NYU

Enhanced MoE with Random Routing based on Transformer

Yuyang Ji (yj2669)

Hechuan Liang (hl5035)

01/23/2020

Executive Summary

- Problem Statement: Traditional Transformer models exhibit high performance across various datasets but are computationally expensive.
- Solution Approach: We propose a novel Transformer architecture that integrates a Mixture of Experts (MoE) and random routing.
- Value/Benefit: This approach effectively enhancing performance with a marginal increase in computational requirements.

Technical Challenges

Complex Integration:

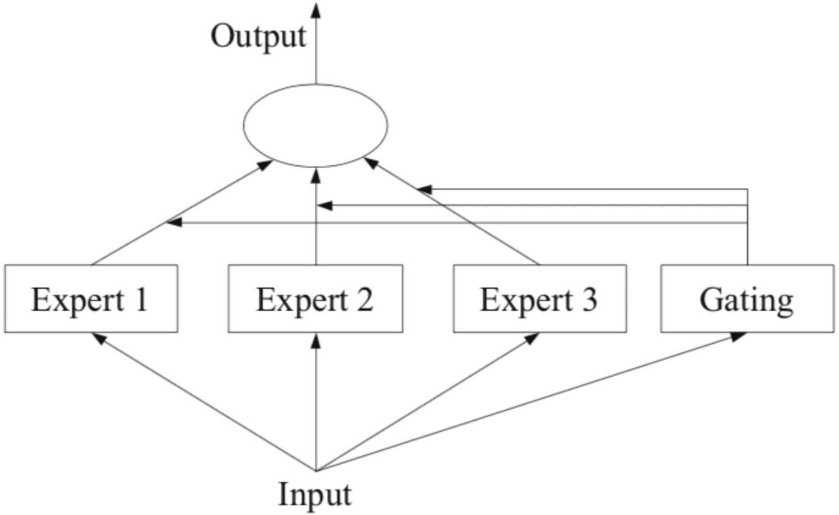
The primary challenge lies in seamlessly integrating MoE and random routing with the Transformer architecture. This requires sophisticated modifications to the standard Transformer model to accommodate MoE layers without disrupting the core functionalities.

Managing Computational Overheads:

While MoE promises enhanced performance, its integration inherently risks increasing computational load. The challenge is to optimize the model to gain performance benefits without proportionally escalating computational costs.

Approach: Mixture of Expert

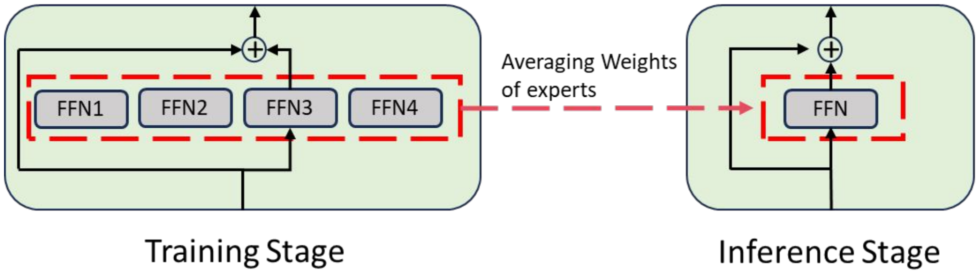
The objective of sparsely-activated model design is to support conditional computation and increase the parameter count of neural models like Transformers while keeping the floating point operations(FLOPs) for each input example constant.



