CS324 - Large Language Models

Modular Architectures

HUMANE Lab

김태균

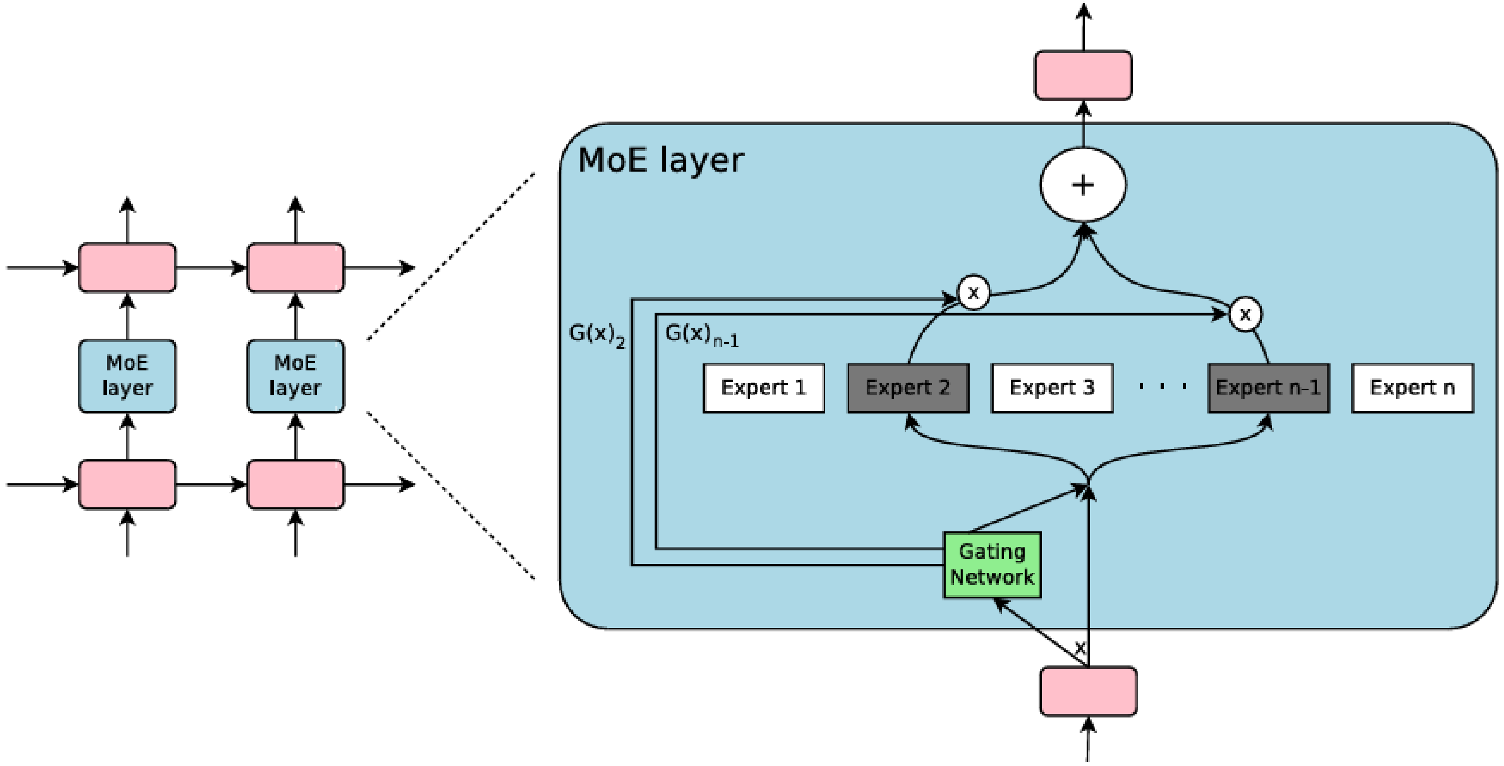
2025.01.31

Introduction

- Scaling dense transformer models up is non-trivial, requiring data, model, and pipeline parallelism
- To address these issues, we explore two different types of selective architectures
  - Mixture-of-Experts (MoE)
  - Retreival

What is a Mixture of Experts (MoE)?

