Next-Level Video Representations

Master Project Winter 2024/2025 – AI and Intelligent Systems Group

Project Description

Motivation. Processing videos is computationally expensive due to the large amount of frames, especially for longer video clips. With the rise of advanced video processing techniques, there is a growing need to develop models that can effectively handle and analyze video data. While Transformer- and CNN-based models for single images have achieved strong performance on a variety of academic benchmarks as well as in industrial deployment, understanding the content of videos still poses a challenge. Previous approaches have mostly utilized 3D-convolutions or augmented commonly used image models to process the additional time dimension. To mitigate the usual large model sizes for video representation encoding, we want to develop a new video foundation model that is parameter-efficient and can achieve state-of-the-art results on video understanding tasks such as action recognition and video question answering.

Goals. This project aims to push the boundaries of video representation learning by training a new video foundation model. Specifically, we will focus on predictions in the embedding space and explore if this can encourage the model to work more on a conceptual level instead of focusing on low-level features.

Approach. A visualization of the proposed method is shown in Figure 1. The model will leverage parallelization first to encode short pieces of a video individually and then contextualize the resulting vector embeddings over the entire range of the video length. The pretraining objective is masked embedding prediction on the encoded sequences, i.e., the top model's objective is to reconstruct masked embeddings of incoming clips based on the other embeddings in the context. Formulating the objective on the embedding space allows the model to abstract away from pixel-level details and outsource the lower-level processing to the pretrained encoder. The resulting embeddings (c_i in the figure) also yield *contextualized vectors*.

We will evaluate the model on several video understanding tasks and compare it to prior video models.

Research Questions

- How do pretrained image models perform in encoding sequences of video frames?
- How can we process videos efficiently, e.g., without processing every frame, under low computation requirements?
- What are the best practices for pretraining and fine-tuning video representation models on downstream datasets with temporal dependencies?
- How effective is the NextLevelModel in masked frame embedding prediction for video data?

Outline and Setup

The following working packages and project phases are envisioned for the course of this project

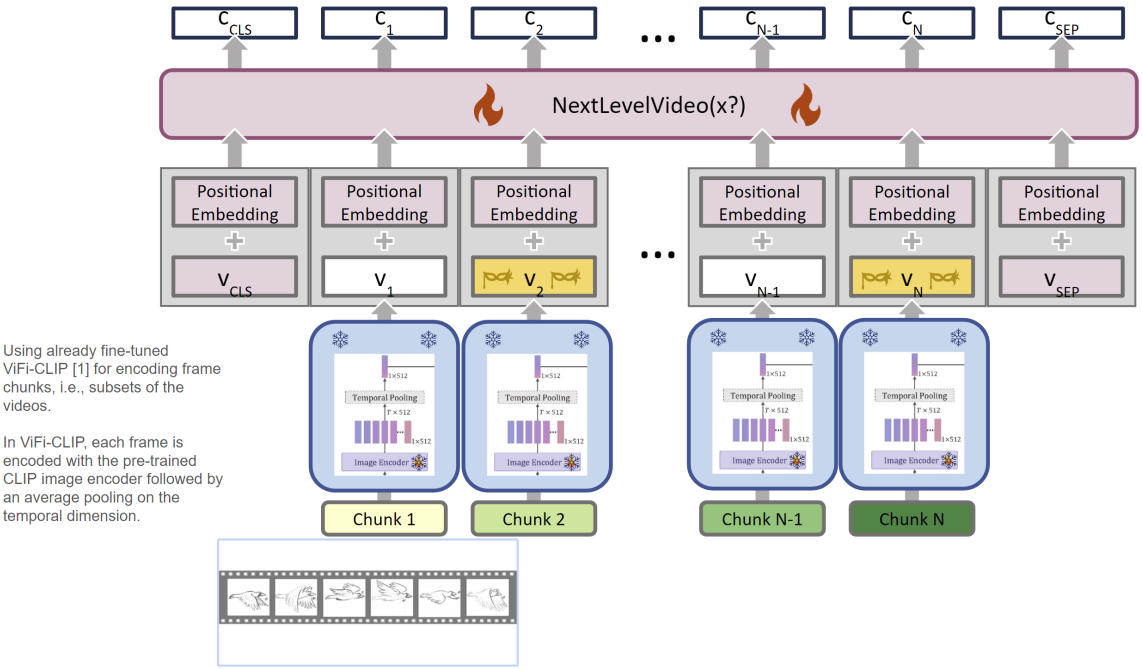

Figure 1: NextLevelVideo: a smaller encoder model embeds shorter video clips in parallel while the top model connects these on a larger semantic level.

