Federated LLM Training with Heterogeneous Mobile Clients

Andrzej Szablewski[1], Lorenzo Sani[1,2], Nicholas D. Lane[1,2]

[1]University of Cambridge, [2]Flower Labs, Correspondence to Andrzej Szablewski: as3623@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Motivation

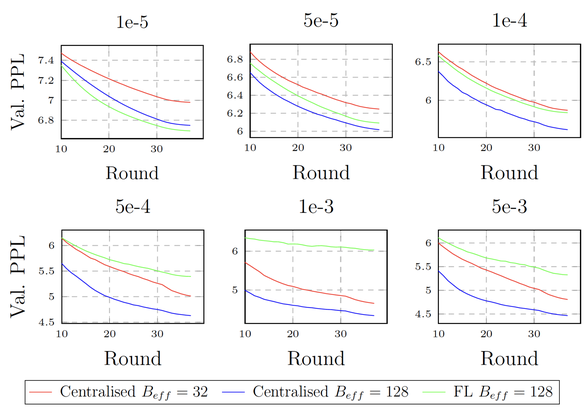
Federated learning has been recently leveraged for collaborative training of LLMs in the cross-silo setting[1]. However, using mobile hardware results in additional challenges:

1. Mobile clients often differ in available hardware, directly influencing the maximum micro-batch size and its corresponding processing time.
2. The size of LLMs and limited network bandwidth allow for only a small number of communication rounds, resulting in sparse updates to the aggregate model[2].

We explore the **optimal choice of training and model hyperparameters** to achieve the **lowest model perplexity** and **minimise client idling time.**

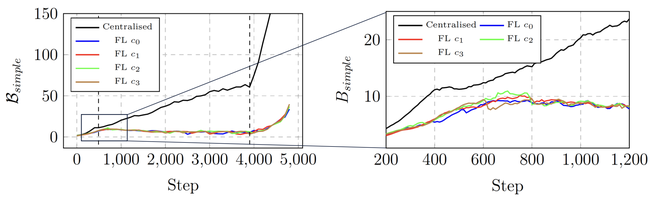
Optimising Federated Hyperparameters

Learning rate does not translate between centralised and federated settings, across varying mini-batch sizes.



The compute-optimal mini-batch size predictions using the Gradient Noise Scale[3] suggest that FL is more efficient with smaller mini-batches, or GNS theory does not hold in FL.

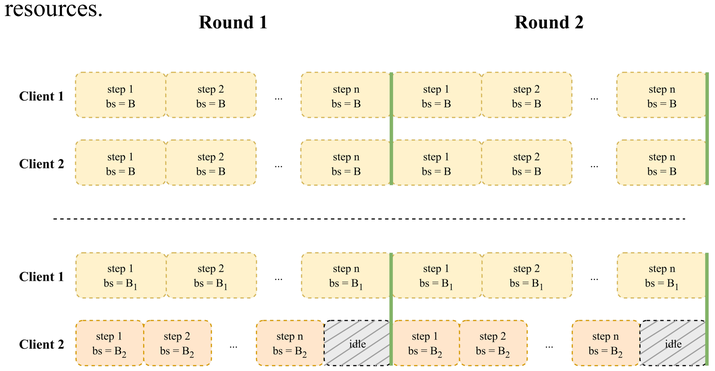
Change of Gradient Noise Scale throughout training of a 124M GPT model



Client Idle Time

The duration of each optimisation step depends on the client's hardware and the chosen training hyperparameters.

When round processing times differ, they lead to client idling, effectively wasting local resources.



