Towards Efficient Federated Multilingual Modeling with LoRA-based Language Family Clustering

Anonymous ACL submission

Abstract

Federated Multilingual Modeling (FMM) plays a crucial role in the applications of natural language processing due to the increasing diversity of languages and the growing demand for data privacy. However, FMM faces limitations stemming from the substantial communication costs in networking and the conflicts arising from parameter interference between different languages. To address these challenges, we introduce a communication-efficient federated learning framework with low-rank adaptation and language family clustering for Multilingual Modeling (MM). In this framework, we maintain the weights of the base model, exclusively updating the lightweight Low-rank adaptation (LoRA) parameters to minimize communication costs. Additionally, we mitigate parameter conflicts by grouping languages based on their language family affiliations, as opposed to aggregating all LoRA parameters. Experiments demonstrate that our proposed model not only surpasses the baseline models in performance but also reduces the communication overhead.

1 Introduction

Multilingual modeling is increasingly important in natural language processing (NLP) as a result of the growing diversity of languages used online (Limisiewicz et al., 2023). However, gathering multilingual data can prove prohibitively expensive due to its distributed nature and data privacy concerns (Wang et al., 2022; Gala et al., 2023). To address this challenge, Federated Learning (FL) is employed to train a multilingual model across various institutions and data sources (Chen et al., 2023; Zhang et al., 2023b; Fu and King, 2023). The fundamental concept of FL revolves around the exchange of model parameters rather than the transmission of sensitive data, thereby preserving data privacy (Zhang et al., 2023c; Xu et al., 2023).

Nevertheless, the increasing size of pre-trained language models (PLMs) presents challenges

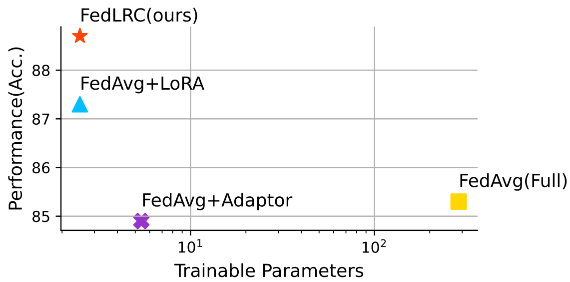

Figure 1: Benchmark Result on Text Classification Task.

when fine-tuning the federated multilingual model (FMM) with a small dataset in the federated setting (Zhang et al., 2023d). This is mainly due to the bottleneck created by transmitting large model parameters through the network (Kim et al., 2023). Beyond the communication cost, FMM naturally encounters non-IID (Non-Independently and Identically Distributed) issues (Zhang et al., 2023a). Owing to differences in linguistic systems and culture, languages such as English and Chinese exhibit significant distribution shifts. When adapting the model towards a specific target language, it can potentially interfere with the modeling of other languages (Xu et al., 2022), resulting in significant Parameter Conflicts (PC) (Liu et al., 2023; Chronopoulou et al., 2023) and damaging the transfer performance (Xu et al., 2022).

To this end, we propose a communication-efficient federated learning framework with a language family clustering for multilingual modeling. Motivated by the parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Ruder et al., 2022; Sung et al., 2022; Hu et al., 2023), as illustrated in Figure 2, we fine-tune on a small set of parameters via Low-Rank adaptation techniques (LoRA), while keeping the parameters of the original PLMs unchanged. To the best of our knowledge, we represent the pioneering application of LoRA on FL. Since the LoRA adapter contains fewer trainable parameters, our approach significantly reduces the communication overhead. To alleviate the interference between different languages, we are further grouping languages into clusters following the language family shown in Figure 3. Experiments are showing that our approach demonstrates superior performance with higher efficiency compared to various baseline models. Below we summarize our contributions as follows:

i. We propose FedLFC, a communication-efficient federated learning framework with PEFT in the setting of Multilingual Modeling. Our work represents the pioneering application of LoRA on FL, resulting in a remarkable reduction of communication overhead by a factor of 100.

ii. We employed the language family clustering strategy to alleviate the parameter conflict in the setting of federated multilingual modeling.

iii. We show the superiority of FedLFC in three downstream tasks, *i.e.*, language modeling, machine translation, and text classification.

