VideoRAG: Scaling the context size and relevance for video question-answering

Shivprasad Sagare, Prashant Ullegaddi, Nachiketh K S,
Navnith R, Kinshuk Sarabhai, Rajeshkumar SA
PhroneticAI
Correspondence: shivprasad.sagare@phronetic.ai, rajesh.kumar@phronetic.ai

Abstract

Recent advancements have led to the adaptation of several multimodal large language models (LLMs) for critical video-related use cases, particularly in Video Question-Answering (QA). However, most of the previous models sample only a limited number of frames from video due to the context size limit of backbone LLM. Another approach of applying temporal pooling to compress multiple frames, is also shown to saturate and does not scale well. These limitations cause videoQA on long videos to perform very poorly. To address this, we present VideoRAG, a system to utilize recently popularized Retrieval Augmented Generation (RAG) pipeline to select the top-k frames from video, relevant to the user query. We have observed a qualitative improvement in our experiments, indicating a promising direction to pursue. Additionally, our findings indicate that VideoRAG demonstrates superior performance when addressing needle-in-the-haystack questions in long videos. Our extensible system allows for trying multiple strategies for indexing, ranking, and adding QA models.

1 Introduction

In the realm of video-based language models (Video-LLMs), the ability to accurately answer questions about long-form video content remains a significant challenge. VideoRAG, a novel system introduced by researchers, aims to address this issue by retrieving more relevant and informative video frames to enhance the context for video-LLMs.

Several recent works adapt the pre-trained image-text multimodal LLMs to video data, by encoding multiple video frames into a sequence of features. However, this approach is limited by the trade-off between the number of frames and the computation cost. PLLaVA (Xu et al., 2024), a SOTA videoQA model, uniformly samples only 16 frames, which is bound to miss the key details in long videos.

Straightforward approach of averaging the spatial and temporal dimensions, as implemented in the VideoChatGPT system (Maaz et al., 2023), leads to the loss of substantial spatial information and fails to achieve optimal performance as the training dataset is scaled (Xu et al., 2024).

In contrast, VideoRAG employs a retrieval-augmented approach, where the system dynamically retrieves the most relevant video frames to supplement the input to the video-LLM. VideoRAG allows indexing of multiple video frames using multiple encoders, and searching the top-k most relevant frames given a query. We also release a human-annotated benchmark dataset along with our system, specially curated for long videos.

The evaluation of VideoRAG on question-answering tasks shows that it outperforms baseline video-LLMs that rely on direct frame averaging or fixed-frame sampling, particularly on long videos.

2 VideoRAG system

Our VideoRAG system leverages standard components of the Retrieval-Augmented Generation (RAG) pipeline and image encoding. It is platform-agnostic, supporting the addition of various strategies at each component level. The system architecture comprises six main components:

Video Processing We extract the candidate frames from the video, focusing on efficient and scalable processing to manage the memory overhead associated with storing frame information. Other aspects of the video are extracted as well, for example the audio signal, and the objects appearing the video. We also process the videos to extract chunks to track the activities happening during these chunks.

Video Encoding Entities from video, such as frames, chunks, and other extracted information

Our demo screencast is available at video link

Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations, pages 7–8
September 23–27, 2024.

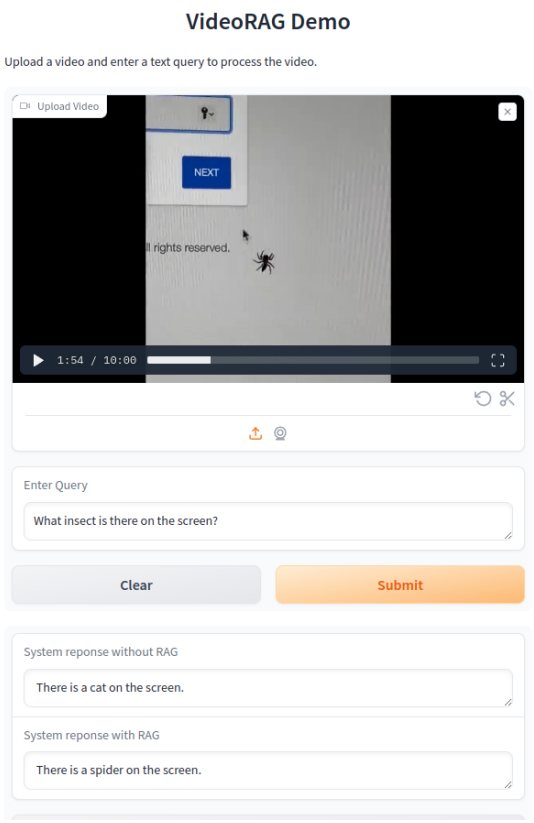

Figure 1: An example of VideoRAG.

is encoded into fixed-size vector semantic representations for further indexing. We explore multiple strategies to encode the videos intelligently, including the features like image frames, audio signal, and the objects appearing in the video. We rely on multiple strategies like SentenceTransformers (Reimers and Gurevych, 2019), (Li et al., 2022), (Zhai et al., 2023) for encoding the text query.

Indexing We employ a vector database, to store the frame embeddings effectively. This is a standard component in the RAG pipeline. These vector databases allow for efficient lookup algorithms for retrieval of top-k samples.

Query Encoding The text query is encoded similarly to the video encoding strategy to perform semantic similarity search over the index.

Retrieval and Ranking The vector database retrieves the top-k entities matching the query by performing efficient similarity search. We combine the results from multimodal indexes to generate the final ranking for the videos.

Generation with Video LLM Finally, we pass the selected entities and the text query to the video-text LLM model. We generate the answers using the state-of-the-art video-text LLM for this phase.

Metrics	Without RAG	With RAG
Correctness	2.00	**2.46 (+23%)**
Detail	2.45	**2.75 (+12%)**
Context	2.48	**2.94 (+18%)**

Table 1: Comparison of VideoRAG against traditional systems denotes substantial gains in accuracy and context while generating the answers. The scores are from LLM-based evaluation with the range of 1(worst) to 5(best).

3 Usage and Evaluation

VideoRAG is planned to be used for easy experimentation, and evaluation of long form video question-answering systems. Our system allows for easy extension to additional indexing mechanisms, ranking strategies, and videoQA models. We compare our VideoRAG outputs with that of non-RAG baseline systems. We curated a dataset of a few long videos and corresponding question-answer pairs. Table 1 showcases the quantitative evaluation results, using metrics and evaluation scripts from VideoChatGPT evaluation framework (Maaz et al., 2023). Our observations indicate that VideoRAG effectively minimizes hallucinations when addressing needle-in-the-haystack questions. We attribute this to its ability to retrieve relevant video entities in response to user queries.

References

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *Preprint*, arXiv:2404.16994.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.