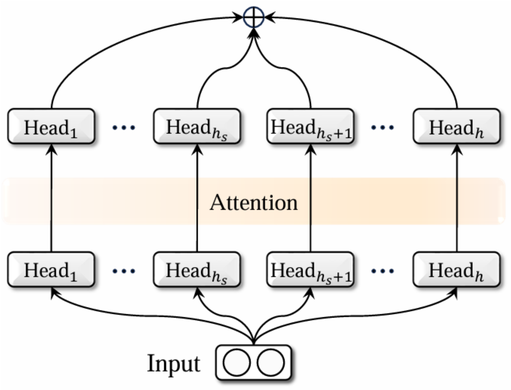
MoH: Multi-Head Attention as Mixture-of-Head Attention

Peng Jin, Bo Zhu, Li Yuan, Shuicheng Yan

Multi-Head Attention



(a) Multi-Head Attention

- The multi-head attention mechanism is typically represented in its concatenation form:

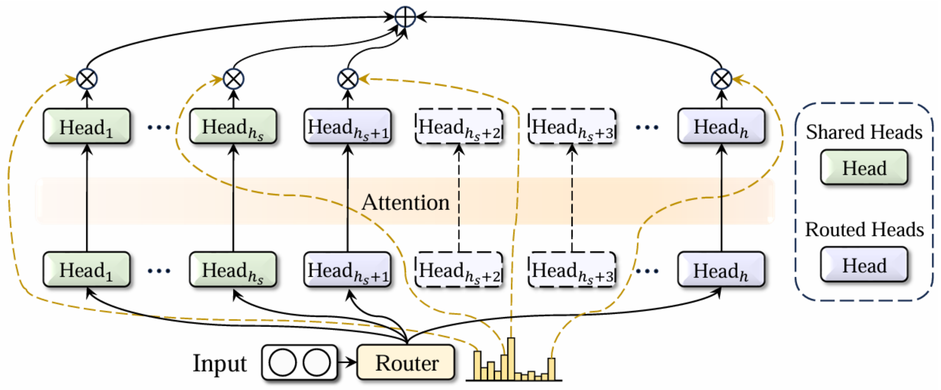
$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{X}') = \text{Concat}(\boldsymbol{H}^1, \boldsymbol{H}^2, ..., \boldsymbol{H}^h)\boldsymbol{W}_O,$$
$$\boldsymbol{H}^i = \text{Attention}(\boldsymbol{X}\boldsymbol{W}_Q^i, \boldsymbol{X}'\boldsymbol{W}_K^i, \boldsymbol{X}'\boldsymbol{W}_V^i),$$

- However, from another perspective, if we decompose $W_O \in R^{d_v \times d_{out}}$ by rows, we can express multi-head attention in a summation form:

$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{X}') = \sum_{i=1}^{h} \boldsymbol{H}^i \boldsymbol{W}_O^i$$

- **Each attention head operates in parallel, and the final output is the sum of all attention heads.**

Multi-Head Attention



(b) Our proposed Mixture-of-Head Attention

- Inspired by the great success of MoE, we propose Mixture-of-Head attention (MoH), which treats attention heads as experts:

$$\text{MoH}(\boldsymbol{X}, \boldsymbol{X}') = \sum_{i=1}^{h} g_i \boldsymbol{H}^i \boldsymbol{W}_O^i, \qquad [\alpha_1, \alpha_2] = \text{Softmax}(\boldsymbol{W}_h \boldsymbol{x}_t),$$

$$g_i = \begin{cases} \alpha_1 \text{Softmax}(\boldsymbol{W}_s \boldsymbol{x}_t)_i, & \text{if } 1 \le i \le h_s, \\ \alpha_2 \text{Softmax}(\boldsymbol{W}_r \boldsymbol{x}_t)_{i-h_s}, & \text{if Head } i \text{ is activated}, \\ 0, & \text{otherwise}, \end{cases}$$

- MoH has two significant advantages:
  - First, MoH **enables each token to select the appropriate attention heads**, enhancing inference efficiency without compromising accuracy or increasing the number of parameters.
  - Second, MoH **replaces the standard summation in multi-head attention with a weighted summation**, introducing flexibility to the attention mechanism and unlocking extra performance potential.