2 Methodology

2.1 Federated Multilingual Modeling.

We begin by introducing the formulation of Federated Multilingual Modeling (FMM) (Weller et al., 2022). Given N language datasets $\{D_j\}_{i=1}^{N}$, The goal of FMM is to collaboratively train a multilingual FL model that achieves high performance in the downstream tasks. Specifically, in the setting of FMM, we assume there are N client $\{C_i\}_{i=1}^{N}$. Each client C_i owns only one language D_i and the different client has different languages. Let $\boldsymbol{\Theta}_i$ be the trainable parameters of the local model in C_i. At each training round l, the clients train the local FL model with parameter $\boldsymbol{\Theta}^{(l)}$ on their own dataset D_i and then send parameters to the server S. The server S then aggregates these parameters to generate the global parameters $\boldsymbol{\Theta}^{(l+1)}$ and sends $\boldsymbol{\Theta}^{(l+1)}$ to all clients for the subsequent training round. FedAvg is employed for aggregation by default (McMahan et al., 2017) and is computed as follows:

$$\boldsymbol{\Theta}^{(l+1)} = \sum_{i=1}^{N} \frac{1}{N} \boldsymbol{\Theta}_i^{(l)}. \quad (1)$$

2.2 Federated Efficient Fintuning with Low-Rank Adaption

In FMM, training the entire FL model incurs substantial communication costs as it involves computing/exchanging a large number of parameters

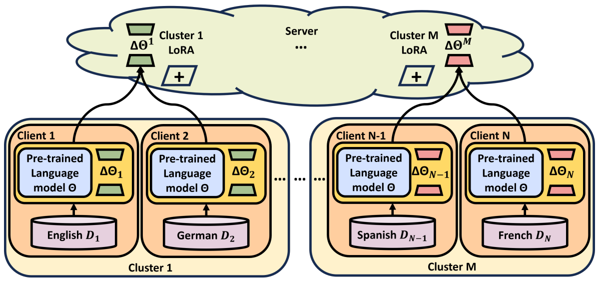

Figure 2: The overall framework of FedLFC.

through the networks. The success of fine-tuning on pre-trained language models (PLMs) motivates us to explore adjustment of the small portion of parameters in the FMM.

FMM with Low-Rank Adaption. It has been shown that PLMs exhibit a low "intrinsic dimension" when adapting to specific tasks (Aghajanyan et al., 2021) and can still learn efficiently despite a random projection to a smaller subspace. Inspired by this, in FMM, we hypothesize the local updates to the weights $\boldsymbol{\Theta}$ for each client also have such low "intrinsic rank" during training. Therefore we employ the Low-Rank Adapter (LoRA) for efficient FMM fine tuning. Specifically, instead of training and exchanging $\boldsymbol{\Theta}$ for each client, we only adjust the parameters of adapter $\Delta\boldsymbol{\Theta}$ in propagation. Specifically, the forward process for the linear layer in the FMM model is computed as follows:

$$\boldsymbol{h} = \boldsymbol{\Theta x} + \Delta\boldsymbol{\Theta x} = \mathbf{BA}\boldsymbol{x}, \quad (2)$$

where $\boldsymbol{x}$ represents the output of the previous layer, $\boldsymbol{h}$ is the hidden state. Note that $\boldsymbol{\Theta} \in \mathbb{R}^{d \times k}$ is parameters of the PLM used in the local model, which is frozen. $\Delta\boldsymbol{\Theta}$ is the parameters of the adapter, which is updated during training rounds. $\Delta\boldsymbol{\Theta}$ can be factorize into two matrix $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times k}$ As the intrinsic rank $r << min(d, k)$ is small, $\Delta\boldsymbol{\Theta} = \mathbf{BA}$ has fewer parameters to communicate.

