Multimodal Machine Learning

A Sparsely-activated model approach

Jianzhu Yao, Jun 24th

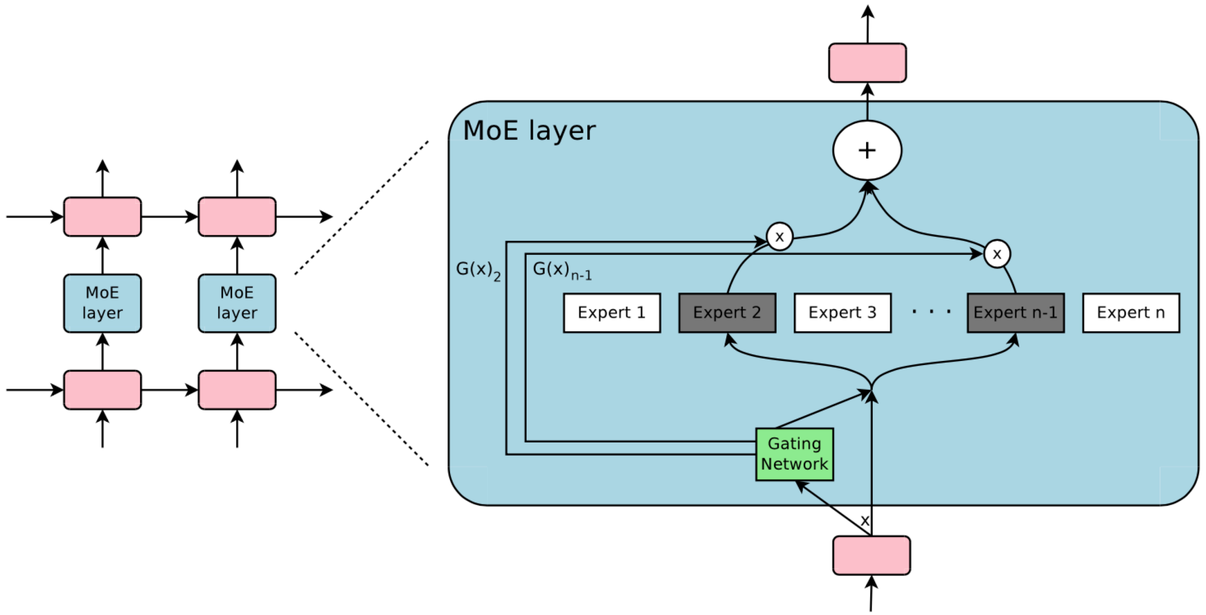
Dept. of Computer Science and Technology, Tsinghua University

Outline

1. Introduction
2. Unimodal Sparsely-activated model
    1. Definition
    2. Special features
3. Multimodal Sparsely-activated model
    1. Definition
    2. Existing work and new challenges
4. What’s more? A new paradigm or just papers?

