Aligning Multi-Modalities as a Unified Vector Space

Ishaan Singh ishaans@usc.edu
Astha Kinra kinra@usc.edu
Abhigyan Singh abhigyan@usc.edu
Kusum P Grandhi kgrandhi@usc.edu
Prajwal Gupta prajwalg@usc.edu

Abstract

This project enables the development of an Intelligent Memory Recall Assistant that enhances users' ability to store, organize, and retrieve multimodal information, such as videos, audio, and text entries.

1 Introduction

This project is focused on multimodal retrieval of **"activities"** by leveraging advanced NLP models.

That said, our model aims to be able to map **"activities"** into a common semantic vector space. An **activity** can be recorded in any of the three modalities. This vector space will support context-based retrieval (across all 3 modalities - video, audio, text).

2 Research Question

It is difficult to remember everything. It is even more difficult to retrieve from memory, with a lot of thought clutter.

While we extend (Hassan Akbari, 2021) to use long data formats, is it still possible to capture a common semantic vector representation for varying modalities?

BASELINE: (Hassan Akbari, 2021)

3 Relation with NLP - CSCI 544

This project focuses on utilizing NLP for information extraction, summarization, and context-based retrieval. NLP models can process and understand user inputs, categorize them, and help in generating relevant responses, making the memory recall process efficient and personalized.

4 Related Work - Models

- **Transformer-Based Models for Multimodal Representation:** (Hassan Akbari, 2021) represents the closest work to our project. VATT uses transformer-based architecture to map video, audio, and text into a shared vector space.
- **Multimodal Embeddings:** Some notable ones include (Alec Radford, 2021) for image-text alignment and (Alexei Baevski, 2020) for speech processing. Our work will draw from these principles to handle videos, audios, and texts in a unified space.

5 Related Work - Datasets

YouTube-8M (Sami Abu-El-Haija, 2016) is a large-scale video dataset with millions of labeled video segments, ranging from 1 to 10 minutes in length. **TED-LIUM v2 (Rousseau et al., 2014)** releases a collection of 1,495 audio talks with corresponding transcripts. **Wikipedia Database (Various, latest)**, filtered on education articles only.

6 Methodology

6.1 Data Preprocessing

In the preprocessing stage, the raw videos, audios, and text entries will be standardized.**Videos** will be segmented into activity-focused clips, and preprocessed to match the (Sami Abu-El-Haija, 2016) dataset. **Raw audio recordings** that are independent of video will not be preprocessed as followed by (Hassan Akbari, 2021).**Text entries**, such as journals or notes, will be handled by tokenization and alignment techniques like WordPiece or Byte-Pair Encoding (BPE) to generate tokens preserving context and structure.

This step is crucial for ensuring that all modalities are in sync, facilitating accurate mapping during the embedding process.

6.2 Model Architecture and Training

The model will employ a transformer-based architecture, drawing heavily on the principles used in (Hassan Akbari, 2021) - Figure 1.

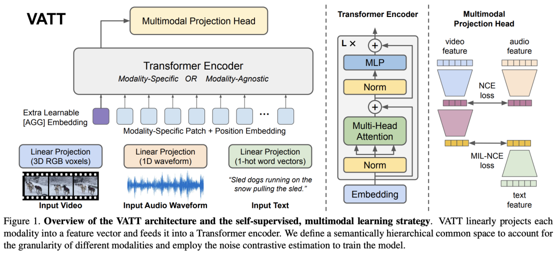

Figure 1: Model architecture in (Hassan Akbari, 2021)

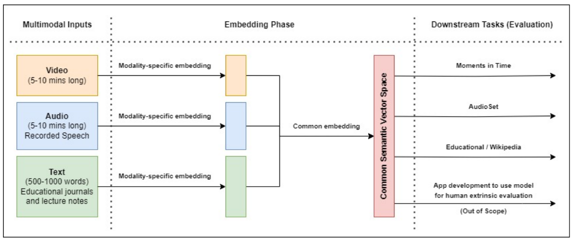

Figure 2: The project procedure -

- **Modality-specific linear projections** will be applied to extract embeddings.
- These embeddings will then pass through **positional encoding** to retain temporal or sequential information as shown in Figure - 2.
- The model will rely on **contrastive learning** to train these embeddings, as in (Tomas Mikolov, 2013).

This architecture will map the inputs to a shared semantic vector space, ensuring that activities with similar semantic meaning are located close to one another in this space.

6.3 Evaluation

There are 2 ways to evaluate this embedding model:

- We will use downstream tasks:
  - Video: (Monfort et al., 2019) releases an open dataset of video+audio clips that have also been used in (Hassan Akbari, 2021) comparing our model against the baseline on precision, recall, and accuracy.
  - For text and audio: we evaluate via other audio-to-transcription libraries or models to see if we do well on those tasks. The baseline (Hassan Akbari, 2021) uses (Gemmeke et al., 2017) and we intend to use the same.
- External evaluation: We will use human evaluation as well to benchmark the retrieval.

6.4 Foreseeing - Computing Resources

This project is a compute-intensive task. Training the (Hassan Akbari, 2021) baseline transformer on short-format media took a lot of resources on Google's end, and we wish to extend that model to train on larger formats. CARC will be needed for our task.

References

Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. 2021. Learning transferable visual models from natural language supervision. *Neural Information Processing Systems*.

Abdelrahman Mohamed Michael Auli Alexei Baevski, Henry Zhou. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Rui Qian Wei-Hong Chuang Shih-Fu Chang Yin Cui Boqing Gon Hassan Akbari, Liangzhe Yuan. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2014. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3935–3939, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joonseok Lee Paul Natsev-George Toderici Balakrishnan Varadarajan Sudheendra Vijayanarasimhan Sami Abu-El-Haija, Nisarg Kothari. 2016. Youtube-8m: A large-scale video classification benchmark.

Greg Corrado Jeffrey Dean Tomas Mikolov, Kai Chen. 2013. Efficient estimation of word representations in vector space.

Various. latest. Wikipedia encyclopedia, filtered on educational articles only.