Expedited Training of Visual Conditioned Language Generation via Redundancy Reduction

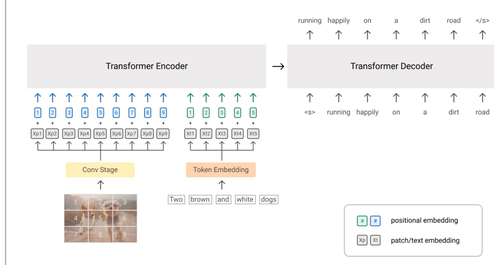
Yiren Jian[1], Tingkai Liu[2], Yunzhe Tao[2], Chunhui Zhang[1], Soroush Vosoughi[1], and Hongxia Yang[2]

(1) Dartmouth College (2) ByteDance Inc.

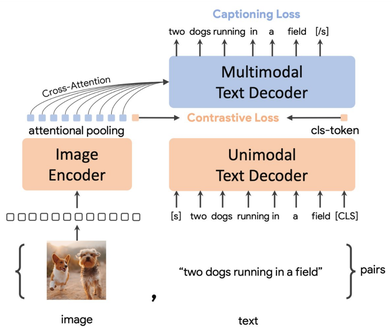
Main Contributions

- ✓ For **reducing vision redundancy** within the vision language connector, we adopt **Token Merging**, initially designed to enhance ViT inference speed without training. Concurrently, we present a novel temporal token contextualization scheme for video modeling.
- ✓ Compared with BLIP-2, while requiring just a fraction of the computational resources
- ✓ We introduce a straightforward spatial attentive temporal modeling technique that allows for the seamless adaptation of pre-trained image-text models to video tasks.

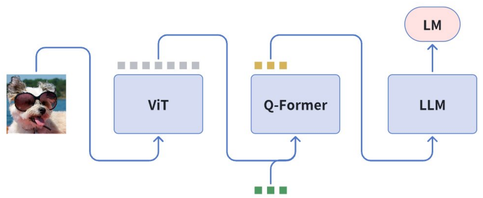
Background



SimVLM [1]: Training from scratch with 500 TPUs, 1.8B image-text pairs.



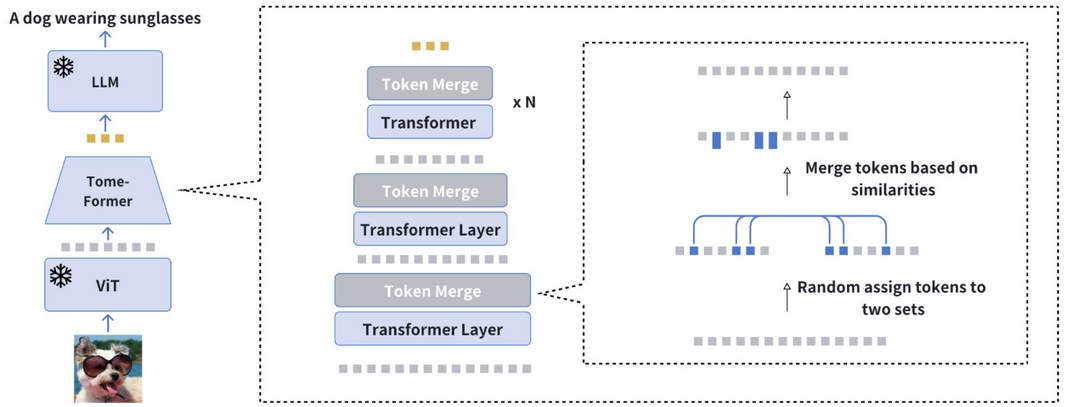
CoCa [2]: Training from scratch with 2048 TPUs, billion-scale data, 5 days.



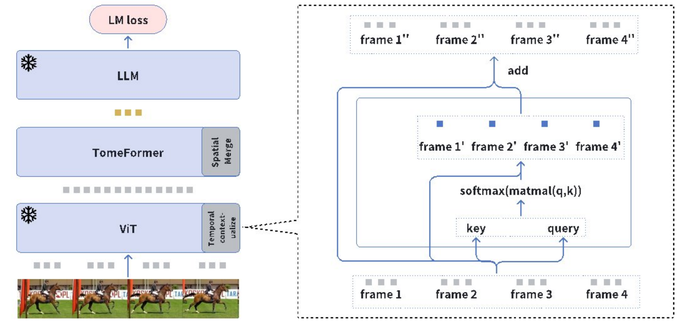
BLIP-2 [3]: Leveraging the pre-trained ViT and LLM, but still requires 10 days on 8x A100.

