

Q-MoE: Connector for MLLMs with Text-Driven Routing

Supplementary Materials

A TRAINING

A.1 Router

PRE-Router is the most general version of Router under the Mixture of Expert setting. The input of Expert FFN is $x \in \mathbb{R}^{N_q \times D}$. It consists of a simple linear layer ($W_o \in \mathbb{R}^{D \times E}$) that maps dimension from dimension of hidden states D to number of experts E . Then it follows by a softmax operation to output probability values.

$$g(x)_i = (\text{softmax}(W_o x + \epsilon))_i \quad (1)$$

POST-Router, specifically operates on the outputs of the Expert FFNs $f(x)_i$. We first reduce the dimension of $f(x)_i$ from $N_q \times D$ to $1 \times D$ through mean pooling. Then we introduce a route vector $W_p \in \mathbb{R}^{D \times 1}$ that, given any $f_i(x)$, determine whether and how much each expert is activated. Then, we apply a softmax function to all concatenated $W_p f_i(x)$, and output the activation probability score of each expert. The motivation behind this design is to make decision of FFN activation based on the processing results of the certain FFN in a posterior manner.

$$m(x) = \text{meanPool}(x), m(x) \in \mathbb{R}^{1 \times D} \quad (2)$$

$$g(x)_i = (\text{softmax}([W_p m(f_0(x)) + \epsilon, \dots, W_p m(f_E(x)) + \epsilon]))_i \quad (3)$$

CLS-Router operates on the embedding of the [CLS] token from the Q-MoE text input. The Router defines an weight matrix $W_c \in \mathbb{R}^{D \times 1}$, and then outputs probability through softmax. This approach aims to leverage textual information to assist in the selection of Expert FFNs.

$$g(x)_i = (\text{softmax}([W_p h_{CLS} + \epsilon, \dots, W_p h_{CLS} + \epsilon]))_i \quad (4)$$

A.2 Training and Hyper-parameters

Details are shown in Table 1.

Config	Parameter
Max training steps	50k
Batch size per GPU	32
Optimizer	AdamW
Beta1	0.9
Beta2	0.999
Weight decay	0.05
Image resolution	224 × 224
Learning rate	1e-6 to 5e-5 in first 600 steps(linear warmup)
Learning rate schedule	cosine
GPU	4 × Nvidia A100 (80G)

Table 1: Hyper-parameters

B DATA

B.1 Datasets

In this section, we show the statistic of training datasets and evaluation datasets. Firstly, for the fine-tuning related datasets, we show how each dataset distributes and what ratio we use for sample balancing in Table 3. For the evaluation datasets, we illustrate the metric of 6 datasets that evolve in fine-tuning and 4 datasets that evolve in zero-shot performance evaluation.

Dataset	Train	Validation	Test	Metric	SampleRatio
GQA	943000	12578	12578	VQA-score	10
OKVQA	9009	5046	5046	VQA-score	1
VQAv2	658104	214354	447793	VQA-score	9
AOKVQA	17056	1145	1145	Accuracy	2
COCOCap	414113	5000	5000	CIDEr	7
TextCap	109765	15830	16445	CIDEr	4

Table 2: Datasets For Fine-tuning Related Statistic.

Dataset	Test	Metric
VizWiz	4319	Accuracy
HatefulMemes	500	Accuracy
Visual Spatial Reasoning	1222	Accuracy
NoCaps	4500	CIDEr

Table 3: Datasets For Zero-shot Evaluation Related Statistic.

B.2 Instruction List

We list three types of instruction format for Generative QA, Multi-choice VQA and Image Captioning respectively in Table 4.

C EXPERIMENTS

C.1 Ablation Experiment Results

In this section, we give full results on all fine-tuning datasets, including the ablation results of main components in Table 5, routing in Table 6, expert setup and combination in Table 7 and training tasks in Table 8.

Task	Dataset	Instruction Format
Generative QA	GQA?	{QUESTION},
	VQAv2?	Q: {QUESTION} A: ,
	OKVQA?	Based on the image, respond to this question with a short answer: {QUESTION}, {QUESTION} A short answer to the question is , Question: {QUESTION} Short answer: ,
Multi-choice VQA	AOKVQA?	{QUESTION} Choose from {CANDIDATES}. , Q: {QUESTION} Multi Choices: {CANDIDATES} A: ,
		question: {QUESTION} Multi Choices: {CANDIDATES} Answer: , "{QUESTION} Choose one from the following possible answers: {CANDIDATES}. ", {QUESTION} Choose from {CANDIDATES}. The answer is ,
Image Captioning	COCO Captions?	A photo of ,
	TextCap?	An image that shows , Write a short description for the image. ,
		Write a description for the photo., A short image description: ,

Table 4: Instruction Templates.

Components	GQA	OKVQA	VQAv2	AOKVQA	COCO Caption	Text Caption	VQA Average	Caption Average	AVERAGE
Q-MoE	63.611	58.640	79.560	75.022	139.215	108.082	69.208	123.648	87.355
\- CROSS	63.563	58.280	79.470	74.585	139.137	108.000	68.975	123.568	87.173
\-Path-Optim	63.778	57.750	79.600	73.799	139.522	106.964	68.732	123.243	86.902

Table 5: Complete Results for Ablation of Main Components.

routing	Strategy	GQA	OKVQA	VQAv2	AOKVQA	COCO Caption	Text Caption	VQA Average	Caption Average	AVERAGE
-	-	62.466	57.520	78.180	75.197	138.595	106.606	68.341	15.000	86.427
PRE	-	62.824	57.250	78.430	74.672	138.890	106.381	68.294	122.636	86.408
CLS	Path	62.673	57.440	78.430	75.459	138.195	105.424	68.500	121.810	86.270
POST	Optim	63.659	57.820	79.530	72.227	139.433	106.847	68.309	123.140	86.586
CROSS	-	63.651	58.500	79.340	72.926	139.789	108.519	68.604	124.154	87.121

Table 6: Complete Results for Ablation of routing.

Expert Combination	GQA	OKVQA	VQAv2	AOKVOA	COCOCap	TextCap	VQA Average	Caption Average	AVERAGE
task expert=1	62.466	57.520	78.180	75.197	138.595	106.606	68.341	122.601	86.427
task expert=3	63.722	58.500	79.240	74.672	139.228	107.729	<u>69.034</u>	123.479	87.182
task expert=4	63.651	58.500	79.340	72.926	139.789	<u>108.519</u>	68.604	124.154	87.121
task expert=3 + general expert	<u>63.611</u>	58.640	<u>79.560</u>	<u>75.022</u>	139.215	108.082	69.208	123.648	87.355
task expert=7 + general expert	63.555	58.320	79.570	73.624	<u>139.229</u>	108.687	68.767	<u>123.958</u>	87.164

Table 7: Complete Results for Expert Setup and Combination..

Tasks	GQA	OKVQA	VQAv2	AOKVOA	COCOCap	TextCap	AVERAGE
VQA	63.45	56.68	76.88				65.67
	63.61	58.50	79.81				67.31(+1.64)
+AOKVQA	63.28	57.99	79.21	72.49			68.24
	63.52	58.72	79.00	74.32			68.89(+0.65)
+COCOCap	63.21	59.04	79.37	73.10	139.77		82.90
	63.74	58.36	79.53	74.93	139.14		82.99(+0.09)
+TextCap	62.466	57.52	78.18	75.20	138.60	106.61	86.43
	63.61	58.64	79.56	75.02	139.21	108.08	87.35(+0.94)

Table 8: Complete Results for Training Tasks.