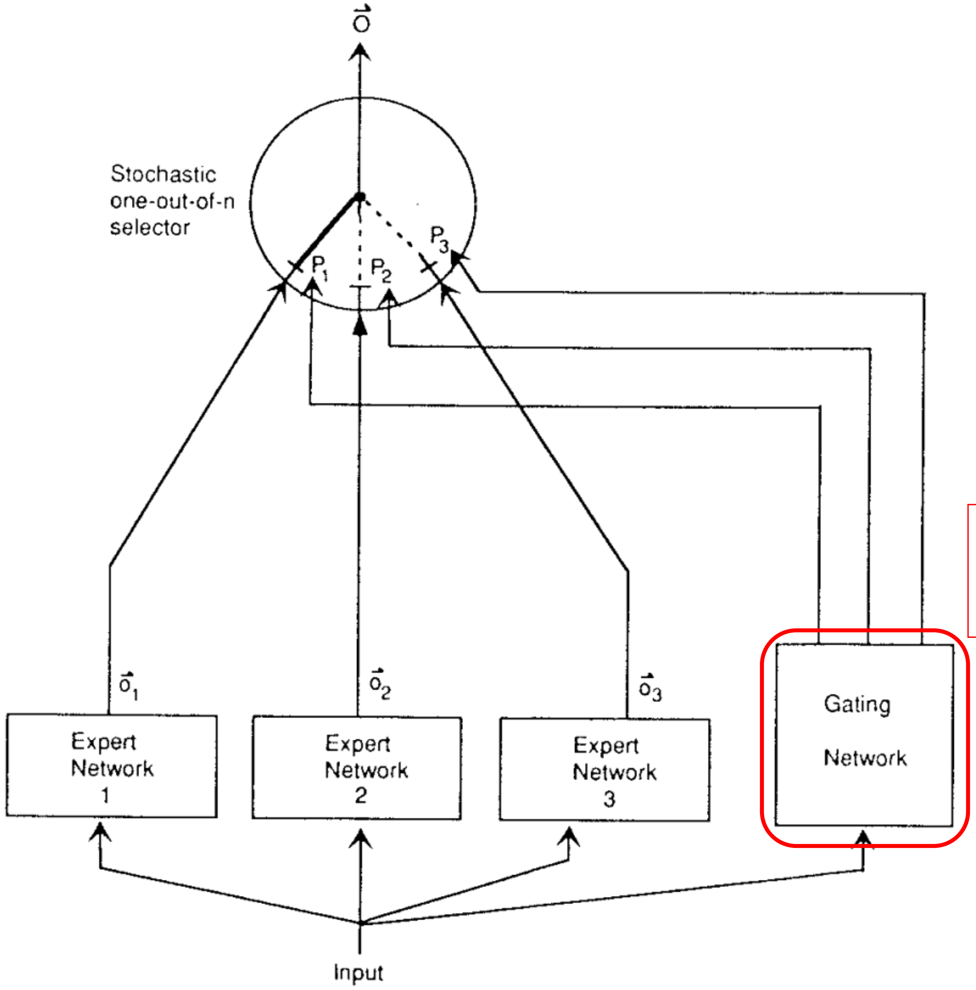
Mixture of Thoughts

Learning to Aggregate What Experts *Think*, Not Just What They *Say*

Jacob Fein-Ashley, Dhruv Parikh, Rajgopal Kannan, Viktor Prasanna

What is an expert system?

Origin of the Mixture of Experts (MoE)



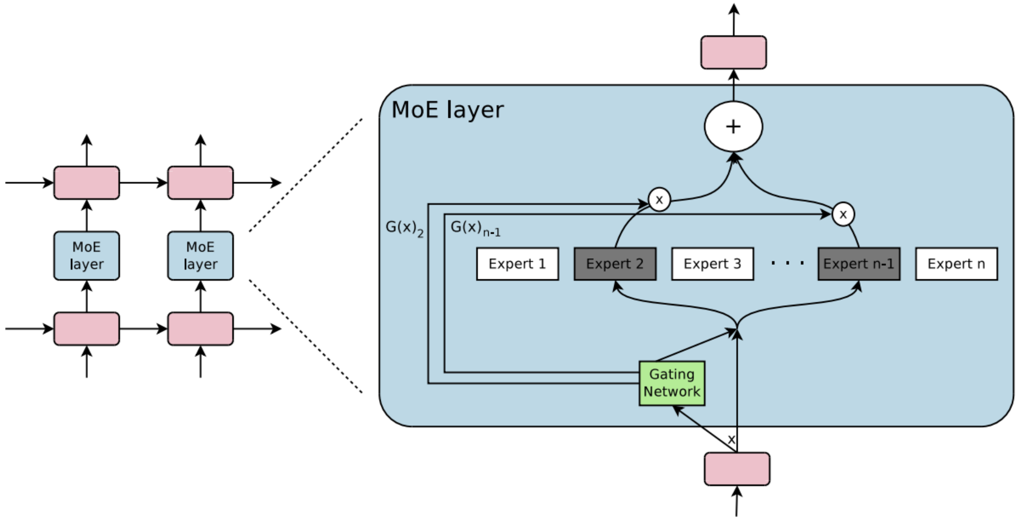
A gating or "routing" network decides which expert is best for the given input

