Mixture of Diverse Size Experts

Manxi Sun, Wei Liu, Jian Luan, Pengzhi Gao, and Bin Wang

Xiaomi AI Lab, Beijing, China

TL;DR

- In this paper, we propose the **Mixture of Diverse Size Experts (MoDSE)**, a new MoE architecture with layers designed to have **experts of different sizes**. Our analysis of difficult token generation tasks shows that experts of various sizes achieve better predictions, and the routing path of the experts tends to be stable after a training period.
- To tackle the uneven workload distribution from diverse sizes experts, we introduce **an expert-pair allocation strategy** to distribute the workload across multiple GPUs evenly.
- Comprehensive evaluations across multiple benchmarks demonstrate the effectiveness of MoDSE, as it **outperforms existing MoEs** by allocating the parameter budget to experts adaptively while **maintaining the same total parameter size** and **inference speed**.

Model

- **Diverse Size Experts:**

  We denote the designed Diverse Size Experts as $\{\hat{E}_1(\cdot), \cdots, \hat{E}_N(\cdot)\}$, and the dimension of the hidden layer for $\hat{E}_i(\cdot)$ is $\hat{h}_i$, h is the dimension of the hidden layer in conventional MoE structure.

  To maintain the overall parameter size, the experts are grouped into pairs (i_k^1, i_k^2), where $k \in 1 \cdots n$ indicates the pair of the experts. The average value of $\hat{h}_i$ within each pair equals h, with one expert being larger than the average size and the other smaller. Typically, the number of experts is even, ensuring the experts can be grouped into pairs, thus the total parameter size of the MoDSE model matches that of the vanilla MoE model.

$$\hat{y} = \sum_{i=1}^{N} \hat{G}_i(x)\hat{E}_i(x) \tag{1}$$

$$(i_1^1, i_1^2), \cdots, (i_n^1, i_n^2), \text{ with } n = \frac{N}{2} \tag{2}$$

$$\hat{h}_{i_k^1} + \hat{h}_{i_k^2} = 2 \times h, \text{, with } k \in 1 \cdots n \tag{3}$$

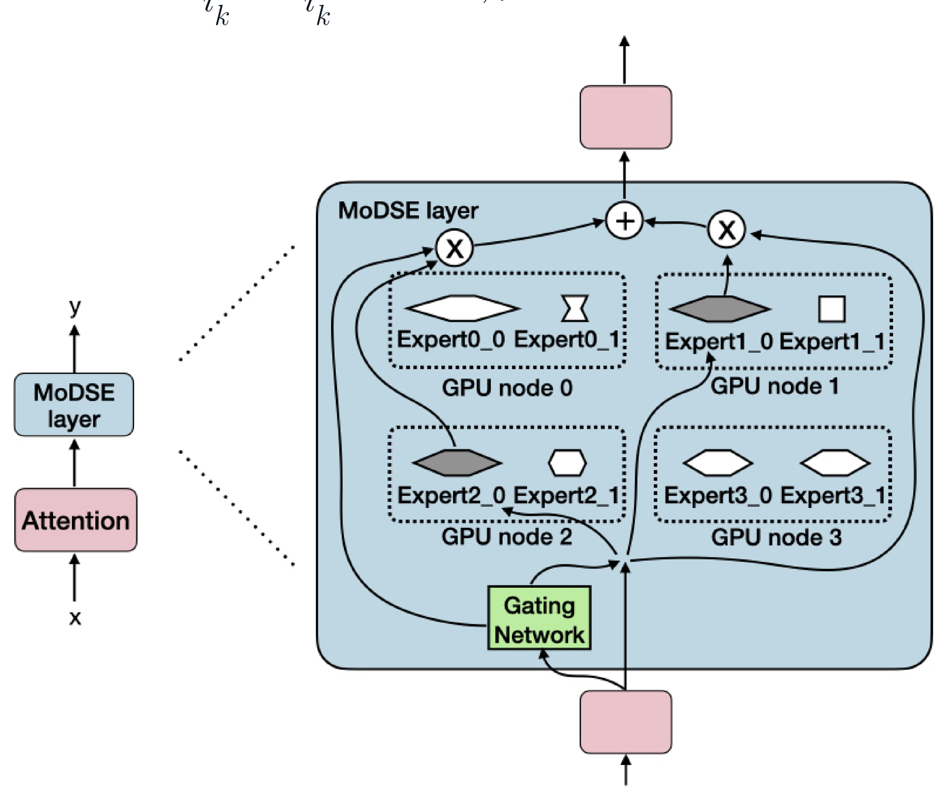

Fig. 1: Overview of a MoDSE layer with different sizes of experts.

