异构设备上的包容性联邦学习

No One Left Behind: Inclusive Federated Learning over Heterogeneous Devices

Introduction

在传统的FL中，相关工作认为全局模型能够在任意终端设备运行的，这显然与现实是不符合的。对于一些复杂模型(BERT)来说，虽然他们能够获得不错的效果，但是在现实场景下，一些终端是没有足够的资源进行模型训练的。一种方法是将那些弱终端剔除，但这种会造成数据公平性问题，同时也会影响模型最终的训练效果。或者将模型复杂度与参与FL的最弱设备进行匹配，但这样会造成水桶效应的限制。

因此，直观上使强设备训练复杂模型，弱设备训练简单模型的方法能够充分利用终端设备的资源。相关工作也采用知识蒸馏的技术，将异构模型的参数进行交换，但他们使用了同一数据源，这对于弱节点来说是困难的。HeteroFL提出共享大小模型中共享参数的方法，从而避免使用同一数据源。但这将引发大小模型不匹配的问题：因为1）对模型修剪将会打破模型的原有结构，使得大模型中的知识无法很好的迁移至小模型，而且小模型反而会影响全局模型的训练进程，2）同时分享参数并不能很好的进行知识迁移。

文章提出一种能够灵活适应设备资源的FL方法(InclusiveFL), 强设备训练复杂模型，弱设备训练简单模型。为了解决不匹配问题，文章提出共享模型底层的方法，同时文章也提出了一种异构模型的自适应汇聚方法。并且文章提出了一种动量知识蒸馏的方法来进行复杂模型的知识迁移，使得小模型的顶部编码层能够和大模型的顶部编码层行为一致。

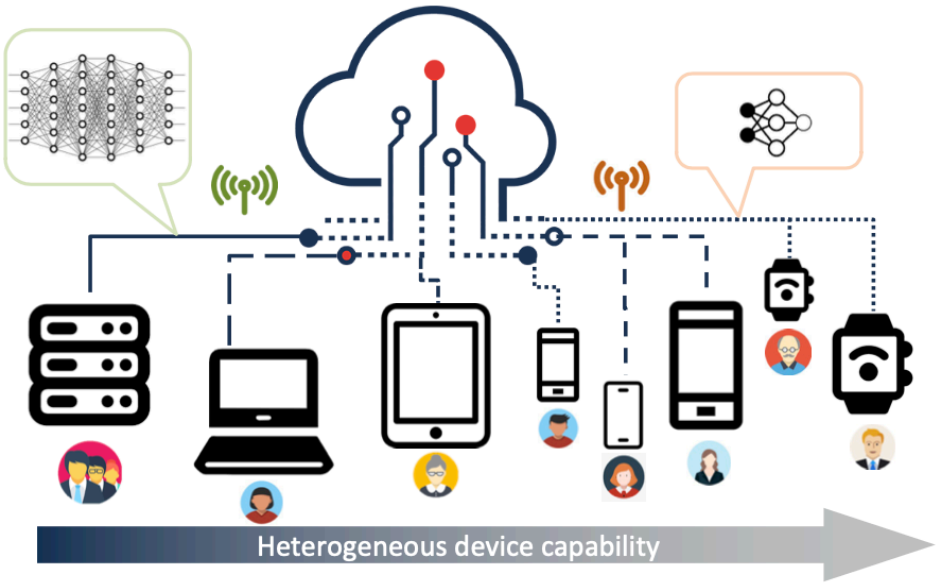


Figure 1: InclusiveFL over heterogeneous devices.

Background and Relation Work

Algorithm

FedAvg是一种传统FL算法，采用平均的方法进行梯度汇聚：

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$$

另外，最近提出的FedAdam采用一种自适应算法：

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$w_{t=1} \leftarrow w_t + \eta \frac{m_t}{\sqrt{v_t} + \tau}$$