Phase 1: Literature Review and Dataset Analysis

- We will dedicate the first few weeks to the literature review on the recent advancements in video understanding and foundation models.
- Candidate datasets and tasks for training and testing the NextLevel model will be finalized.
- Exploratory data analysis will be performed on the datasets.

Phase 2: Preprocessing

- Down-sampling developments: To have a more efficient processing pipeline, we aim at reducing the number of frames in videos, e.g., consecutive frames with very few differences. Different down-sampling strategies will be visited and experimented with at this stage of the project. The goal is to find a down-sampling technique that preserves the performance in relevant tasks such as action recognition.
- Finalizing Chunking Strategies: As illustrated in Figure 1, videos are firstly divided into chunks of sequential frames. At this stage, we explore different chunking strategies, e.g., based on some frames or changes in the distribution of pixel values, etc.

Phase 3: NextLevel Training and Testing

- Training the NextLevel model described in Figure 1 using a masked embedding prediction regression loss function.
- Measuring the performance on a range of video tasks: From action recognition to video clustering to the possible connection with LLMs to enable video question answering.
- Measuring computational requirements in terms of a number of trainable parameters, FLOPs, etc., in the training and testing time.

Phase 4: Iterative Refinement

- Based on the performance of the first iteration on the downstream tasks, we apply refinements for adjusting the loss function.

What will you gain?

By participating in this project, you will gain valuable hands-on experience in deep learning for video representation learning. Furthermore, you will experience a complete research pipeline scenario, starting from ideation, exploration, implementation, result analysis, and iterating over the pipeline. Lastly, you will be part of a fun working environment in our group with social activities and —since recently— an excellent coffee machine!

Contact

The Artificial Intelligence and Intelligent Systems group will supervise this project. For more information, contact:

- Prof. Gerard de Melo: gerard.demelo@hpi.de
- Sarah: sedigheh.eslami@hpi.de
- Tamara: tamara.czinczoll@hpi.de
- Babajide: babajide.owoyele@hpi.de

Related Work

Video foundation models have been used in various tasks such as video question answering [3], retrieval [5] and captioning [13]. Previous work has aimed at developing foundation models for video understanding tasks by refining or adapting the pre-trained CLIP [9] vision-language model. There are numerous commonly known examples of this line [10, 6, 12, 5, 7], but also others like VideoMamba, TubeViT and VideoMAE [11, 8, 4] adopting new strategies. Furthermore, the masked embedding prediction of the NextLevelModel has been tested in the natural language domain [2] and showed strong performance while being very parameter-efficient. Similar approaches also exist in the vision domain for image encoding, [1, among others], but not yet for video.

Requirements

This project is suitable for students with previous experience in deep learning, ideally computer vision. The knowledge of python and at least one deep learning framework, e.g., PyTorch, HuggingFace, TensorFlow, Jax, etc, is strongly desired. Ideally, we are looking for a group of 3–5 students.

References

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

[2] T. Czinczoll, C. Hönes, M. Schall, and G. de Melo. Nextlevelbert: Investigating masked language modeling with higher-level representations for long documents. *arXiv preprint arXiv:2402.17682*, 2024.

[3] D. Ko, J. S. Lee, W.-Y. Kang, B. Roh, and H. J. Kim. Large language models are temporal and causal reasoners for video question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[4] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao. Videomamba: State space model for efficient video understanding, 2024.

[5] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[6] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery.

[7] N. Madan, A. Møgelmose, R. Modi, Y. S. Rawat, and T. B. Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.

[8] A. Piergiovanni, W. Kuo, and A. Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning, 2022.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[10] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6545–6554, June 2023.

[11] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, June 2023.

[12] M. Wang, J. Xing, and Y. Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[13] Y. Wang, D. Gao, L. Yu, W. Lei, M. Feiszli, and M. Z. Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pages 709–725. Springer, 2022.