Federated Parameter-Efficient Funin Tuning. Our approach involves freezing a pre-trained model and solely training adapters, which is more parameter-efficient. For each client C_i, we add a LoRA module with trainable parameter $\Delta\boldsymbol{\Theta}_i$ in parallel to the PLMs parameter $\boldsymbol{\Theta}_i$. In each training round l, we freeze the parameters of the PLM, $\boldsymbol{\Theta}_i^{(l)}$ and only update LoRA parameters $\Delta\boldsymbol{\Theta}_i^{(l)}$. At the end of each training round, clients transfer their updated LoRA parameters to the server. When the server receives the parameters of all clients, it aggregates LoRA parameters as

$$\Delta\boldsymbol{\Theta}^{(l+1)} = \sum_{i=1}^{N} \frac{1}{N} \Delta\boldsymbol{\Theta}_i^{(l)}. \tag{3}$$

2.3 Updating LoRA Parameters with Language Family Clustering

The parameter conflict (PC) issue is common in FMM. As language from different sources exists in diverse distributions, such non-i.i.d. nature causes conflict when aggregating the parameters trained on different D_i. The update of the parameter $\boldsymbol{\Theta}_i$ from one client may have an adversarial effect on the others, yielding suboptimal performance.

Language Family Clustering (LFC). To alleviate PC in FMM, we introduce LFC. Research related to FL has shown that clustering a subset of clients that share a similar distribution strategy can reduce the PC (Vahidian et al., 2023; Ruan and Joe-Wong, 2022; Liu et al., 2023). Typical methods employ heuristic prior knowledge to determine the group of parameter aggregation. In language modeling, languages can be categorized together based on linguistic information, forming language families. Following the language family clustering in (Paul et al., 2009). We aggregate LoRA parameters using language family clusters as shown in Figure 3, *i.e.*, Germanic (including English and German), Italic (including Spanish, French, and Portuguese), Balto-Slavic (including Russia, Polish, Czech and Lithuanian), Sino-Tibetan (including Chinese), Uralic (including Finnish), Afro-Asiatic (including Arabic), and Japonic (including Japanese).

Let $\{\mathcal{G}_m\}_{m=1}^{M}, (M \leq N)$ denotes the set of family in taxonomy. Each $\mathcal{G}_m$ contains a set of index i indicating the i-th clients with datasets D_i belong to the m-th language family. The aggregation in Equation 3 then change to

$$\Delta\boldsymbol{\Theta}^{m,(l+1)} = \sum_{i\in\mathcal{G}_m} \frac{1}{|\mathcal{G}_m|} \Delta\boldsymbol{\Theta}_i^{(l)}. \tag{4}$$

Note that we have M LoRA adapters associated with different language families $\mathcal{G}_m$. We use corresponding $\Delta\boldsymbol{\Theta}^{m,(l+1)}$ for inference in downstream tasks with specific language. The overall algorithm is shown in Algorihtm 1.

3 Experiment

Tasks and Datasets. We evaluate our model in three takes *i.e.*, Language Modeling (LM), Machine Translation (MT), and Text Classification

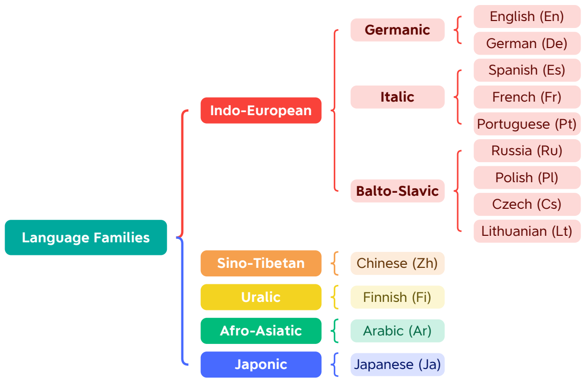

Figure 3: Language families form (Paul et al., 2009).

(TC) using four datasets *i.e.*, Europarl, MTNT, UN Corpus, and News Classification. The statistics of each dataset are shown in Table 4. We detail the description of each dataset in Appendix 4.

