Solving Token Gradient Conflict in Mixture-of-Experts for Large Vision-Language Model

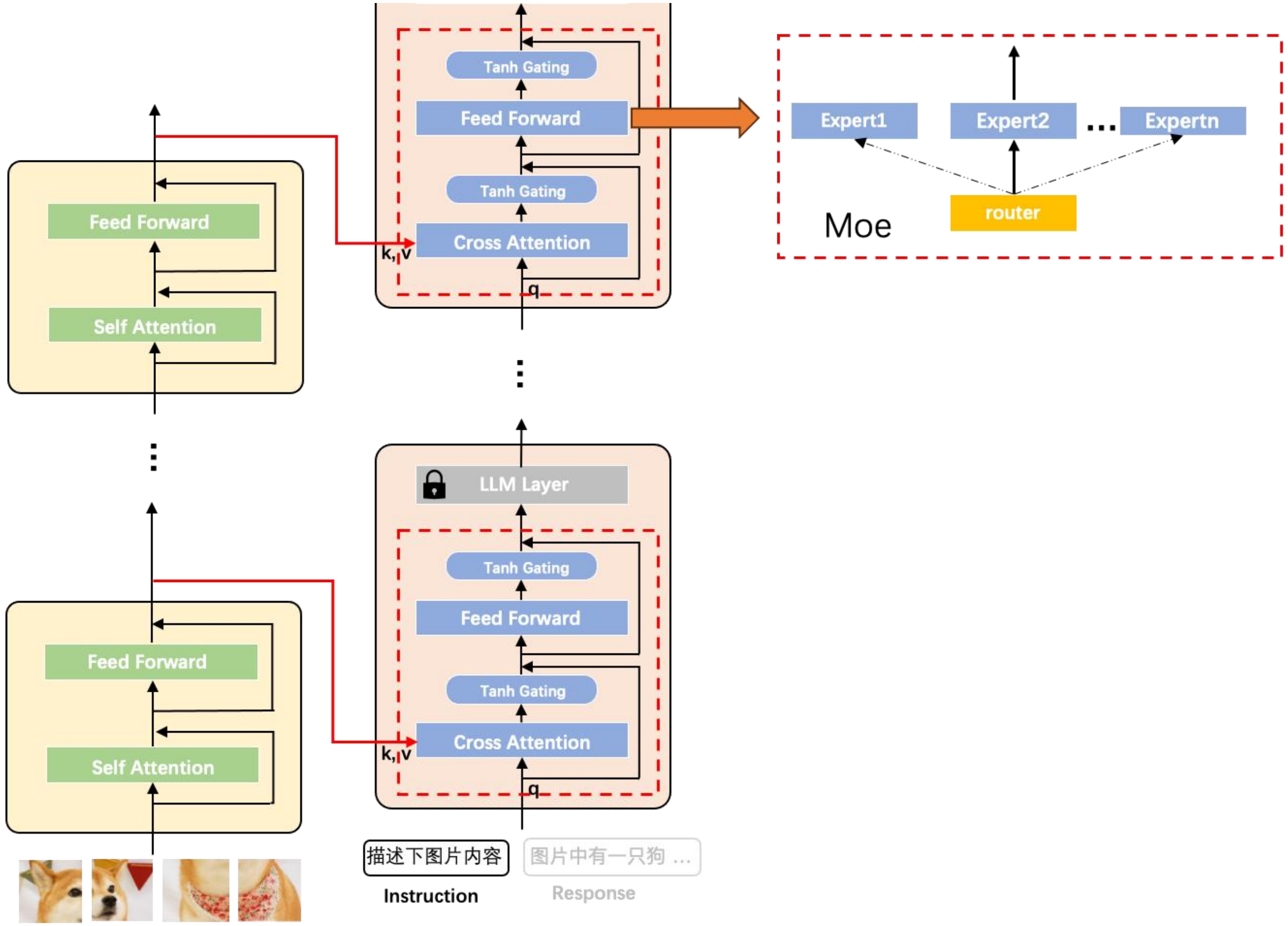
ICLR 2025

Codes will be released in https://github.com/longrongyang/STGC

Mixture-of-Experts (MoE)