NYU

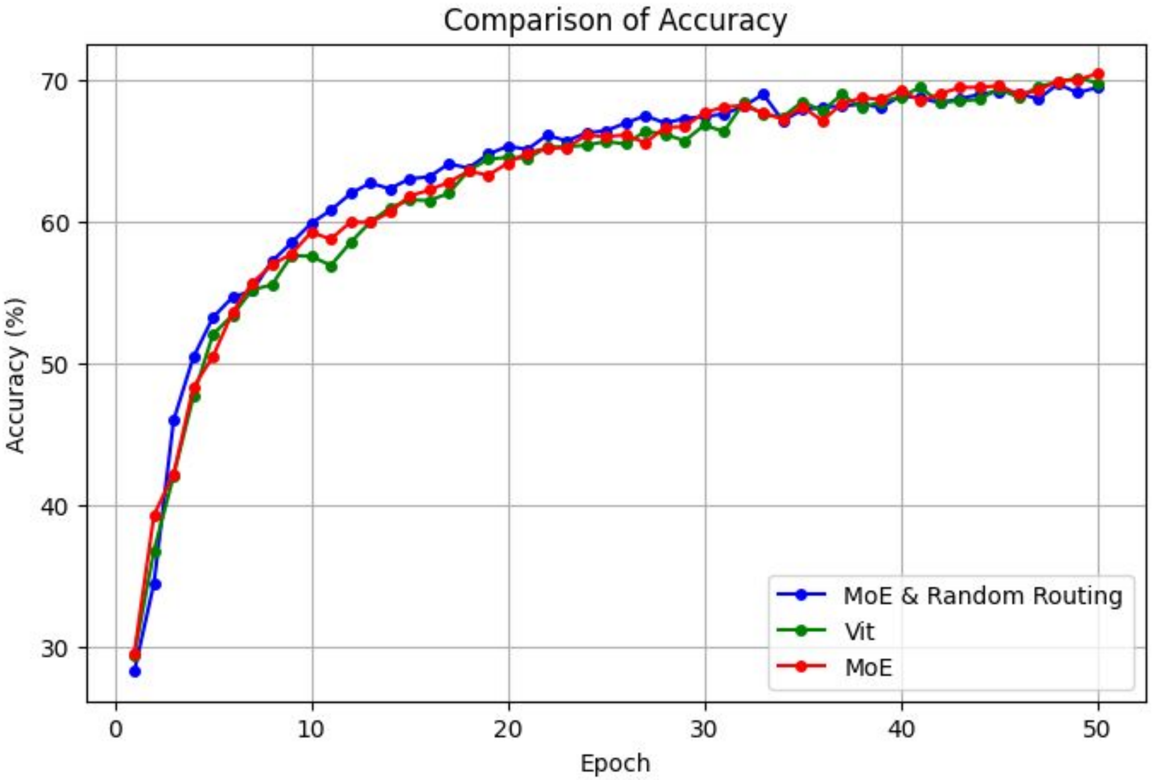
Approach: Random routing

Stochastic routing policy like random routing to work as well as classical routing mechanism like Switch routing with the following benefits. Since input examples are randomly routed to different experts, there is no requirement for additional load balancing as each expert has an equal opportunity of being activated simplifying the framework. Further, there are no added parameters, and therefore no additional computation, at the Switch layer for expert selection.



NYU

Summary of Main Results



NYU

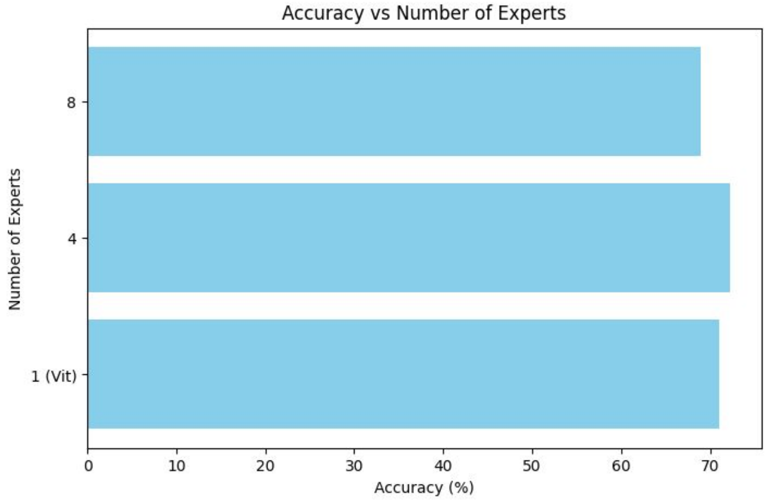
Evaluation: Parameters

In the “MoE & random routing” approach, the inputs are randomly assigned to different experts within the model. This random routing process does not introduce additional parameters. Instead, this randomness ensures that all experts are equally likely to be chosen, maintaining the overall parameter count of the model. The lack of a sophisticated routing algorithm, which might otherwise increase the parameter load, is a key factor in keeping the parameter count unchanged.