- **Load Balance Consideration:**

  In MoDSE, experts with hidden layer sizes larger than the average have a higher workload due to increased parameters, both during training and inference phrases. To address this load imbalance problem, we propose the **expert-pair allocation** strategy, which places each pair of experts on the same GPU and ensures that each GPU contains an equal number of parameters.

Analysis on Token Routing

- The ratio between the largest and the smallest number of tokens routed to the experts in the baseline model ranges from 1.2 to 3.0. The statistics for the MoDSE setting show a non-uniform distribution, with ratios larger than 3.0 appearing, particularly in the first 2 layers of the model and for the experts with the second largest probability.
- But in the last epoch, only one ratio remains larger than 3.0, with the others ranging from 1.5 to 3.0, indicating that the token distribution among experts becomes more balanced by the end of the training.
- As shown in Figure 2(c, d), it is notable that the experts chosen by the most tokens are not always the ones with larger sizes. Conversely, experts with larger sizes can sometimes be the least visited by the tokens.

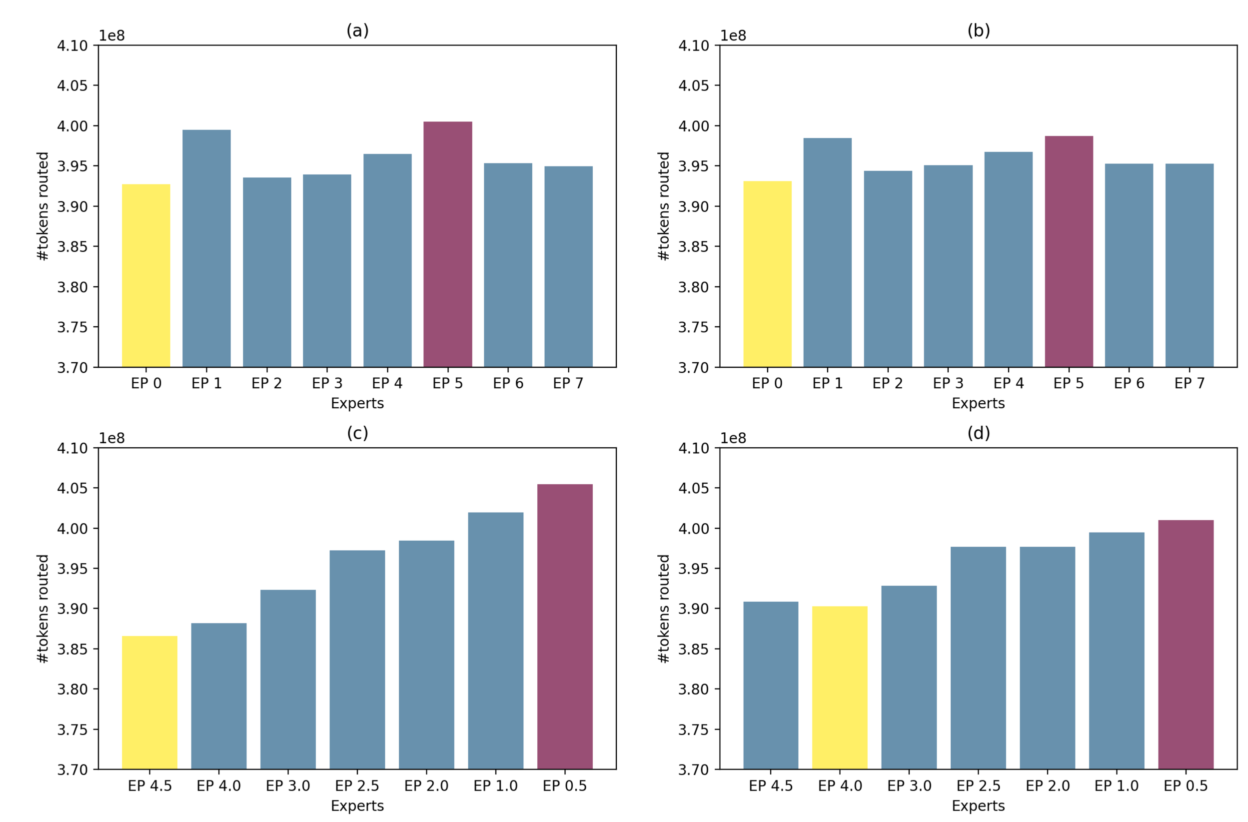

Fig. 2: The number of tokens routed to each expert. The bar is the sum of the number across the layers. Figure (a) shows results in Baseline in epoch 2, and (b) in the last epoch. Figure (c) shows results in MoDSE in epoch 2, and (d) in the last epoch. The purple bar indicates the most routed expert, and the yellow indicates the least.

