FedSpaLLM: Federated Pruning of Large Language Models

Scientific Achievement

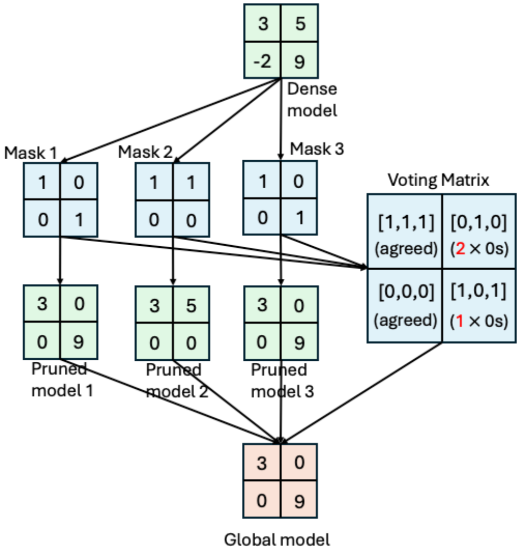
- FedSpaLLM is a federated learning (FL) framework to enable collaborative pruning of pre-trained Large Language Models (LLMs)
- The final global model achieving up over 30% improvements in a key performance metric over client local models.

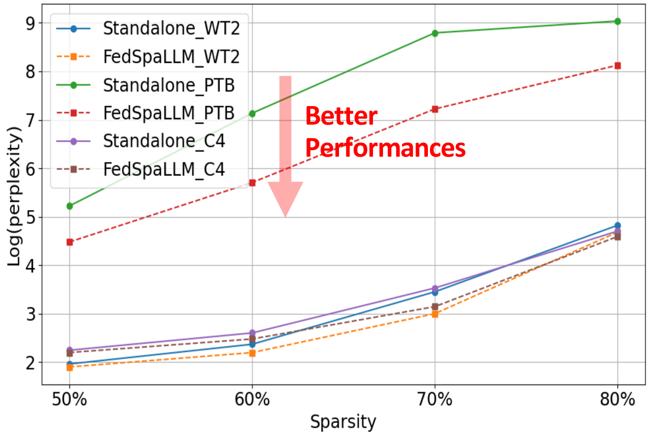
Significance and Impact

- FedSpaLLM is the first FL framework for pruning large language models over a set clients that leads to a sparse global model with improved performances and safeguards the privacy of the clients' data.
- The framework can also be applied to other large ML models, e.g., foundation models, in a wide range of applications and settings.

Technical Approach

- FedSpaLLM proposes two major modifications to the standard FL algorithm to address disagreements in mask selection in pruning and ensure the sparsity of the global model.
- FedSpaLLM allows flexible integration of various pruning algorithms, making it adaptable to different model architectures, including emerging large-scale ML models.
- Introduces a sparsity-aware aggregation scheme that coordinates pruning decisions across clients while preserving local data privacy and computational efficiency.





The image on the left shows the innovative approaches in FedSpaLLM to address the mask disagreements and global model sparsity discrepancy with a voting matrix and novel aggregation schemes. The image on the right shows the performance improvements in perplexity, a key evaluation metric, from FedSpaLLM over Standalone models from the clients.

PI(s)/Facility Lead(s): Kibaek Kim
Collaborating Institutions: Emory University
ASCR Program: ASCR - AI for Science
ASCR PM: Steven Lee
Publication(s) for this work: G. Bai, Y. Li, Z. Li, L. Zhao, K. Kim. “FedSpaLLM: Federated Pruning of Large Language Models” NAACL 2025.