CS 195: CLIP Zero Shot

Syed Ashal Ali

PAC Lab, Spring Quarter 2024

1 Introduction: Purpose

There are two types of model approaches that we are considering for the VLM architecture. The first type involves only image-captioning (so have image and text pairs as input) with potential models including BLIP-2, TinyLlava, and VILA. This is the type of model architecture that is being explored by Shreyas. The second type of model architecture involves image and video captioning (images and videos can be used as inputs along with the associated text captions) which is implemented through the "Frozen in Time" architecture. I am responsible for exploring the potential of this architecture.

2 Our Approach

For this quarter, my primary task was understanding the Frozen in Time architecture and learning how to implement it by establishing a pipeline that performs zero-shot inference on the PubMed dataset that we are using for training the VLM.

2.1 What is Frozen in Time?

Bain et al. created the Frozen in Time architecture to allow for image and video inputs. Zane worked on this architecture during his internship at Microsoft and understanding his code base thoroughly was what I spent most of my time on this quarter so that implementing future steps becomes easier with the groundwork laid. Fig. 1 displays a brief overview of the architecture provided by the Frozen in Time framework. The architecture is trainable on video and image datasets (which are the medium echo datasets exist in). Frozen in Time is able to achieve this by treating images as 'frozen' snapshots of a video and then learns to increasing temporal context and learn from the video input received.

The architecture has a visual and text encoder. Input for visual is an image or a video clip and the text encoder takes in a tokenized sequence of words. The video clip inputs are divided into non-overlapping spatiotemporal patches which are processed through a 2D Convolutional Layer. The resultant embedded sequences are used as transformer input. Both text and video encodings are projected to a common dimension via single linear layers. Finally, cosine similarity computed between text and video to get the zero shot alignment score.

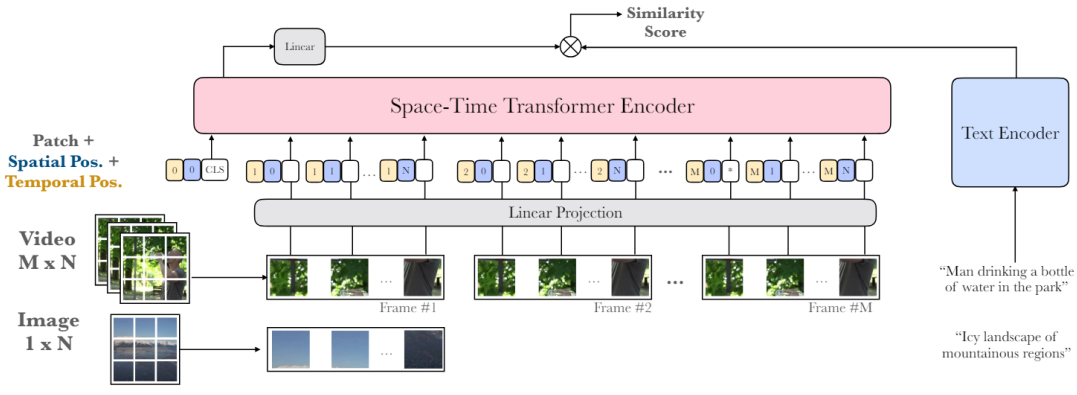

Figure 1: Frozen in Time model architecture from Bain et al.

2.2 Implementation/Methodology

The first part of my task involved developing a strategy to set up Zane's code base in a way that it could be used on our dataset. In order to achieve this, Zane and I met several times in order to go through any challenges that I am facing in understanding his code. Zane was a pivotal resource to get unstuck since he had the best understanding of the different parts of the code that he had wrote.

Firstly, I loaded in the pre-trained CLIP model that the code base was using from Hugging Face. Next, I coded two different data loaders. The first data loader was responsible for reading in and storing the different image paths. The second data loader was responsible for reading in and storing the text captions associated with each image. After this, I converted the loaded in data in torch tensors so that I could work towards generating embeddings for the images and the text captions. Using the pre-trained CLIP model, I generated the matrices representing the image and video embeddings using the provided functions in the CLIP model. Lastly, I computed the cosine similarity in order to receive the zero-shot inference or alignment score.

2.3 Results

The zero-shot obtained was 67.73%. However, the most significant result of this task was having the pipeline set up correctly for the steps that are to come.

2.4 Challenges

This was the first time where my code was entirely dependant on someone else's. At first, my task was to directly change the code base with Biomed-CLIP. This task was incredibly challenging as I struggled to understand what was going on in the code base. Over the 3 weeks that I worked on this older task, it was difficult to tell where everything was located and where my code was going wrong. As I was not making much progress with this, I consulted with Chieh-Ju and broke down my older task into first getting the unchanged code base to work (performing the zero shot that I discuss in this paper), and then using the established pipeline to integrate Biomed-CLIP into the pipleine (performing the zero shot that I discuss in future steps). To overcome the challenge of understanding the code base completely, I tried being very active with my communication. On top of the weekly meeting that Chieh-Ju, Shreyas, and I have, I would meet separately with Zane and Chieh-Ju to discuss my progress and resolve any questions relating to parts of the code base that were giving me errors.

3 Future Steps and Conclusion

With the basic pipeline and primary groundwork established this quarter, I have started making progress on my summer goals. Currently, I am working on replacing the visual and text encoders in the Frozen in Time architecture to Biomed-CLIP. After I perform the zero-shot inference with the new encoders in place, the future steps will depend on the results. If optimal, I will proceed with fine-tuning on real echo images and videos. If sub optimal (considering that the linear projection layers were initiated with random weights), I will fine tune Biomed-CLIP on the PMC-OA subset to align the linear projection layers. Finally, the results of the different approaches we are using (image captioning vs image and video captioning) are going to be compared. Based on our results, we will finalize the model architecture for our VLM and continue making progress on the project.

4 References

1. Bain, M. et al. (2022). Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. arXiv. https://arxiv.org/abs/2104.00650