PruFed: Federated Finetuning of Pruned LLMs

SouthNLP

Shrenik Bhansali, Alwin Jin, Tyler Lizzo, Larry Heck
AI Virtual Assistant (AVA) Lab, Georgia Institute of Technology

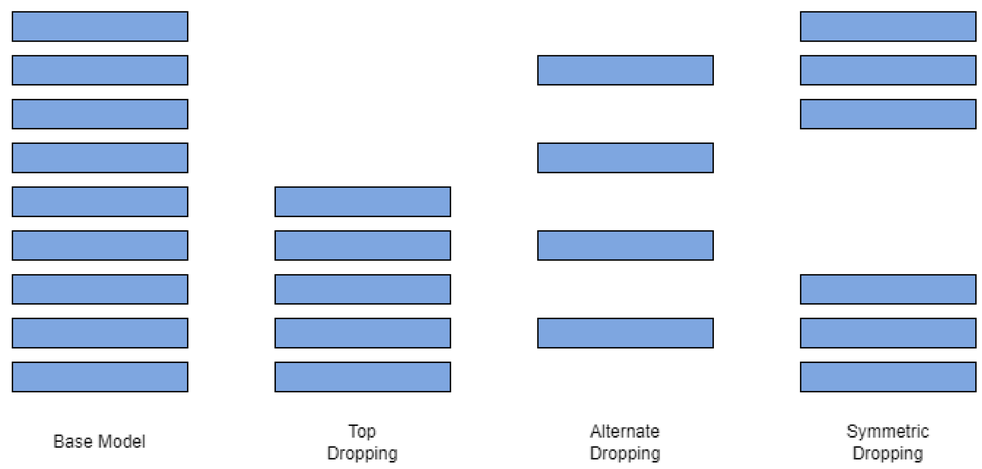
Georgia Institute of Technology

Motivation: Data Sensitivity & Computational Inadequacy

Traditionally, Large Language Models are fine-tuned in a centralized manner, requiring the collection of vast amounts of data, a process associated with significant time, labor, and accessibility challenges. Instead, we look to develop a federated learning methodology that retains user privacy, bypasses traditional data collection, and distributes finetuning to resource constrained devices. Thus, we introduce PruFed, a model-heterogenous federated finetuning approach.

Method: Pruning and Federating

- Parameter pruning approaches can be compute intensive
- Therefore, we look to primitive, low-cost, task-agnostic layer pruning techniques

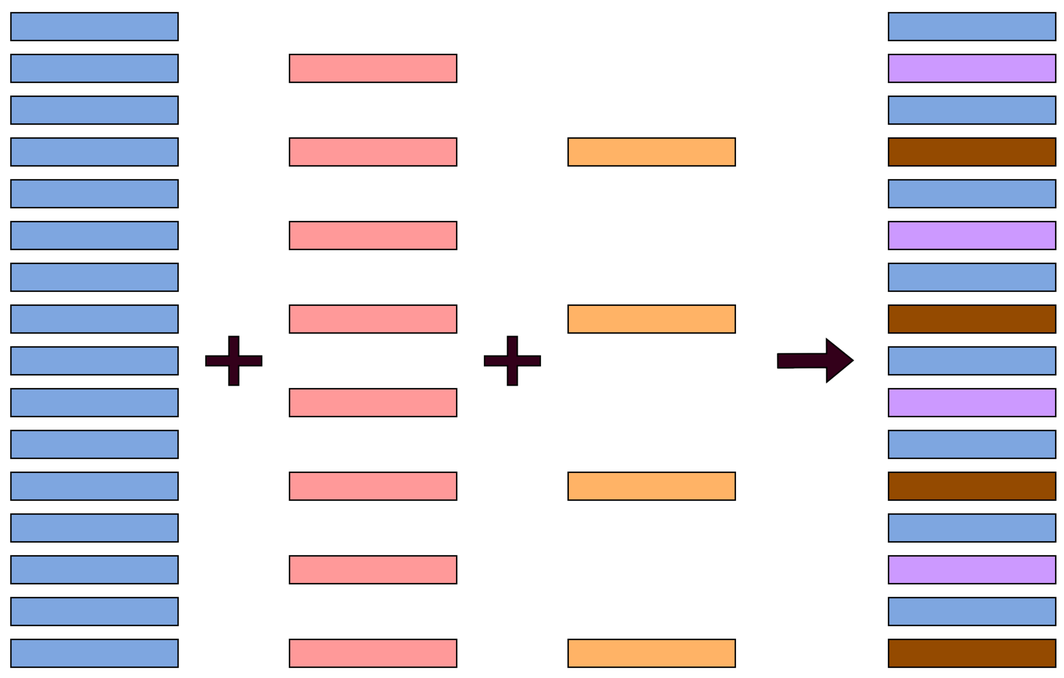


We investigate various layer-dropping strategies

- We create a federated learning system to finetune and aggregate pruned models of different depths
- We aggregate across shared layers of pruned models to create robust global models

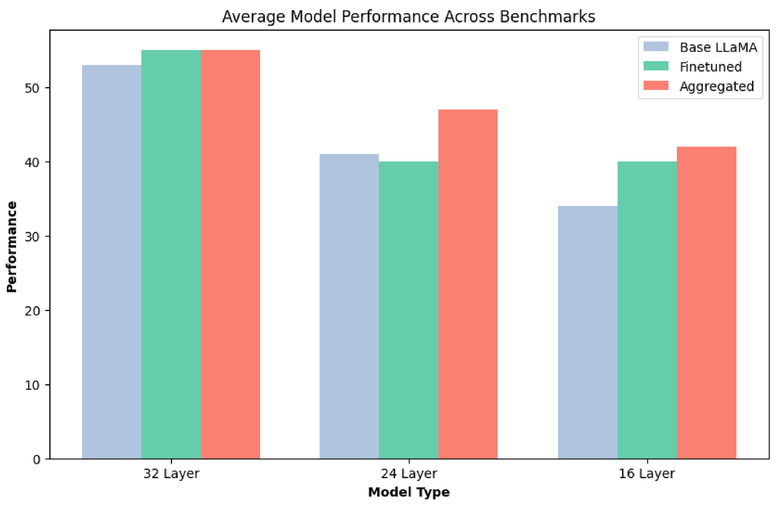
Initialize θ^a, θ^b, θ^c
for n clients **do**:
  Initialize $U_i = (\theta^i, \Delta w)$
end for
while $k \leq K$ **do**:
  U_k // sample portion of users
  U_k^a, U_k^b, U_k^c // Group U_k by model depth
  for device type $i \in \{a, b, c\}$ **do**
    for client $c \in U_k^i$ with adapter weights Δw **do**
      $c = \text{InstructionTuning}(\Delta w)$
    end for
  end for
  $U_k = \text{HeteAgg}(U_k)$
end while

Algorithm for model-heterogenous federated finetuning of pruned language models



Visualization of layer-wise heterogenous aggregation (HeteAgg)

Results



- Ideal Layer-Pruning strategies are task-dependent
- On average, Top-Alternate Dropping performs best
- For more sparse models, PruFed matches or outperforms their non-federated counterparts (regardless of pruning strategy)