CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering

Supplementary Materials

This supplementary material provides more details to support the main paper, including the implementation of our proposed method, the dataset and baseline details, complete experimental results, and visualization analysis to support its effectiveness.

1. Implementation Details

The structure of our complete model is shown in Figure 1. The input includes X_v and X_q, where X_v represents the image and X_q represents the question and instruction. We employ the pre-trained CLIP visual encoder ViT/14 [6], capable of extracting image features $F_v = g(X_v)$. In our experiments, we consider grid features from both before and after the final layer of the Transformer. We apply a trainable projection matrix W to transform F_v into the language embedding tokens E_v, which have the same dimension as the word embedding space in the LLM. We employ Vicuna-7B [4] as LLM, we froze its FFN parameters and used a trainable CL-MoE for continual fine-tuning. We divide the VQA v2 dataset into 10 sub-tasks: recognition, location, judge, commonsense, count, action, color, type, subcategory and causal. The detailed information is shown in Table 1, and we list examples of questions and answers for each task.

2. Details of the Baselines

In the continual tuning MLLM experiments, we re-implement several representative continuous learning methods, including regularization-based methods EWC [5] and MAS [1], and rehearsal-based methods ER [3], DER [2], VS [7], and VQACL [8]. For fairness, we conduct experiments on both VL-T5 and LLaVA, with all implementations based on official code.

- **EWC** [5] identifies some crucial parameters in the neural network and by reducing changes to these parameters during continued training, the memory of past skills or knowledge is preserved. Fisher information is used to measure the importance of each parameter to the training set data. Once the importance of each parameter to the mastered tasks is measured, the EWC adds a parameter shift regularization term to the loss function according to the importance level.

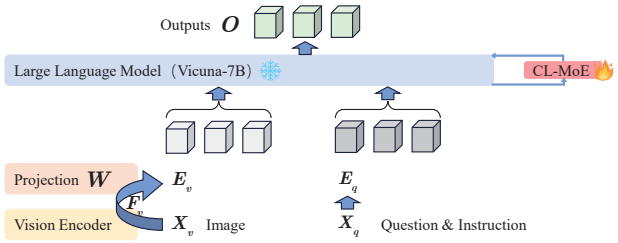

Figure 1. The complete structure and workflow of the model.

- **MAS** [1] calculates the importance of neural network parameters in an unsupervised and online manner. Given a new sample input to the network, MAS accumulates an importance measure for each parameter of the network based on the sensitivity of the predicted output function to changes in that parameter. When learning a new task, changes to important parameters can be penalized, effectively preventing the overwriting of crucial knowledge related to previous tasks.
- **ER** [3] is a rehearsal-based method that stores a subset of visited samples in a fixed-size memory. During subsequent training steps, it randomly selects these stored samples for retraining to avoid forgetting.
- **DER** [2] combines Experience Replay and Knowledge Distillation methods by sampling the network's logits along the optimization trajectory to ensure the model's consistency with its past performance during training. It uses a replay buffer to store past training samples and their corresponding logits, rather than just storing labels. Reservoir sampling is employed to avoid reliance on task boundaries, thus handling situations where task boundaries are blurred.
- **VS** [7] introduces a new cross-task feature embedding consistency constraint to ensure compatibility between features learned currently and those from a previous image library. A metric-based knowledge distillation method is developed to achieve consolidation of old knowledge and feature consistency by bringing embeddings from different feature spaces closer. In addition, VS proposes a new CL scenario, simulating real-world visual search application systems, including previously disjoint and blurry settings, and handling situations where new

Tasks	Train	Test	Examples
Recognition	131,478	5,628	Q: What is this photo taken looking through? A: net.
Location	12,580	611	Q: Where is the giraffe? A: near tree.
Judge	160,179	7,194	Q: Is this man a professional baseball player? A: yes.
Commonsense	25,211	1,100	Q: Has the sheep recently been shaved? A: no.
Count	62,156	2,658	Q: How many tattoos can be seen on this man's body? A: 1.
Action	33,633	1,373	Q: What is the person doing? A: skiing.
Color	50,872	2,196	Q: What color is the players' shirt? A: orange.
Type	23,932	1,089	Q: What type of ice cream is on the plate? A: vanilla.
Subcategory	31,594	1,416	Q: What brand of beer is visible? A: sierra nevada.
Causal	5,868	200	Q: Why is the man on the street? A: homeless.

