UNIVERSITÄT BIELEFELD
Faculty of Technology

Master Thesis

Video-based Retrieval-Augmented Generation

Semantic Computing Group
Jan-Philipp Töberg
jtoeberg@techfak.uni-bielefeld.de

Intelligent agents are challenged by unknown situations in open worlds. They cannot perform everyday tasks like cutting food or pouring drinks without encountering unknown motions, objects or environments. To mitigate this problem, LLMs have been investigated as potential resources to provide these robots with knowledge is a possible way to increase their world understanding and support their planning capabilities [1]. However, these LLMs struggle with hallucinations and a providing knowledge not explicitly represented in their training data.

As a solution, we are investigating **Retrieval-Augmented Generation** (RAG) techniques [2], in which the prompt of an LLM is enhanced by chunks extracted from external resources based on their similarity and relevance for the initial user command. In general, these external resources are text-based, but for providing knowledge relevant for robots solving manipulation problems, we investigated the usage of tutorial video transcripts as possible resources.

In this thesis, you will extend this investigation by developing, evaluating and comparing different RAG pipelines that incorporate knowledge embedded in videos. These pipelines generally follow these three ideas:

1. Using the audio of the video as textual information
2. Using Video-text-to-text models to generate a textual description of the video content
3. Creating a multi-modal embedding of the complete video with all its audio & visual information

No prior knowledge regarding is required. Regarding the programming language, it is advised to use Python. The thesis can be taken in English or German.

Related literature

[1] Y. Ding et al., 'Integrating Action Knowledge and LLMs for Task Planning and Situation Handling in Open Worlds', Auton Robot, vol. 47, no. Special Issue on Large Language Models in Robotics, pp. 981–997, 2023, doi: https://doi.org/10.1007/s10514-023-10133-5.

[2] P. Lewis et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', in Annual Conference on Neural Information Processing Systems (NeurIPS), Curran Associates, Inc., 2020, pp. 9459–9474. doi: 10.48550/arXiv.2005.11401.

The Semantic Computing Group researches and develops methods that enable machines to acquire relevant knowledge as well as linguistic capabilities. Using methods from *natural language understanding* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading unstructured textual data. In particular, the group focuses on methods for information extraction, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at: `http://www.sc.cit-ec.uni-bielefeld.de/`

Interested? @mail to `jtoeberg@techfak.uni-bielefeld.de`