"Adaptive Mixtures of Local Experts" by Jacobs, Jordan, Nowlan, & Hinton (1991)

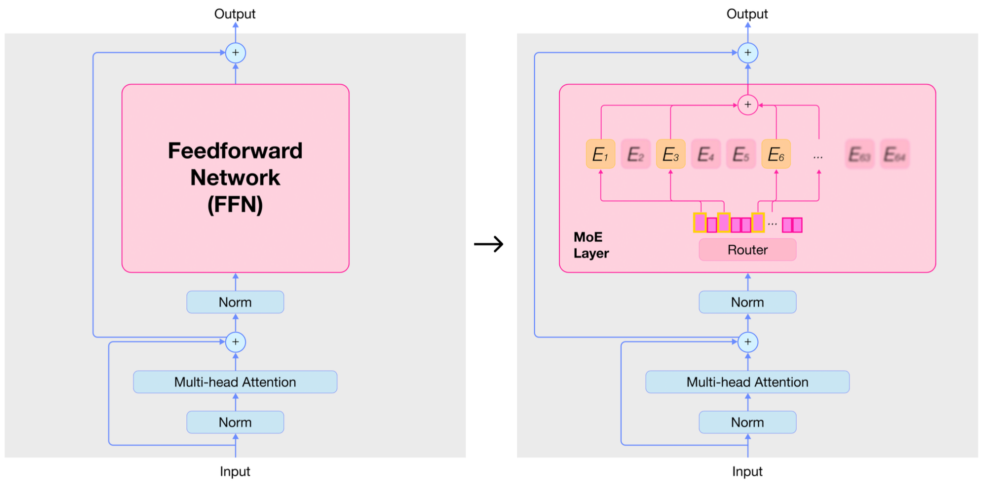
The “Modern” MoE

- MoE is popular now because of a trick
- Trick: use the gating network to compute a score for each expert for a given token. Only pick the top-K experts for that token

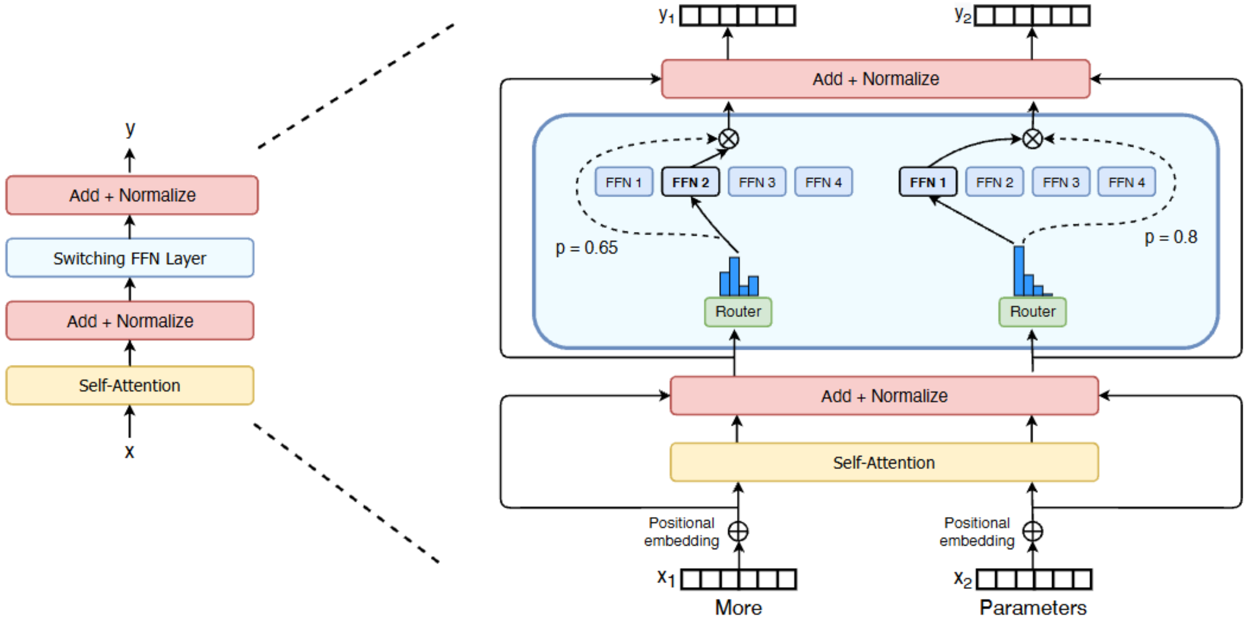
- Only those selected experts run a forward pass and get gradients
- Can get insanely large networks with little cost