ICLR

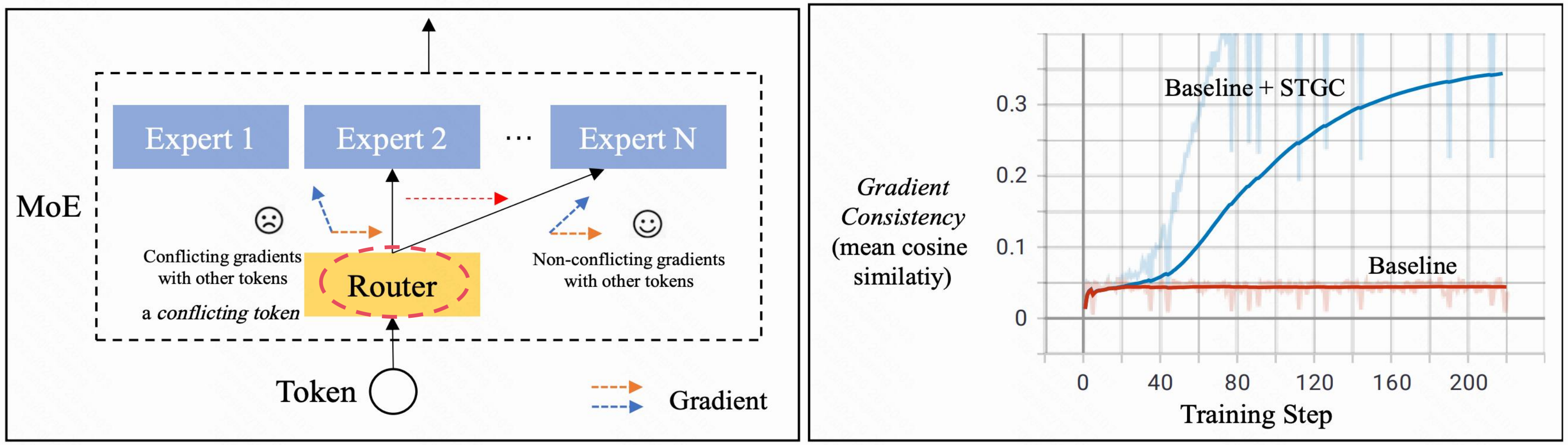
- The MoE system replaces the FFN layer in LLM with multiple experts
- The router predicts the probability of each token dispatched to different experts. Tokens are then dispatched to the experts with the Top-*k* predicted probability



浙江大學 ZHEJIANG UNIVERSITY

The routing of tokens

- A critical goal of token routing is to reduce interference between diverse data
- *How to define conflicting tokens?*
  - Related LoRA-MoE studies: Sample-level instruction features or embeddings
  - These techniques suffer from optimization interference risk and token-level interference within a sample
  - This study models data interference through the lens of token-level gradients
- *How to solve conflicting tokens?*
  - We propose the STGC: A novel loss to move *conflicting tokens* to other experts to reduce conflicts



(a) Our goal: reduce gradient conflicts of tokens within an expert

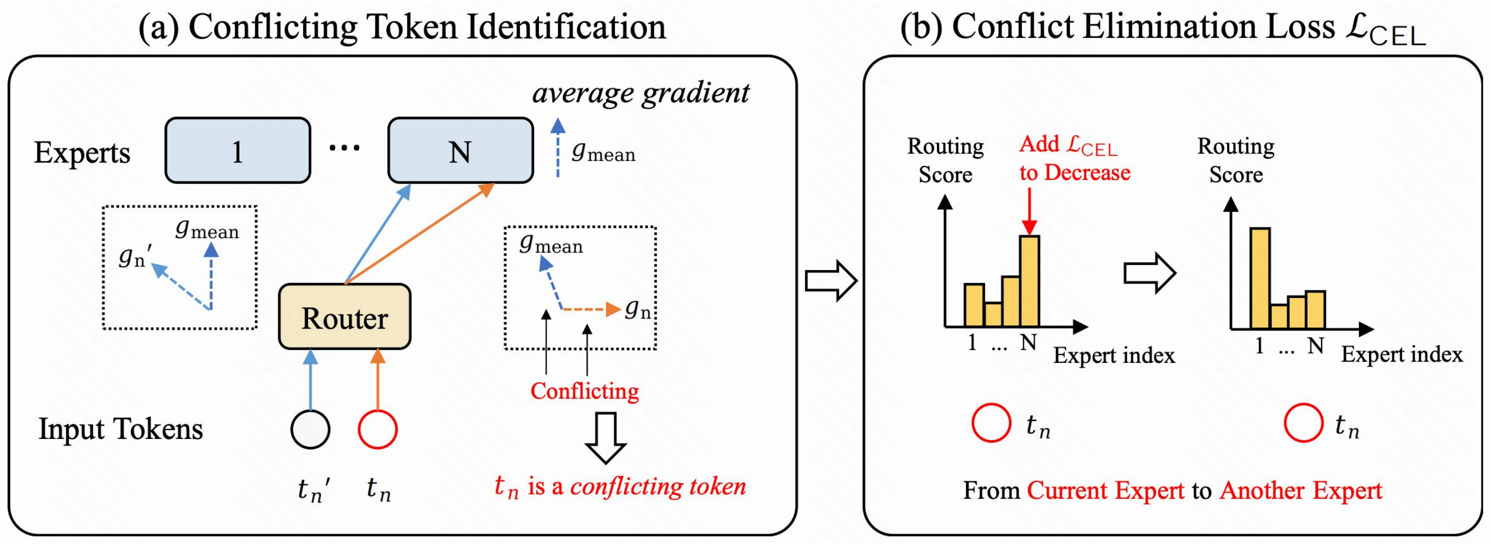
(b) *Gradient consistency* of tokens within an expert before and after using STGC

ZHEJIANG UNIVERSITY

ICLR

Eliminate conflicts

- The goal is to reduce the interference between different data: Use token-level gradients to depict the relationships (*conflict or no conflict*) between data within an expert.