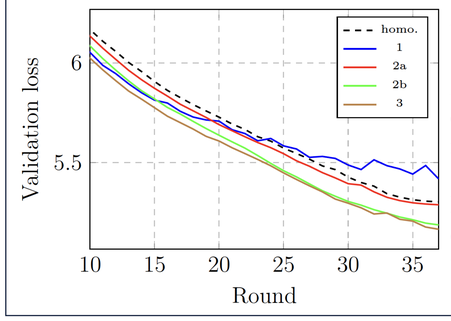
Hardware-Heterogeneous Federated Learning

Fixing the number of communication rounds and the training budget, we propose the following hardware-heterogeneity resolution **strategies** with the corresponding aims**:**

- Strategy 1 – Identical number of optimisation steps per client
- Strategies 2a and 2b – Identical number of samples per client (with or without gradient accumulation, respectively)
- Strategy 3 – Minimal client idling time

Strategy	Training Wall-Time	Avg. Idling Ratio/client	PPL
1	2,078s	20.11%	5.454
2a	2,344s	26.85%	5.290
2b	3,376s	37.10%	5.238
3	**1,876s**	**2.62%**	**5.185**
homo.	1,318s	1.64%	5.315

Aggregate Model Perplexity (124M GPT)



Strategy 3 minimises the client idling time and achieves the lowest perplexity.

Aggregating gradient updates of different fidelities may lead to an increased convergence rate.

In strategy 1, the top-performing client dominates the other contributors, leading to worse performance.

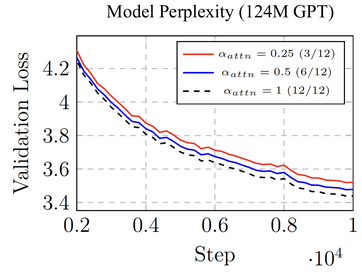
Selective Multi-Head Attention (SMHA)

To further balance the computation across clients, we train only a selection of attention layer parameters in GPT – Selective Multi-Head Attention (SMHA).

Selective training of attention layers meaningfully decreases training wall-time at the expense of a slight increase in perplexity.

The number of actively trained heads can be dynamically varied during training, implementing head redundancy metrics.

Model Perplexity (124M GPT)



Fraction of Trained Heads	Change in Wall-Time ↓	Change in Perplexity ↓
¼	-21.89%	+2.3%
½	-15.37%	+1.1%

Partial Model Training in FL

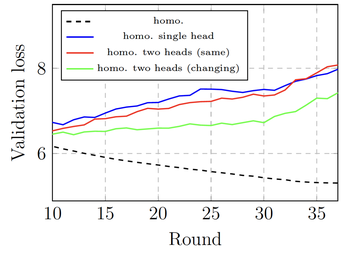
Attention layers can be split across clients by associating only selected attention head parameters with each client.

At least one client needs to train all model parameters for convergence.

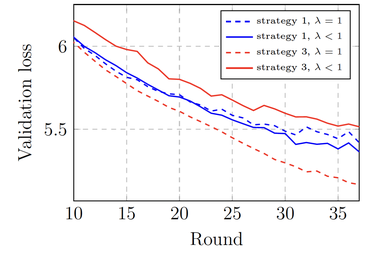
Strategy 3 with SMHA ($\lambda < 1$) decreases training time as expected, but results in worse perplexity compared to complete model training ($\lambda = 1$).

Strategy	Change in Wall-Time ↓	Change in Perplexity ↓
1 with SMHA	+0.96%	-1.02%
3 with SMHA	-4.15%	+6.74%

Hardware-homogeneous, with each client training a fraction of attention heads



Hardware-heterogeneous, with at least one client training all attention heads



References

[1]Iacob, Alex, Lorenzo Sani, Bill Marino, Preslav Aleksandrov, William F. Shen, and Nicholas Donald Lane. 'Worldwide Federated Training of Language Models'. arXiv, 27 May 2024. https://doi.org/10.48550/arXiv.2405.14446.

[2]Sani, Lorenzo, Alex Iacob, Zeyu Cao, Royson Lee, Bill Marino, Yan Gao, Dongqi Cai, et al. 'Photon: Federated LLM Pre-Training'. arXiv, 5 November 2024. http://arxiv.org/abs/2411.02908.

[3]McCandlish, Sam, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 'An Empirical Model of Large-Batch Training'. arXiv, 14 December 2018. https://doi.org/10.48550/arXiv.1812.06162.