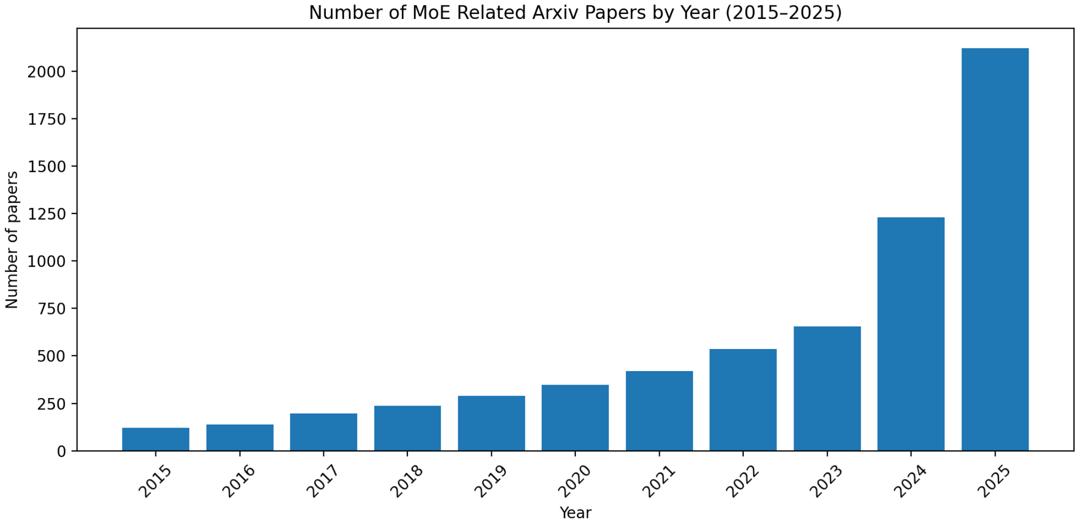
Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. Shazeer et al. 2017



MoE Now



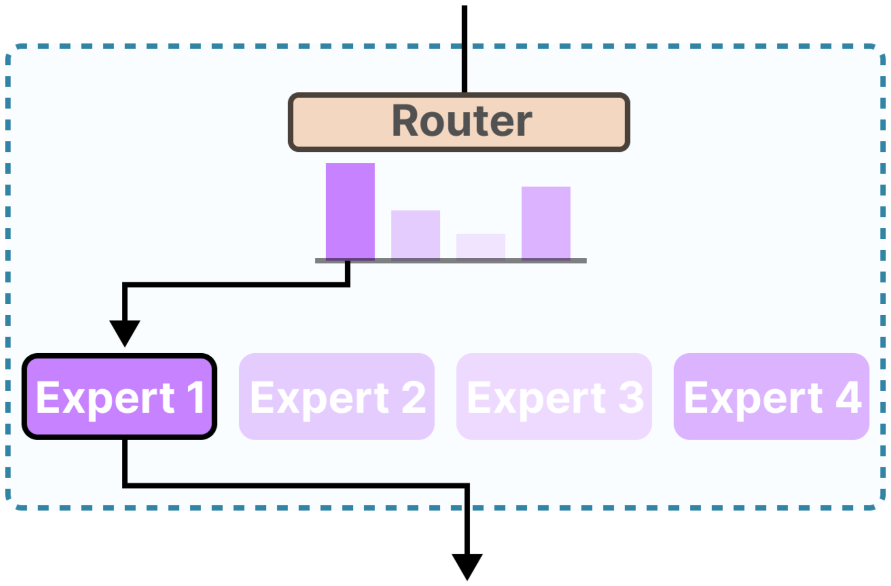
Switch Transformer: another sparsity trick