Training cost is the challenge!

Our Solutions: Token Merging Transformer (TomeFormer) in EVLGen-image and EVLGen-video



- **EVLGen-image** employs a streamlined, single-stage training mechanism with a unified loss.
- Visual tokens (in grey) are progressively aggregated based on their inherent similarities [4] at each layer of the TomeFormer architecture.
- The final set of merged tokens (in orange) serves as semantically rich but computationally efficient soft prompts, guiding the LLM to generate a corresponding caption for the input image.



EVLGen-video: For more spatial redundancy, temporal contextualize can pool multiple frames, then add back to each original frame.

Visualization of Merged Tokens in TomeFormer

Pre- and post-training visualization of merged tokens in EVLGen. The visual features compressed via token merging exhibit semantic informativeness even prior to training. This inherent characteristic facilitates EVLGen's ability to converge quickly in an end-to-end training setup.

Experimental Results

Models	# pre-train image-text	# trainable params	# stage-1 steps	# stage-2 steps	VQAv2 val	GQA test-dev	OK-VQA test	COCO val	Clock time
VL-T5	9.2M	224M	-	-	13.5	6.3	5.8	-	-
FewVLM	9.2M	740M	-	-	47.7	29.3	16.5	-	-
Frozen	3M	40M	-	-	29.6	-	5.9	-	-
VLKD	3M	406M	-	-	42.6	-	13.3	-	-
BLIP-2	104M†	110M+‡	-	80k/250k*	✗	✗	✗	✗	✗
BLIP-2	104M	110M+	250k	80k	44.6	30.6	26.0	137.7	234 hrs
EVL_{Gen}	104M	55M	-	90k	45.9	30.6	25.8	134.0	47 hrs
EVL_{Gen}	11M†	110M	-	150k	46.3	30.0	23.0	135.1	80 hrs
EVL_{Gen}	104M	110M	-	150k	46.9	30.8	24.8	137.0	80 hrs
EVL_{Gen}	104M	110M	-	250k	**48.4**	**30.9**	**27.2**	**139.1**	133 hrs

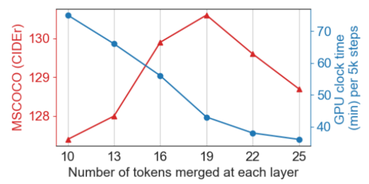
Comparison of methods on zero-shot VQA and MSCOCO captioning (CIDEr) tasks without additional fine-tuning. Both BLIP-2 and EVLGen use OPT-2.7b as the LLM decoder. ∗ : BLIP-2 without extensive stage-1 pre-training will collapse.

	LLM	Model	C	B4	M	R
NoCaps	OPT	BLIP-2	112.2	44.4	29.5	59.7
		EVL_{Gen}	**117.4**	**45.9**	**30.3**	**61.1**
	Vicuna	BLIP-2	115.6	45.3	30.3	60.6
		EVL_{Gen}	**119.0**	**45.9**	**30.6**	**61.5**
Flickr30K	OPT	BLIP-2	77.1	28.7	23.9	51.6
		EVL_{Gen}	**82.0**	**30.0**	**24.5**	**52.4**
	Vicuna	BLIP-2	80.0	30.1	**24.8**	52.1
		EVL_{Gen}	**81.8**	**30.3**	24.5	**52.2**

Comparison of different models' performance on zero-shot NoCaps and Flickr30K Captioning.

Models	C	B4	M	R
Baseline (concat)	65.5	44.4	31.9	64.1
Baseline (mean)	67.8	47.3	32.2	65.0
EVL_{Gen}-image	68.4	47.6	32.4	65.3
EVL_{Gen}-video	69.8	48.3	32.6	65.8
EVL_{Gen}-video-scst	**74.0**	**49.2**	**33.0**	**66.5**
Video-LLaMA	59.3	47.7	29.6	63.7
VideoChat	58.0	46.5	29.5	63.4
VideoCoCa (open)	63.0	48.5	31.4	64.8

Comparison of different models' performance on MSR-VTT video captioning.



Trade-off between MSCOCO captioning scores (depicted in red) and GPU training time (depicted in blue) as a function of the number of tokens merged (r) in TomeFormer.

References

[1] Wang, Zirui, et al. "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision." ICLR 2022
[2] Yu, Jiahui, et al. "Coca: Contrastive captioners are image-text foundation models." TMLR 2022
[3] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.
[4] Bolya, Daniel, et al. "Token Merging: Your ViT But Faster." ICLR 2023.