大型語言模型推論優化：基於動態路由與負載平衡之 MLP 稀疏化研究
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

Introduction

1. **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.
2. **Focal Point:** MLP Layers

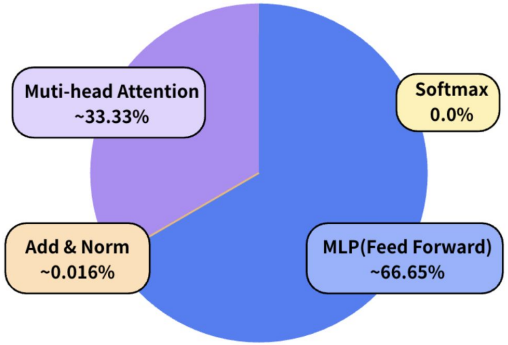


Fig 1: 66.7% of core weights in models.

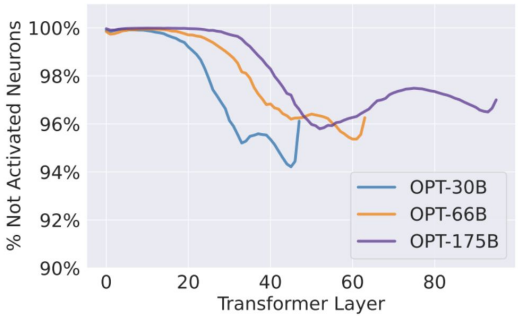


Fig 2: 95% of neurons remain inactive.

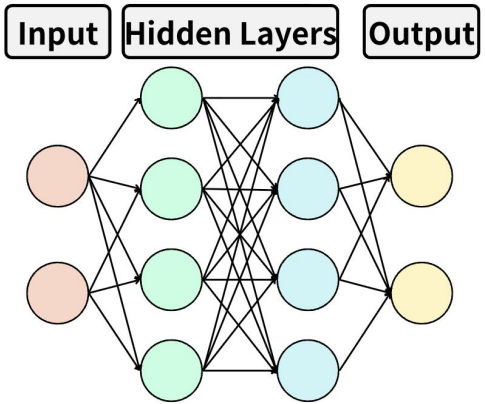


Fig 3: Fully connected architecture.

3. **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)**

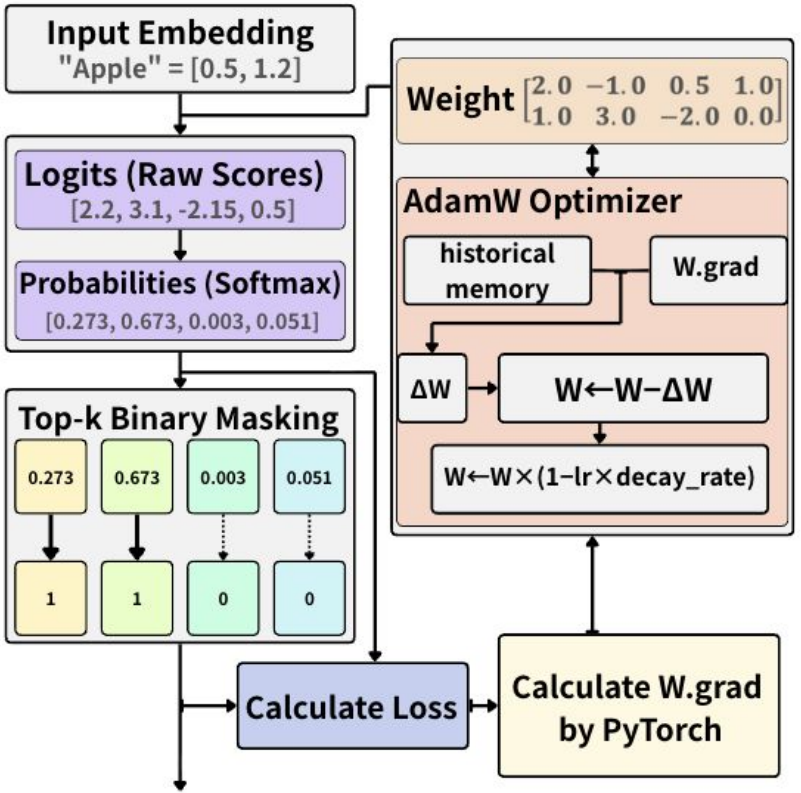
- Action: Project hidden states to generate activation probabilities.
- Result: Identify redundant neurons with low computational overhead.