Ensemble LLM

- These previous papers put the “experts” all inside of one network
- Another similar line of work uses a single LLM as an “expert”, an LLM ensemble

- Ensemble LLM works focus on routing (gating)-based methods

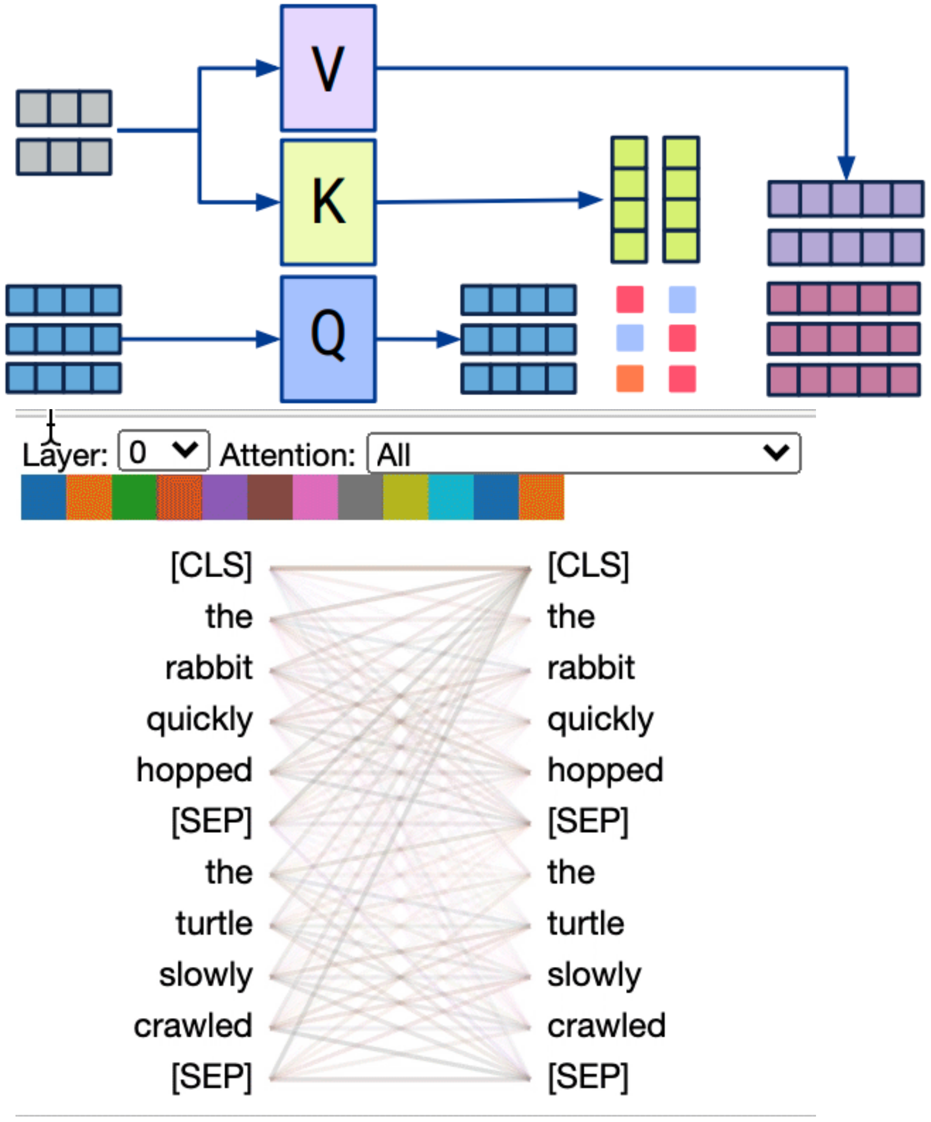


Our Method: Mixture of Thoughts (MoT)

- Previous expert systems combine what the experts "*say*" at the end: at the output level
- Motivation: Learn what each expert in the ensemble *thinks* and *says* for a more robust system

Attention and Cross Attention

- Self-attention is the foundation of the Transformer
- Self-attention compares a token to every other token in the same sequence
- Cross-attention is slightly different, Q comes from one sequence and the KV values come from a different one.
- Strong way of letting one source of information learn or "look" at another



Our Method cont.

An “Interaction Layer”

What the “stack” looks like

Results

Results cont.

