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

Jiwan Chung and Youngjae Yu

jiwan.chung.research@gmail.com

BMVC 2023

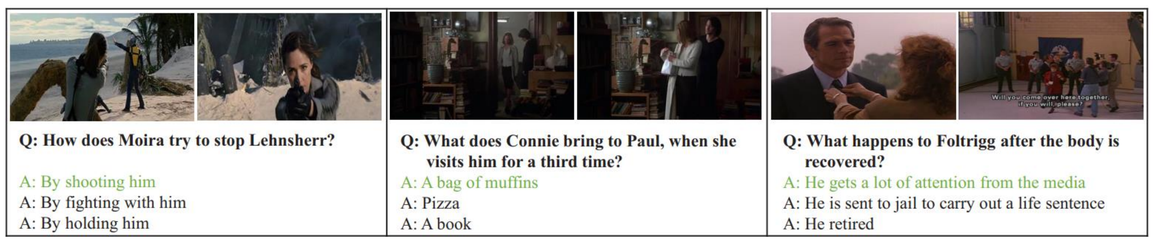
Code github.com/JiwanChung/long-story-short

Summary

- **Task**: Extending large language models to long video QAs
- **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 benchmarks (MovieQA & DramaQA)

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 models can perform multimodal reasoning by transforming the visual context to text forms and using them as inputs.
- **Problems**
  - Long video QAs (e.g. MovieQA) are long unsolved problem since they require machines to model both the long narratives and visual contexts.

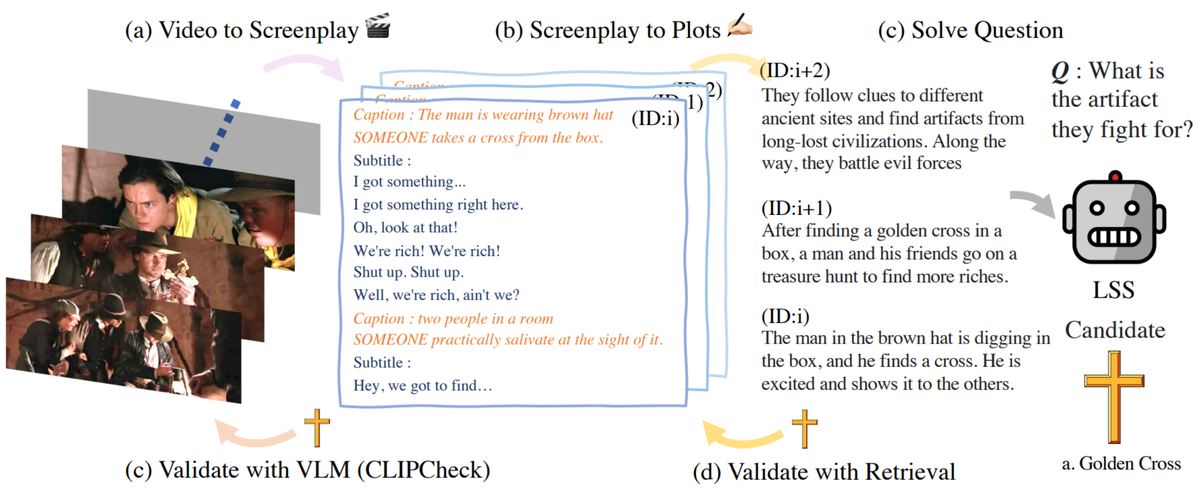


MovieQA (2015)

- 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.

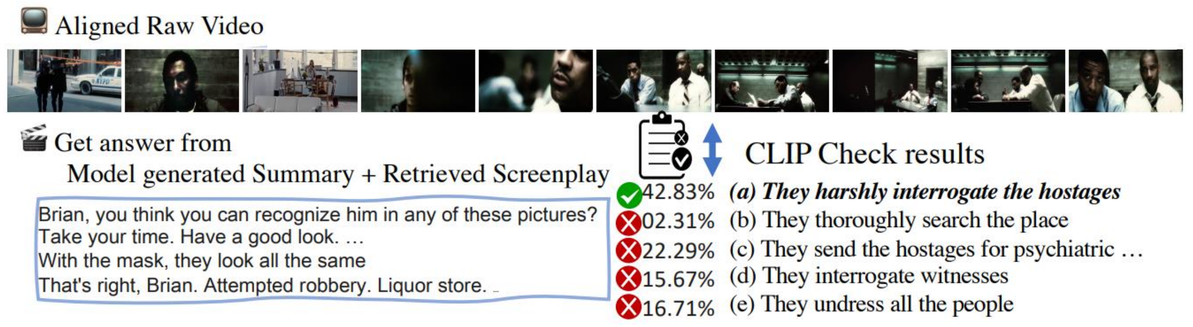
Long-Story-Short Framework

- **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 piece using an LLM,
  3. Given the question and answer, **search** for the relevant clip using the plot 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),.



CLIPCheck

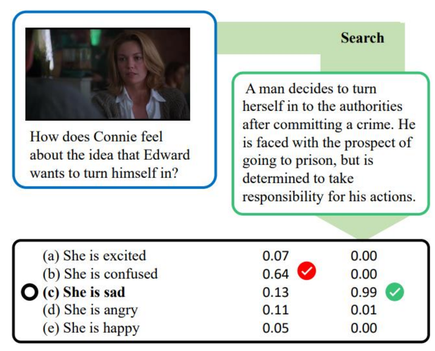
- **Motivation**
  - LLM-based methods such as our Long-Story-Short relies on the feature extractor to correctly convert the visual details.
  - However, visual feature extractors are imperfect, incurring errors in visual grounding of the video QA system.
- **Method**
  - We introduce **CLIPCheck**, a post-inference likelihood modification method for LLM-based multimodal QA frameworks.
  - Given the question and possible answer candidates, we compare the visual 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.



Experiments

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

	Model	Aligned	V + S	V Only	S Only
Supervised	A2A	✓	41.66	40.28	41.05
	PAMN	✓	43.34	42.33	42.56
	UniversalQA	✓	48.87	50.67	47.62
	DHTCN	✓	49.60	47.38	48.43
zeroshot	No Context	✗	36.36	34.28	38.07
	LSS	✓	53.44	49.83	56.42
	LSS-Search	✗	51.24	49.00	53.09
	LSS-Search+CLIPCheck	✗	51.49	49.55	53.09



MovieQA (2015)

Model	Level3	Level4
CharacterAttention	60.82	65.62
Kim *et al.* [14]	70.00	70.00
LSS	72.20	75.23
+Caption	73.54	75.68
+CLIPCheck	**75.78**	**79.28**
+Caption+CLIPCheck	75.34	77.93
+CLIPCheck-Shuffle	71.74	73.87

DramaQA (2020)

Model		Labels: Plot	Labels: Aligned	Acc
Supervised [13]		✓	✓	68.00
GPT3 w\o Context		✗	✗	36.90
LSS	Base	✓	✓	66.76
	+ Search	✓	✗	48.98
	+ Plot	✗	✓	65.80
	+ Plot + Search	✗	✗	53.34

PororoQA (2017)

Qualitative Samples

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

WikiPedia

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.