Introduction

A simple review on Mixture of Experts



$$G(x) = Softmax(KeepTopK(H(x), k))$$

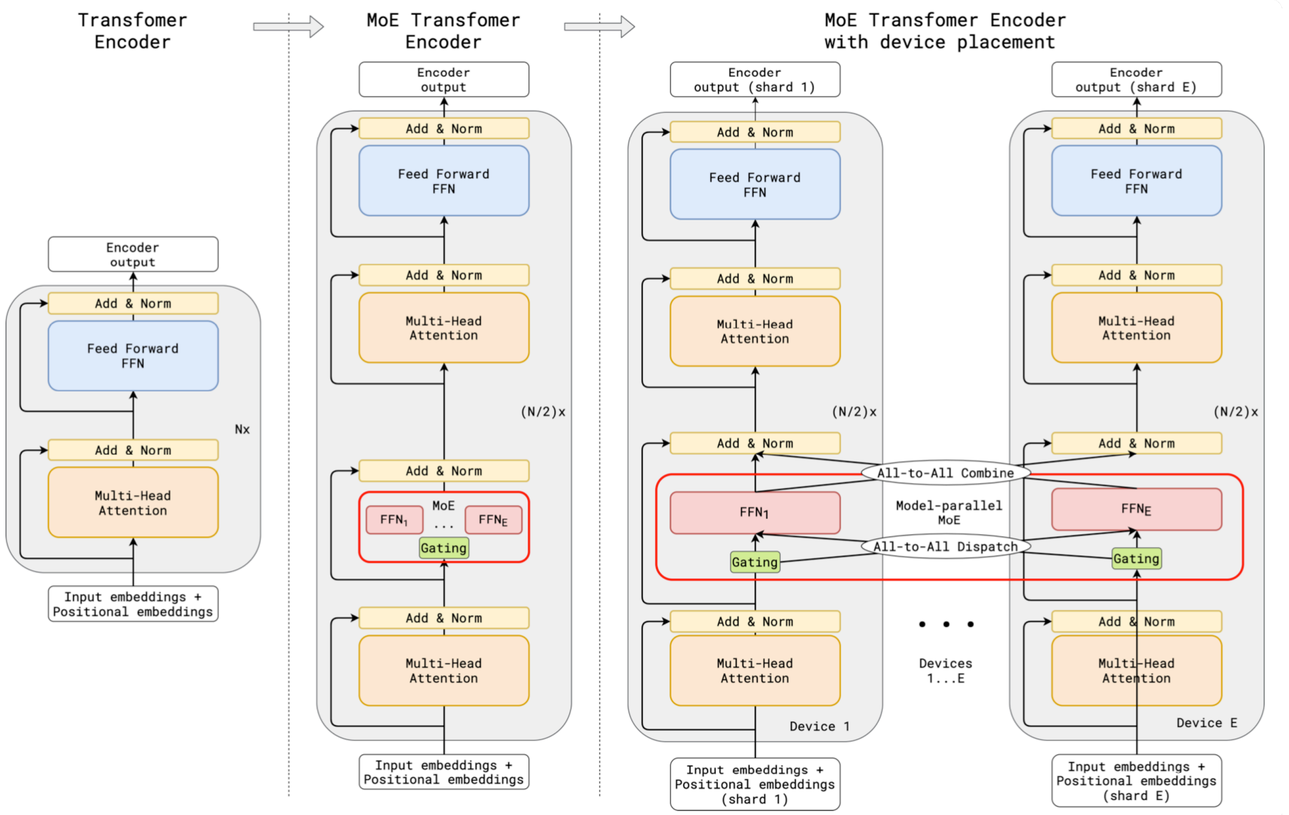
$$KeepTopK(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v. \\ -\infty & \text{otherwise.} \end{cases}$$

