WORLDWIDE EDGE-SILO FEDERATED LEARNING OF LANGUAGE MODELS

Alex Iacob [1] **Lorenzo Sani** [1] **Bill Marino** [1] **Preslav Aleksandrov** [1] **William F. Shen** [1] **Nicholas Donald Lane** [1]

As the societal impact of Language Models (LMs) extends, their reliance on large amounts of computation and multi-terabyte scrapped datasets that may be low-quality, copyrighted, or otherwise risk-laden raises practical, legal, and ethical questions. While medium-sized organizations may have access to valuable data and moderately capable hardware, the synchronization constraints of data-parallel SGD [2] upon Language Model (LM) training make collaboration unfeasible. Similarly, besides its server, an organization may have access to a fleet of edge devices holding private data without the ability to train an LM on their own.

Federated Learning (FL) has been recently proposed [1] as an alternative paradigm to centralized LM training due to its ability to relax synchronization requirements, avoid the movement of data, and address privacy and security issues. However, when scaled globally, federated learning requires collaboration across potentially incompatible legal jurisdictions, security requirements, and privacy regimes while accounting for the inherent locality of language data.

This work addresses the challenges that emerge once federated learning is scaled to a global level.

Governance The first challenge is determining how to create federated systems spanning actors with different legal, privacy, and security concerns or how to combine already existing federated systems seamlessly.

Edge-silo Collaboration The second challenge is enabling collaboration between low-resource edge devices and their governing silo, allowing access to otherwise private data.

Locality of Data The final challenge is the heterogeneity of naturally distributed data, as organizations may have datasets that differ in terms of language (e.g., for federations spanning multiple countries), genres, or complexity.

Our proposed Worldwide Federated Language Model (WorldLM) training system addresses these challenges by building federations of federations (Fig. 1).

1. WorldLM allows each sub-federation to adapt to its unique requirements and permits easy integration between previously isolated groups of actors, with potential secure aggregation or differential privacy protocols at any level. Crucially, WorldLM naturally transitions from cross-device settings at the edge-server level to a cross-silo setting for a regional or central server. This is achieved by allowing edge clients to train a restricted section of the model backbone via early exits.

2. To allow for personalization at the server level, each model is partitioned into *backbone* and *key* layers of the model architecture, with the *key* layers used as a means of both personalization and *residual* layer sharing. Personalization is achieved via an attention mechanism combining key layers from either client descendants or ancestors based on their similarity.

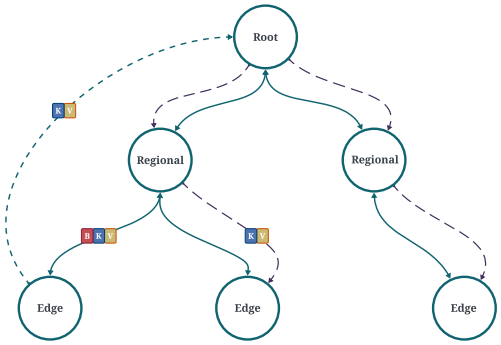

Figure 1. WorldLM structure, nodes may be both clients and servers for entire sub-federations. Nodes exchange information in the form of full models and lower-dimensional *residual* layer embeddings, which get aggregated via an attention mechanism, where they serve as *key: value* pairs. While full models are exchanged in a fixed manner, residual layer embeddings are dynamically routed to the most appropriate sub-federation as decided by a similarity function.

3. *Residual* layer sharing allows edge servers to select key layers highly dissimilar to the server model from the mobile devices under their control. These layers can then be routed to edge servers belonging to another federation with more relevant data. Since communication happens between servers, this comes at zero extra communication cost to the mobile devices and only requires transferring a limited number of model layers.

WorldLM represents a significant departure from previous federated approaches for training language models while maintaining a practical and communication-efficient implementation. Early results show that it can reduce communication frequency relative to centralized data-parallel training by $500\times$ from the perspective of the participating servers, while allowing mobile devices to participate by fine-tuning the first block of the model. WorldLM also effectively reduces the impact of language-induced data heterogeneity through *residual* layer sharing at an additional communication cost of $\mathcal{O}(MdK)$ for the servers where d is the depth of the federation tree, K is the total number of parameters of the key layers, and M is a hyperparameter specifying the number of residuals selected for transmission. Where M is much smaller than the number of edge clients selected per round and K is between 1 to 4 layers of the model. Given its communication efficiency and low-comms mechanism for tackling heterogeneity, we believe that WorldLM will provide an excellent starting point for training federated language models across geographic, legal, and cultural boundaries.

REFERENCES

[1] A. Douillard et al. Diloco: Distributed low-communication training of language models. *CoRR*, abs/2311.08105, 2023.

[2] S. Rajbhandari et al. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, November 9-19, 2020*, page 20. IEEE/ACM, 2020.

[1]Cambridge: {aai30, ls985,wlm27,pa511,fs604,ndl32}@cam.ac.uk. Work conducted with the support of the Ministry of Education of Romania, through the Credit and Scholarship Agency.