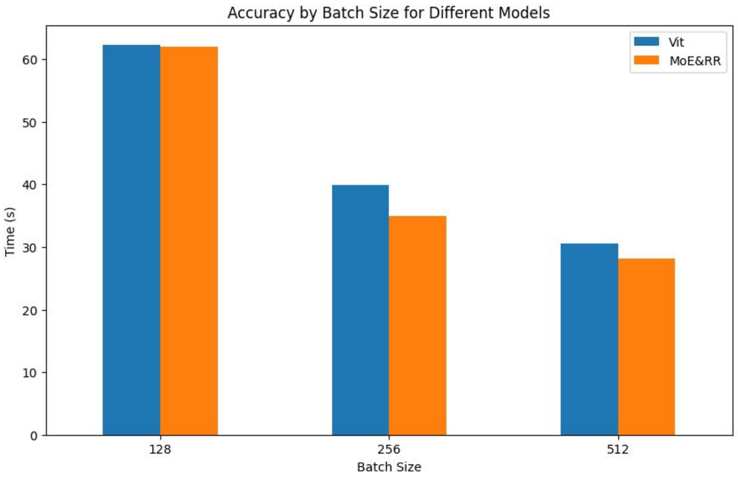
	MoE & random routing	MoE	Vit
Parameters	12,798,490	12,812,938	12,798,490

Evaluation: Number of Experts

Number of Experts	1(Vit)	4	8
accuracy(%)	71.06	72.18	68.96



Evaluation: Batch Size

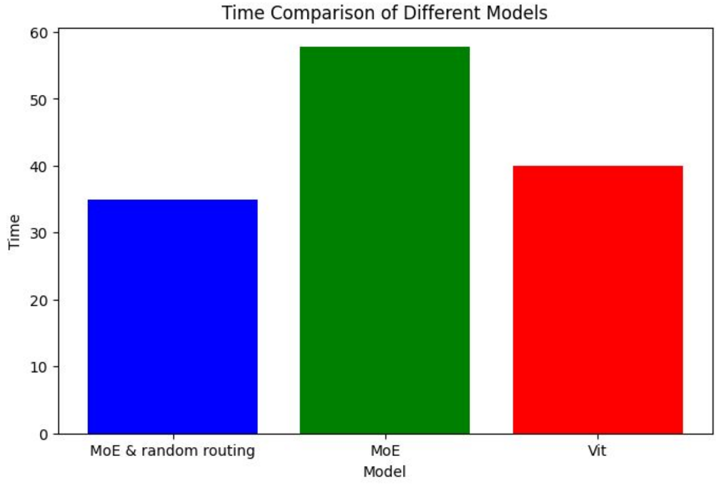


Batch size model	128	256	512
Vit	62.33	39.91	30.53
MoE&RR	62.01	34.89	28.18

NYU

Evaluation: Training Time

	MoE & random routing	MoE	Vit
Time(s)	34.89	57.78	39.91



Conclusions/Observations

In our study, we observed that while Mixture of Experts (MoE) offers a slight improvement in performance, the increase is not substantial.

Random routing, on the other hand, significantly speeds up the training process compared to traditional MoE.

Additionally, our experiments show that an optimal number of experts maximizes performance, with four experts providing the best results. Increasing the number of experts beyond this does not necessarily lead to better performance and may introduce additional complexity.

Thank you!

GitHub Link:

https://github.com/gagi0911/hpml_final_project

NYU