COSE474-2024F: Final Project Proposal: Scene Understanding of Household Activities

Andi Analta Dwiyanto Palalangan Tunru [1]

1. Introduction

Interpreting household activities in real-time has significant implications for various applications such as smart home systems and elderly care. With this project I aim to develop a system that can recognize and describe household activities, enabling more intuitive human-machine interaction. The motivation behind this project is to address the unique needs of households and provide a system that enhances safety and well-being in home environments.

2. Problem Definition & Challenges

The problem is to develop a real-time scene understanding model capable of recognizing and describing household activities. This involves detecting and interpreting various actions and objects present in the scene, such as cooking, cleaning, or interacting with household items. The goal is to generate accurate textual descriptions or labels for these activities, providing contextually relevant insights

2.1. Challenges

1. **Complexity of Household Activities** : Household scenes can be cluttered and involve multiple overlapping objects and actions, making activity recognition non-trivial.
2. **Generalization to Diverse Environments** : The model needs to generalize across various home setups, lighting conditions, and different types of furniture and objects.
3. **Temporal Coherence** : Understanding activities requires not only recognizing objects but also understanding their interaction over time, which is crucial for activities like ”entering the room” and ”turned on the television”
4. **Hallucination** : In video captioning, the ”hallucination” problem occurs when a model generates descriptions that include objects, actions, or scenes that are not present in the video. This happens when the model relies heavily on its training data and learns correlations that might not be directly applicable to the specific video being captioned. As a result, it may introduce elements into the captions that seem reasonable based on previous patterns but are inaccurate for the given input, making them misleading.

3. Related Works

1. **Temporal Reasoning Graph for Activity Recognition** (Jingran Zhang, 2019) Introduces a novel approach to video-based activity recognition that addresses the crucial need for effective temporal reasoning. The authors propose a Temporal Reasoning Graph (TRG) framework, which constructs learnable graphs to capture multi-scale temporal relations in videos. Evaluated on datasets like Something-Something and Charades, the TRG achieves state-of-the-art performance, demonstrating its ability to extract highly discriminative features for activity recognition.

4. Datasets

Charades is a dataset designed for video-based understanding of human activities, with a focus on activities that occur in **indoor household environments**. It contains a large collection of **video clips** that depict people performing everyday activities in homes, such as cooking, cleaning, watching TV, reading, and interacting with objects.

5. State-of-the-art Methods and Baselines

1. **VideoCLIP** (Hu Xu, 2021) extends the capabilities of CLIP to video data, allowing it to comprehend and align video frames with text descriptions. By learning temporal relationships between sequences of frames and their associated textual context, VideoCLIP can recognize complex actions and interactions over time, making it ideal for tasks like action recognition and video retrieval. This ability to understand sequences of events makes VideoCLIP effective for scene understanding in dynamic environments, such as analyzing activities in videos or understanding interactions in household settings.

2. **LLaVA-NeXT** (Bo Li, 2024) is a large multimodal model (LMM) that enhances scene understanding and video captioning by processing visual data from images, videos, and 3D environments. Using the ”AnyRes” technique, it breaks down high-resolution inputs into smaller patches, making it ideal for capturing details in dynamic scenes. Its ability to generalize across longer video sequences enables the model to produce more accurate and descriptive captions, which

is crucial for understanding the temporal progression of activities in videos. Additionally, LLaVA-NeXT uses Direct Preference Optimization (DPO) to refine its output with AI feedback, leading to better alignment with user instructions in tasks like video analysis and captioning

References

Bo Li, Yuanhan Zhang, D. G. R. Z. F. L. H. Z. K. Z. Y. L. Z. L. C. L. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Hu Xu, Gargi Ghosh, P.-Y. H. D. O. A. A. F. M. L. Z. C. F. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Jingran Zhang, Fumin Shen, X. X. H. T. S. Temporal reasoning graph for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 123–132, 2019.