Table 1: In-distribution results (accuracy %, higher is better). "Time" is average end-to-end evaluation minutes.

Method	MMLU	GSM8K	CMMLU	ARC-C	HEval	Avg	Time
Base models							
Mistral-7B	62.1	36.7	43.8	49.4	29.0	44.2	38.8
MetaMath-Mistral-7B	59.9	69.6	43.8	48.3	29.8	50.3	40.5
Zephyr-7B-Beta	59.8	33.0	42.8	58.0	22.0	43.1	40.6
Chinese-Mistral-7B	57.4	41.0	49.7	43.5	21.4	42.6	40.2
Dolphin-2.6-Mistral-7B	60.5	52.4	43.7	52.6	45.1	50.9	42.1
Meta-LLaMA-3-8B	**64.6**	47.8	51.8	49.4	26.7	48.1	41.2
Dolphin-2.9-LLaMA-3-8B	59.5	69.8	44.7	49.4	49.4	54.6	38.6
Ensembles / routers							
Voting	63.3	67.4	47.5	50.9	42.9	54.4	343.8
CosineClassifier	59.7	69.0	45.5	50.6	46.3	54.2	49.7
ZOOTER	60.5	66.7	45.3	53.1	44.3	54.0	47.3
LoraRetriever	63.3	66.6	51.8	57.1	40.0	55.8	46.2
RouterDC	61.1	70.3	51.8	58.5	51.0	58.5	46.8
Avengers	<u>62.8</u>	<u>71.6</u>	<u>52.6</u>	**60.9**	<u>53.7</u>	<u>60.3</u>	51.3
Ours							
MoT (ours)	63.1	**72.2**	**53.0**	<u>60.4</u>	**54.1**	**60.5**	51.8

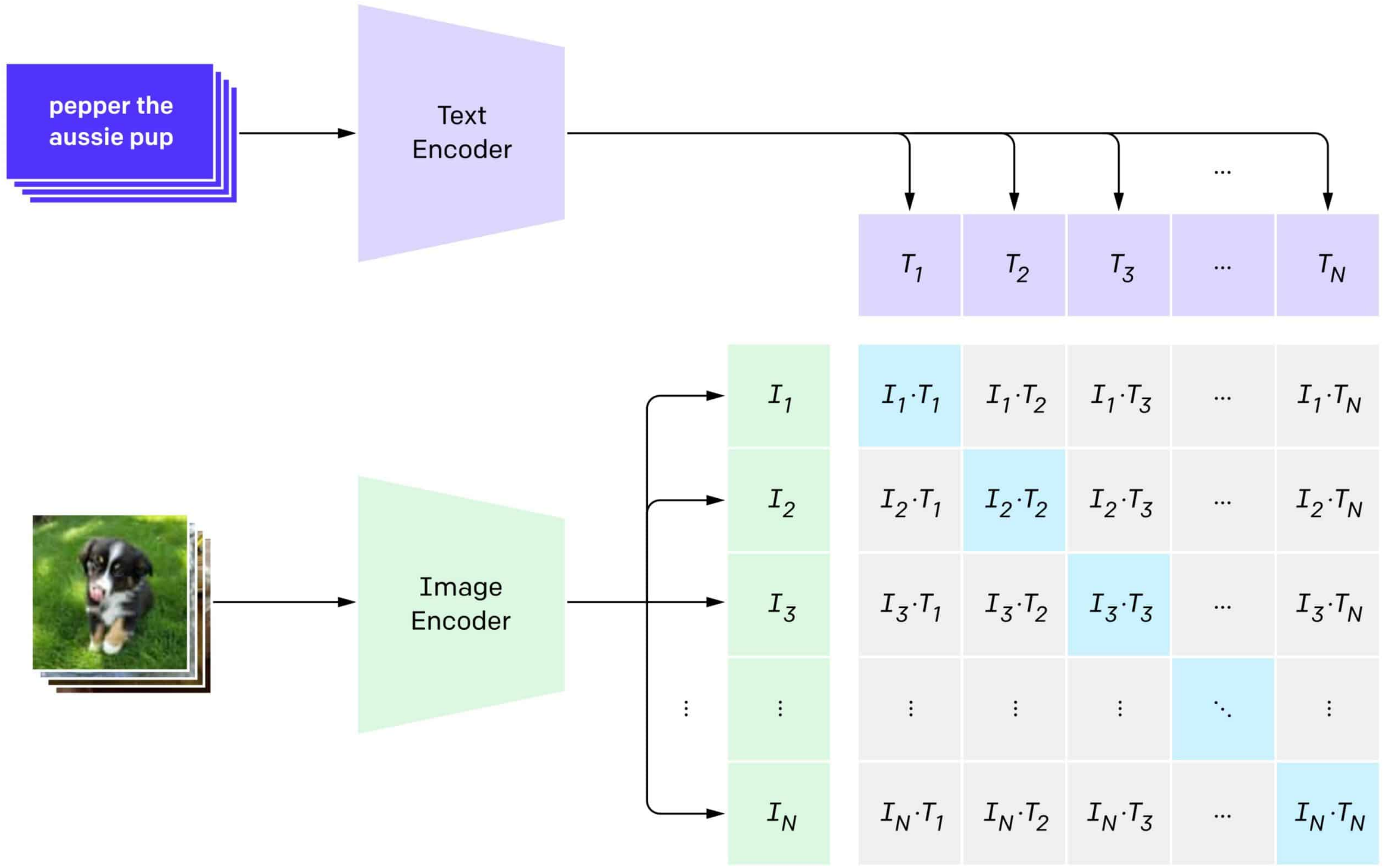
Table 2: Out-of-distribution results (accuracy %).

