Analysis of Video Data Using Multimodal Embeddings

Jonathan Doughty (02042009) William King (01974615) Carlo Reyes (02024264) Kevin Tran (01770539)

Abstract

This project presents an interactive visualization for video embeddings generated from a video dataset. UMAP was used for dimensionality reduction to map high-dimensional video embeddings into a 2D space. HDBSCAN was used for clustering to aid in a meaningful visualization of the semantic relationships between videos. In this interactive display, qualitative analysis can be made on the effectiveness of the embedding model. Additionally, we analyze distance-related metrics between videos and their corresponding human descriptions for a quantitative analysis.

1. Introduction

Normally a large set of video data would take a lot of time and resources to analyze and understand. Traditional methods often struggle with the efficiency needed to handle datasets on a large scale. Our project addresses this by demonstrating methods of analysis on a video dataset including clustering, 2D latent space visualization, and semantic search functionality. The objective is to demonstrate the power of embedding-based representation for video data. By converting videos into embeddings, these models can support analytical use cases like intel gathering and event detection.

1.1. External Applications

Additionally, these models have academic and recreational uses like content recommendations and tasks that require semantic understanding. Recommendation systems on YouTube or TikTok rely on similar embedding models to group similar content. The key challenge lies in creating meaningful embeddings that accurately represent the content. Our analysis focuses on Google's multimodal embedding model, aiming to determine how well it creates semantical embeddings of multimodal data. We found that the model performed fairly well classifying videos in their expected clusters.

2. Method

To first process our data, we passed our videos through Google's multimodal embedding model to produce 1408-dimension embeddings. We also made embeddings from the video's text descriptions to later use in our distance metrics analysis.

2.1. Evaluation Approach

Our evaluation approach started with embedding videos along with their corresponding text descriptions in a common vector space. This would allow for us to then observe distance metrics between videos and their corresponding text embeddings. We also used an MLP classifier to analyze the efficacy of our model. This model would classify our embeddings into gender categories, displaying at how effective our model was at creating informative embeddings. Since we were also making a interactive 2D model, we could analyze the embedding's locality compared to other embeddings.

After creating the initial embeddings, UMAP dimensionality reduction was used so our data could be mapped onto a scatter plot. This would allow us to look for trends and groupings in data and to check if their corresponding placements and groups fit with their content. Along with this, we used HDBSCAN to supplement our dimensionality reduction to better visualize semantic relationships between videos. Using GPT-4o-mini generated description of texts associated with video clusters, additionally helping with visualization of content clusters and verification of model efficacy. With cosine similarity, we evaluated how close videos and texts were to other videos and their corresponding video.

3. Data

The original data consisted of 1 million videos found on the HuggingFace database. Due to the sheer quantity, the amount was reduced to 17,000 videos, which were 4-15 seconds long. Shorter videos were chosen to allow less time spent on grabbing embeddings. These embeddings also had human descriptions of each video, which we used for comparisions later on.

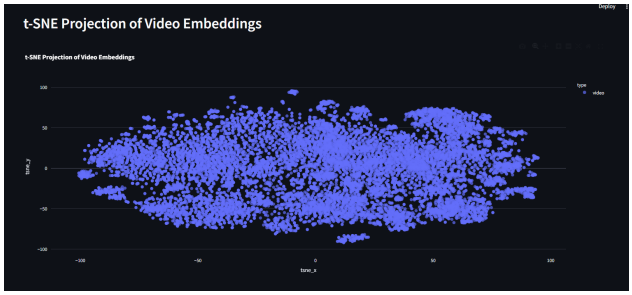

Figure 1. The inital 2D projection of the 17,000 embeddings using t-SNE dimensionality reduction

After using t-SNE to reduce the dimensionality of the embeddings down to 2 dimensions, the data was plotted on an initial, interactive display. The data appeared fairly distributed; small clusters do appear but upon inspection videos in these clusters contain very similar content.

