^{N_\mathcal{U}}$ 进行训练。这里，$\mathcal{A}s$ 表示强数据增强函数 [6]，，而 τ是一个阈值。无监督损失表示为：

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

数据增强数据增强是最先进半监督学习方法的重要组成部分。以往的关键词检测研究采用了多种数据增强技术，例如对音频流进行随机移位和重采样、添加背景噪声，以及掩蔽频率和时间分量 [4,5]。为了更好地利用半监督学习方法，我们将梅尔频谱图视为图像，并采用基于图像的数据增强技术，例如 RandAugment [9]。Mixup数据增强 [7, 10], 通过随机凸组合混合有标签数据和无标签数据，是另一种半监督学习方法。我们进行了广泛的消融研究，并证明，恰当运用数据增强策略能够利用丰富的无标签数据，提升关键词唤醒模型的性能。

从预训练模型迁移 经典的半监督学习和半监督联邦学习方法假设有一个小型有标签数据集和一个大型无标签数据集，即 $N_\mathcal{L} \ll N_\mathcal{U}$。当有大型有标签数据集可用时，半监督联邦学习可以将预训练关键词检测模型适配到新的˙数据域。我们提出在服务器上使用小型有标签数据集、在设备上使用大型无标签数据集，对预训练模型进行微调 θ。迁移后的关键词检测模型参数可表示为 $\theta = \text{argmin}\ \ell(f_\theta(\mathcal{A}_w(x_l))\ y_l) + \ell(f_\theta(\mathcal{A}_s(x_u))\ y_u)$，其中 $\mathcal{A}_w$ 和 $\mathcal{A}_s$ 分别是弱数据增强和强数据增强的函数，ℓ 是交叉熵损失函数。

4. 实验

4.1. 实验设置

Speech Commands 数据集 我们使用公开基准数据集 Speech Commands V1 和 V2 数据集 [11]开展实验。这些数据集包含十二个类别标签，包括十个单词（‘yes’、‘no’、‘up’、‘down’、‘left’、‘right’、‘on’、‘off’、‘stop’和‘go’），以及两个类别（‘silence’和‘unknown’）。“silence”类别从背景噪声中随机采样，而“unknown”类别由数据集中其余关键词构成，各关键词数量相等。我们采用时序卷积 ResNet18（TC-ResNet18） [12]作为骨干模型。实验涉及 100 台设备，并且在所有实验中，我们保持每个通信轮次的活跃设备比例恒定 $C = 0.1$ [2]。对于独立同分布数据划分，我们对数据样本进行均等分配。对于平衡的非独立同分布数据划分，我们采用早期工作 [3,7]中概述的方法，并设置 $K = 2$ 以模拟标签偏斜的数据异质性。对于不平衡的非独立同分布数据划分，我们从狄利克雷分布 Dir(α) [7] 中采样数据。我们使用不同的随机种子进行四次随机实验，并在表格中以括号内的数值、在图中以误差条报告标准差。

唤醒词数据集 在本研究中，我们在一个已去标识化的内部音频数据集上评估了半监督学习和半监督联邦学习技术。该数据集包含约 1.4 万小时的训练数据和 1,500 小时的测试数据，用于唤醒词检测。唤醒词

关键词检测是一项利用端侧模型识别指定关键词，并据此启动设备的任务。具体而言，我们的实验聚焦于一个二元分类任务，包含两个输出类别：检测到唤醒词和未检测到唤醒词。为此，我们将音频流表示为梅尔频谱图，并采用具有多个卷积层和最大池化层的卷积神经网络（CNN）。为了评估所得模型的性能，我们采用给定虚警率下的相对拒识率（Relative FRR@FAR）指标。该指标定义为：在基线模型的虚警率固定的情况下，测试模型的拒识率与基线模型的拒识率之比。因此，Relative FRR@FAR 的值越小于 1 越理想。我们使用不同的随机种子进行了四次随机实验，标准差在 0.01 以内。

4.2. 实验结果

半监督联邦学习 在本研究中，我们将研究结果与集中式设置下的两种基线方法——“全监督”和“部分监督”进行比较。“全监督”是指使用全部有标签数据训练模型的方法，而“部分监督”则表示使用部分有标签数据训练模型。为评估结果，我们在表 1 中采用准确率，在表 2 中采用相对 FRR@FAR。分析结果表明，半监督学习和半监督联邦学习技术优于“部分监督”情形，其表现与“全监督”情形相当。具体而言，使用语音命令数据集中的 250 条和 2500 条有标签数据时，与“部分监督”情形相比，我们在准确率上分别取得了约 40% 和 10% 的提升。使用 2500 条有标签唤醒词语音流时，与预训练模型和“部分监督”情形相比，我们在相对 FRR@FAR 上的性能分别提升了约 0.2 和 0.1。总体而言，研究结果证明，利用充足的无标签设备端数据从头训练关键词唤醒模型，或基于预训练模型进行迁移训练，能够有效提升性能。