Experiment Settings. We use different pre-trained models for different tasks *i.e.*, mBERT (Sanh et al., 2019) for language modeling, M2M100 (Fan et al., 2021) for machine translation, and XLM-Roberta (Conneau et al., 2019) for text classification. A detailed setting including system and hyperparameters is in Appendix A.2.

Baselines. We perform the experiment on three different settings *i.e.*, *Centralized Model*, *FedAvg*, and *Standalone*. The centralized model employs centralized training (Weller et al., 2022), where all data is collected in one place. FedAvg employs Federated Averaging (McMahan et al., 2017) training within the federated learning framework, dividing data across different clients. Both of them train a conventional multilingual model with all parameters. Standalone setting trains data exclusively in one language and tests its performance across all languages, demonstrating a scenario where a model is trained using data from a single client (Weller et al., 2022). To show the superiority of LFC and LoRA, we further freeze parameters of PLMs in the setting of Centralized and FedAvg. We train *LoRA* (Hu et al., 2022) and typical *Adapter* (Houlsby et al., 2019) without LFC.

Evaluation Metric. For the language modeling task, we use perplexity (PPL) as the evaluation metric (Weller et al., 2022). For neural machine translation task, we use BLEU as evaluation metrics, using ScareBLEU package (Post, 2018). For the text classification task, we use accuracy as an evaluation metric.

Table 1: Results for FL experiments on the LM task. The standard deviation (std) is reported Table 6.

Method	# TP ↓	UN ↓ En	Es	Zh	Ru	Ar	Fr	Avg	Europarl ↓ En	Cs	Lt	Es	Pl	Fi	Pt	De	Avg
Centralized	-	7.4	4.8	6.9	3.9	5.2	4.6	5.6	9.8	3.8	4.8	6.0	3.9	5.8	9.2	8.4	5.9
+ Adapter	-	10.4	6.2	9.0	4.7	7.2	5.9	7.0	10.6	7.1	8.2	7.3	5.8	7.6	7.6	7.9	7.7
+ LoRA	-	11.3	6.7	9.7	5.0	7.6	6.4	7.5	10.7	6.9	8.0	7.3	5.7	7.4	7.5	8.0	7.6
Standalone	-	33.0	16.1	43.0	10.3	10.8	14.0	25.4	9.4	2.8	2.6	4.3	2.8	3.0	3.7	3.5	4.0
FedAvg	135.4M	**8.7**	**4.2**	**5.4**	4.1	**4.2**	5.1	**5.1**	10.4	6.4	9.2	**5.9**	5.9	7.8	7.5	7.9	7.7
+ Adapter	2.5M	22.8	14.9	17.0	9.9	17.2	14.3	15.5	12.0	10.6	14.2	8.3	7.5	10.7	9.4	9.2	10.1
+ LoRA	1.2M	10.8	6.6	9.3	5.0	8.1	6.3	7.5	11.4	8.8	11.3	7.8	6.6	9.3	8.5	8.8	8.9
FedLFC	**1.2M**	9.4	5.6	8.0	**4.0**	6.1	**5.1**	6.4	**10.4**	**6.1**	**6.3**	7.1	**5.4**	**6.4**	**7.2**	**7.7**	**7.1**

Table 2: Results for FL experiments on the machine translation task.

