DLCV FINAL PROJECT – Group 11

Enhanced Real-World VQA : A Selective-Based Approach

Ren-Hau Shiue (薛仁豪) , Jing-En Huang (黃靖恩) , Kai Chen (陳凱)

Abstract

In this project, we address video question-answering (VQA) challenges within the STAR dataset [1]. We present a modified version of the Flipped VQA 7B model [2], enhancing it by implementing a trainable frame selector and utilizing Llama-adapter [3] for fine-tuning. Also, we conduct an in-depth analysis of failed predictions and fine-tune hyper-parameters for improved accuracy. Finally, we secured the first place in our class and third place on the leaderboard at the submission deadline.

Backbone Flipped-VQA

The introduction of LLM can sometimes result in suboptimal answers when the model overly relies on inaccurate linguistic priors. In response to this challenge, Flipped-VQA framework, as illustrated in Fig. 1, aimed at encouraging the model to predict all possible combinations of (V, Q, A) triplets by reversing the source pair and target label, thereby gaining a deeper understanding of their intricate relationships.

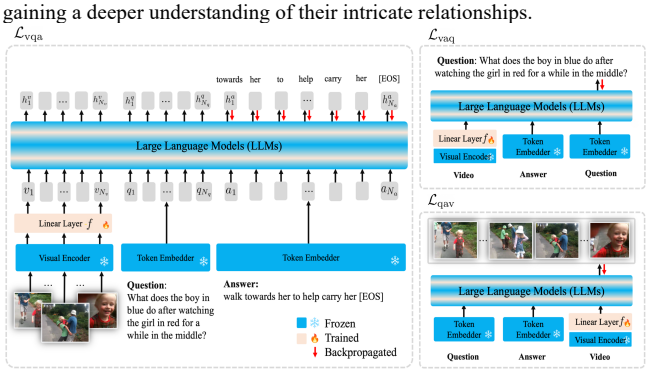


Fig. 1: Illustration of LLMs with Flipped-VQA [2].

Selector

The original sampling techniques employed in Flipped-VQA involves uniform sampling. However, with a limited sampling frequency, some critical frames may be omitted, while numerous redundant frames are included. To address this concern, we introduce a trainable selector into the model. This selector dynamically samples frames from the video using an attention block with question features as the query. This concept is inspired by [4].

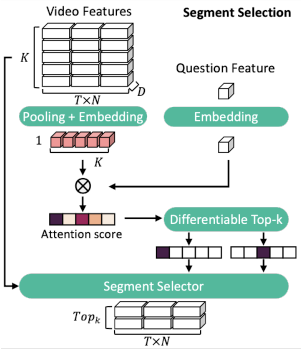


Fig. 2: Illustration of Selector in [4].

Video Frame Resampling

The original sampling frequency was set at 1 frame per second in Flipped-VQA framework, potentially resulting in the loss of information for longer video clips. Consequently, we have redone the sampling procedure by increasing the sampling frequency tenfold, now capturing 10 frames per second.

Ensemble Models

Due to the complexity of the problem dataset and our model, the models trained under various settings exhibit substantial diversity and independence. To leverage this diversity, we employ a voting strategy across all well-performing trained models.

Experience Results (Produced on a RTX-4090 GPU)

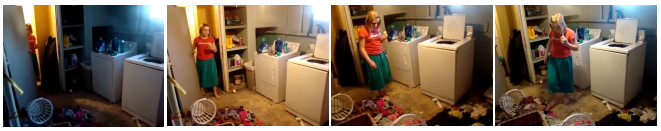
Resample	Ensemble	Selector	Int	Seq	Pre	Fea	Mean
			64.01	68.13	57.96	47.48	59.50
			63.45	67.05	59.78	47.65	59.48
	✓		65.40	68.21	60.61	49.91	61.03
✓			68.04	71.88	61.87	47.65	62.36
✓	✓		**69.98**	**72.97**	**63.27**	**50.78**	**64.25**
		✓	60.13	62.44	55.77	49.84	57.55
✓	✓	✓	58.38	61.85	54.01	48.04	55.57

Analysis and Discussion

Successful Selections

Question: What did the person do with the sandwich?

Uniform Samples: Wrong Prediction: Took



Selected Samples: Correct Prediction: Ate

⇨ Correct prediction results from a better sampling.

Failed Selections I

Question: Which object was opened by the person?

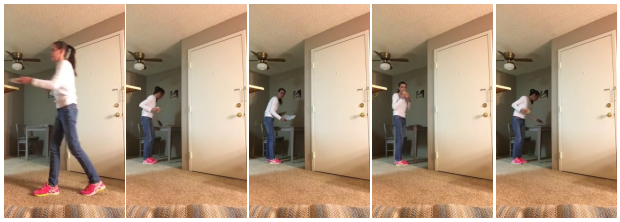
Answer: The laptop. Wrong Prediction: The door.

⇨ Selector incorrectly focuses on the frames the man opens his jacket. Options should also be used for selection.

Failed Selections II

Question: Which object did the person put down before they held the food?

Answer: The dish. Wrong Prediction: The phone/camera.



Not Included. Included in Selected Samples

⇨ We could opt for the combination of uniform samples and selected samples.

Conclusion

Regarding the suboptimal performance of the experiment with the selector, we posit several conjectures: Firstly, the training parameters for the selector should not be concurrently utilized with the Llama finetuning parameters. Secondly, it is advisable to complete the Llama finetuning prior to integrating the selector for joint training. Thirdly, the question options should be incorporated along with the question itself into the selector's input. Lastly, the sampling should not rely entirely on the selector due to the inherent risk of malfunction.

Reference

[1] Bo Wu, et el. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In NeurIPS 2021.

[2] Dohwan Ko, et el. Large Language Models are Temporal and Causal Reasoners for Video Question Answering. In EMNLP 2023.

[3] R Zhang, et al. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. arXiv:2303.16199, 2023

[4] Gao, Difei, et al. "MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering." In IEEE/CVF 2023.