Heterogeneous Devices

为了解决异构设备的问题，相关工作采用知识蒸馏的方法来对异构模型的知识进行迁移，但这些工作没有考虑弱节点的存储资源限制。

HeteroFL对模型进行拆分，通过共享参数的方法来摆脱上述方法遇到的问题，但这种方法会对模型造成影响，稳定性不足。

Optimization

其他工作对FL的通讯代价给出了优化方案，但没有考虑设备异构的问题。

Methodology

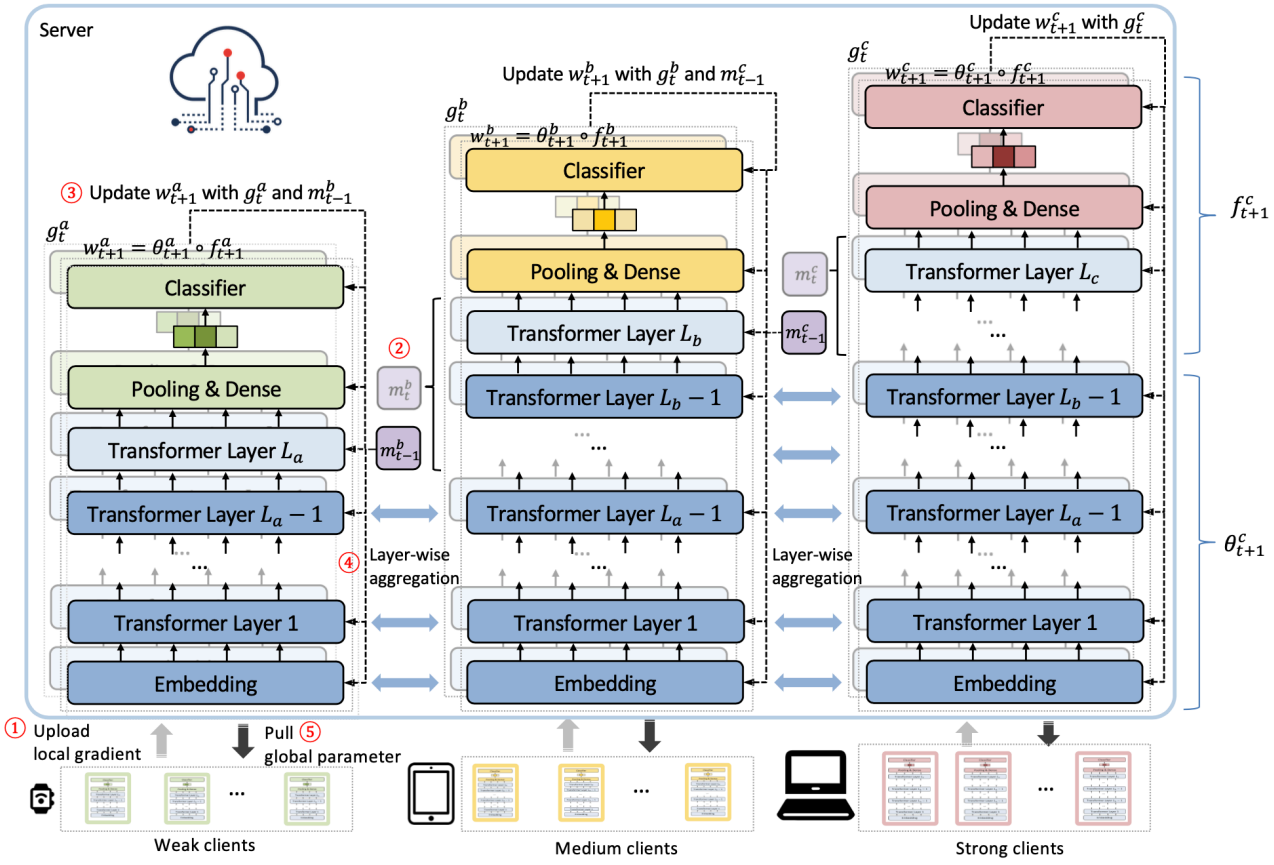


Figure 2: Overview of *InclusiveFL*.

Layer-wise Heterogeneous Aggregation Framework

自适应异构模型的大致思想是通过减少模型层数来降低模型参数，从而适应设备资源。另外模型底层能够保留更多的数据细节，因此为不同模型分配不同的模型层数。在进行梯度汇聚时，相同模型进行梯度汇聚，然后不同模型间进行梯度汇聚。

在进行异构模型梯度汇聚时，本地模型分为两部分1）参数共享部分θ和2）独立部分f (独行编码层/特定子任务层)。InclusiveFL是一种仅改变顶层f的通用框架，因此在梯度汇聚时可以使用FedAvg对θ进行汇聚。由于异构模型底层特征具有相似的结构，因此可以保留模型特征。独立部分出于两点考虑：1）能够缓解不同模型间的不匹配问题，由于更深层的模型能够进一步抽取数据特征，如果单纯进行f_{t+1}的合并将影响模型最终表现；2）f能够使得模型能够得到不同的训练结果。