loss threshold	avg. loss red.	#tokens
2.0	0.58	180
1.8	0.46	222
1.6	0.36	337
1.4	0.32	730
1.2	0.22	1991
1.05	0.18	3633

Table 3. Average CE loss reduction across different intervals. The higher the initial CE loss, the more significant the improvement demonstrated by the MoDSE model. The avg. loss red. stands for the average CE loss decrease from baseline to MoDSE.

Difficult Tokens Routing Distribution

- To identify which experts handle the difficult tokens, further analysis is conducted on the 180 tokens with a CE loss greater than 2.0 in the baseline setting.
- For these difficult tokens, as shown in Figure 4, more tokens choose the larger experts, while fewer tokens select the smaller experts. This phenomenon is even more pronounced when only considering the top one expert. More than twice as many tokens (6215) chose the larger experts compared to the smaller ones (3085).
- This result indicates that the larger experts, with capabilities to handle tokens with more difficult prediction tasks, are more frequently chosen by tokens facing more challenging next-token generation tasks.

Results

- **Evaluations & Decoding Efficiency:**

Benchmark	MoE	MoDSE	MoE	MoDSE
AGIEval [Acc., 5 shots, 615]	26.2	**28.1**	48s	59s
MMLU [Acc., 5 shots, 2341]	26.5	**29.9**	3min 26s	3min 27s
INTENT [Acc., 5 shots, 741]	13.6	**16.5**	1min 31s	1min 34s
GSM8K [EM, 8 shots, 100]	5.9	**7.7**	20min 26s	20min 43s
LAMBADA [EM, 5 shots, 100]	36.8	**38.9**	40min44s	40min48
MATH [EM, 5 shots, 100]	0.8	**2.6**	21min 21s	21min 34s
TriviaQA [EM, 5 shots, 100]	5.2	**8.3**	46min 53s	48min 55s
PIQA [EM, 5 shots, 100]	53.1	**57.6**	44min56s	43min34s
SIQA [EM, 5 shots, 100]	42.9	**60.9**	2min35s	2min36s

The second and third columns compare the MoE baseline and MoDSE on a size of $700M \times 8$. With the same amount of parameters, MoDSE achieves better performance than the baseline.

The fourth and fifth columns show the inference duration of the baseline and MoDSE models on downstream tasks.