Method	PreAlg.	MBPP	C-EVAL	Avg	Time
Base models					
Mistral-7B	24.8	37.9	46.4	36.4	31.3
MetaMath-Mistral-7B	39.2	37.7	45.2	40.7	30.6
Zephyr-7B-Beta	20.8	31.1	44.9	32.3	32.7
Chinese-Mistral-7B	18.5	29.6	48.4	32.2	32.9
Dolphin-2.6-Mistral-7B	29.3	44.9	45.1	39.8	28.4
Meta-LLaMA-3-8B	27.7	43.0	52.0	40.9	27.9
Dolphin-2.9-LLaMA-3-8B	39.7	47.3	44.8	44.0	27.6
Ensembles / routers					
Voting	39.0	41.6	48.5	43.0	205.4
CosineClassifier	37.0	38.5	47.8	41.1	33.0
ZOOTER	34.4	41.1	45.0	40.2	31.6
LoraRetriever	35.4	43.1	52.0	43.5	31.2
RouterDC	38.8	46.8	51.9	45.9	32.6
Avengers	<u>39.0</u>	<u>48.1</u>	<u>52.6</u>	<u>46.6</u>	37.9
Ours					
MoT (ours)	**39.9**	**48.6**	**55.3**	**47.9**	38.1

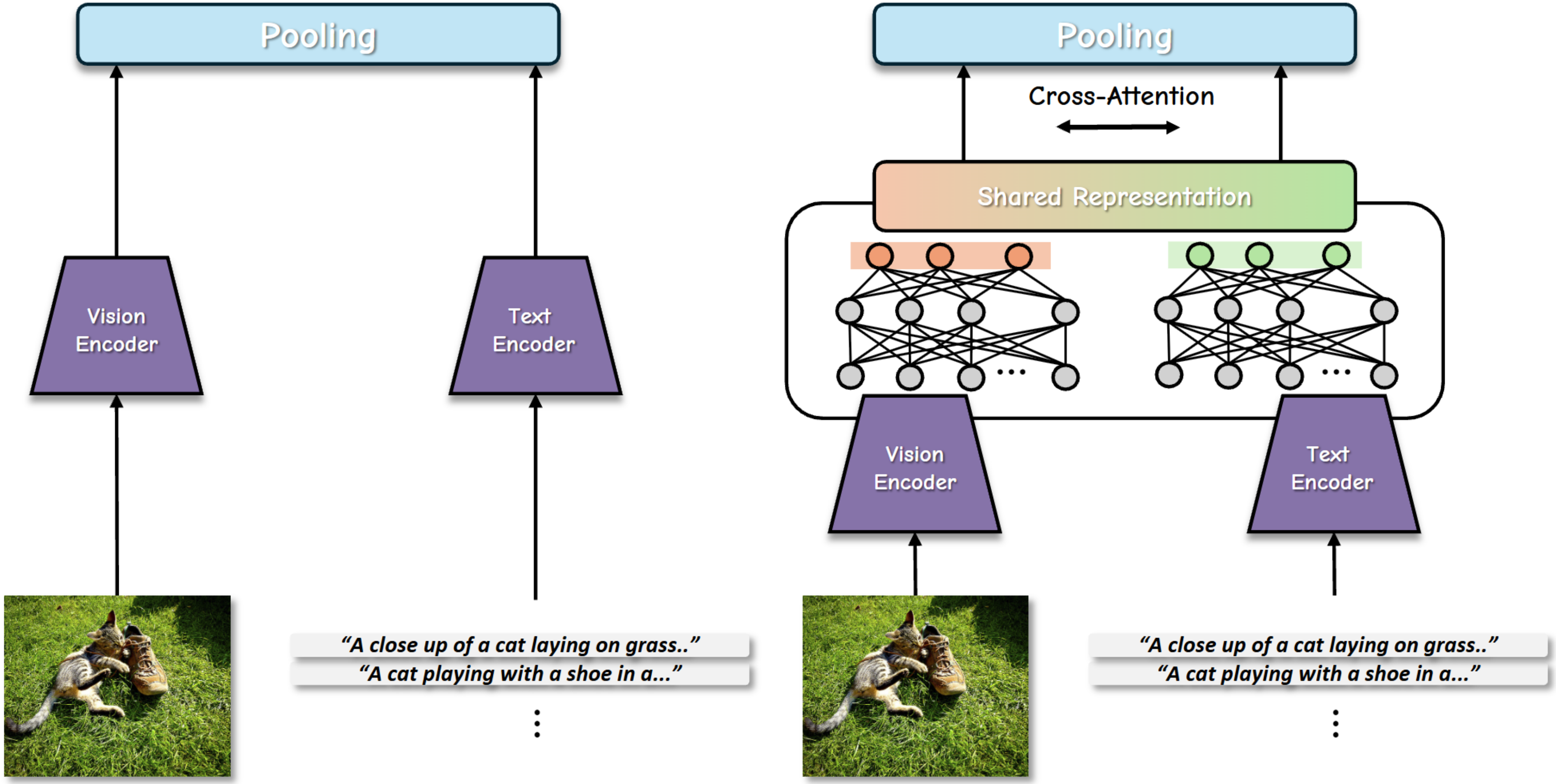
Bridging Hidden States in Vision-Language Models

Benjamin Fein-Ashley, Jacob Fein-Ashley

Vision Language Models (VLMs)



Similar Idea as Before Applied to VLMs



CLIP-style contrastive framework

Align Encoder Hidden States (Ours)

