SITCON 2024

Enhanced Real-World Video Question-Answering : A Selective-Based Approach

Jing-En Huang 黃靖恩 – National Taiwan University (40840210s@gapps.ntnu.edu.tw)

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 have successfully increased the accuracy of the original Flipped-VQA from 59.50 to 65.50, achieving a 10% improvement in accuracy.

STAR - A Benchmark for Situated Reasoning in Real-World Videos

STAR is a novel benchmark for Situated Reasoning, which provides challenging question-answering tasks, symbolic situation descriptions and logic-grounded diagnosis via real-world video situations. Reasoning in the real world is not divorced from situations. A key challenge is to capture the present knowledge from surrounding situations and reason accordingly.

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.

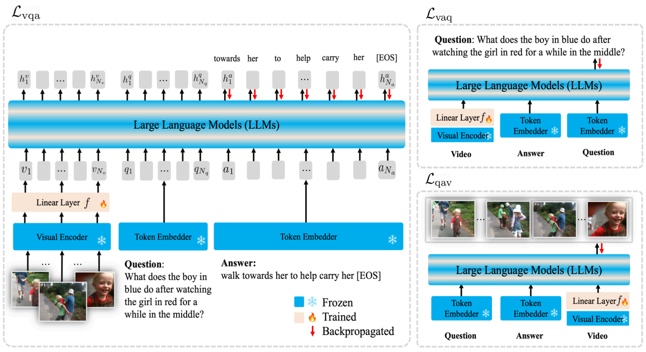

$\mathcal{L}_{\text{vqa}}$
towards
her
to
help
carry
her
[EOS]
h_1^v
$h_{N_v}^v$
h_1^q
$h_{N_q}^q$
h_1^a
$h_{N_a}^a$
Large Language Models (LLMs)
v_1
v_{N_v}
q_1
q_{N_q}
a_1
a_{N_a}
Linear Layer f
Visual Encoder
Token Embedder
Token Embedder
Question:
What does the boy in blue do after watching the girl in red for a while in the middle?
Answer:
walk towards her to help carry her [EOS]
Frozen
Trained
Backpropagated
$\mathcal{L}_{\text{vaq}}$
Question: What does the boy in blue do after watching the girl in red for a while in the middle?
Large Language Models (LLMs)
Linear Layer f
Visual Encoder
Token Embedder
Token Embedder
Video
Answer
Question
$\mathcal{L}_{\text{qav}}$
Large Language Models (LLMs)
Token Embedder
Token Embedder
Linear Layer f
Visual Encoder
Question
Answer
Video

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].

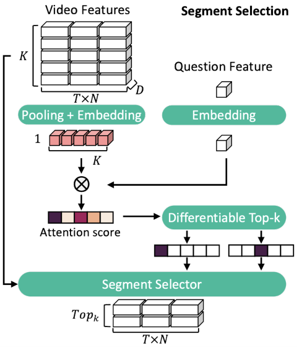


Fig. 2: Illustration of Selector in [4].

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.

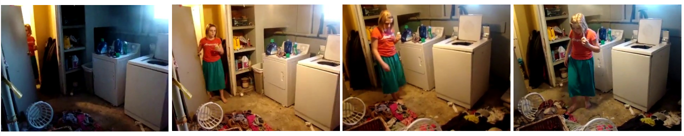
Experience Results (Accuracy produced on a RTX-4090 GPU)

Resample	Ensemble	Selector	Int	Seq	Pre	Fea	Mean
(Provide by Flipped-VQA Author)							
			64.01	68.13	57.96	47.48	59.50
(Re-produced Flipped-VQA)							
			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
		✓	67.32	69.81	61.48	48.52	61.78
✓	✓	✓	**71.58**	**73.67**	**64.70**	**52.01**	**65.50**

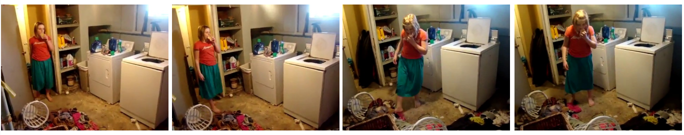
Selector Analysis

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.

Conclusion

Firstly, during the training process, adjusting the ratio of selector model training iterations to Llama model finetuning iterations to 10:1 significantly enhanced the efficacy of the selector.

Secondly, reintegrating the selector architecture for finetuning after completing the finetuning of the Llama model resulted in improved model performance.

Thirdly, inputting both the question options and the question itself into the selector led to an increase in its effectiveness.

Lastly, relying entirely on the selector for sampling posed a risk of sample corruption. Modifying the sampling approach to a uniform distribution in a 1:5 ratio with the selector yielded better results for the selector's performance.

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.