- A method of using multiple expert networks to select the most suitable expert for processing a specific input
- Two components
    1. Experts
    2. Gate network (or Router)



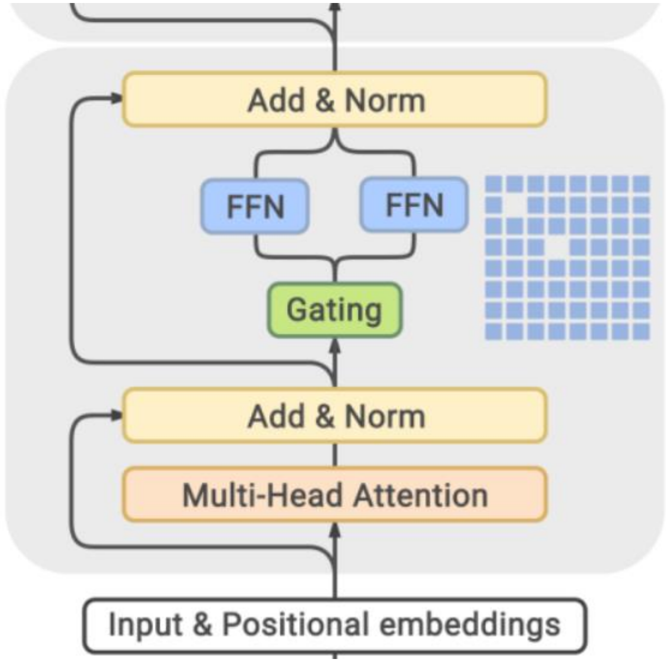
What is a Mixture of Experts (MoE)?

$$f(x) = \sum_{e=1}^{E} \underbrace{g_e(x)}_{\text{gating}} \underbrace{h_{\theta_e}(x)}_{\text{expert}}.$$

- Training
- Saving compute
  - If gating function which places zero on most experts, then we only have to evaluate the experts with nonzero gating function
  - e.g. [0.04, 0.8, 0.01, 0.15] -> [0, 0.84, 0, 0.16]
- Balancing experts
  - MoE is only effective if all experts pitch in
- Parallelism
  - MoE is very conducive to parallelization

Sparsely-gated mixture of experts

- Apply MoE idea to each token and each transformer block
- Turn each feed-forward network into a MoE feed-forward network

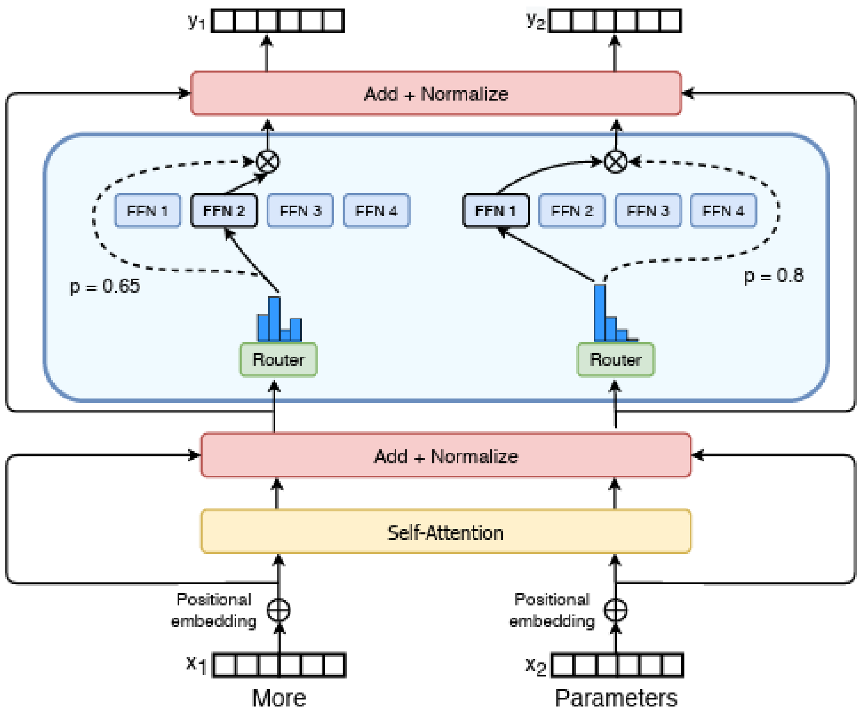


Sparsely-gated mixture of experts

- Apply MoE idea to each token and each transformer block
- Turn each feed-forward network into a MoE feed-forward network
- Top-2 experts
- Balancing experts
  - Add load-balancing loss

Switch Transformer

- Top-1 expert (to get even more sparsity)
- Tricks
  - Selective casting from FP32 to FP16
  - Smaller parameters for initialization
  - Expert dropout
  - Expert parallelism
- Trained a 1.6T parameter model
- Improved pre-training speed compared to T5-XXL by 4x