Method	# TP ↓	MTNT ↑ En-Fr	En-Ja	Avg	UN ↑ En-Fr	Ar-Es	Ru-Zh	Avg
Centralized	-	32.2±0.5	32.3±0.2	32.1±0.7	39.3±0.6	37.5±0.9	24.0±0.2	33.8±0.6
+ Adapter	-	31.9±0.5	30.4±0.3	31.7±0.1	36.9±0.9	34.0±0.6	20.3±0.2	30.4±0.3
+ LoRA	-	32.3±0.6	32.5±0.2	32.2±0.6	37.6±0.3	34.9±0.3	20.2±0.2	31.3±0.6
Standalone	-	27.1±0.5	28.1±0.7	27.6±0.6	34.6±0.5	33.8±0.5	18.5±0.6	29.0±0.4
FedAvg	483.9M	32.9±0.2	33.3±0.8	32.9±0.6	38.2±0.4	35.9±0.3	21.1±0.1	31.1±0.7
+ Adapter	12.7M	32.6±0.4	33.0±0.2	32.6±0.6	35.8±0.9	31.9±0.6	19.2±0.8	29.2±0.4
+ LoRA	9.4M	33.3±0.6	32.5±0.5	33.2±0.8	36.3±0.6	32.7±0.5	19.8±0.7	29.5±0.7
FedLFC	**9.4M**	**34.0**±0.2	**33.6**±0.1	**33.8**±0.4	**38.7**±0.7	**37.9**±0.5	**22.1**±0.2	**32.9**±0.1

Table 3: Results for FL experiments on the text classification task.

Method	# TP ↓	En ↑	Es ↑	Fr ↑	De ↑	Ru ↑	Avg ↑
Centralized	-	93.5±0.7	86.3±0.5	82.9±0.3	89.6±0.1	88.5±0.4	88.1±0.2
+ Adapter	-	92.7±0.4	86.7±0.6	81.7±0.1	88.5±1.0	87.4±0.5	87.4±0.3
+ LoRA	-	91.8±0.4	83.7±0.3	80.4±0.5	86.4±0.4	85.3±0.1	85.5±0.1
Standalone	-	22.8±1.2	40.8±0.7	40.8±0.1	40.8±0.5	77.1±0.2	44.5±0.3
FedAvg	278.1M	90.7±0.4	84.3±0.2	80.5±0.3	87.6±0.1	83.4±0.5	85.3±0.2
+ Adapter	5.4M	91.5±0.5	85.7±0.7	79.1±0.2	86.9±0.7	81.3±0.8	84.9±0.7
+ LoRA	2.5M	**93.8**±0.3	85.8±0.6	80.7±0.3	89.4±0.7	86.7±0.3	87.3±0.2
FedLFC	**2.5M**	93.5±0.1	**86.6**±0.1	**82.7**±0.5	**90.1**±0.1	**91.0**±0.1	**88.7**±0.1

3.1 Main Results

In this section, we discuss the results and observations in Table 1, 2, and 3 respectively. Overall, our approach demonstrates superior performance compared to other FL methods in most tasks. Following are several key observations.

FMM Model Outperform Standalone. The standalone model serves as the lower performance bound for each task. Our experimental results demonstrate that a majority of FedAvg models outperform the standalone model. This observation highlights the necessity of FMM for language model training in real-world scenarios, as it enables the using the training data without data barriers.

Parameters Efficient FT *vs*. Full-Parameters FT. We observe that the parameters efficient finetuning model outperforms the full fine-tuning models. This shows the effectiveness of LoRA in FMM.

Lower Communication Costs. Being consistent in three tasks, the introduction of LoRA led to a remarkable reduction in the number of trainable parameters by a factor of 100 which is shown in Table 1, 2, and 3 respectively. In comparison to full fine-tuning and adapters, LoRA utilizes the fewest training parameters and GPU memory across the three tasks.

Clustering Strategy Improves Performance. By incorporating an LFC strategy, the performance improvement varies significantly across different languages. Notably, the clustering strategy proves to be more beneficial for languages with limited resources. In Table 1, we observe that compared to other languages, Ar (8.1→6.1), Cs (8.8→6.1), Lt (11.3→6.3), and Fi (9.3→6.4) exhibit a greater decrease in perplexity (PPL). These languages are typically associated with medium or low-resource datasets in real-world scenarios. This confirms that LFC is more effective in low-source languages.

4 Conclusion

In the paper, we propose, FedLFC, a communication efficient federated learning framework for Multilingual Modeling. Two crucial techniques, *i.e.*, Federated Efficient-Finetning with LoRA and Language Family Clustering are introduced to solve the problem of communication overhead and parameter conflict caused by language interference. Experiments show that our proposed model is both efficient and effective.