在本研究中，我们展示了图3所示的半监督学习（SSL）和半监督联邦学习（SSFL）的学习曲线。除测试准确率外，我们还以可视化形式呈现了“全监督”和“部分监督”方法的结果、应用置信度阈值前（标签准确率）和后（阈值准确率）的伪标签准确率，以及置信度阈值以上的无标签数据比例（标签比例）。结果表明，使用置信度阈值筛选伪标签后，生成标签的准确率有所提高。此外，标签比例也有所增加，表明预测熵有所降低。

针对异构端侧数据的交替训练 我们使用 Speech Commands V1 和 V2 数据集，对联邦学习（FL）中的并行训练和交替训练进行了比较，结果如表 1 和表 2 所示。“并行”训练指半监督学习（SSL）与联邦学习（FL）的传统结合方式，而“交替”训练则涉及服务器和设备。

表1。使用准确率（↑）比较半监督学习（SSL）和半监督联邦学习（SSFL）与公开 Speech Commands 数据集上的基线方法。

数据集			Speech Commands V1		Speech Commands V2	
有监督数据比例			≈1%	≈10%	≈1%	≈ 10%
全监督			99.7(0.0)		99.6(0.0)	
部分监督			62.4(0.7)	90.3(0.2)	60.2(0.9)	89.7(0.1)
半监督			95.7(0.2)	97.8(0.0)	96.5(0.1)	98.1(0.1)
非独立同分布，$K=2$	并行	FL	68.2(1.7)		77.9(2.4)	
	交替	FL	98.3(0.1)		97.6(0.0)	
		半监督联邦学习 88.6(0.3)		94.8(0.1)	86.9(0.4)	93.0(0.1)
非独立同分布，狄利克雷分布(0。1)	并行	FL	97.6(0.2)		97.1(0.2)	
	交替	FL	98.3(0.2)		98.1(0.0)	
		半监督联邦学习 90.1(0.0)		95.6(0.1)	89.1(0.2)	94.2(0.2)
非独立同分布，狄利克雷分布(0.3)	并行	FL	99.0(0.1)		98.7(0.1)	
	交替	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	并行	FL	99.6(0.0)		99.3(0.0)	
	交替	FL	99.5(0.0)		99.3(0.0)	
		半监督联邦学习 95.3(0.2)		97.8(0.0)	95.9(0.1)	97.7(0.0)

表 2. 使用预训练模型，在唤醒词数据集上比较半监督学习和半监督联邦学习与基线方法的相对拒识率@虚警率（↓）。

监督数据占比			≈ 3%
全监督			0.80
部分监督			0.90
半监督			0.81
不平衡	并行	FL	0.85
	交替	FL	0.81
		SSFL	0.85

以交替方式进行训练 [7]。我们的结果表明，交替训练显著提升了联邦学习在非独立同分布数据分布下的性能。具体而言，在异质性最高的情况下（非独立同分布，$K=2$），对于 Speech Commands V1 和 V2，交替训练的准确率分别比并行训练高出 30% 和 20%。此外，当从具有高度不平衡设备端数据的预训练模型进行迁移时，我们观察到，在给定误接受率（FAR）的情况下，交替训练的相对拒识率（FRR）比并行训练低 0.04。这些发现表明，在服务器端使用少量带标签的独立同分布数据，即可有效解决联邦学习场景中的非独立同分布数据划分问题。

数据增强 我们的研究聚焦于五种数据增强策略，即“BasicAugment”、“SpecAugment”、“RandAugment”、“RandAugment（精选）”和“MixAugment”。“BasicAugment”包括时间偏移、重采样和随机背景噪声，由 Rybakov 等人提出 [5]。“SpecAugment”采用掩蔽方法处理梅尔频谱图的频率分量和时间分量，由 Park 等人提出 [4]。“RandAugment”采用 Cubuk 等人提出的基于图像的数据增强方法 [9]，而“RandAugment（精选）”则采用精选的基于图像的数据增强策略。“MixAugment”采用随机

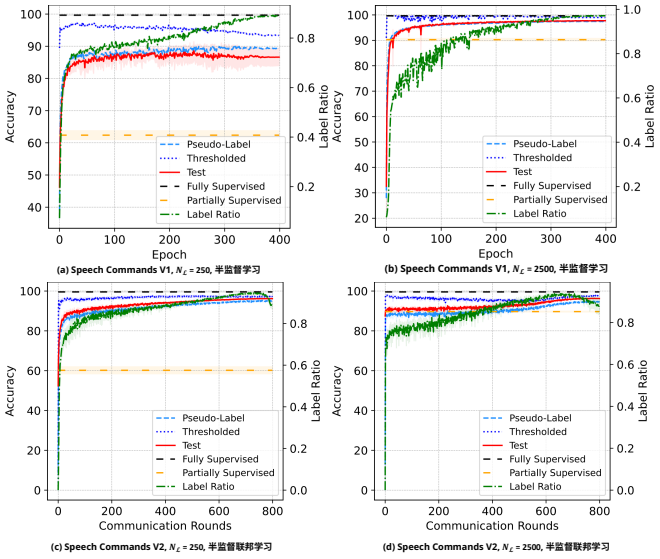

图 3。使用 Speech Commands 数据集和 $N_{\mathcal{L}} = \{250, 2500\}$ 时，TC-ResNet18 的半监督学习与半监督联邦学习的学习曲线。