Main Results

Methods	#Params (M)	#Activated Heads (%)	Acc (%)
Focal-B (Yang et al., 2021)	90	100	83.8
FocalNet-B (Yang et al., 2022b)	89	100	83.9
CoAtNet-2 (Dai et al., 2021)	75	100	84.1
MViTv2-B (Li et al., 2022)	52	100	84.4
MOAT-2 (Yang et al., 2022a)	73	100	84.7
iFormer-L (Si et al., 2022)	87	100	84.8
TransNeXt-B (Shi, 2024)	90	100	84.8
MoH-ViT-B	90	75	**84.9**
MoH-ViT-B	90	50	84.7

Resources

Paper

Code

Methods	#Params (M)	#Activated Heads (%)	FID↓	sFID↓	IS↑	Precision↑	Recall↑
DiT-S/2 400K (Peebles & Xie, 2023)	33	100	68.40	-	-	-	-
MoH-DiT-S/2 400K	33	90	**67.25**	**12.15**	**20.52**	**0.37**	**0.58**
MoH-DiT-S/2 400K	33	75	69.42	12.85	19.96	0.36	0.55
DiT-B/2 400K (Peebles & Xie, 2023)	130	100	43.47	-	-	-	-
MoH-DiT-B/2 400K	131	90	**43.40**	**8.40**	**33.51**	**0.49**	**0.63**
MoH-DiT-B/2 400K	131	75	43.61	8.48	33.43	**0.49**	0.62
DiT-L/2 400K (Peebles & Xie, 2023)	458	100	23.33	-	-	-	-
MoH-DiT-L/2 400K	459	90	**23.17**	**6.16**	**58.92**	**0.61**	**0.63**
MoH-DiT-L/2 400K	459	75	24.29	6.38	57.75	0.60	**0.63**
DiT-XL/2 7,000K (Peebles & Xie, 2023)	675	100	9.62	6.85	121.50	0.67	**0.67**
DiT-XL/2 7,000K (cfg=1.25)	675	100	3.22	5.28	201.77	0.76	0.62
MoH-DiT-XL/2 2,000K	676	75	10.95	6.19	106.69	0.67	0.66
MoH-DiT-XL/2 2,000K	676	90	10.67	6.15	107.80	0.67	0.65
MoH-DiT-XL/2 7,000K	676	90	8.56	6.61	129.54	0.68	**0.67**
MoH-DiT-XL/2 7,000K (cfg=1.25)	676	90	**2.94**	**5.17**	**207.25**	**0.77**	0.63

Methods	#Activated Heads (%)	MMLU (5)	CEVAL (5)	CMMLU (5)	GSM8K(8)	TruthfulQA
LLaMA3-8B (Dubey et al., 2024)	100	65.2	52.3	50.7	49.5	35.4
MoH-LLaMA3-8B	75	65.8	61.5	64.4	56.9	44.0

Methods	#Activated Heads (%)	MMLU (5)	CEVAL (5)	CMMLU (5)	GSM8K(8)	TruthfulQA
LLaMA3-8B (Dubey et al., 2024)	100	81.9	30.0	83.9	75.5	94.0
MoH-LLaMA3-8B	75	80.1	30.3	84.0	76.4	92.2

Methods	#Activated Heads (%)	MMLU (5)	CEVAL (5)	CMMLU (5)	GSM8K(8)	TruthfulQA
LLaMA3-8B (Dubey et al., 2024)	100	81.0	72.5	31.5	59.0	61.6
MoH-LLaMA3-8B	75	78.8	72.9	28.3	60.1	**64.0**