Limitations

In this paper, we only test the approach on Bert, M2M100 and xlm-roberta PLMs. In the future, we will conduct research on applying the approach to Large Language Models (LLM). Secondly, we only use the same number of data in each language for fine-tuning. The data partition is different from the real-world. We will validate the effectiveness of the model on datasets with varying quantities of different languages. Thirdly, there are other kinds of clustering strategy, such as gradients clustering, random clustering. Following Liu et al. (2023), we only choose language family clustering strategy. We will test other clustering strategy.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

M Saiful Bari, Batool Haider, and Saab Mansour. 2021. Nearest neighbour few-shot learning for cross-lingual classification. *arXiv preprint arXiv:2109.02221*.

Gaode Chen, Xinghua Zhang, Yijun Su, Yantong Lai, Ji Xiang, Junbo Zhang, and Yu Zheng. 2023. Win-win: A privacy-preserving federated framework for dual-target cross-domain recommendation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*. AAAI Press.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Xinyu Fu and Irwin King. 2023. Fedhgn: A federated framework for heterogeneous graph neural networks. *arXiv preprint arXiv:2305.09729*.

Jay Gala, Deep Gandhi, Jash Mehta, and Zeerak Talat. 2023. A federated approach for hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3248–3259, Dubrovnik, Croatia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Yeachan Kim, Junho Kim, Wing-Lam Mok, Jun-Hyung Park, and SangKeun Lee. 2023. Client-customized adaptation for parameter-efficient federated learning. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Yi Liu, Xiaohan Bi, Lei Li, Sishuo Chen, Wenkai Yang, and Xu Sun. 2023. Communication efficient federated learning for multilingual neural machine translation with adapter. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5315–5328, Toronto, Canada. Association for Computational Linguistics.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Lewis M Paul, Gary F Simons, Charles D Fennig, et al. 2009. Ethnologue: Languages of the world. *Dallas, TX: SIL International. Available online at www. ethnologue. com/. Retrieved June*, 19:2011.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Yichen Ruan and Carlee Joe-Wong. 2022. Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8124–8131.

Sebastian Ruder, Jonas Pfeiffer, and Ivan Vulić. 2022. Modular and parameter-efficient fine-tuning for nlp models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 23–29.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005.

Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. 2023. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Haoyu Wang, Handong Zhao, Yaqing Wang, Tong Yu, Jiuxiang Gu, and Jing Gao. 2022. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *The ACM Web Conference 2022*.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. 2022. Pretrained models for multilingual federated learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1413–1421, Seattle, United States. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Baobao Chang, Songfang Huang, and Fei Huang. 2022. S4-tuning: A simple cross-lingual sub-network tuning method-tuning: A simple cross-lingual sub-network tuning method. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–537.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. 2023. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*.

Dun Zeng, Siqi Liang, Xiangjing Hu, Hui Wang, and Zenglin Xu. 2023. Fedlab: A flexible federated learning framework. *Journal of Machine Learning Research*.

Tianshu Zhang, Changchang Liu, Wei-Han Lee, Yu Su, and Huan Sun. 2023a. Federated learning for semantic parsing: Task formulation, evaluation setup, new algorithms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12149–12163, Toronto, Canada. Association for Computational Linguistics.

Yifei Zhang, Dun Zeng, Jinglong Luo, Zenglin Xu, and Irwin King. 2023b. A survey of trustworthy federated learning with perspectives on security, robustness and privacy. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1167–1176, New York, NY, USA. Association for Computing Machinery.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023c. FEDLEGAL: The first real-world federated learning benchmark for legal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023d. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

A.1 Description of Datasets

Below is a detailed description of three datasets:

News Classification. The News Classification (NC) dataset from the XGLUE benchmark (Liang et al., 2020) is utilized for the text classification (TC) task. This dataset includes five languages: English, Spanish, French, German, and Russian. Our objective is to predict the 10 kinds of article categories based on the article title and body, such as finance, sports, or travel. We sample 8,000 instances for training and 1,000 for evaluation or testing.