Momentum Distillation

由于大模型能够从数据中获取更多的知识特征，因此文章采用知识蒸馏的办法来将知识从大模型迁移至小模型。集中式的知识蒸馏在小模型上加一层隐藏层，从而使小模型的训练效果与大模型对齐；然而这种方法在FL上不可行。

相反，文章指出在底层共享的异构模型中，应该是得小模型的隐藏层与大模型的顶层的行为对齐。因此，文章采用L_a到L_b的平均梯度梯度计算梯度动量m_{t-1}^b，并将梯度动量引入L_a来弥补与模型b中的差距。这种方法不限制模型的种类；另外要注意的是梯度动量的初始值为0，其应用效果与超参数β相关。

Experment

文章使用GLUE进行分类实验，并在三种不同的数据集上进行NER训练。

Table 1: Statistics of the medical NER datasets.

Dataset	# Sentences	Entity Types	# Entity
SMM4H	3,824	ADE (1707)	1,707
ADE	4,483	ADE (5678), Drug (5076), Dosage (222)	10,976
CADEC	7,683	ADE (5937), Drug(1796), Disease(282), Finding(425), Symptom(268)	8,535

并且文章设置了5个对比实验：

- AllLarge: 取消资源限制，每个设备使用最大模型
- AllSmall: 每个设备使用最小模型
- DropWeak: 丢弃弱节点
- Local: 使用符合设备资源的训练数据进行训练
- HeteroFL

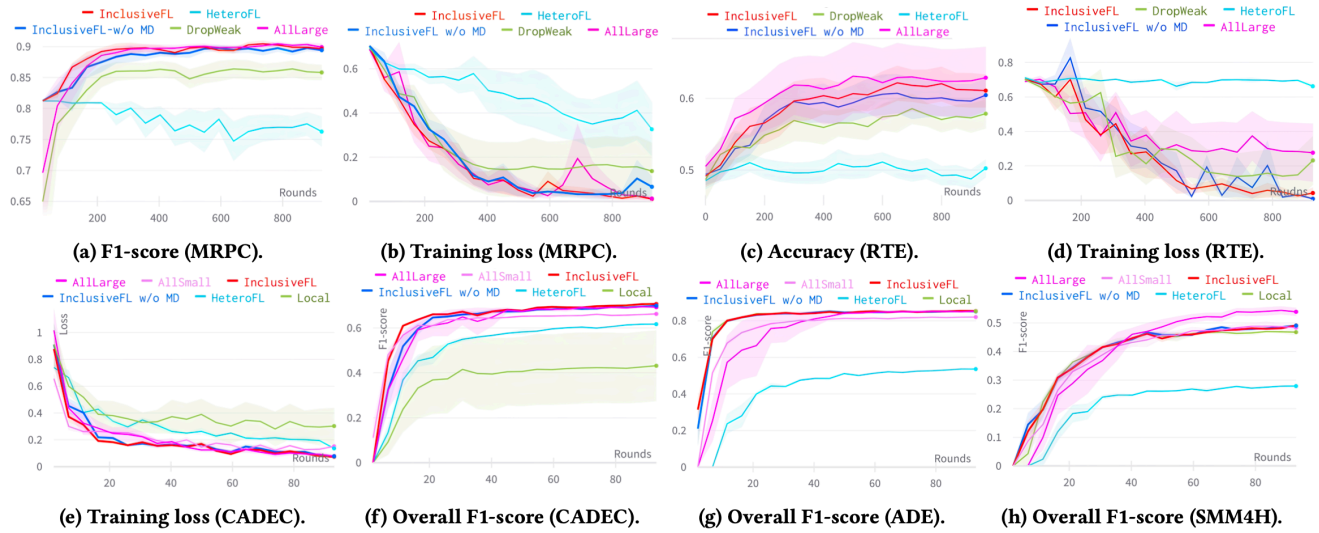
实验结果显示出InclusiveFL极大提高了原有的训练效果，并接近于AllLarge。

Table 2: Performance comparison of federated learning methods on the GLUE benchmark for text classification tasks. Performance is evaluated on the final large model in a global inference scenario. Underlines indicate the performance is superior to baseline methods. Bold faces indicate the best method for training with heterogeneous devices.