- **Training Convergence:**

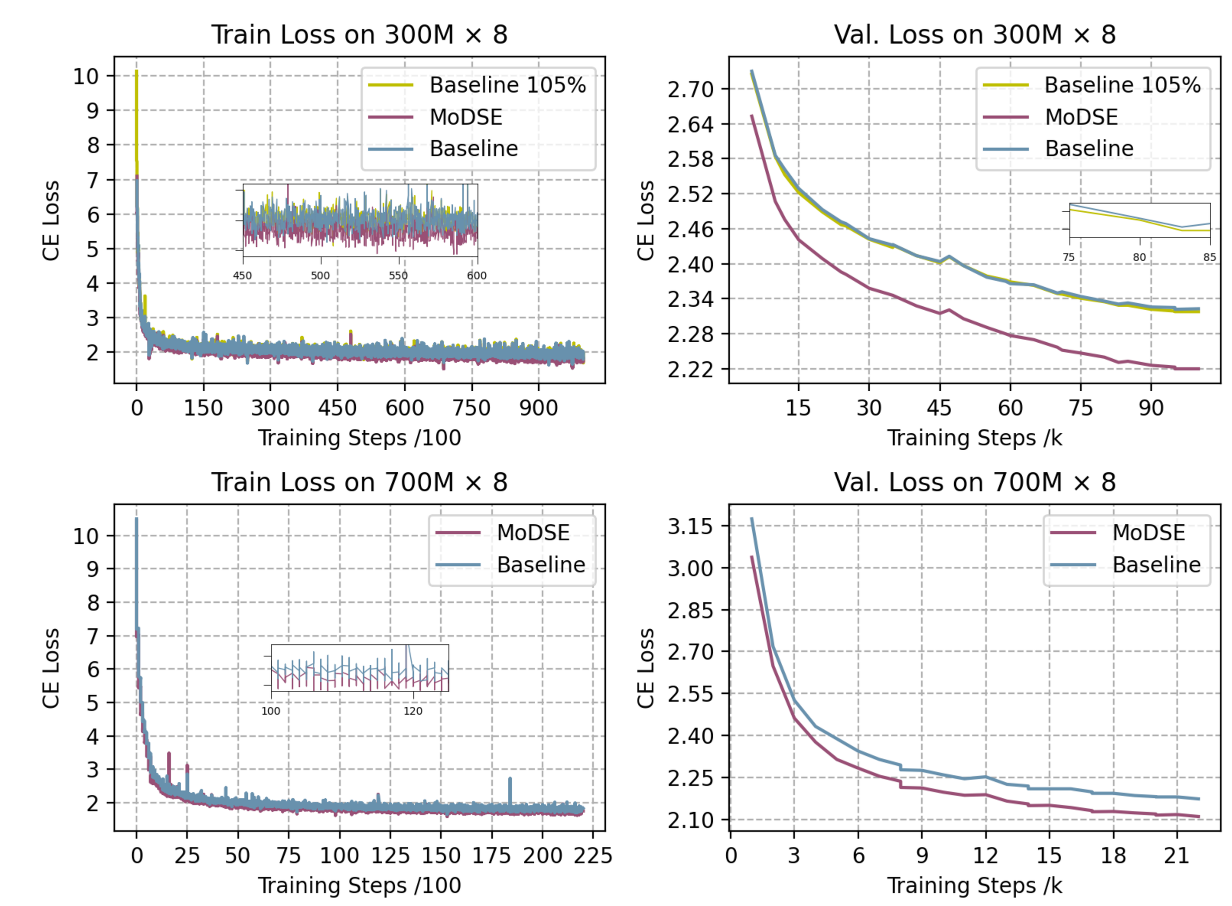

Fig. 3: Training and validation loss curves, with cross-entropy loss values indicated on the curves.

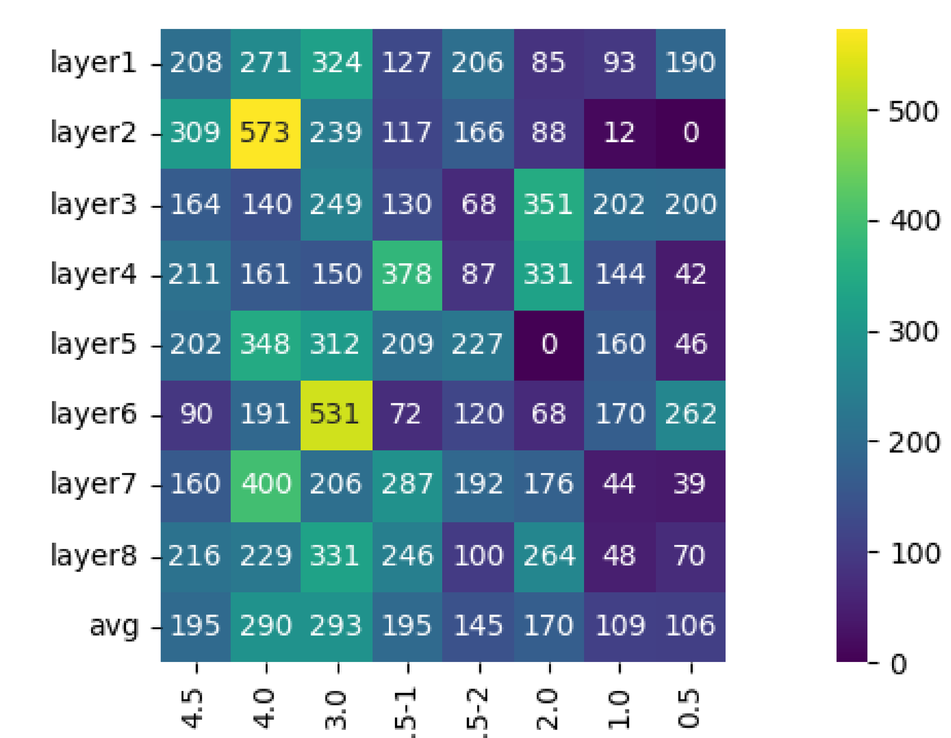

Fig. 4: The top one expert choice of difficult tokens across eight layers. More tokens are routed to larger experts, distributed on the left half of the heat map.