LIFE: Leader-driven Hierarchical & Inclusive Federated Learning

Wenhao Li, Christos Anagnostopoulos,
Shameem Puthiya Parambath, Kevin Bryson

University of Glasgow, UK

IEEE BigData 2024
Special Session on Federated Learning on Big Data

DEC 15-18,2024 @ Washington DC, USA

Contents

Introduction

Federated Learning (FL) enables privacy-preserving distributed training by exchanging minimal information between clients and a central server [1].

In FL, clients receive a *global* model, train locally, and send updates back to the server for aggregation over several rounds.

Challenges

Challenges in Heterogeneous FL:

- **Resource Variability**: Different clients (e.g., smartphones, edge servers) have varying computational resources, complicating model aggregation.
- **Model Fairness**: Weak clients struggle to train large models, risking exclusion and unfairness.

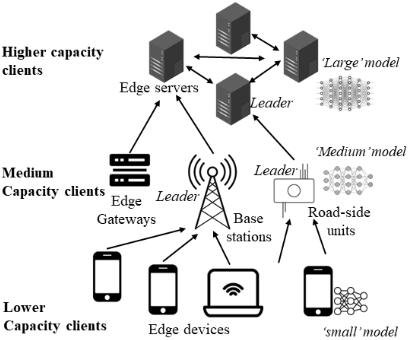


Fig. 1: FL Framework

Contents

Methodology

* **Layer-wise Model Aggregation**

$$\theta^{t+1}_{|\mathcal{C}|,\ell} = \sum_{c\in\mathcal{C}} \frac{n_c}{n}\theta^{t}_{c,\ell} \text{ and } \theta^{t+1}_{|\mathcal{C}|,\ell_v} = \sum_{c\in\mathcal{C}_v} \frac{n_c}{n_v}\theta^{t}_{c,\ell_v}$$

Calculate momentum $m^t_{l,k}$ using

$$m^t_{l,k} = \frac{1}{d_l - d_{l-1} + 1} \sum_{\ell\in[d_{l-1}:d_l]} \Delta^t_{l,k,\ell}$$

* **Knowledge distillation** using

$$\Delta^{t+1}_{l-1,k,d_{l-1}} = \beta \cdot m^t_l + (1-\beta)\cdot \Delta^t_{l-1,k,d_{l-1}}$$

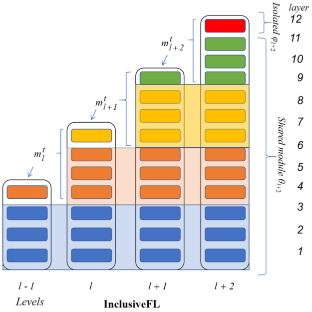


Fig. 2: InclusiveFL [2]

Methodology

Aggregate fragment with different weights:

$$\theta^{t+1}_{L,1:d_L-1} = \underbrace{\underbrace{\underbrace{\sum_{l=1}^{L}\frac{N_l}{N}\theta^t_{l,\ell_1}}_{\theta^t_{1,1:d_1-1}} + \cdots + \sum_{l=v}^{L}\frac{N_l}{N^*_v}\theta^t_{l,\ell_v}}_{\theta^t_{v,1:d_v-1}} + \cdots + \sum_{l=L-1}^{L}\frac{N_l}{N^*_{L-1}}\theta^t_{l,\ell_{L-1}}}_{\theta^t_{L-1,1:d_{L-1}-1}}$$

LIFE Basic Structure

- Leaders at different levels
- Leaders fine-tune their subordinate models
- Aggregation takes place only in neighbouring levels
- Layer-wise Model Aggregation + Knowledge distillation

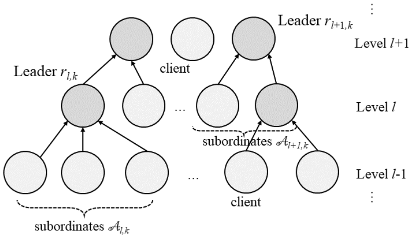
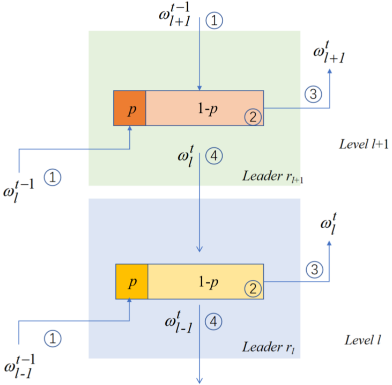


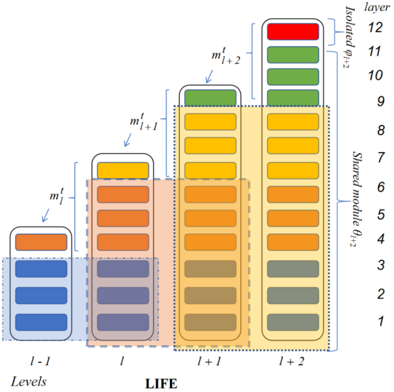
Fig. 3: LIFE Basic Structure

LIFE : 1. Leader-driven fine-tuning: from higher hierarchical levels (TOP) to lower ones (BOTTOM)



2. Leader driven aggregation

$$\theta^{t+1}_{l,m,\ell} = \frac{n_{l,m}}{n_{l,m}+1}\theta^{t}_{l,m,\ell} + \frac{1}{n_{l,m}+1}\theta^{t}_{l+1,m,\ell}$$



Contents

Experimental Results

Datasets & models : RoBERTa-based transformer model (12-layer, 8-layer, and 4-layer) on GLUE (General Language Understanding Evaluation) benchmark.