Balanced Assignment of Sparse Experts layers

- Joint optimization over all the tokens in the batch
- Assign each token 1 expert, but load balancing is a constraint
  - B : the number of tokens in the batch
  - E : the number of experts
  - a : assignment vector

$$a = [a_1, \ldots, a_B] \in \{1, \ldots, E\}^B$$

$$\text{maximize} \sum_{i=1}^{B} w_{a_i} \cdot x_i \quad \text{subject to} \quad \forall e : \sum_{i=1}^{B} \mathbf{1}[a_i = e] = \frac{B}{E}.$$

Decentralized mixture-of-experts

- So far, the MoE was motivated from a perspective of a central organization (e.g. Google or Facebook) scaling up a massive LLM
- However, MoE suggests a much more radical decentralization
  - e.g. Harness the hundreds of millions of consumer PCs
- Main consideration
  - Many nodes
  - Frequent node failures
  - Home-Internet communication bandwidth
- Distributed hash tables (DHT)

Mixtral 8x7B

- An open-source language model based on the MoE architecture, developed by Mistral AI
- Includes an expert network of eight 7B parameter models
- Activates two experts per token for computation
- The actual computation is at the level of a 14B model, but the performance is close to a 56B model

Mixtral 8x7B

	LLaMA 2 70B	GPT - 3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	87.1%	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

Summary

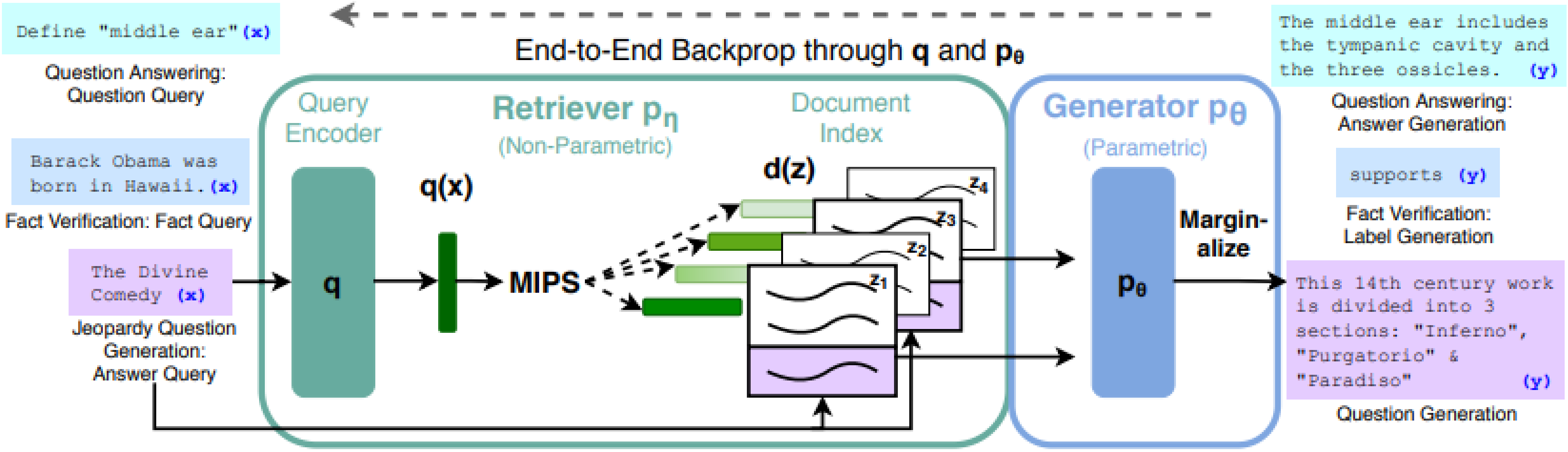
- MoE : classic idea of applying different experts to different inputs
- Allows for training much larger language models
- Much more efficient per input than dense transformer models

Retrieval-based models

- Model that retrieves relevant information from an external database and utilizes it
- Operation Principle
  - Retrieve a relevant sequence z based on input x
  - Generate the output y given the retrieved sequence z and input x

Retrieval-augmented generation (RAG)

- By utilizing external knowledge, the model can remain relatively small in size while still leveraging a wide range of information
- For knowledge updates, the external knowledge base can be updated to reflect the latest information without requiring model retraining



Retrieval-augmented generation (RAG)

- The retrieval-based models are highly geared towards knowledge-intensive, question answering tasks
- Beyond scalability, retrieval-based models provide interpretability

Conclusion

- In order to scale, MoE and retrieval-based methods are more efficient than dense transformer