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$

Unimodal Sparsely-activated Model

Text Models

GShard

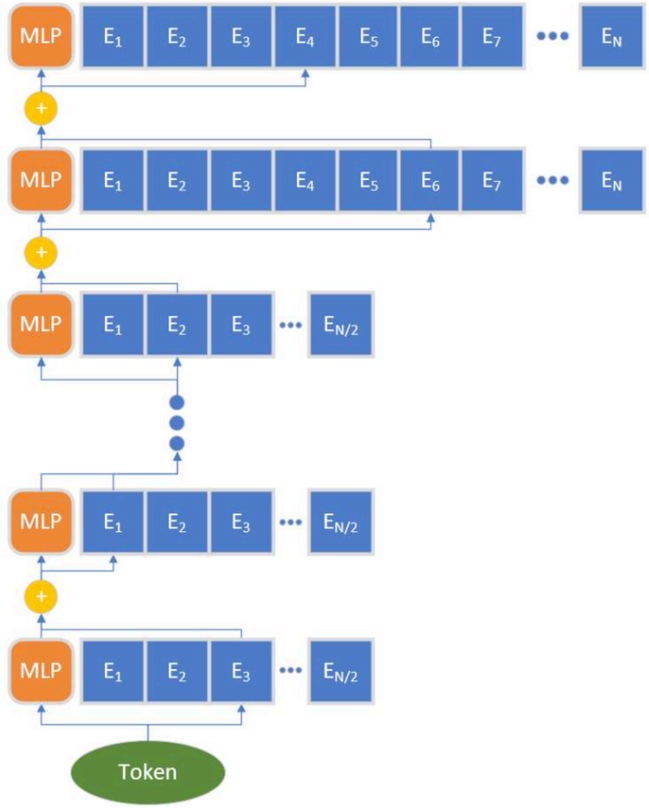


$$\mathcal{G}_{s,E} = \text{GATE}(x_s)$$

$$\text{FFN}_e(x_s) = wo_e \cdot \text{ReLU}(wi_e \cdot x_s)$$

$$y_s = \sum_{e=1}^{E} \mathcal{G}_{s,e} \cdot \text{FFN}_e(x_s)$$

PR-MoE

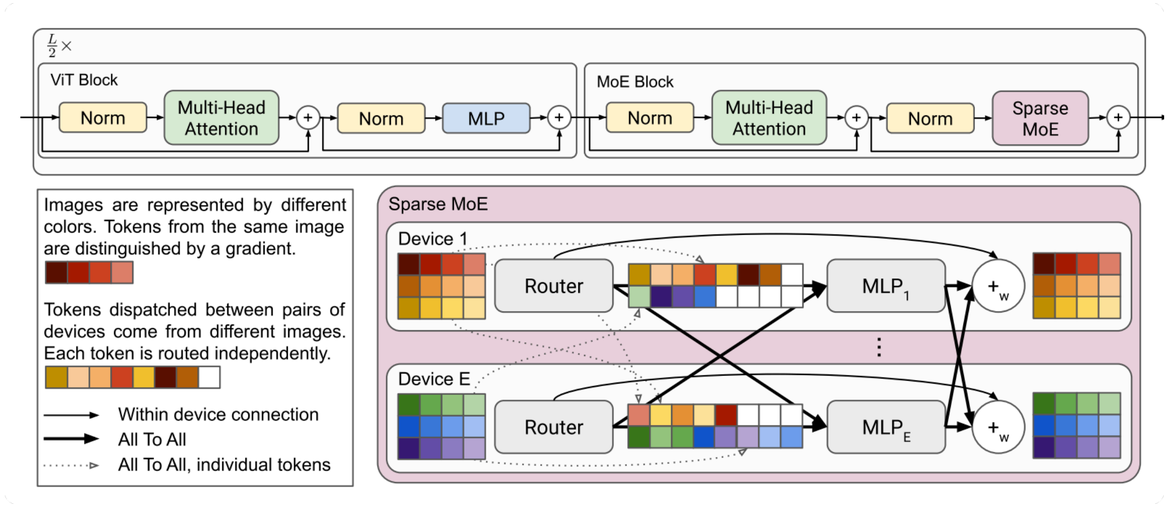


Pyramid-MoE + Residual-MoE

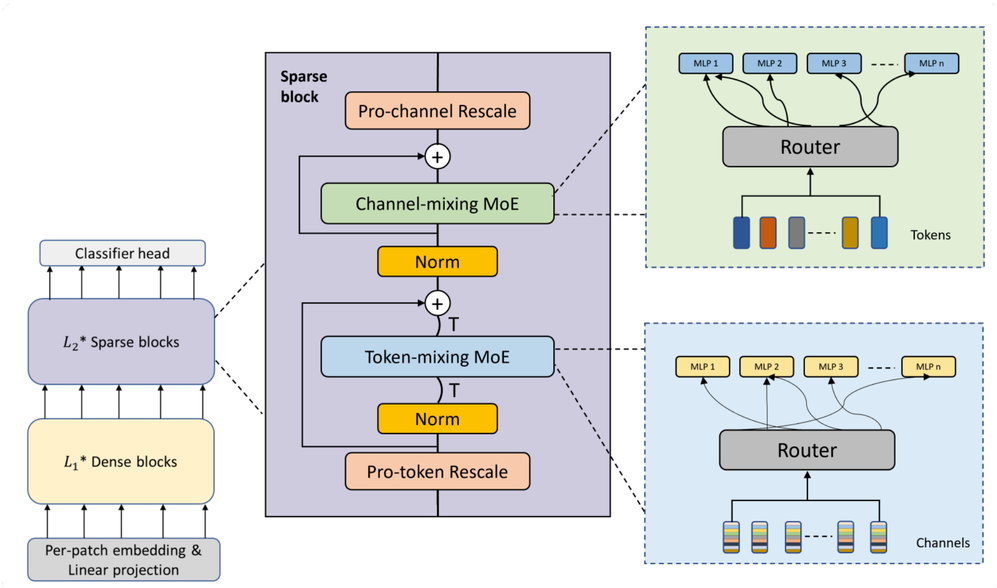
Unimodal Sparsely-activated model

Vision models

Vision MoE (V-MoE)



