Inference Optimization for Large Language Models: A Study of MLP Sparsity Based on Dynamic Routing and Load Balancing

Advisor: Prof. Chien-Chung Ho
An Li, Lee-Hsin Feng
Department of Computer Science and Information Engineering, National Cheng Kung University

Introduction

- **Background:** Edge deployment of Large Language Models (LLMs) is bottlenecked by massive parameter sizes and high inference costs. Effective compression requires identifying and eliminating parameter redundancy.
- **Focal Point: MLP Layers:**

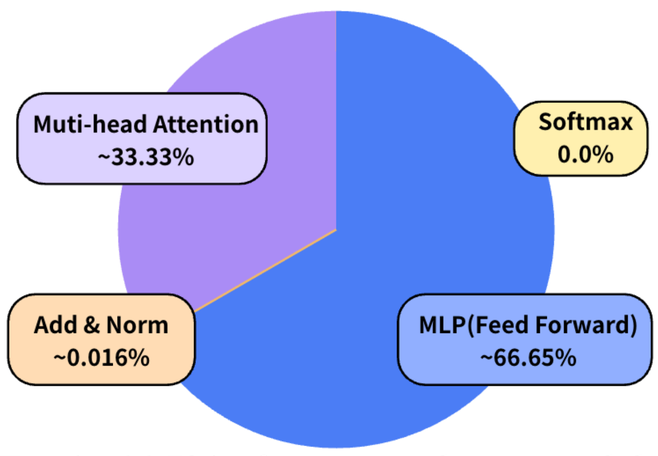


Fig 1: 66.7% of core weights in models.

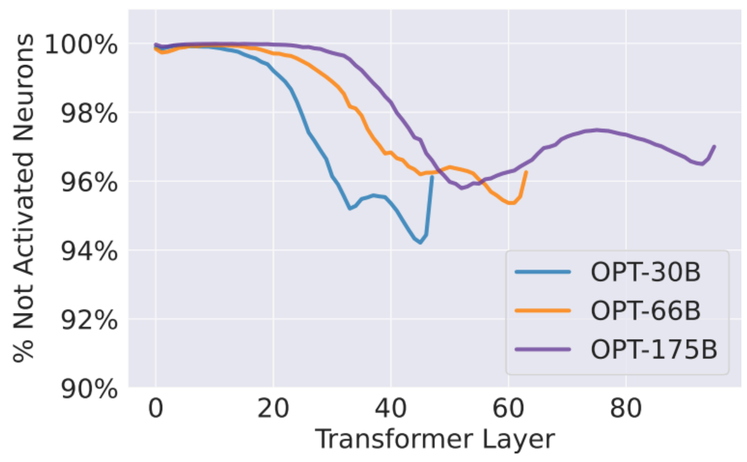


Fig 2: 95% of neurons remain inactive.