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

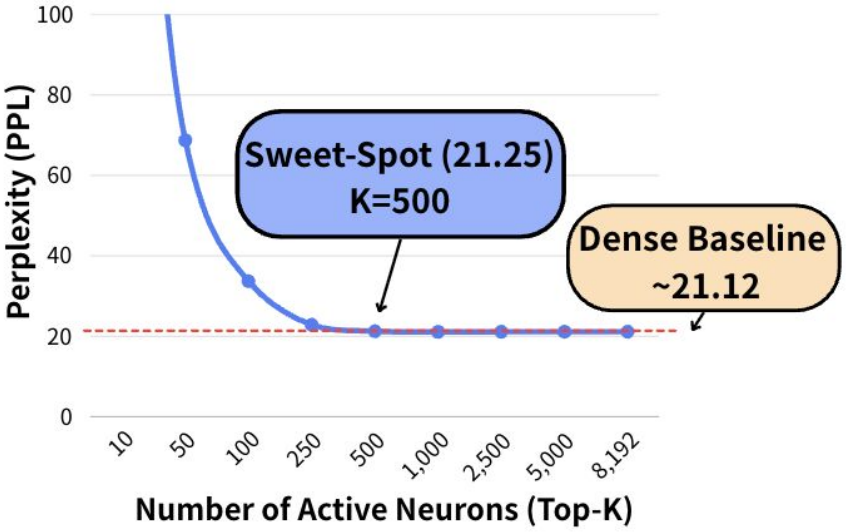


Fig 4: K=500 achieves performance comparable to the baseline.

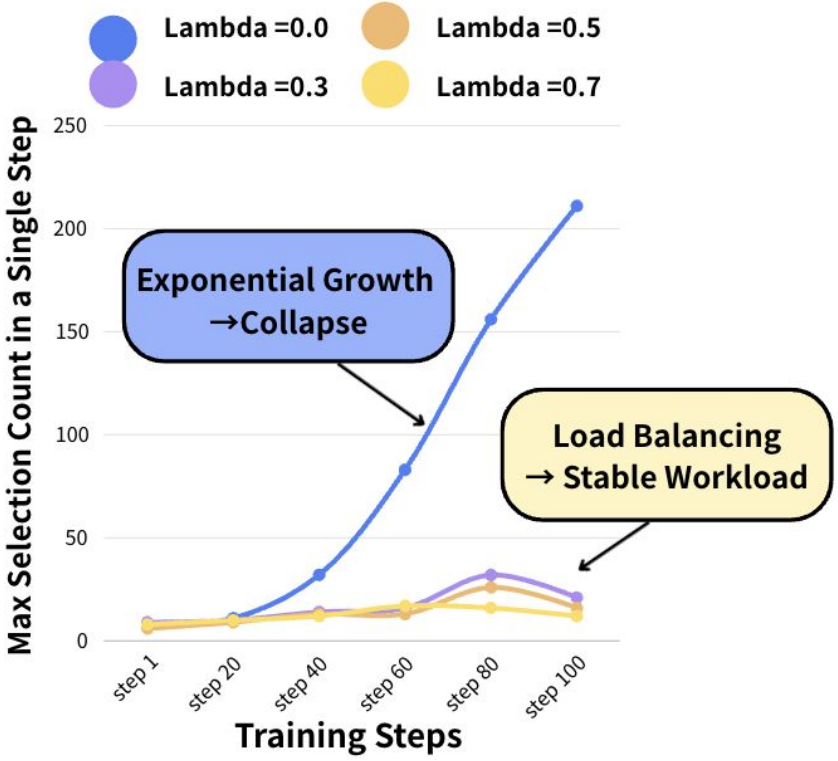


Fig 5: Load balancing stabilize the workload across steps.

Conclusions

1. Empirical Validation of Dynamic Sparsity： 僅保留約 6% 活躍神經元時，模型 Perplexity 仍維持與全運算基準高度一致。

2. Establishing Theoretical Bounds： 從架構層面支持 MLP 層中高達 94% 的參數在推論時具備數學冗餘性的假設，為邊緣設備的模型壓縮提供了理論上限的參考依據。

3. Proof of Concept & Future Work

- 本專題已於軟體層面成功實作動態路由與遮罩，阻斷無效邏輯路徑。
- 下一步可結合底層稀疏矩陣運算核心 (Sparse Kernels) 或專用 AI 晶片，將此理論冗餘實質轉化為 VRAM 的節省與推論加速。