Sparse MLP



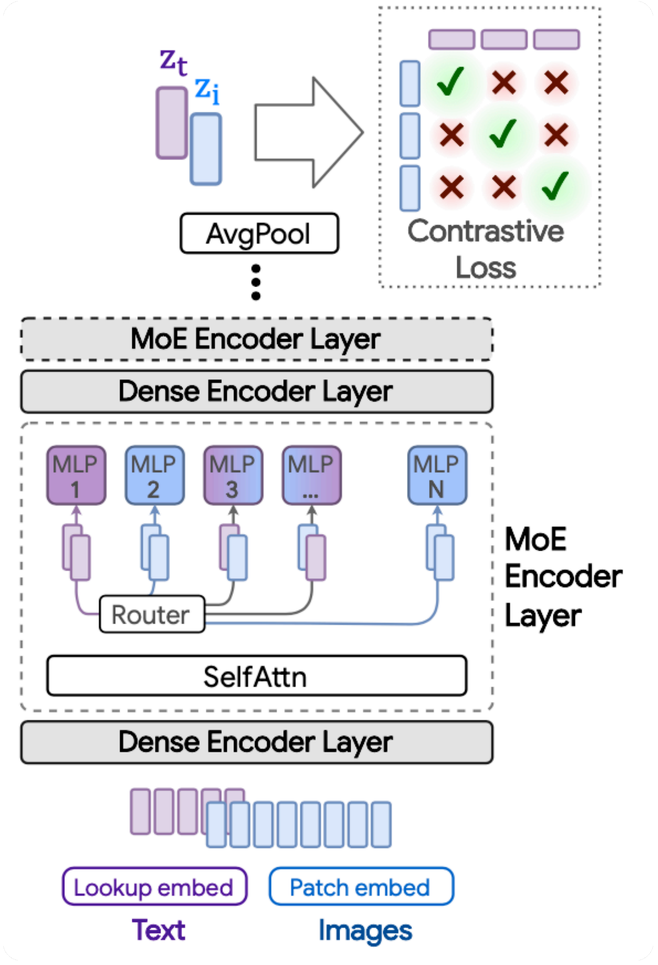
Multimodal Sparsely-activated model

Framework

Multimodal Sparsely-activated model

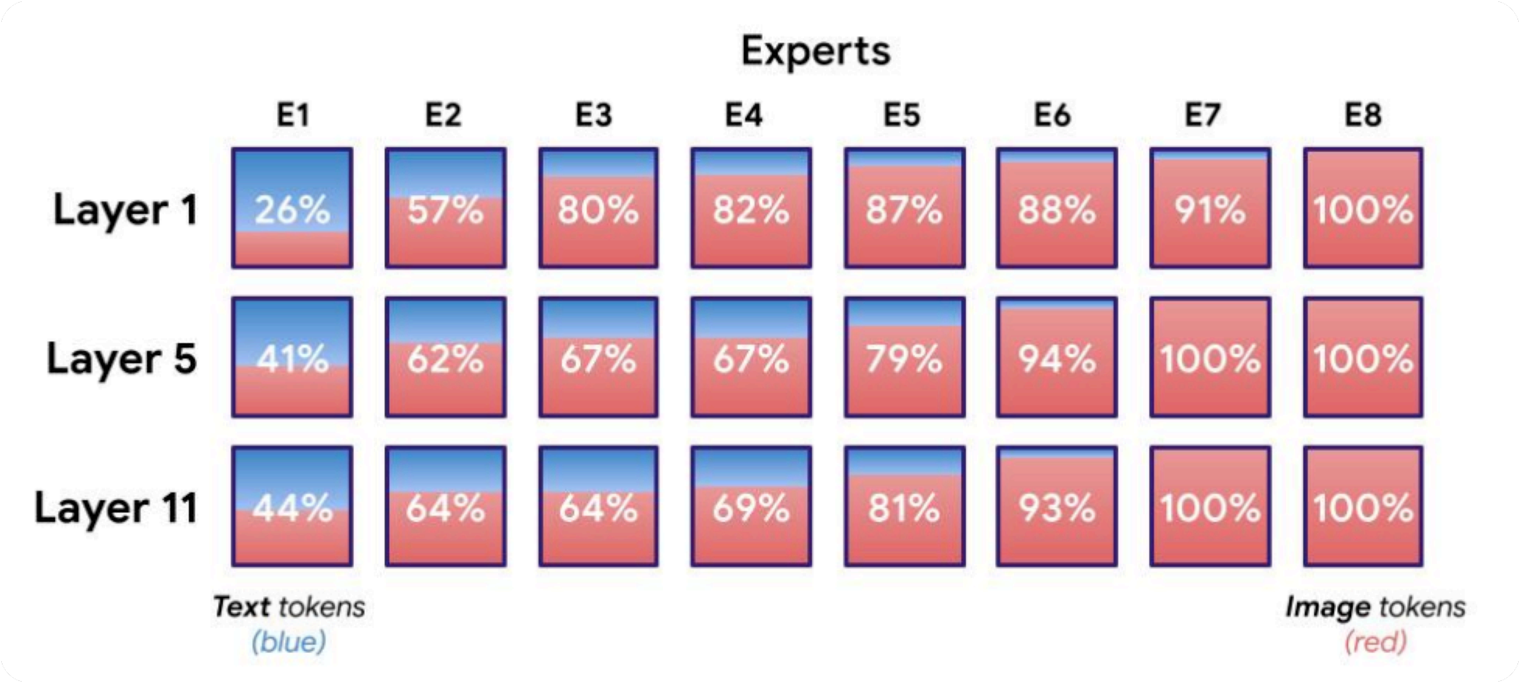
LIMoE: What's new?