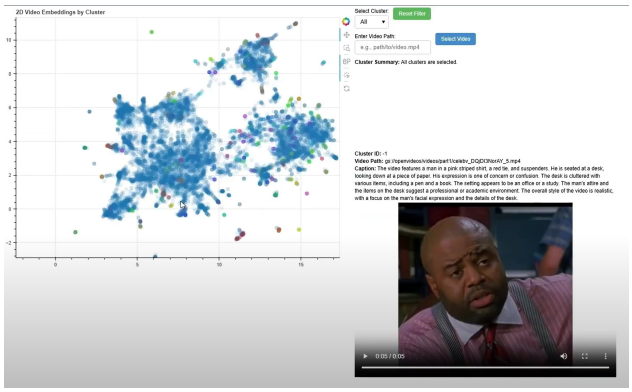

Figure 2. The final 2D projection of the embeddings using UMAP dimensionality reduction

Looking at the interactive display that uses UMAP and HDBSCAN, we can better visualize the clustering between the data.

4. Results

We used an interactive model and cosine similarity for evaluation metrics. Using the interactive model and semantic search function, we can visually verify if videos with similar content were in the same region. In tandem with this, using cosine similarity to find the relation between videos and their corresponding text embeddings can help display model efficacy. The MLP classifier mentioned earlier would categorize videos into four different gender labels based on the text description: male, female, both, and neither. This model would parse through the text descriptions and categorize them based on what it thinks is the gender of the people featured in the video.

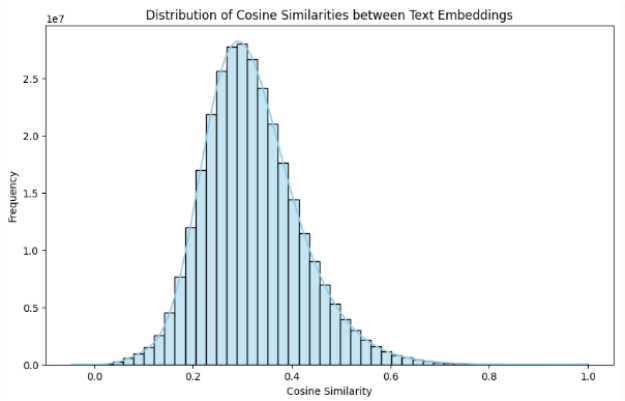

Figure 3. The average cosine similarity between texts is 0.32

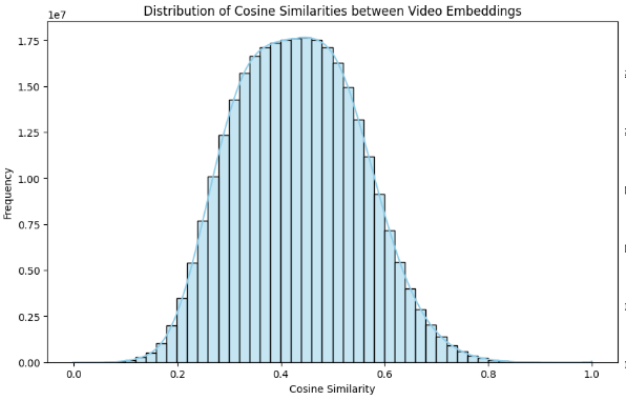

Figure 4. The average cosine similarity between videos is 0.43

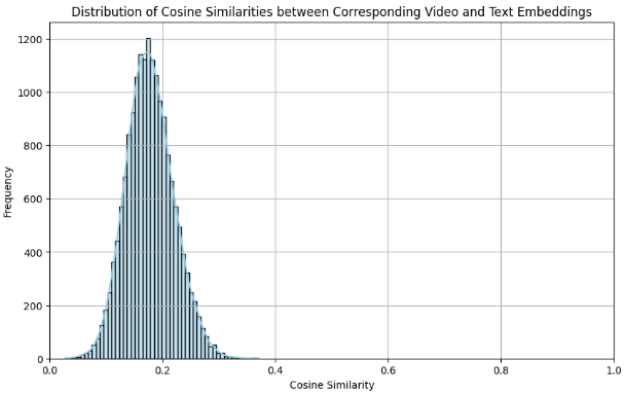

Figure 5. The average cosine similarity between texts and their corresponding videos is 0.18