MTNT. The Machine Translation of Noisy Text (MTNT) dataset (Michel and Neubig, 2018) is one of widely adopted datasets. It consists of noisy comments on Reddit and professionally sourced translations. <English, French> and <English, Japanese> language pairs are utilized in our experiments. Previous research has utilized this dataset to assess the robustness of machine translation (MT) systems against domain shifts (Li et al., 2019). Given that FL inherently deals with client data that exhibits inherent shifts from centralized data, our study is well-suited to leverage this dataset.

UN Corpus. The UN Corpus (Ziemski et al., 2016) is the initial parallel corpus comprised of United Nations documents provided by the original creator. It consists of UN documents manually translated over the past 25 years (1990 to 2014) and encompasses the six official UN languages: Arabic, Chinese, English, French, Russian, and Spanish. We make use of this dataset for language modeling (LM) and machine translation (MT) tasks. In the LM task, we employ 50,000 instances per language for training data and allocate 5,000 instances for validation or testing. As for the MT task, we have three language pairs: <English, French>, <Arabic, Spanish>, and <Russian, Chinese>. During training, we sample 10,000 instances, while 5,000 instances are set aside for evaluation purposes.

Europarl. We utilize the Europarl corpus (Koehn, 2005), which comprises transcripts from European Union meetings, as our data source. The dataset comprises parallel text in 11 languages, from which we gather data samples for the language modeling (LM) task. Specifically, we collect data samples from 8 languages: English, Spanish, Portuguese, French, German, Finnish, Polish, Lithuanian, and Czech. To facilitate training, we extract 20,000 instances, while reserving 5,000 instances for validation or testing.

Table 4: Datasets related to three tasks.

Task	Dataset	# Train	# Dev	# Test	Metric
LM	Europarl	160,000	40,000	40,000	PPL
	UN	300,000	30,000	30,000	PPL
MT	MTNT	11,210	1,798	2,019	sacreBleu
	UN	30,000	15,000	15,000	sacreBleu
TC	NC	40,000	5,000	5,000	Accuracy

Algorithm 1: Cluster Aggregation

Input: The clusters set G;
Initial LoRA parameters $\mathbf{\Theta}^0$;
Clients set $\{C_i\}_{i=1}^N$;
The clients id list in each cluster g;
Training round L.
Output: LoRA Parameters $\{\Theta_i^L\}_{i=1}^N$.

for *i from* 1 *to* N **do**
    Initialize $\mathbf{\Theta}_i^0$ with $\mathbf{\Theta}^0$;
for *l from* 1 *to* L **do**
    for *i from* 1 *to* N **do**
        // local update of client i
        update $\mathbf{\Theta}_i^{l-1}$ with local data;
    // cluster aggregation of LoRA parameters
    foreach g *in* G **do**
        $\mathbf{\Theta}_g^l = \sum_{id \in g} \frac{1}{|g|} \Theta_{id}^{l-1}$;
        foreach *id in g* **do**
            $\mathbf{\Theta}_{id}^l = \Theta_g^l$;

A.2 Training Details

We have employed FedLab (Zeng et al., 2023) [1] as our federated framework. The training methodology outlined in (Weller et al., 2022) was followed. The maximum sequence length was set to 512. These experiments were conducted on a 4 GPU cluster comprising A100 GPUs, with each GPU having 80GB of memory. The AdamW optimizer was employed. Each client completed a full epoch of local learning before synchronizing with the server. To enhance performance, four different learning rates (1e-4, 5e-4, 1e-3, 5e-3) were utilized, with 5e-4 yielding the best results. The model was trained for 20 epochs for the language modeling task, 25 epochs for the machine translation task, and 30 epochs for the text classification task. In FL training, FedAvg was used as the learning algorithm. The adapter bottleneck was set to 128.

[1] https://github.com/SMILELab-FL/FedLab/

Table 5: Results for LM experiments on the UN Corpus.

