Generating Audio Descriptions for Videos Using Contrastive Language-Image Pre-Training (CLIP) Interrogation

Julia Chen
Mobile Web Application Development Research Lab

01

Introduction

Context

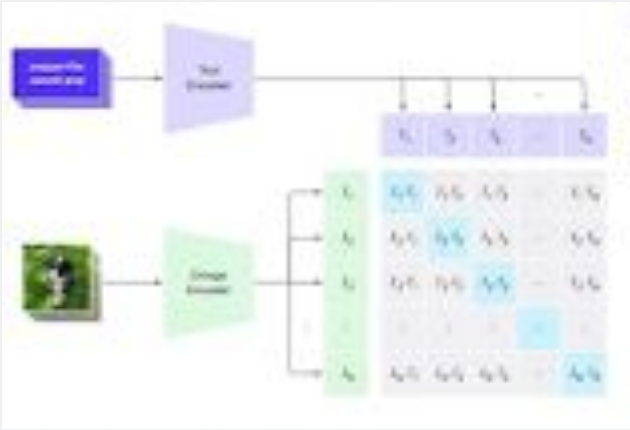
- Audio descriptions
  - Provide blind and low vision individuals information about key visual elements of a video
  - Typically manually transcribed

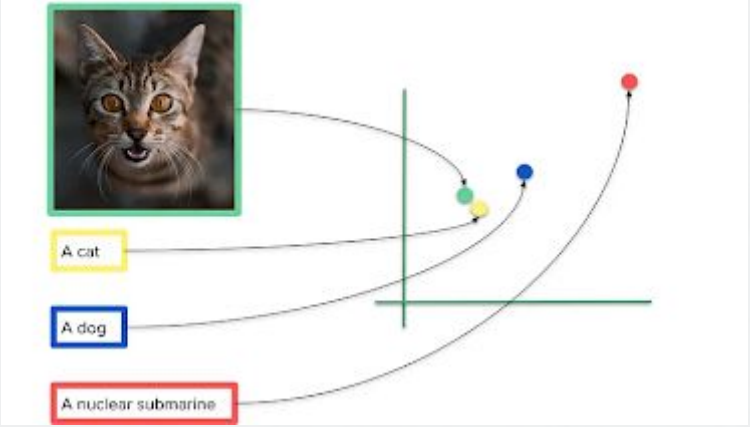
Objective

- Create an algorithm that generates audio descriptions of videos for blind and low vision individuals
- Audio descriptions should be
  - Concise
  - Accurate
  - Generated efficiently

Background

- Contrastive Language-Image Pre-training (CLIP)
  - Embeds images and text to a shared vector space
  - Embeddings are optimized during training





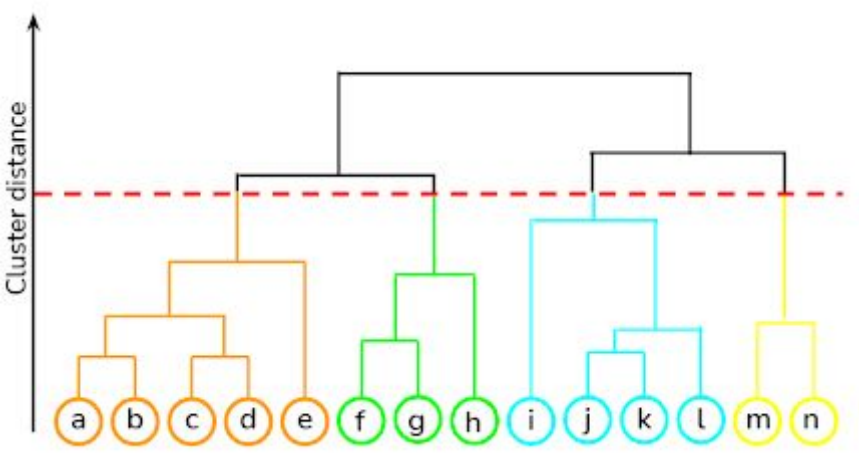
Background

- Image Interrogation
  - Generates prompt that would have produced the given image

a cliff with a waterfall in the middle, imax 7. 0 mm footage, imax photography 4 k, ocean cliff view, cinematic lens flare, beautiful cinematography, anamorphic lens flares, stunning cinematography, anamorphic lens flare, neil blomkamp film landscape, shot on anamorphic lenses, anamorphic cinematography, imax cinematography