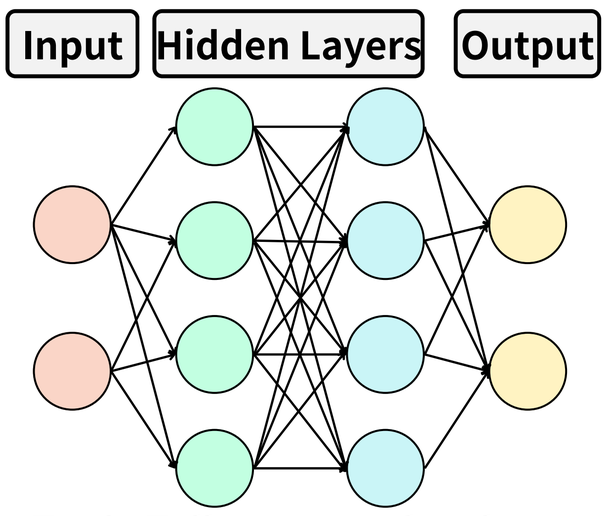
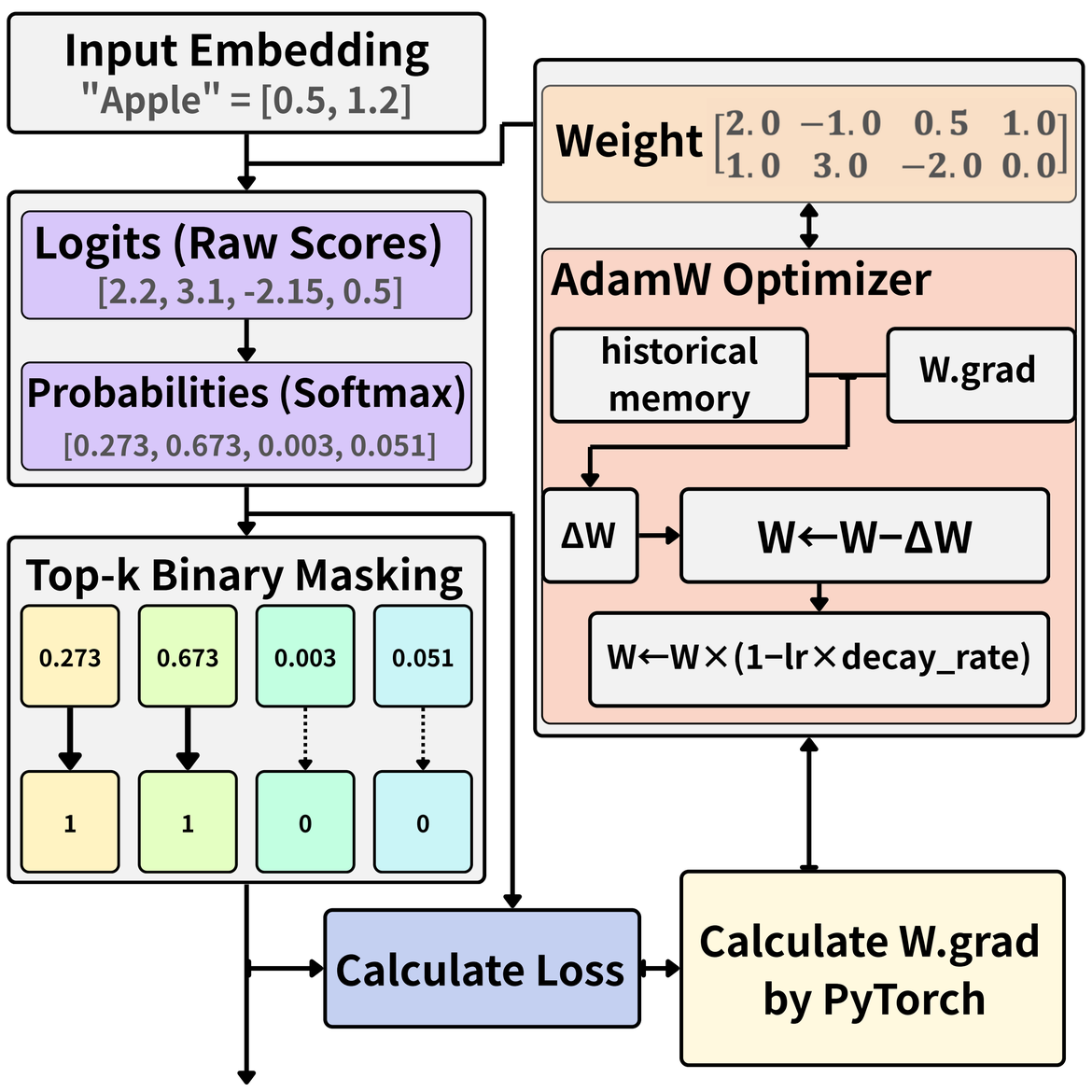


Fig 3: Fully connected architecture.

- **Objective:** To prove smaller models, which possess denser parameters maintain baseline accuracy despite pruning the majority of inactive MLP neurons.

Architecture



1. Dynamic Routing (Prediction)

- Action: Project hidden states to generate activation probabilities.
- Result: Identify redundant neurons with low computational overhead.

2. Top-K Masking (Execution)

- Action: Apply binary masks to force unselected neurons to zero.
- Result: Bypass invalid pathways at the software level.

3. Load Balancing (Optimization)

- Action: Apply auxiliary loss to penalize over-activated neurons.
- Result: Ensure uniform utilization of all model parameters.