For input, we should project the different modalities to the desired width.



Multimodal Sparsely-activated model

LIMoE: What's new?



In the training setup, there are many more image tokens than text tokens, so all experts tend to process at least some images.

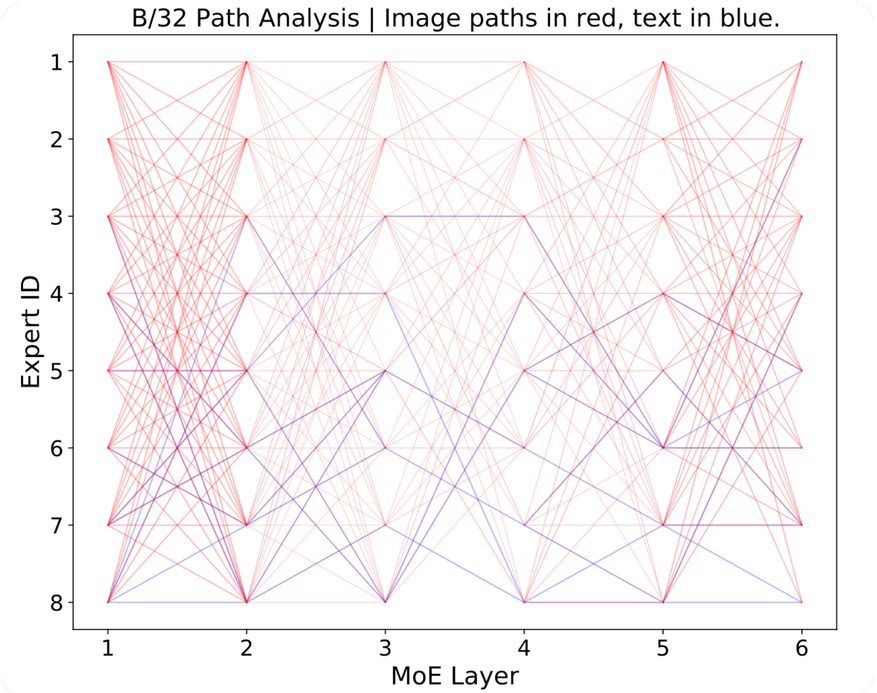
Multimodal Sparsely-activated model

New challenges —Load Balance Problem

New load balance problem:

Due to the shortcomings of existing algorithms, perfectly balancing image tokens but dropping all text tokens can also lead to a small auxiliary loss.

So for multimodal MoE, we need something more than the classic auxiliary losses.



Multimodal Sparsely-activated model

New challenges — Loss function

Local Entropy Loss

$$\Omega_{\text{local}}(\mathbf{G}_m) := \frac{1}{n_m} \sum_{i=1}^{n_m} \mathcal{H}(p_m(\texttt{experts}|\mathbf{x}_i))$$

Motivation:

Centralizing the router's preference

Function:

Local Entropy encourages **concentrated** router weights, but at the expense of the ***diversity*** of the text experts.