Background

- Hierarchical clustering
  - Creates clusters of different granularities



Background

- JPEG file size
  - Quantify image complexity

185 KB

422 KB

02

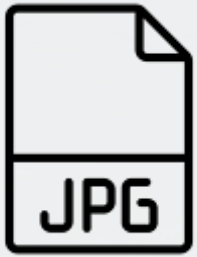
Methodology

Libraries

- OpenCV 2 (cv2)
- scipy
- gTTS
- pydub

Algorithm Design

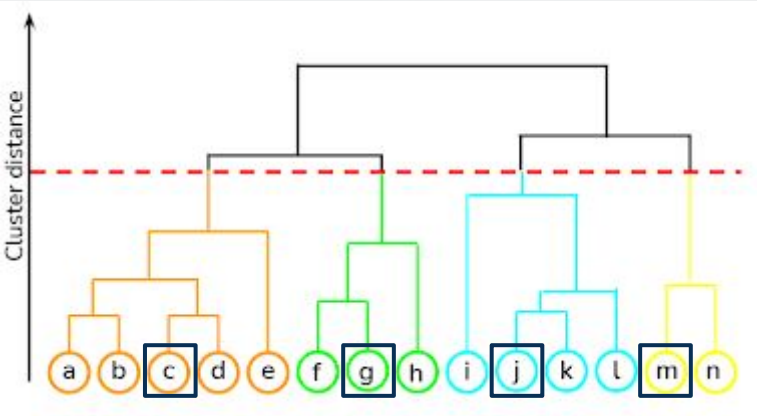
- Frame selection
  - Separate video into frames
  - Quantify using JPEG image size



→ 493 KB

Algorithm Design

- Frame selection
  - Perform hierarchical clustering
  - Select frame in center of each cluster



Algorithm Design

- Description generation
  - CLIP Interrogator analyzes selected frames
  - Description generated is associated with the cluster the frame represents

→ a cliff with a waterfall in the middle

Challenges

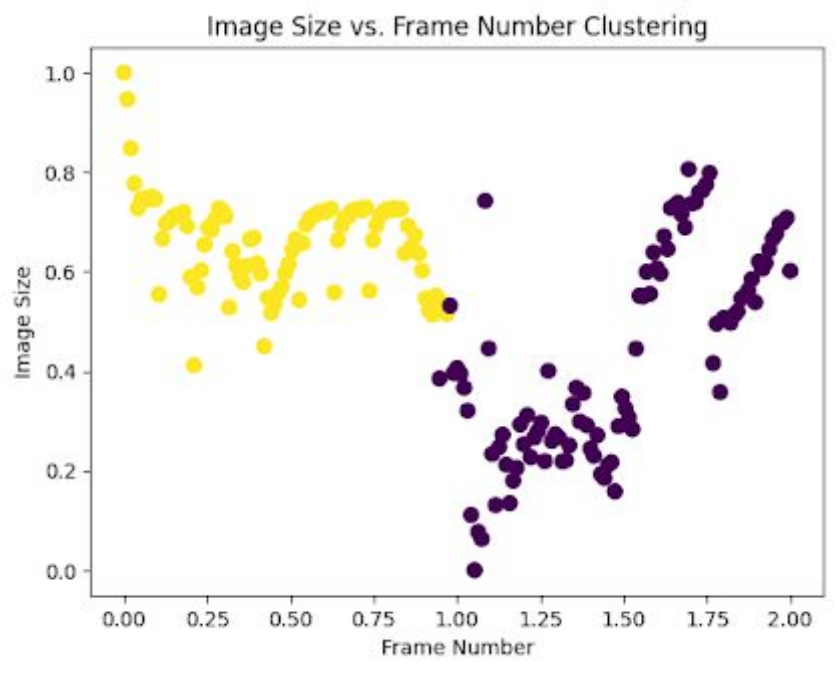
- You Only Look Once Version 11 (YOLOv11)
  - As frame selection metric