Architecture

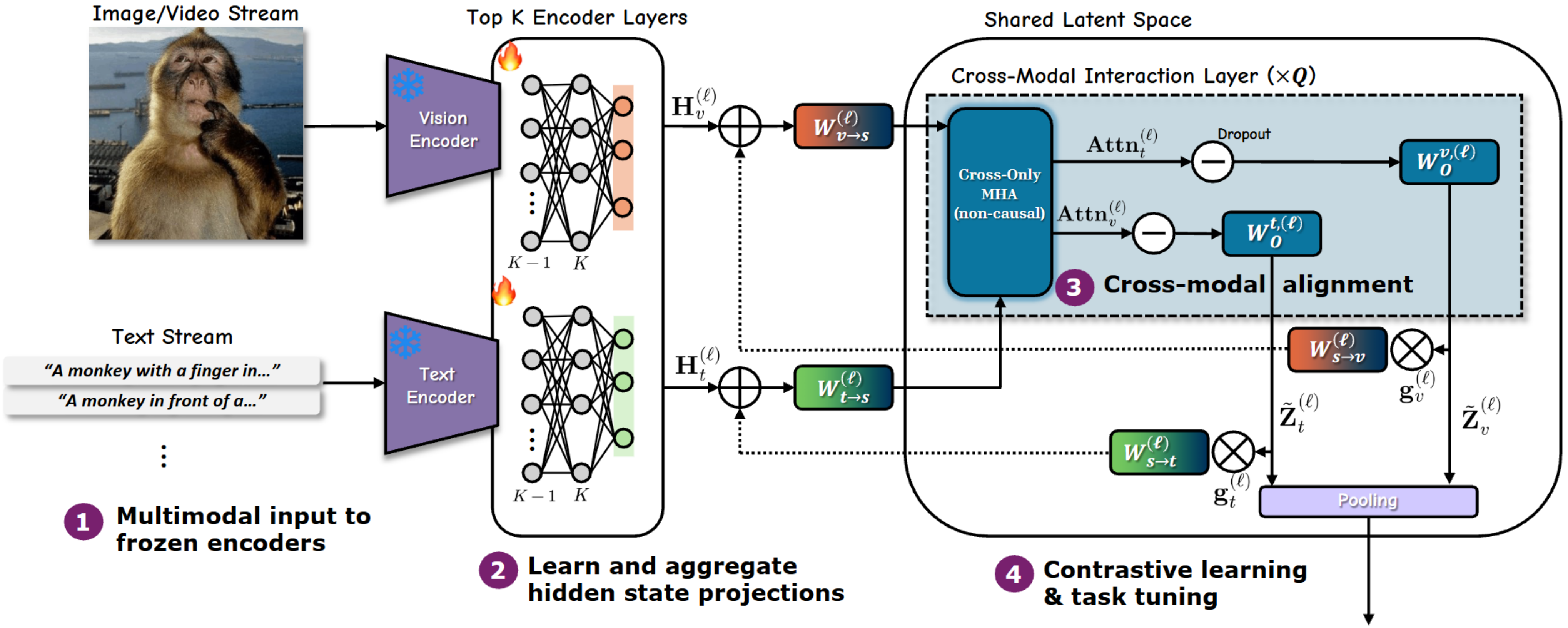


Figure 2. Architecture framework for **BRIDGE**. We propose an architecture where the hidden states of text and vision encoders are aligned directly rather than through pooled embeddings and contrastive loss. In a shared latent space, cross-only MHA is applied with residuals reverse-projected to respective embedding spaces.

Results

Model	Backbone		# Params	MSCOCO (Karpathy 5K)				Flickr30K (1K test)			
	Image	Text		TR@1	TR@5	IR@1	IR@5	TR@1	TR@5	IR@1	IR@5
CLIP [26]	ViT-B/32	Transformer	151M	37.8	62.4	58.4	81.5	86.5	98.0	67.0	88.9
ALBEF [16]	ViT-B/16	BERT-Base	203M	77.6	94.3	60.7	84.3	77.6	94.1	61.0	84.5
BLIP (14M) [17]	ViT-B/16	BERT-Base	213M	80.6	95.2	63.1	85.3	96.9	99.9	87.5	97.6
BRIDGE (Ours)											
2 interaction layers	ViT-B/16	BERT-Base	236M	*81.3*	*96.3*	*66.9*	*86.4*	*97.2*	*99.9*	*88.2*	*97.8*
4 interaction layers	ViT-B/16	BERT-Base	250M	*81.5*	*96.5*	*67.2*	*86.7*	*97.4*	*99.9*	*88.5*	*97.9*
6 interaction layers	ViT-B/16	BERT-Base	264M	***81.6***	***96.6***	***67.5***	***86.9***	***97.5***	***99.9***	***88.8***	***98.0***