Methods	Client-inclusive	Agg. type	COLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
AllLarge	N/A	Homo.	63.03	86.48	91.5	92.09	91.49	76.12	94.43	90.6	85.45
DropWeak	No	Homo.	37.77	85.98	88.81	91.24	89.47	62.17	94.06	89.26	79.85
AllSmall	Yes	Homo.	34.91	78.83	82.5	85.93	79.37	58.94	90.14	83.68	74.29
HeteroFL	Yes	Hete.	8.15	31.83	81.51	62.7	73.79	52.71	84.98	30.54	53.28
InclusiveFL-w/o MD	Yes	Hete.	52.69	86.28	91.2	91.59	**90.34**	63.9	94.04	89.91	82.49
InclusiveFL	Yes	Hete.	**54.85**	**86.36**	**91.42**	**91.76**	90.32	**65.85**	**94.17**	**89.94**	**83.08**

Table 3: Performance comparison of federated learning methods on medical datasets for named entity recognition. Performances are evaluated for each named entity on each local model (SMM4H with 4-layer model, ADE with 9-layer model and CADEC with 12-layer model) in the local inference scenario. Underlines and bold faces have the same meaning as in Table 2.

Methods	SMM4H	ADE			CADEC					Avg.
	ADE	Drug	ADE	Dose	ADE	Symptom	Drug	Disease	Finding	
AllLarge	55.08	95.21	80.91	18.97	71.23	43.37	90.43	33.55	29.85	57.62
Local	46.77	95.12	80.14	11.25	44.15	19.9	53.53	20.58	16.53	43.11
AllSmall	49.56	92.76	75.67	13.51	64.82	26.71	87.91	21.07	20.23	50.25
HeteroFL	27,94	73.9	37.57	11.57	59.9	26.92	84.58	25.77	19.88	42.51
InclusiveFL-w/o MD	49.14	95.45	79.63	12	69.98	34.42	90.16	38.31	20.41	54.39
InclusiveFL	**49.9**	**95.49**	**80.34**	**13.91**	**71.1**	**40.91**	**90.21**	**39.34**	**22.97**	**56.02**

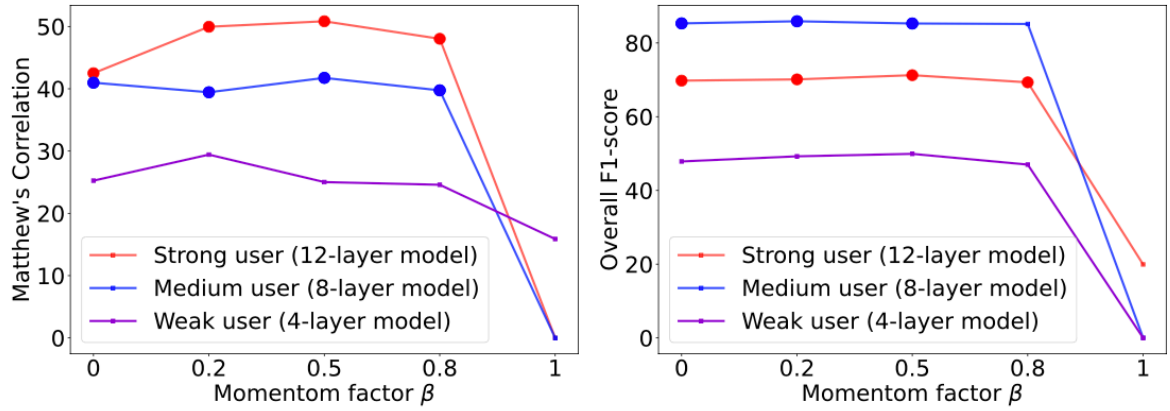


(a) F1-score (MRPC). (b) Training loss (MRPC). (c) Accuracy (RTE). (d) Training loss (RTE).

(e) Training loss (CADEC). (f) Overall F1-score (CADEC). (g) Overall F1-score (ADE). (h) Overall F1-score (SMM4H).

Figure 3: Federated training convergence and best performance. Error band shows the standard error of 5 independent repeats.

同时，文章对于超参数β进行了实验探索：



(a) CoLA with *InclusiveFL.** **(b) Medical NER with *InclusiveFL***

Figure 6: Influence of momentum distillation β.

文章指出，模型在$\beta = 0.2\ or\ 0.5$左右能够去得良好的表现，当$\beta > 0.8$时，将会由于引入过量的蒸馏知识而破坏模型性能。

Thinking

- 文章所提出的方法简单直观有效，能够充分利用异构设备中的计算资源。

- 实验仅考虑了自然语言处理这个领域，其他深度模型的任务特点没有提及。
- 没有体现是否会引入额外的计算开销。