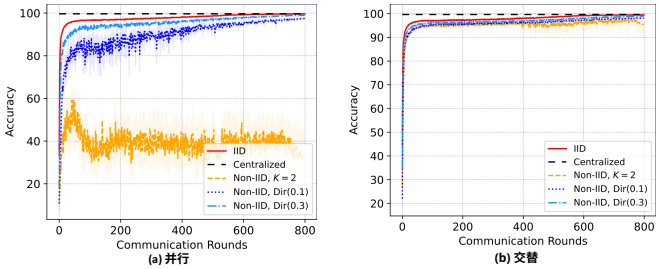

图 4。异构端侧数据下“并行”与“交替”训练的比较。

凸组合来混合有标签数据和无标签数据，由 Berthelot 等人提出 [10]。我们针对强数据增强和损失函数开展了消融研究，如表 3 所示。具体而言，我们使用“BasicAugment”进行弱数据增强 $\mathcal{A}_w$，并试验了强数据增强方法的各种组合 $\mathcal{A}_s$。研究结果表明，采用强数据增强可以显著提升半监督学习方法的性能。此外，我们观察到，基于图像的数据增强方法并不优于“SpecAugment”，而“MixAugment”可与“SpecAugment”结合使用，进一步提升性能。

表 3。使用 $N_{\mathcal{L}} = 250$ 时，关键词唤醒模型半监督学习中所用强数据增强方法的消融研究。

强数据增强 $\mathcal{A}_s$	SpeechCommandsV1	Speech Commands V2
BasicAug	86.1(0.6)	85.2(0.6)
BasicAug, RandAug	90.4(0.2)	89.1(0.2)
BasicAug, RandAug（已选）	89.8(0.8)	89.9(0.3)
BasicAug、SpecAug	95.2(0.3)	95.8(0.2)
BasicAug、SpecAug、RandAug（已选）	94.4(0.3)	94.0(0.1)
BasicAug、SpecAug、MixAug	**95.7(0.2)**	**96.5(0.1)**

5. 结论

本研究探讨了半监督联邦学习方法在关键词检测中的有效性，以及不同强数据增强技术的效果。我们的实验表明，少量有标签数据在服务器上即可支持从头开始训练或从预训练模型迁移，并利用异构的设备端无标签数据。

6. 参考文献

[1] Iv´an L´opez-Espejo，《深度语音关键词检测：综述》，*IEEE Access*，2021年。[2]Brendan McMahan、Eider Moore、Daniel Ram- age、Seth Hampson 和 Blaise Aguera y Arcas，《从去中心化数据中高效通信地学习深度网络》，载于*AISTATS*。PMLR，2017年，第1273–1282页。[3]Enmao Diao、Jie Ding 和 Vahid Tarokh，《Het- eroFL：面向异构客户端的计算与通信高效联邦学习》，载于*ICLR*，2021年。[4] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le，《Specaugment：一种用于自动语音识别的简单数据增强方法》，*arXiv*预印本 *arXiv:1904.08779*，2019年。[5] Oleg Rybakov、Natasha Kononenko、Niranjan Subrah- manya、Mirko Visontai 和 Stella Laurenzo，《移动设备上的流式关键词检测》，*arXiv*预印本 *arXiv:2005.06720*，2020年。[6] Kihyuk Sohn、David Berthelot、Chun-Liang Li、Zizhao Zhang、Nicholas Carlini、Ekin D Cubuk、Alex Kurakin、Han Zhang 和 Colin Raffel，《Fixmatch：利用一致性与置信度简化半监督学习》，*arXiv*预印本 *arXiv:2001.07685*，2020年。[7] Enmao Diao、Jie Ding 和 Vahid Tarokh，《SemiFL：面向未标注客户端的通信高效半监督联邦学习》，*NeurIPS*，2022年。[8] David Leroy、Alice Coucke、Thibaut Lavril、Thibault Gisselbrecht 和 Joseph Dureau，《用于关键词检出的联邦学习》，载于*ICASSP*。IEEE，2019年，第6341–6345页。[9]Ekin D Cubuk、Barret Zoph、Jonathon Shlens 和 Quoc V Le，《Randaugment：通过缩小搜索空间实现实用的自动数据增强》，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会，2020年，第702–703页。[10]David Berthelot、Nicholas Carlini、Ian Goodfellow、Nicolas Papernot、Avital Oliver 和 Colin Raffel，《Mix- match：一种全面的半监督学习方法》，*arXiv*预印本 *arXiv:1905.02249*，2019年。[11]Pete Warden，《Speech Commands：一个用于有限词汇语音识别的数据集》，*arXiv*预印本 *arXiv:1804.03209*，2018年。[12] Seungwoo Choi、Seokjun Seo、Beomjun Shin、Hyeong- min Byun、Martin Kersner、Beomsu Kim、Dongyoung Kim 和 Sungjoo Ha，《面向移动设备实时关键词检测的时间卷积》，*arXiv*预印本 *arXiv:1904.03814*，2019年。