Global Entropy loss

$$\Omega_{\text{global}}(\mathbf{G}_m) := -\mathcal{H}(\tilde{p}_m(\texttt{experts}))$$

Motivation:

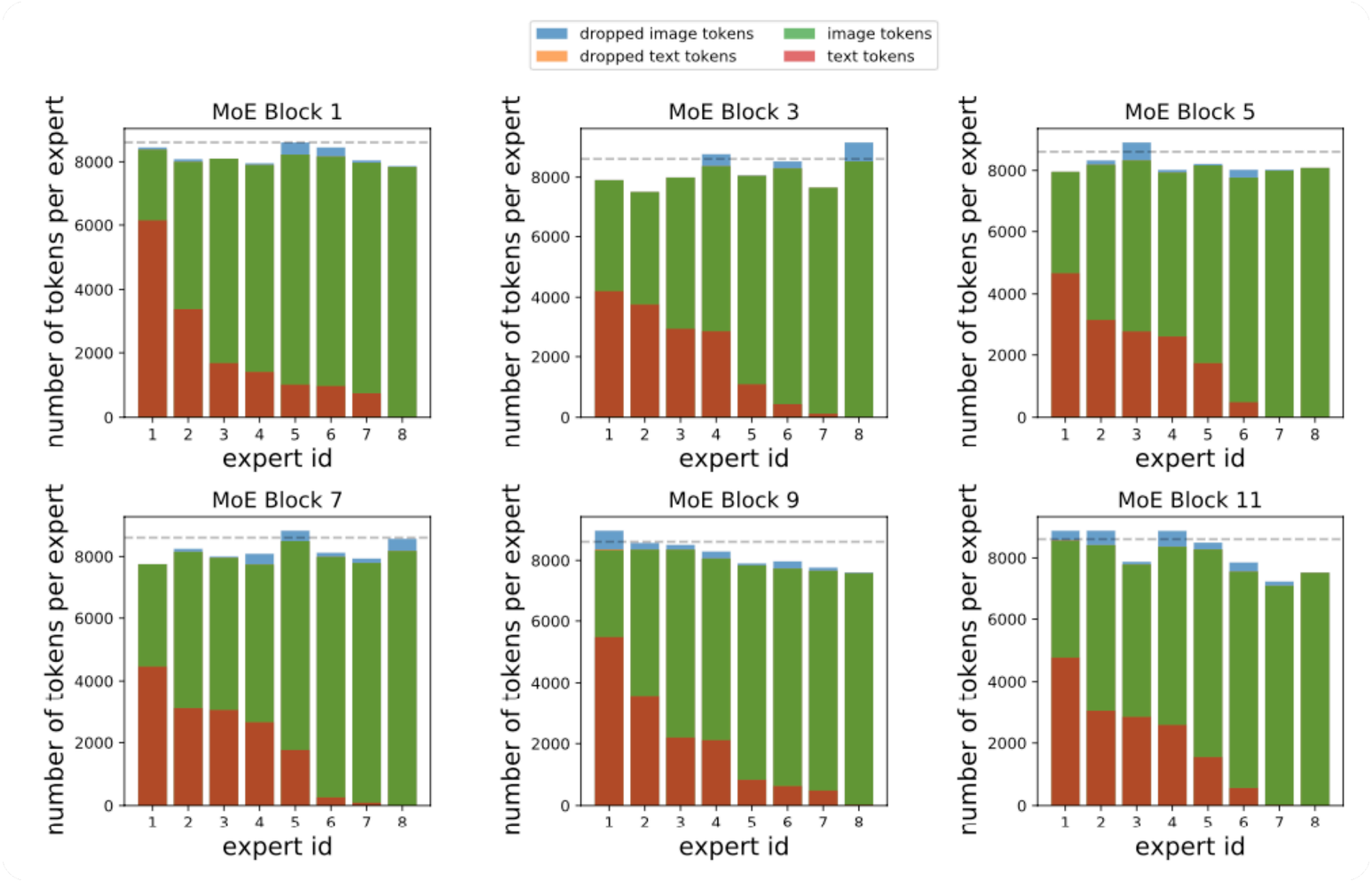
Don't use single expert

Function:

Global Entropy encourages maximization of the marginal entropy, a more **uniform** expert distribution.

Multimodal Sparsely-activated model

Loss function — Better performance and success rate



What’s more

A new paradigm or just more papers?

- Dynamically increase the number of experts when supplementing data?
- How to choose expert initialization parameters for more stable training? (Lottery Ticket Hypothesis)
- Previous aux losses are too complicated
- How to deal with the problem of multimodal data imbalance? A heuristic expert training strategy?
- Try to implement high-level multimodal information for controllable MoE training?

Reference

1. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts
2. Cross-token Modeling with Conditional Computation
3. Scaling Vision with Sparse Mixture of Experts
4. LIMoE: Learning Multiple Modalities with One Sparse Mixture-of-Experts Model (Google AI Blog)
5. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale
6. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding
7. Mixture of experts: A literature survey