Language as the Medium:

Multimodal Video Classification through Text only

Laura Hanu, Anita L. Verő, James Thewlis

{laura,anita,james}@unitary.ai

Introduction

Hypothesis: Can we use textual descriptions alone as the medium to convey visual and audio information to an LLM?

Motivation:

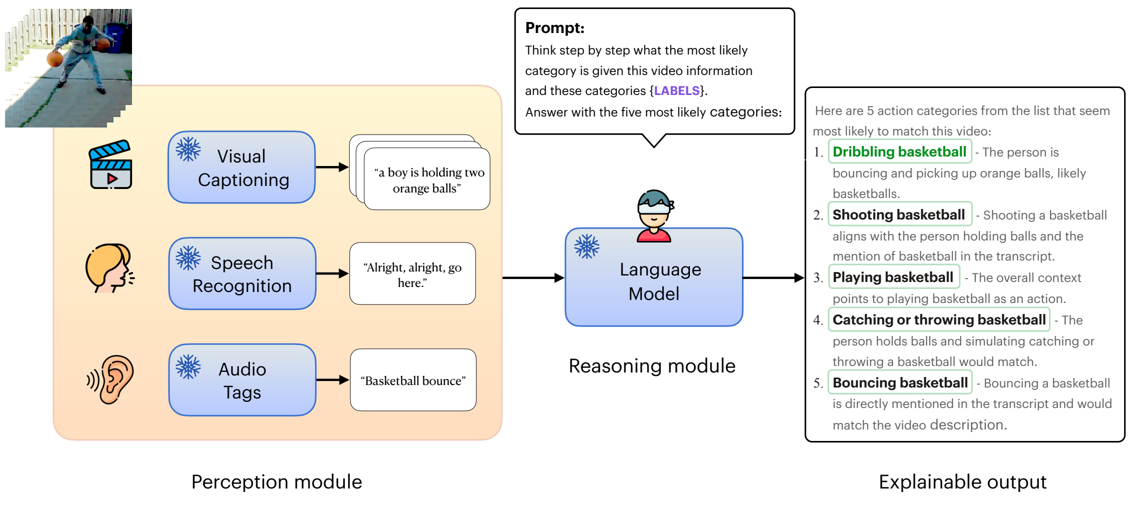
- Leverage the general knowledge of LLMs for better contextual video understanding
- Plug & play different perception or reasoning models
- No training needed

Contributions

1. We introduce a **new multimodal zero-shot video classification approach** consisting of:
  a. a **"perception"** phase where specialised models act as sensory proxies
  b. a **"reasoning"** phase where an LLM is used to analyse these multimodal textual clues in order to classify a video.

2. We demonstrate that LLMs can use these **multimodal textual clues** as proxies for "sight" or "hearing" and classify videos **in-context**.

Method



Perception models

Video: We extract visual captions from video frames, with BLIP-2 [1].

Audio:

- We use **Faster Whisper** [2] to obtain audio transcripts.
- We leverage **ImageBind** [3] to get audio embeddings and compute the similarity with the textual embeddings of the AudioSet labels.

Reasoning models

- **GPT3.5-turbo**
- **Claude-instant-1**
- **Llama2 - Llama-2-13b-chat variant** (13B parameters) [4]

Structured Output

To convert free-flowing natural language outputs to 5 ranked class names:

- **GPT API: JSON Schema** feature
- **Claude**: ask for the results to be returned as JSON
- **LLama2:** Parse the observed numbered list in the output

Experiments

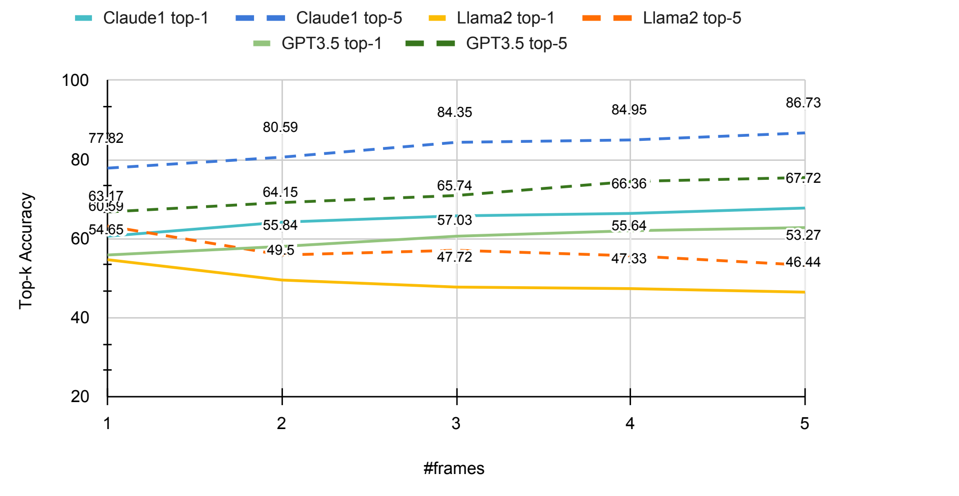
Comparing different levels of context using Claude-instant-1