Experimental Results

- Model: OPT-1.3B
- Environment: Python, NVIDIA RTX 5080

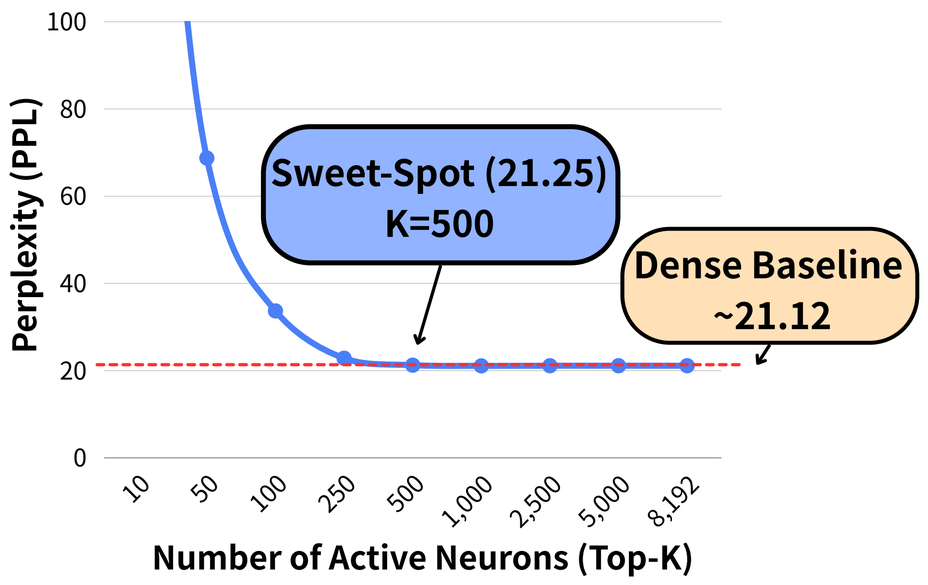


Fig 4: K=500 achieves performance comparable to the baseline.

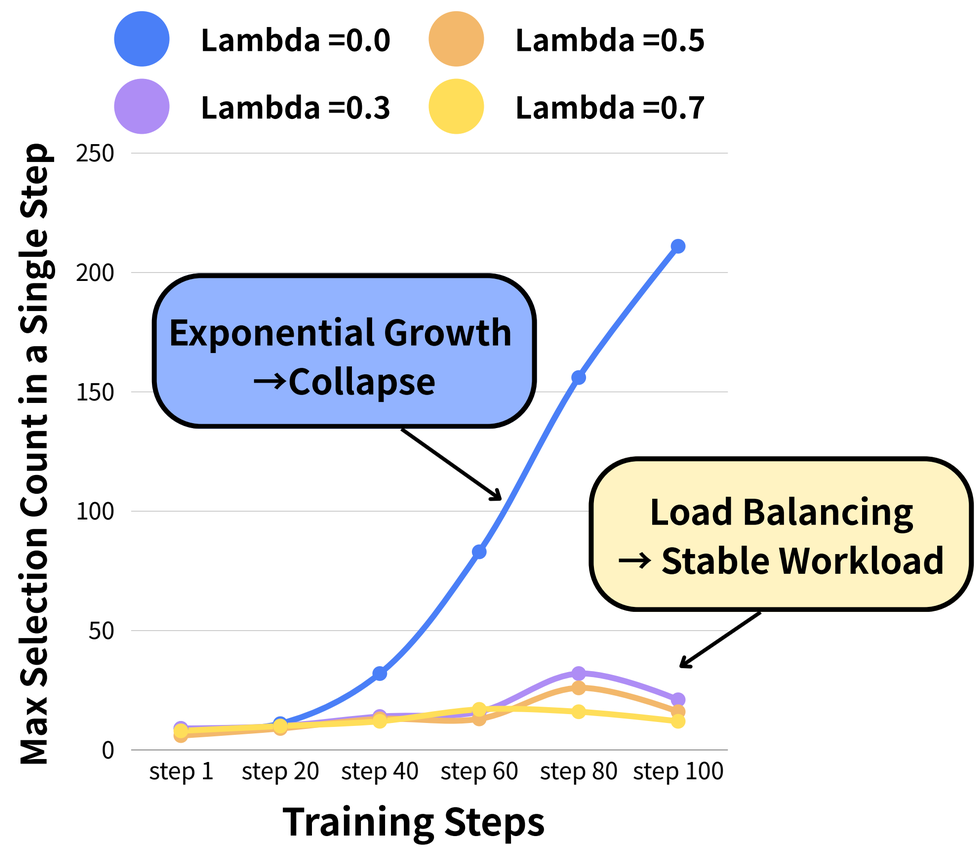


Fig 5: Load balancing stabilize the workload across steps.

Conclusions

This project empirically validates that up to 94% of MLP parameters possess mathematical redundancy during inference, establishing a theoretical upper bound for model compression on edge devices. We have successfully proven the concept of dynamic routing and masking at the software level. Future work will focus on integrating hardware-level sparse matrix computation cores (e.g., CUDA Sparse Kernels) to translate this theoretical redundancy into substantive VRAM savings and inference acceleration.