Table 1. The 10 sub-tasks statistics of VQA v2 in the instruction tuning MLLM for continual VQA tasks.

Method	n	Various task in VQA v2										AP	AF
		Rec.	Loc.	Jud.	Com.	Cou.	Act.	Col.	Typ.	Sub.	Cau.		
CL-MoE	1	19.25	14.81	54.59	56.97	24.23	46.20	27.58	26.09	36.47	18.89	32.51	20.69
	2	34.32	32.48	67.08	64.96	32.94	62.22	34.68	40.13	48.00	24.42	44.12	8.40
	4	45.00	36.89	74.54	70.58	38.31	69.06	39.98	52.85	55.80	18.89	50.19	1.15
	8	46.50	37.18	75.22	71.39	40.90	69.54	43.66	52.68	55.55	20.74	51.34	-0.02

Table 2. Influence of various experts number n in our CL-MoE on VQA v2.

Method	K	Various task in VQA v2										AP	AF
		Rec.	Loc.	Jud.	Com.	Cou.	Act.	Col.	Typ.	Sub.	Cau.		
CL-MoE	1	46.50	35.90	73.59	69.27	40.14	70.17	42.76	52.52	55.30	20.28	50.64	0.69
	2	46.50	37.18	75.22	71.39	40.90	69.54	43.66	52.68	55.55	20.74	51.34	-0.02
	3	46.48	36.47	74.73	71.07	39.66	70.03	41.37	53.59	56.23	22.58	51.22	0.30
	4	46.11	37.18	75.01	72.21	38.90	70.17	43.08	51.94	56.30	18.43	50.93	0.17

Table 3. Influence of top K experts in our CL-MoE on VQA v2.

Methods	Various task in VQA v2										AP	AF
	Rec.	Loc.	Jud.	Com.	Cou.	Act.	Col.	Typ.	Sub.	Cau.		
VQACL	49.90	34.32	70.50	61.33	42.71	71.55	61.66	50.56	46.31	18.43	50.73	4.91
CL-MoE	53.90	38.89	78.46	75.22	46.51	74.79	64.67	57.80	57.48	23.04	57.08	-1.14

Table 4. Influence of task orders with reverse task order in our CL-MoE on VQA v2.

categories can be seen or unseen.

- **VQACL** [8] proposes a novel representation learning strategy that can distinguish sample-specific and sample-invariant features, utilizing compositional tests to evaluate the model's capability to generalize new skill and concept combinations, use rehearsal methods to alleviate the forgetting problem.

3. Experimental Result

Due to space constraints, we present the complete experimental results in the supplementary materials, which contain the influence of various experts number n, top K experts, and reverse task order, as shown in Table 2, Table 3 and Table 4. The results in Table 4 show that our CL-MoE achieves the best performance on every sub-task, proving that our method is robust to the task order and can effectively alleviate the forgetting problem.

4. Visualization Analysis

We visualized some cases to demonstrate the effectiveness of our method (Figure 2). The results show that VQACL forgets the learned knowledge after continual learning,

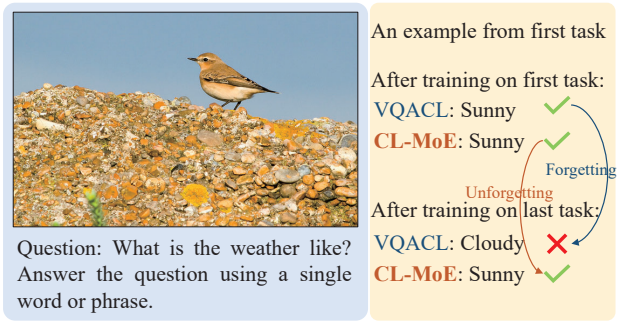

Figure 2. The visualization result of our CL-MoE compared to VQACL on VQA v2.

while our CL-MoE can alleviate the catastrophic forgetting problem.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, pages 139–154, 2018. 1

[2] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33: 15920–15930, 2020. 1

[3] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019. 1

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6, 2023. 1

[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[7] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *CVPR*, pages 16702–16711, 2022. 1

[8] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *CVPR*, pages 19102–19112, 2023. 1, 2