Experimental setup:

- 1:6:19 and 1:1:1 client ratios
- different skewed label distributions in the non-iid case
  - small clients (Left,a=0.1/a=0.3) & medium clients (Left,a=0.5)
  - small clients (Left,a=0.3) & medium clients (right,a=0.5)

Experimental Results

Table 1: Model configuration study: Leader fine-tuning ratio p and T2B vs. B2T Directions (SST2)

Leader Fine-tuning ratio p				
	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$
Accuracy	$86.39\% \pm 0.0156$	$86.43\% \pm 0.0193$	$84.72\% \pm 0.0218$	$81.68\% \pm 0.0281$
LMA rounds T_A				
	$T_A = 1$	$T_A = 3$	$T_A = 5$	$T_A = 10$
Accuracy	$87.12\% \pm 0.0156$	$87.97\% \pm 0.0102$	$87.73\% \pm 0.0085$	$87.69\% \pm 0.0169$
T2B & B2T Directions				
	T2B-w	T2B-w/o	B2T-w	B2T-w/o
Accuracy	$82.05\% \pm 0.0897$	$81.98\% \pm 0.0104$	$81.59\% \pm 0.0845$	$81.97\% \pm 0.0102$

Table 2: Structure comparison (accuracy); SST2

	IID 1:1:1	IID 1:6:19	a=0.1 (S-L/M-L)	a=0.3 (S-L/M-L)	a=0.3 (S-L/M-R)
LIFE	90.02	**86.81**	**90.02**	**88.53**	**91.51**
InclusiveFL	**90.60**	81.54	90.00	84.98	91.47
LIFE-M	**91.51**	**85.55**	**89.99**	89.91	**91.06**
InclusiveFL-M	89.11	82.57	89.07	**90.37**	90.83
LIFE-S	**83.37**	**81.19**	**50.92**	**68.69**	**80.73**
InclusiveFL-S	78.78	79.70	50.89	62.16	77.64

- LIFE's model enhancement on medium and lower levels is outstanding
- Enhancement is more effective when the opposite label distribution is skewed

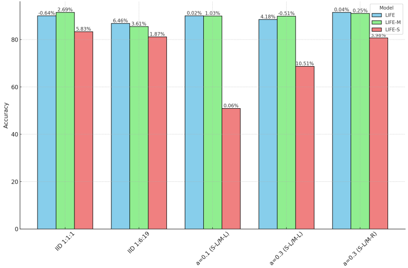


Fig. 4: Visualise Results

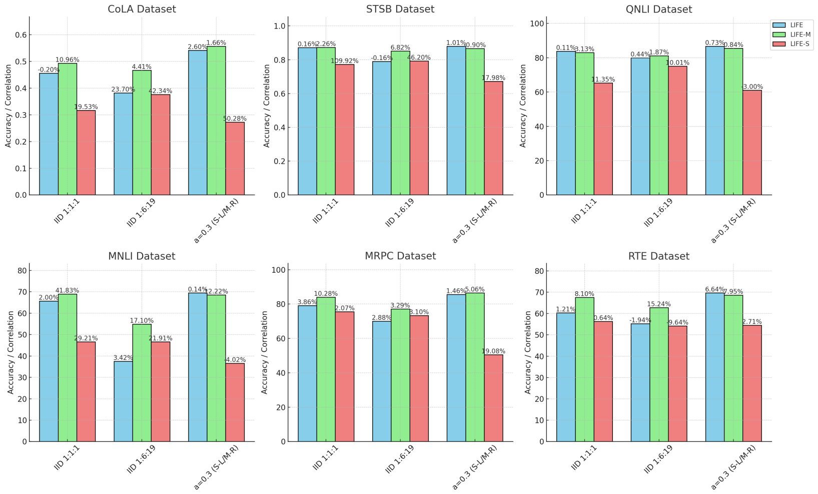


Fig. 5: Comparison. Leader-driven improvement is more effective at the subordinate level

Overall improvement on top, medium, and bottom levels up to 37.4% for CoLA, 58.22% for STSB, 6.8% for QNLI, 23.06% for MNLI, 7.88% for MRPC and 10.43% for RTE.

Contents

Conclusion

Impact: LIFE uses leader-driven aggregation and a top-down aggregation order that enhances the overall accuracy of the model, especially in helping weak clients.

Future Work

- Hierarchical model pruning across levels
- Leader election and leader-member matching

Contents

Q&A

Thank you!

References I

B. e. a. McMahan, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, vol. 54, 2017, pp. 1273–1282.

R. Liu, F. Wu, C. Wu, Y. Wang, L. Lyu, H. Chen, and X. Xie, "No one left behind: Inclusive federated learning over heterogeneous devices," in *28th ACM SIGKDD*, 2022, p. 3398–3406.