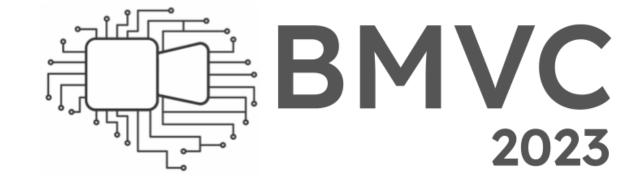
Long Story Short: a Summarize-then-Search Method for Long Video Question Answering



Jiwan Chung and Youngjae Yu

jiwan.chung.research@gmail.com



Code C github.com/JiwanChung/long-story-short

Summary

- **Task:** Extending large language models to long video QAs \bullet
- **Problems**:
 - 1. LLMs do not speak multimodality
 - 2. Videos induce extreme long context understanding
- **Solution**: Long-Story-Short, a Summarize & search method
- Output: Achieving state-of-the-art in standard long video narrative QA

CLIPCheck

- Motivation
 - LLM-based methods such as our Long-Story-Short relies on the feature extractor to correctly convert the visual details.
 - However, visual feature extractors are imperfect, incurring errors in visual
 - grounding of the video QA system.
- Method

Introduction

- Background

- Large Language Models (LLMs) can understand long-context narratives and generate adaptive outputs.
- Socratic Models (Zeng et al., 2022) showed that the large language mod els can perform multimodal reasoning by transforming the visual context t o text forms and using them as inputs.

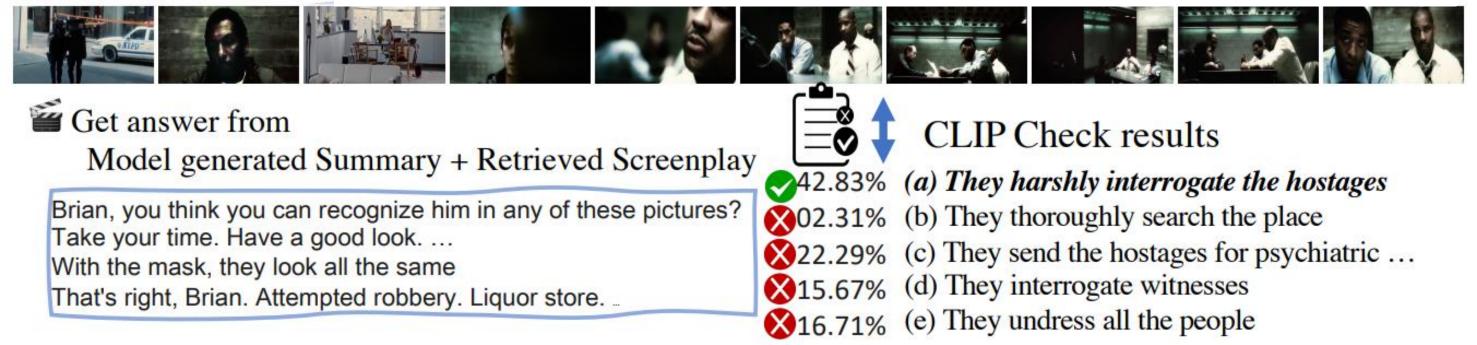
- Problems

- Long video QAs (e.g. MovieQA) are long unsolved problem since they re quire machines to model both the long narratives and visual contexts.



- We introduce **CLIPCheck**, a post-inference likelihood modification metho d for LLM-based multimodal QA frameworks.
- Given the question and possible answer candidates, we compare the visu al alignment strength of each candidate using CLIP.
- This CLIP-based alignment score is added to the LLM output likelihood to build the final score over the answer candidates.

Aligned Raw Video



Experiments

- We achieved state-of-the-art in two standard long video QA benchmarks.
- Further, our zero-shot framework outperforms supervised methods as well.

MovieQA (2015)

- LLMs are known to struggle with extremely long context. First, there is a hard token length limit, typically spanning 4,000 to 30,000. Second, LLMs tend to ignore some portions of context as it gets longer.

- Long video understanding usually requires modelling ~500,000 tokens.