Table 1. **Image–Text Retrieval on MSCOCO and Flickr30K.** Comparison of recent VLMs on the MSCOCO Karpathy 5K split [13] and the Flickr30K 1K test set [25]. TR: text-to-image retrieval; IR: image-to-text retrieval. All values are Recall (%).

Results cont.

Model	Backbone		VQAv2 [3]	
	Image	Text	test-dev	test-std
UNITER [7]	Faster R-CNN	BERT-Base	73.8	74.0
OSCAR [20]	Faster R-CNN	BERT-Base	73.6	73.8
VinVL [39]	Faster R-CNN	BERT-Base	76.5	76.6
ALBEF [16]	ViT-B/16	BERT-Base	75.8	76.0
BLIP (14M) [17]	ViT-B/16	BERT-Base	78.3	78.3
SimVLM [32]	Transformer	Transformer	80.0	80.3
BRIDGE (Ours)	ViT-B/16	BERT-Base	***80.6***	***80.7***

Table 2. **VQA on VQAv2.** Comparison of BRIDGE with prior vision–language models on the VQAv2 benchmark [3]. All values are overall VQA accuracy (%).

Model	Backbone		NLVR2 [28]	
	Image	Text	dev	test-P
ALBEF (4M) [16]	ViT-B/16	BERT-Base	80.24	80.50
ALBEF (14M) [16]	ViT-B/16	BERT-Base	82.55	83.14
TCL [35]	ViT-B/16	BERT-Base	80.54	81.33
BLIP (14M) [17]	ViT-B/16	BERT-Base	82.67	82.50
BRIDGE (Ours)	ViT-B/16	BERT-Base	***83.04***	***82.87***

Table 3. **Natural language visual reasoning on NLVR2.** Accuracy (%) on the NLVR2 dev and public test set (Test-P) for models with ViT-B/16 and BERT-Base backbones.