A Second Pair of Eyes: Composing Foundation Models for Egocentric Video QA

By: Ronak Malde, Arjun Karanam

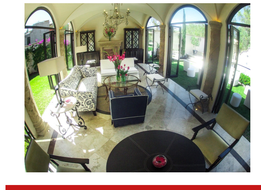
CS 231n

Introduction

Today, Augmented Reality hardware appears just on the horizon, with devices like the Meta Quest Pro and the AppleVision Pro available to consumers. These devices all have cameras pointing out into the world, but as of now, they're just used for features such as hand and environment tracking. We'd like to see if these cameras can act as a "second pair of eyes", using new developments in LLMs and Vision models to help you process and learn about the world around you. To achieve this goal, we plan on building a model that, given an egocentric video, can answer questions about the scene that it sees.More explicitly, we wish to build a model such that given a question and a video, it can output the correct answer to the question given a set of answer choices.

Problem Statement

At its core, our problem can be framed as a Video Question-Answering problem. We are seeking to take as input the video feed of what the user is seeing, as well as questions that the user is posing, and return a valid answer to that question. Framed a bit more precisely, the goal of the Video QA algorithm is to predict an answer a* given a video V and a question Q. When predicting a*, the model is given model multiple choices to choose from, such that a* is in A (where A is a set of potential answers). Correctness is computed based on the model's output answer a' to the correct answer a*.



What am I looking at?

What color is the couch?

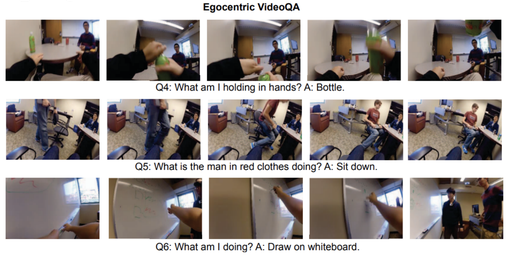
Am I inside or outside?

Dataset

Egocentric Data

Characterized by: First-person Perspective, User's Context, and Lack of Data

EgoVQA Dataset[1]



500+ Videos

500+ questions, w/ 5 answer choices

Train / Val / Test
250 / 150 / 120

Related Works

Traditional Approaches

TGIF, 2017[2]: State of the art model of the time in Video-QA, used ResNet + LSTM with attention architecture

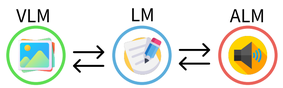
Data augmentation techniques for Video QA, 2020[3]: Saw large improvements in video question benchmarks using data augmentation techniques

(2.5+1)D Spatio-Temporal Scene Graph, 2022[4]: Used a 2.5D scene graph to encode motion features and a transformer-based reasoning pipeline, surpassed previous SOTA results

Foundation Model Approaches

OFA (Once-For-All) Model, 2022[5]: Multimodal transformer-based foundation model trained on many tasks, performs at or above SOTA for all cross-modal tasks

Socratic Models, 2022[6]: Framework for chaining together foundation models for different domains to run zero-shot inference, achieves results comparable to models specially trained on those tasks. Uses natural language is an intermediate embedding between models



Methods

Baseline

For the baseline we simply used a pre-trained ViT Model, to generate a caption, and then fed that caption + the question into GPT-3.5, and prompted it to pick an answer.

OFA Model - PromptCap

Disadvantage of baseline: Question not taken into account when generating the caption
Solution: Once-For-All Model (OFA) multimodal, encoder-decoder model. Specifically, we use a variation called PromptCap[7] which is trained on the downstream task of generating captions

Finetuning PromptCap

To account for Egocentric data, we fine tune PromptCap by freezing all layers except for the transformer layer in the Encoder, as we hypothesize this will allow the model to learn to better attend to first person relations.

Pipeline Variations

Caption Generation

Caption Only

Input to VLM: What does this image describe?

Caption + Question

Input to VLM: Describe the image in a way that would help someone answer the question: [question]?

Caption + Question + Choices

Input to VLM: Describe the image in a way that would help someone answer the question:[question]? Where the answer choices for the question are: [choices]

Direct Question

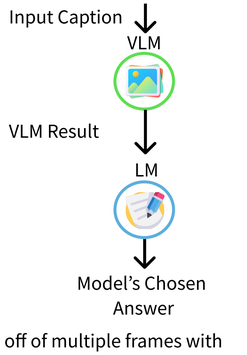
Question Only

Input to VLM: [question]?

Question + Choices

Input to VLM: [question]? Your choices are: [choices]

Generalized Pipeline



Augmenting World State

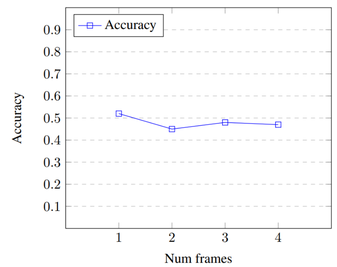
Finally, we explore adapting our methods to video, by constructing the world state off of multiple frames with timestamps, to encode more detail and information about motion throughout the video.

Experiments

Model Variations

Model	Accuracy
Baseline	0.12
Prompt (Unfinetuned)	0.17
Prompt	0.22
Prompt + Question	0.29
Prompt + Question + Choices	0.30
Question	0.38
Question + Choices	**0.52**
ST-VQA (original paper SOTA)	0.38

Augmenting World State



Analysis

We see that all variations of the PromptCap model per-formed better than the baseline and better than random guessing (20%). Giving the Prompt-Cap model the questions and the answers but no additional prompt yielded the best results of 52%. As an interesting comparison, our best-performing model achieves significantly higher results on the task compared to the original SOTA model from the dataset paper, ST-VQA. The results show that giving the PromptCap model the most amount of relevant information about the question andthe answer choices yields the best results.
Additionally, its eems that additional semantics in the prompt actually decrease the overall accuracy. This also aligns with the way the PromptCap model finetuned OFA, in which it was given just the barebones question.
Finally, our experiments showed that our methods for augmenting the world state actually decreased model performance, and using one frame achieved better results.

Future Work

1 - Explore temporal reasoning to better answer questions dealing with motion (i.e I am sitting down, or I am using my phone)

2 - Build a demo using real hardware, and explore additional egocentric challenges that arise (noisy data, long time sequences, memory, etc.)

References

[1] - C. Fan, "EgoVQA - An Egocentric Video Question Answering Benchmark Dataset," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019
[2] - Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1359-1367.
[3] - Falcon, A., Lanz, O., & Serra, G. (2020). Data augmentation techniques for the Video Question Answering task. ECCV Workshops.
[4] - Cherian, A., Hori, C., Marks, T.K., & Roux, J.L. (2022). (2.5+1)D Spatio-Temporal Scene Graphs for Video Question Answering. ArXiv, abs/2202.09277.
[5] - Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., … Yang, H. (2022). OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. ICML 2022.
[6] - Zeng, A., Wong, A., Welker, S., Choromanski, K., Tombari, F., Purohit, A., ... & Florence, P. (2022). Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598.
[7] - Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N. A., & Luo, J. (2022). PromptCap: Prompt-Guided Task-Aware Image Captioning. arXiv preprint arXiv:2211.09699.