VideoLLM for Understanding Activities of Daily Living in Elderly Care

Sindhu Gadiraju, UNC Charlotte
Dr. Srijan Das, College of Computing and Informatics

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

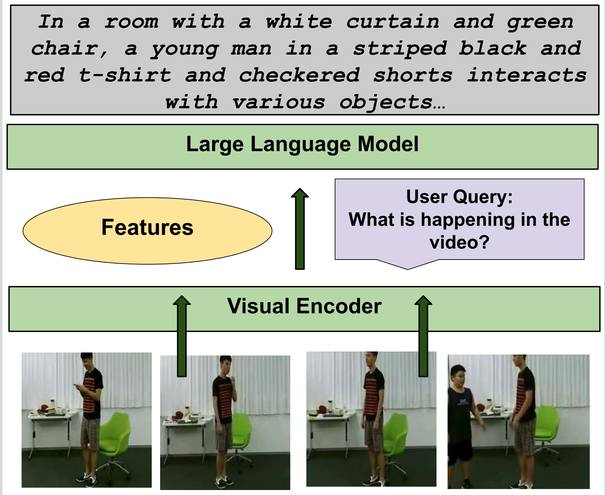
Introduction

- It is anticipated that by the year 2050, one-sixth of the global population will surpass the age of 65, leading to an escalating demand for healthcare services.
- ADL (Activities of Daily Living) refers to routine activities that individuals perform independently to take care of themselves, including tasks like bathing, dressing, eating, toileting, and mobility.
- ***Challenges in ADL/Current Models:***
  - Modeling time - Spontaneous behavior
  - Fine-grained activities (subtle motion)
  - Video conversational models are trained on web videos

Objectives

- Develop a visual model that can accurately classify and understand human actions performed in videos
- Videos particularly focus on depicting daily activities of elderly individuals such as walking, eating, drinking, and other routine tasks.
- The model incorporates both visual and large language models to enhance its capacity for generalization
- Integrate the model into smart home environments for continuous monitoring and assistance.

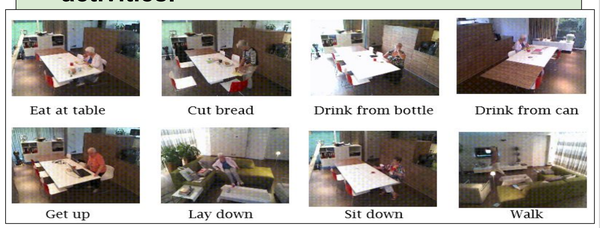
Method



- Input short video clips into the visual encoder
- Utilize features and process user query into the Large Language Model
- Output a detailed description of the video

Collected Data

- Public datasets utilized for training the model include NTU 120, Charades, and PKU-MMVID.
- Train model to recognize and understand fine-grained actions, particularly those related to Activities of Daily Living (ADL).
  - Fine-grained action recognition involves distinguishing between very similar activities.
- Many existing models struggle with recognizing fine-grained actions due to their focus on more distinguishable activities.



Results

SAMPLE VIDEOS

BASELINE RESULTS | ***SUMMARIZE THE VIDEO*** | **OUR MODEL**

The video shows a man standing in a bedroom and opening a closet door. He then pulls out a book from the closet and begins to read it. The video also show him putting the book back in the closet and closing the door.

What actions are performed in the video?

In the video, the man is seen opening a closet door, pulling out a book, reading it, and then putting the book back in the closet.

Does not understand fine-grained actions

Confuses objects like medicine and phone as book

The video captures a scene set in a bedroom, where a man, possibly dressed in a gray shirt and pants, is engaged in various activities involving an open cabinet or wardrobe. He is seen bending over, reaching into, and examining items within the cabinet, which contains an assortment of objects including books, a vase, decorative figurines, and possibly some electronic devices or a television set. In one moment, he holds a bottle, possibly pouring a liquid into a container on the shelf, and in another, he is seen holding and looking at a remote control, suggesting he might be interacting with an electronic device inside the cabinet. The room is characterized by its beige walls, a blue curtain, and a bed adorned with a blue and white checkered bedspread. On the floor next to the bed lies a red bag, adding a splash of color to the scene. Miscellaneous items, such as a black dresser and personal belongings, are scattered around, contributing to the lived-in feel of the space.

Conclusions

- Developed a first-ever video conversational model for ADL, integrating LLM and visual understanding techniques.
- Evaluated the effectiveness of the model on a public ADL dataset, marking a significant advancement in understanding and analyzing daily activities.

Future Research

- Refinement of Fine-Grained Action Recognition
- Real-Time Monitoring in Smart Environments
- Integration with Telemedicine and Remote Monitoring

References

Das, S., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., & Francesca, G. (2022). Toyota Smarthome: Real-World Activities of Daily Living. https://arxiv.org/abs/2010.14982
Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., & Kot, A. (2019). NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. https://arxiv.org/abs/1905.04757
Maaz, M., Rasheed, H., Khan, S., & Khan, F. (2023). Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. https://arxiv.org/abs/2306.05424