3 chairs, 2 tvs

4 chairs, 3 tvs

Challenges

- K-means clustering
  - As frame selection algorithm



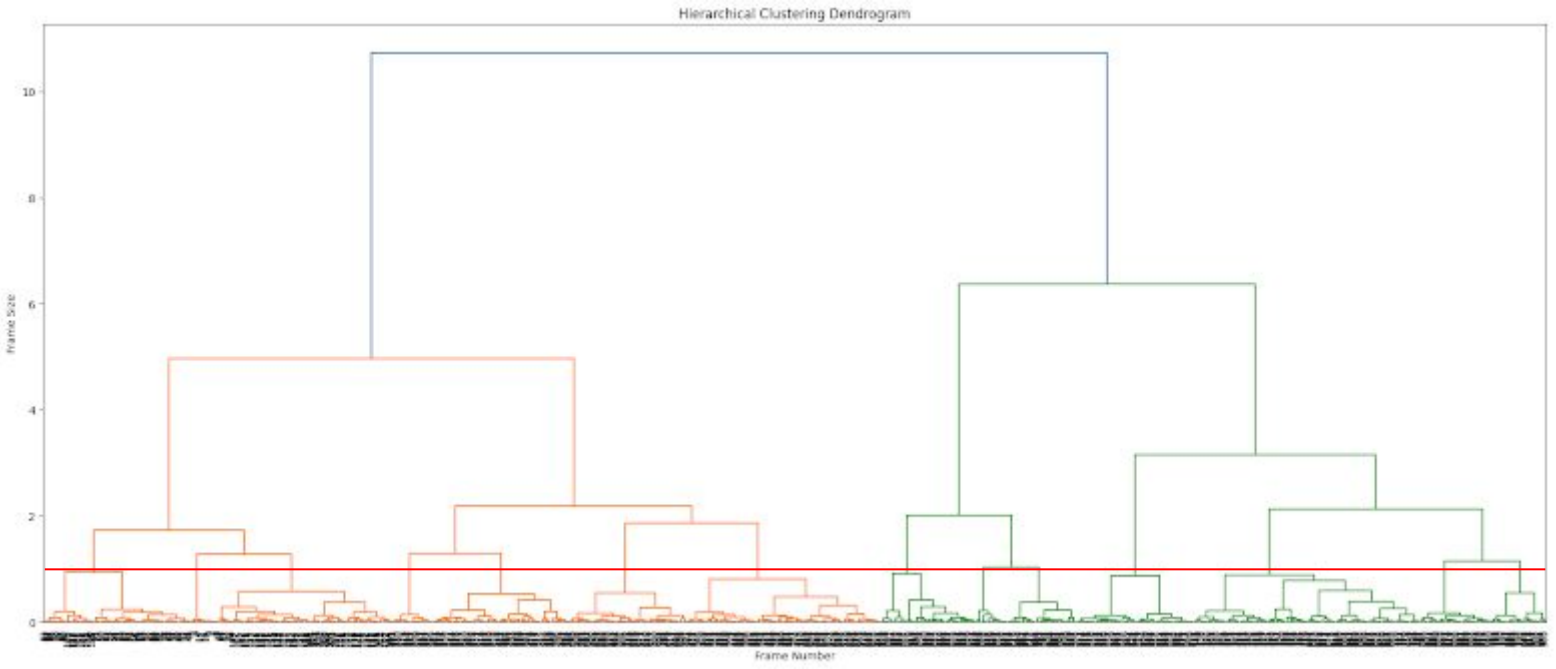
03

Results and Discussion

Input Video

- Clip from *Antarctica: Home at the End of the Earth*
  - Concrete subjects and landscapes
  - Camera pans, moving subjects, and jump cuts

Results


Hierarchical Clustering Dendrogram
Frame Size
Frame Number
0
2
4
6
8
10

Results















a close up of a penguin with a yellow beak

Successes

- Descriptions are concise
- Frequency of descriptions
- Some accurate descriptions
- Frame selection algorithm
  - Selects representative frames

Limitations

- Inaccurate descriptions
  - Unable to consider context
- Audio description alignment
- Expensive runtime
  - 12 seconds of processing per second of video

A seal swimming in the ocean

04

Conclusion

Significance

- Developed novel algorithm to generate audio descriptions
  - Frame selection
    - ↳ Hierarchical clustering on JPEG image sizes
  - Description generation
    - ↳ CLIP interrogation

Future Work

- Optimizing runtime
- Improving CLIP model
  - Description generation that considers previous context
- Improving clusters to align with video jump cuts

Acknowledgements

Special thanks to Mr. Kosek for guiding and advising me for this project and thank you to my peers for providing feedback and support.

Thank you!

Any questions?