BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Large Language Models for Traffic Surveillance Video Understanding

Michal Pyšík*

Abstract
Traditional automated systems for searching and analyzing objects or events in traffic footage usually cannot handle natural language queries and often rely on predefined event detection. Recent advances in large language models (LLMs), particularly in the area of multimodal learning, offer new opportunities to overcome these limitations. The main aim of this thesis is creating a system that integrates multimodal large language models (MLLMs) and related technologies (multimodal embedding models), which allows users to efficiently search for events/objects in hours of traffic footage and analyze selected video segments in detail. The system also allows users to choose between different models of both types, and the key scientific contributions of the thesis are multiple benchmarks of these models on domain-specific datasets.

*xpysik00@stud.fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The deployment of traffic cameras has been steadily increasing in recent years, driven by advancements in technology and a growing emphasis on road safety and traffic management. Existing automated systems for searching and analyzing traffic footage often rely on predefined event detection methods and don't support natual language user interaction [1].

In the recent years, large language models (LLMs) have gained significant attention and achieved remarkable progress in performance and capabilities. Notably, the capabilities of multimodal LLMs (MLLMs) go beyond text processing, as they also integrate information from other modalities such as images, videos, or audio. The aim of this thesis is utilizing video-MLLMs and related technologies to create a system for searching and analyzing traffic footage.

The system utilizes existing non-specialized AI models and the thesis is not concerned with fine-tuning models for the problem domain. For this reason, a crucial part of the work was researching which models are best suited for the system. Since multiple models were selected for both roles, the system allows the user to switch between them and the thesis contains multiple benchmarks of these models in the domain of traffic videos (see section 4).

2. Proposed solution

Despite the very rapid advancements in their capabilities, current MLLM-based models specialized for videos usually have to make significant optimization compromises, and effectively understanding and navigating long-form videos remains a highly relevant challenge [2].

Multimodal embedding models map inputs from different modalities into a shared embedding space. Some people have already used an image-text multimodal embedding model (e.g., CLIP [3]) in combination with a vector database to create a system for efficient semantic search of events inside a collection of videos [4].

For the target system, a two-stage pipeline was proposed, as shown in Figure 1. In the first stage, a user enters a natural language query, which is converted to an embedding by the multimodal embedding model. This embedding is then compared to those stored in a vector database, which contains image embeddings of frames uniformly sampled from uploaded videos, and the best matches are retrieved.

In the second stage, a short video segment corresponding to a search result (or manually selected) is used as input to the MLLM, which is then used to analyze the footage in the form of an interactive chat

with the user.

3. The implemented system

The system is implemented as a web application using Python, Poetry[1], FastAPI[2], FFmpeg[3], and other technologies. The frontend is implemented using Vue.js[4] as a single-page application (SPA). Additionally, the Milvus vector database is used to store frame embeddings and corresponding metadata (video name, timestamp), and a Minio bucket is used to store the uploaded video files.

As shown in Figure 2, the user can search for sampled video frames by entering a text query and selecting the number of top-k results to retrieve. The embedding search is handled by the Milvus database, and the corresponding images are sampled from a corresponding video stored in the Minio bucket. The user can choose one of four available multimodal embedding models running locally: CLIP [3], ALIGN [5], SigLIP [6] and BLIP [7] (or rather just BLIP's feature extractor).

To start a video analysis using the configured MLLM, the user can either enter the analysis parameters (video name, start and end timestamp of the video segment, sampling FPS) manually, or be automatically redirected by clicking the "analyze this segment" button of a search result (see Figure 2). Once the analysis is started, a video player appears (automatically redirected to the start timestamp) along with an interactive, rather familiar chat interface through which he can interact with the model, as shown in Figure 3. The user can choose one of four available MLLMs, most of which are lightweight versions running locally: LLaVA-OneVision [8], GPT-4o [9] (the only flagship, remotely run model), VideoLLaMA 3 [10], and Qwen2.5-VL [11].

Additionally, the system needs to maintain consistency between video data stored in the Milvus vector database and the video files located in the Minio bucket. For this reason, a synchronization mechanism was implemented, which allows the user to easily view, manage, and synchronize all data present in the system.

4. Benchmarks of the available AI models

The image-text retrieval performance of the multimodal embedding models was benchmarked on the CARS196 dataset [12]. For each name of the 196 unique car models, k best-matching embeddings of the images from the dataset were retrieved. The benchmark was evaluated twice, as the second "sub-benchmark" was focused on whether (at least) the car brand was correct. The used metrics include mean Precision@k, mean Recall@k and mean AveragePrecision@k, all for multiple values of k. Plot of the values of the mean Precision@5 metric for both car models and car brands is shown in Figure 4.

The second benchmark was of the same type as the first one (including the metrics), but with Czech traffic signs and their categories instead of car models and car brands. The images and corresponding annotations come from a dataset provided by the thesis' supervisor, which contains images of traffic signs taken in the Czech Republic.

The MLLMs were benchmarked in the form of video question answering (VQA) on a subset of the SUTD-TrafficQA dataset [13]. The models had to answer 2,000 different questions about corresponding video segments, by choosing one of two to four available answers. Apart from correctly describing what's happening in the given video segment, the questions also included event forecasting, reverse reasoning, and other complex reasoning tasks. The accuracies of the models are shown in Figure 5.

5. Conclusions

A two-step pipeline balancing the strong but expensive capabilities of video MLLMs, with the effectivity of multimodal embedding models was proposed. The system, which allows users to search, analyze, and manage traffic video content was implemented. The implemented system also supports the use of multiple different models of both types.

The performance of the available models was benchmarked in the domain of traffic footage, providing valuable insights into their effectiveness in this domain. The multimodal embedding models were benchmarked on image-text retrieval of car models (plus car brands) and Czech traffic signs (plus their categories). The MLLMs were benchmarked through traffic video question answering.

Acknowledgements

I would like to thank my supervisor Ing. Ondřej Klíma, Ph.D. for their help and for providing (and partially preprocessing) the dataset which contains images of traffic signs captured in the Czech Republic.

[1] https://python-poetry.org/
[2] https://fastapi.tiangolo.com/
[3] https://ffmpeg.org/
[4] https://vuejs.org/

References

[1] Trung Nguyen, Tuan Nguyen, Dinh Nguyen, and Minh-Triet Tran. Traffic video event retrieval via text query using vehicle appearance and motion attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2506–2515, 2021.

[2] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. *VideoAgent: Long-Form Video Understanding with Large Language Model as Agent*, pages 58–76. 10 2024.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[4] @Jesus. Contextual search engine with llm. Medium, August 2023. Accessed: 2025-02-17.

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

[6] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Los Alamitos, CA, USA, October 2023. IEEE Computer Society.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

[9] OpenAI Et al. Gpt-4o system card, 2024.

[10] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025.

[11] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[12] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[13] Li Xu, He Huang, and Jun Liu. SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Video Reasoning Over Traffic Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888, June 2021.