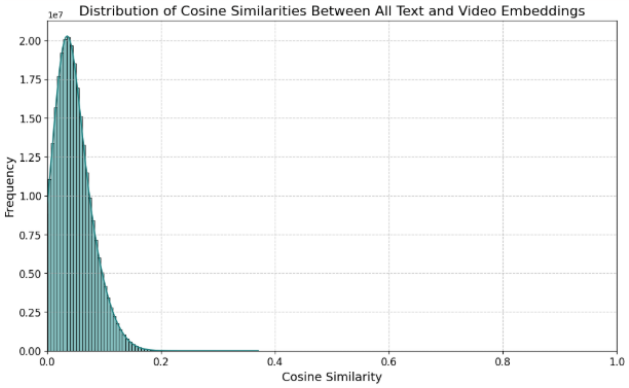

Figure 6. The average cosine similarity between all texts and videos is 0.04

On average, a video's corresponding text description is its 70th nearest text embedding. Using cosine similarity, it was found that texts had an average similarity of 0.18 with their corresponding videos (which is fairly low) , compared to the cosine similarity between all texts and videos, which was 0.04. This displays a 450% increased similarity between texts and corresponding videos compared to texts and non-related videos. It was also found that videos had an average similarity of 0.43 with other videos and texts had an average similarity of 0.32 with other texts. It should be noted that the text embeddings we got from Google's model were fairly far from their related video due to the 30-token cap for text inputs, but it was still closer to its own video compared to unrelated videos. This shows that the descriptions aren't being fully represented.

4.1. MLP Model

For our MLP model, we split the data into a 80/20 train/test split. The results we found was it had a balanced accuracy of 100%, showing that it could successfully classify gender(s) based on the text embeddings.

5. Conclusion

This project demonstrates the utility of embedding-based representation. We were able to use a multimodal model along with dimensionality reduction to plot a variety of short videos and group them based on content. The model somewhat demonstrated their semantic relationships with their corresponding videos.

5.1. Future Direction

In the future, this project could go in the direction of parsing through longer videos or analyzing more relevant datasets with important implications.

5.2. Limitations

One big limitation encountered was the 30-token limit from Google's multimodal model. This caused our embeddings to be fairly far from their corresponding videos, which led to some inaccuracies down the line. Another limitation we encountered was creating a quantifiable metric besides eyeballing our data. This is the main reason we employed an MLP model to assign gender based on text embeddings. We felt this would provide us the best results on the accuracy of our model.

6. Citations and References

[1] "Get Multimodal Embeddings." Generative AI on Vertex AI, 2024, cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings. Accessed 4 Nov. 2024.

[2] nkp37. "OpenVid-1M." Huggingface.co, 2024, huggingface.co/datasets/nkp37/OpenVid-1M. Accessed 4 Nov. 2024.

[3] "TikTok Video Sample — Atlas Map." Nomic.ai, 2024, atlas.nomic.ai/map/36c0185a-c9b3-4e44-9a9d-4b2a84e767b3/3ff50c32-882c-4 25-a2fc-020958bb952c. Accessed 4 Nov. 2024.

[4] OpenAI. "Introduction to the OpenAI API." OpenAI.com, 2024, platform.openai.com/docs/introduction. Accessed 9 Dec. 2024.

7. Contribution Chart

Table 1. Contribution chart for the project

TASK/SUB-TASK	STUDENT ID	COMMENTARY ON CONTRIBUTION
REPORT	02024264	WROTE MOST OF THE REPORT
REPORT	02042009	HELPED WITH REVISION
REPORT	01974615	HELPED WITH REVISION
REPORT	01770539	HELPED WITH REVISION
POWERPOINT	02024264	WROTE POWERPOINT
POWERPOINT	02042009	WROTE POWERPOINT
POWERPOINT	01974615	HELPED WITH REVISION
POWERPOINT	01770539	HELPED WITH REVISION
INTERACTIVE MODEL	02024264	WROTE ORIGINAL INTERACTIVE MODEL THAT GOT SCRAPPED
INTERACTIVE MODEL	02042009	WROTE FINAL VERSION OF INTERACTIVE MODEL
MLP MODEL	02042009	WROTE THE MOST OF THE MODEL
MLP MODEL	01770539	HELPED WITH MODEL INFERENCING
EMBEDDING GATHERING & STORING	01974615	HELPED GRAB AND STORE EMBEDDINGS
EMBEDDING GATHERING & STORING	02042009	HELPED GRAB AND STORE EMBEDDINGS
EMBEDDING GATHERING & STORING	01770539	HELPED GRAB AND STORE EMBEDDINGS
EMBEDDING GATHERING & STORING	02024264	HELPED GRAB AND STORE EMBEDDINGS