Method	# TP ↓	En ↓	Es ↓	Zh ↓	Ru ↓	Ar ↓	Fr ↓	Avg ↓
Standalone	-	33.0±0.8	16.1±1.2	43.0±1.5	10.3±0.8	10.8±0.2	14.0±0.3	25.4±0.9
Centralized	-	7.4±0.2	4.8±0.4	6.9±0.2	3.9±0.1	5.2±0.3	4.6±0.3	5.6±0.3
+ Adapter	-	10.4±0.6	6.2±0.5	9.0±0.2	4.7±0.5	7.2±0.4	5.9±0.2	7.0±0.3
+ LoRA	-	11.3±0.5	6.7±.7	9.7±1.0	5.0±0.5	7.6±0.3	6.4±0.1	7.5±0.6
FedAvg	135.4M	**8.7**±0.2	**4.2**±0.5	**5.4**±0.1	4.1±0.2	**4.2**±0.7	5.1±0.5	**5.1**±0.6
+ Adapter	2.5M	22.8±0.5	14.9±0.5	17.0±0.4	9.9±0.5	17.2±0.1	14.3±0.7	15.5±0.6
+ LoRA	1.2M	10.8±0.9	6.6±0.3	9.3±0.5	5.0±0.6	8.1±0.5	6.3±0.6	7.5±0.8
FedLFC	**1.2M**	9.4±0.3	5.6±0.2	8.0±0.4	**4.0**±0.1	6.1±0.2	**5.1**±0.1	6.4±0.2

Table 6: Results for LM experiments on the Europarl.

Method	# TP ↓	En	Cs	Lt	Es	Pl	Fi	Pt	De	Avg
Standalone	-	9.4±0.9	2.8±0.4	2.6±1.2	4.3±0.6	2.8±0.5	3.0±0.2	3.7±0.6	3.5±0.8	4.0±0.2
Centralized	-	9.8±0.5	3.8±0.6	4.8±0.1	6.0±0.2	3.9±0.8	5.8±0.4	9.2±0.6	8.4±0.5	5.9±0.5
+ Adapter	-	10.6±0.6	7.1±0.5	8.2±0.5	7.3±0.2	5.8±0.8	7.6±0.8	7.6±0.5	7.9±0.5	7.7±0.2
+ LoRA	-	10.7±0.8	6.9±0.9	8.0±0.2	7.3±0.2	5.7±0.6	7.4±0.4	7.5±0.5	8.0±0.8	7.6±0.6
FedAvg	135.4M	10.4±0.6	6.4±0.5	9.2±0.2	**5.9**±0.1	5.9±0.3	7.8±0.6	7.5±0.5	7.9±0.8	7.7±0.6
+ Adapter	2.5M	12.0±0.8	10.6±0.2	14.2±0.6	8.3±0.4	7.5±0.8	10.7±0.2	9.4±0.4	9.2±0.6	10.1±0.5
+ LoRA	1.2M	11.4±0.8	8.8±0.6	11.3±0.4	7.8±0.5	6.6±0.2	9.3±0.5	8.5±0.8	8.8±0.6	8.9±0.4
FedLFC	**1.2M**	**10.4**±0.3	**6.1**±0.4	**6.3**±0.2	7.1±0.1	**5.4**±0.5	**6.4**±0.2	**7.2**±0.7	**7.7**±0.5	**7.1**±0.4

Within the LoRA module, the rank was set to 64, alpha to 32, and dropout to 0.1.

A.3 Extra Observation in the Experiment.

FL Methods Outperforms Centralized methods. In general, centralized models are considered as the upper bound of each task. However, Weller et al. (2022) show that FedNLP, FedAvg-model outperforms centralized-model. We hypothesize that the phenomenon is a result by parameter conflict. While there are shared commonalities, different languages also have distinct characteristics. Consequently, the aggregation of parameters from all languages can potentially interfere with the specific parameters of a particular language (Bari et al., 2021), resulting in a negative impact on transfer performance. The phenomenon is also observed in three tasks of our experiments.