Conflict elimination loss

$$z'_{\text{moe}}(t_n) = -z_{\text{moe}}(t_n),$$

$$p'_{\text{moe}}(t_n)_i = \frac{e^{z'_{\text{moe}}(t_n)_i}}{\sum_{j=1}^{E} e^{z'_{\text{moe}}(t_n)_j}},$$

$$\mathcal{L}_{\text{CEL}} = \frac{1}{N_{all} \cdot E} \sum_{n=1}^{N_{all}} \sum_{i=1}^{E} \log(p'_{\text{moe}}(t_n)_i) \cdot q_{\text{moe}}(t_n)_i,$$

Encourage the decrease of scores

- Conflicting token identification
  - Use the token-level gradients within each expert to identify "*conflicting tokens*"
- Conflict elimination loss
  - Add a novel loss to optimize token routing, and move the "*conflicting tokens*" from their current experts to other experts for processing

浙江大學 ZHEJIANG UNIVERSITY

STGC as a plug-in

Method	LLM	Act.	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet	Avg
MoE-LLaVA-4Top1	S-1.6B	1.6B	74.5*	58.6*	25.7	55.8	45.0	85.2	1245.3	56.2	27.2	53.5
+STGC	S-1.6B	1.6B	74.9*	59.4*	27.4	57.5	46.5	85.8	1276.8	56.8	28.5	54.6
MoE-LLaVA-4Top2	S-1.6B	2.0B	76.7*	60.3*	36.2	62.6	50.1	85.7	1318.2	60.2	26.9	57.3
+STGC	S-1.6B	2.0B	76.9*	60.9*	37.7	62.6	50.7	85.9	1355.1	60.7	28.2	58.0
MoE-LLaVA-4Top2	P-2.7B	3.6B	77.6*	61.4*	43.9	68.5	51.4	86.3	1423.0	65.2	34.3	61.1
+STGC	P-2.7B	3.6B	78.0*	62.1*	47.2	68.1	52.3	86.9	1429.2	66.7	33.3	61.8
MoE-LLaVA-4Top2†	P-2.7B	3.6B	79.9*	62.6*	43.7	70.3	57.0	85.7	1431.3	68.0	35.9	62.9
+STGC	P-2.7B	3.6B	80.3*	63.2*	45.1	70.3	57.4	86.1	1447.6	69.7	35.7	63.5

Method	LLM	Act.	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet	AI2D	ChartQA	DocVQA	Avg
Dense Model															
LLaVA-1.5	V-13B	13B	80.0*	63.3*	53.6	71.6	61.3	85.9	1531.3	67.7	35.4	49.6	18.1	24.0	55.5
Sparse Model															
MoE-LLaVA	S-1.6B	2.0B	76.7*	60.3*	36.2	62.6	50.1	85.7	1318.2	60.2	26.9	48.8	15.3	18.4	49.2
MoE-LLaVA	P-2.7B	3.6B	77.6*	61.4*	43.9	68.5	51.4	86.3	1423.0	65.2	34.3	58.8	19.9	21.5	53.5
DYNMOE-LLaVA	P-2.7B	3.4B	77.9*	61.6*	45.1	68.0	51.8	86.0	1429.6	66.6	33.6	-	-	-	-
MoE-LLaVA†	P-2.7B	3.6B	79.9*	62.6*	43.7	70.3	57.0	85.7	1431.3	68.0	35.9	59.5	15.4	25.6	54.9
Our Method†	P-2.7B	3.6B	80.0*	63.0*	48.6	70.9	58.8	86.5	1481.7	71.0	40.7	64.5	44.7	42.1	61.0

- As a plug-in, STGC consistently brings reliable model performance improvements
- During inference, activating 3.6B parameters performs better than a dense model activating 13B parameters

Scalability

Method	LLM	Data	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet	Avg
MoE-LLaVA-4Top2†	P-2.7B	665K	79.9*	62.6*	43.7	70.3	57.0	85.7	1431.3	68.0	35.9	62.9
+STGC	P-2.7B	665K	80.3*	63.2*	45.1	70.3	57.4	86.1	1447.6	69.7	35.7	63.5
MoE-LLaVA-4Top2†	P-2.7B	1021K	79.7*	63.0*	42.7	71.1	56.9	84.3	1439.9	70.4	42.2	63.8
+STGC	P-2.7B	1021K	80.0*	63.0*	48.6	70.9	58.8	86.5	1481.7	71.0	40.7	64.9

More Training Data

	COLA	MRPC	QNLI	MNLI	RTE	Avg
MoE-8Top2 (Guo et al., 2024)	64.5	90.2	92.4	86.7	74.9	81.7
DYNMOE (Guo et al., 2024)	65.2	90.6	92.6	86.4	73.4	81.6
MoE-8Top2*	64.5	90.0	93.4	86.9	72.9	81.5
+STGC	66.8	91.2	93.8	87.6	74.7	82.8

Language Tasks

- The larger the training data size, the more significant the performance gain brought by STGC
- STGC can bring more obvious performance gains on large language models