Applying LLM to question/answer problem based on sequence of frames

Proposed Enhancements & Recent Advances

Semenov Vadim, MIPT, Cognitive Modeling Lab, March 2025

Core research question:

How can we leverage LLMs to enhance question answering over sequences of frames by incorporating temporal and spatial context for robust place understanding in dynamic or large-scale environments?

My personal task:

1) Filter key frames.
2) Construct relationships between relevant nodes for the question using a multi-modal large language model (MLLM).
3) Predict the answer to the question using the graph with a large language model (LLM).

Context:

Answering questions based on sequences of frames is challenging due to dynamic scene changes and complex spatial-temporal relationships. Traditional graph-based methods offer structure but lack deep semantic reasoning. With the rise of multi-modal and large language models, there's potential to better understand and reason over visual content across time.

Motivation:

This work aims to improve question answering in dynamic environments by combining MLLMs for extracting visual-semantic relations and LLMs for reasoning over graph structures. The approach supports more accurate, context-aware answers, enabling smarter systems for video understanding, navigation, and real-world interaction.

Planned results:

- **Paper Submission**: Aim to write and submit a scientific paper by the end of the semester
- **Technical Innovations**: Enhance the visual scene graph construction pipeline using MLLMs to extract semantic relations between key entities. Explore adding spatial/temporal edges to better capture context, and consider handling dynamic objects or incorporating richer semantic categories.
- **Evaluation:** Test the approach on benchmark datasets (GQA, Visual Genome), and evaluate using standard metrics such as EM, BLEU ROUGE METEOR BERTScore

Current status:

- **Literature Gathered:** created a table summarizing article title, link, publication date, authors' affiliations, citations, code availability, GitHub stars, key differences & possible weaknesses, datasets, metrics, etc.
- **Initial Summaries:** Ferret, Shikra, Kosmos-2.5, Osprey, Sphinx, etc..
- **Datasets Gathered:** GQA(Graph Question Answering), 3DSRBanch, GRIT, Visual Genome

Q/A

References:

- MLLM Review