Long-Story-Short Framework

- Long-Story-Short is a long video QA framework composed of four stages:
 - 1. Extract features from videos and converting them all to text forms.
 - 2. Split the video into shorter clips and **summarize** each clip to a plot piec e using an LLM,
 - 3. Given the question and answer, **search** for the relevant clip using the pl ot pieces as keys (with an LLM as well),
 - 4. Given the question, answer, and the text form features for the retrieved clip, **answer** the question (with an LLM as well),.

	Model	Aligned	V + S	V Only	S Only				Search	
Supervised	A2A	\checkmark	41.66	40.28	41.05	How does Connie feel about the idea that Edward wants to turn himself in?			A man decides to turn	
	PAMN	\checkmark	43.34	42.33	42.56			h	erself in to the authoritie fter committing a crime.	
	UniversalQA	\checkmark	48.87	50.67	47.62			l is	s faced with the prospec	
	DHTCN	\checkmark	49.60	47.38	48.43			in?	oing to prison, but is etermined to take	
	No Context	×	36.36	34.28	38.07				esponsibility for his acti	
shot	LSS	\checkmark	53.44	49.83	56.42	 (a) She is excited (b) She is confused (c) She is sad (d) She is angry (e) She is happy 			0.07 0.00	
zeroshot	LSS-Search	X	51.24	49.00	53.09				0.64 0.00 0.13 0.99 0.11 0.01 0.05 0.00	
	LSS-Search+CLIPCheck	× ×	51.49	49.55	53.09					
			Mov	ieQA (2	2015)					
Model		Level3	Level4					1 1		
CharacterAttention Kim <i>et al</i> . [14]			65.62 70.00	Mode	Model			abels. Aligne	d Acc	
				Supe	rvised [13]		\checkmark	\checkmark	68.00	
LSS +Caption +CLIPCheck +Caption+CLIPCheck		72.20	75.23	GPT	GPT3 w\o Context		×	X	36.90	
		73.54 75.78 75.34	75.68 79.28 77.93	LSS	Base + Search + Plot		<u> </u>	\checkmark	66.76	
							\checkmark	×	48.98	
							X	\checkmark	65.80	
+	CLIPCheck-Shuffle	71.74	73.87		+ Plot + S	Search	×	×	53.34	
DramaQA (2020)				-	PororoQA (2017)					

- Here, we show plot summary samples generated as an intermediate prod

(a) Video to Screenplay 🞬

(b) Screenplay to Plots 🚄

(ID:i+2)

(ID:i+1)

(ID:i)

(d) Validate with Retrieval

The man in the brown hat is digging in

the box, and he finds a cross. He is

excited and shows it to the others.

 \sim 2**----**1) Caption : The man is wearing brown hat (ID:i) SOMEONE takes a cross from the box. Subtitle : I got something.. I got something right here. Oh, look at that! We're rich! We're rich! Shut up. Shut up. Well, we're rich, ain't we? *Caption : two people in a room* SOMEONE practically salivate at the sight of it. Subtitle : Hey, we got to find...

(c) Validate with VLM (CLIPCheck)

(c) Solve Question

 \boldsymbol{Q} : What is They follow clues to different the artifact ancient sites and find artifacts from they fight for? long-lost civilizations. Along the way, they battle evil forces 0 0 After finding a golden cross in a box, a man and his friends go on a treasure hunt to find more riches. LSS

Candidate

a. Golden Cross

uct of Long-Story-Short.

Long Story Short

Harry Potter is being moved to a safe house on the 30th of the month, just before his 17th **birthday**. However, Voldemort and his followers are aware of the move and plan to attack Harry en route. Snape volunteers to kill Harry, but due to the fact that their wands are twinned, he is unable to do so. Bellatrix Lestrange then volunteers and is given the task. The **Death Eaters have ambushed** Harry, Ron, and Fred, and they are nowhere to be found. Hagrid is the only one who made it back safely.



At the beginning of the book, Harry is about to turn seventeen and will lose his deceased mother's protection. Members of the Order of the Phoenix relocate the Dursleys, and prepare to move Harry to The Burrow by flying him there, using Harry's friends as decoys. **Death** Eaters attack them upon departure, and in the ensuing battle, "Mad-Eye" Moody and Hedwig are killed while George Weasley is wounded. Voldemort arrives to kill Harry, but Harry's wand fends him off on its own.

MIR Lab @ Yonsei University