	UCF-101		Kinetics400 (subset)	
Model	**Top-1 Acc.**	**Top-5 Acc.**	**Top-1 Acc.**	**Top-5 Acc.**
BLIP2(FlanT5-XXL)+Claude-1(caps)	63.01	85.35	38.90	54.20
BLIP2(FlanT5-XXL)+Claude-1(caps, speech)	67.06	86.13	**41.20**	57.00
BLIP2(FlanT5-XXL)+Claude-1(caps, speech, audio)	**67.13**	**86.15**	**41.20**	**57.35**

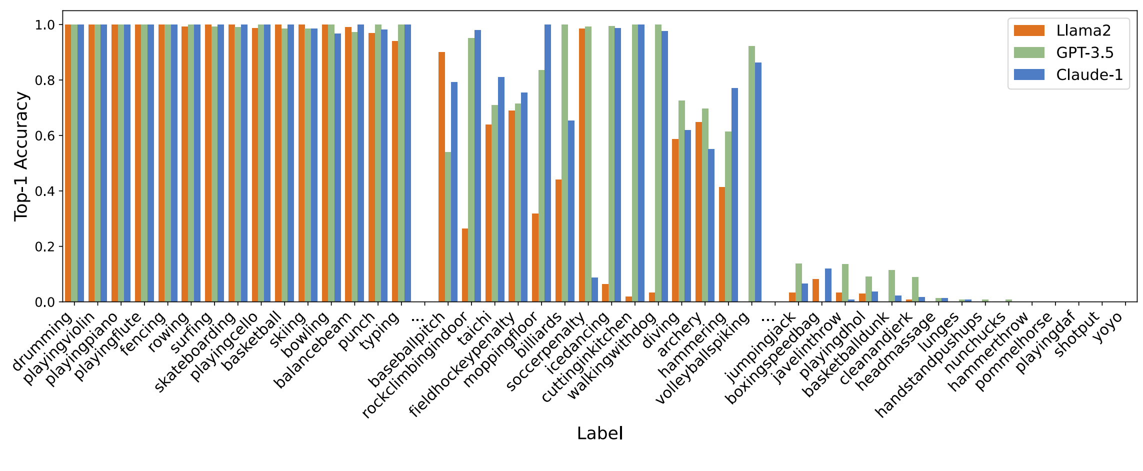
Comparing varying LLMs on the UCF101 test set

Model	Top-1 Acc.	Top-3 Acc.	Top-5 Acc.
BLIP2(FlanT5-XXL)+Llama2-13B	49.56	56.70	58.51
BLIP2(FlanT5-XXL)+GPT3.5	66.37	79.27	82.04
BLIP2(FlanT5-XXL)+Claude-1	**63.01**	**81.49**	**85.35**

Varying number of frames per video



Best and worst performing UCF101 classes



Discussion and Future Work

Limitations:

1. Separate models for vision and speech might **not capture inter-modal interactions**.
2. Frame-by-frame image analysis doesn't account for **temporal relationships** or persistent identities.
3. Generative models can produce **hallucinations** and **unreliable outputs**.
4. Performance **not yet on par with state-of-the-art** zero-shot benchmarks.

Future work:

1. Leveraging **additional video context**, such as user comments
2. Try a **chat-based approach** where the "reasoning" module can ask the "perception" module for clarification to get more information

References

1. Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *arXiv:2301.12597* (2023).
2. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." ICML, 2023.
3. Girdhar, Rohit, et al. "Imagebind: One embedding space to bind them all." IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
4. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv:2307.09288* (2023).