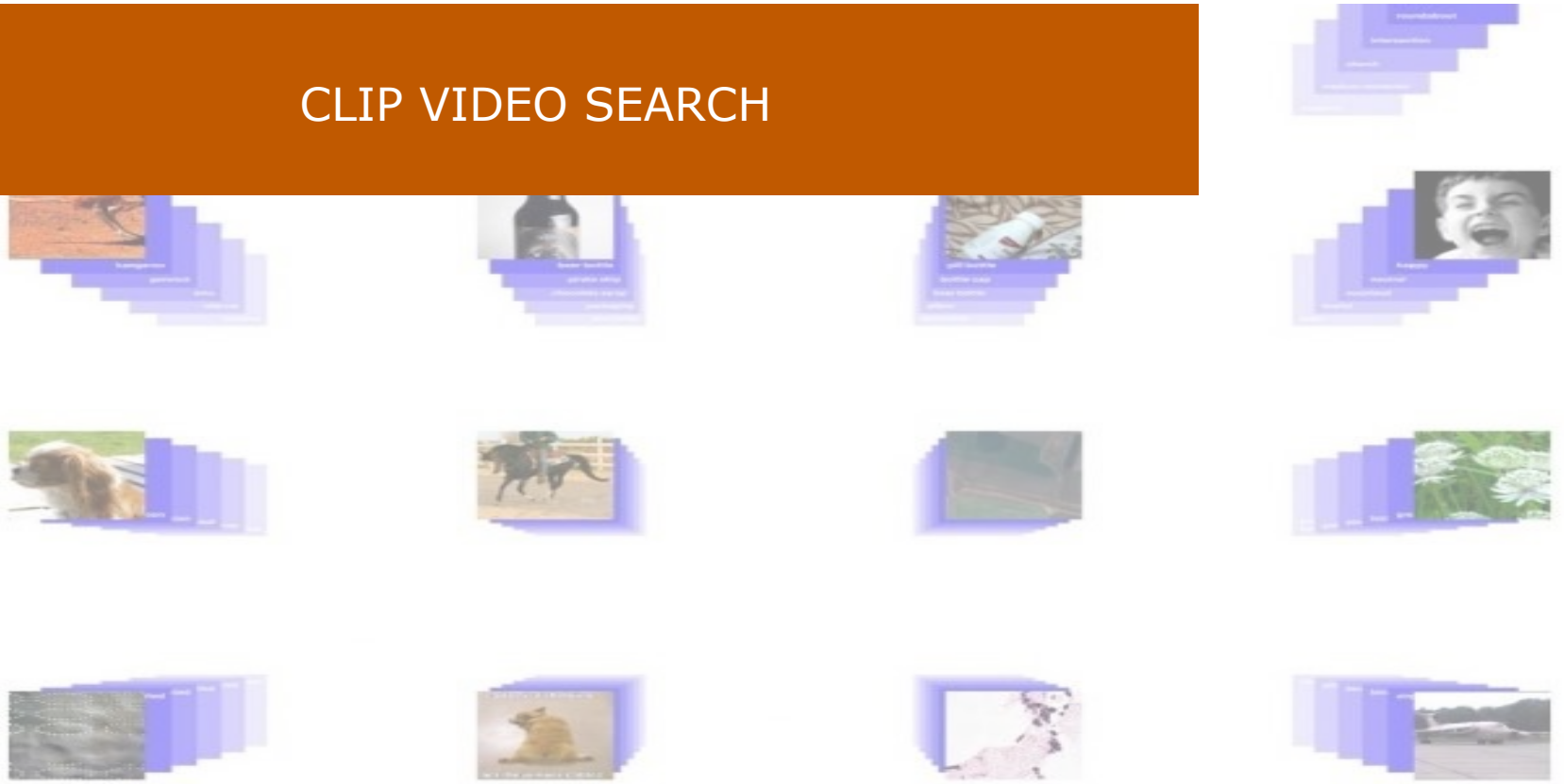
CLIP VIDEO SEARCH



MIS 382N: Introduction to Deep Learning

Brett Bartol, JT Flume, Jackson Hittner, Sam LaPlatney, Garrett Sooter

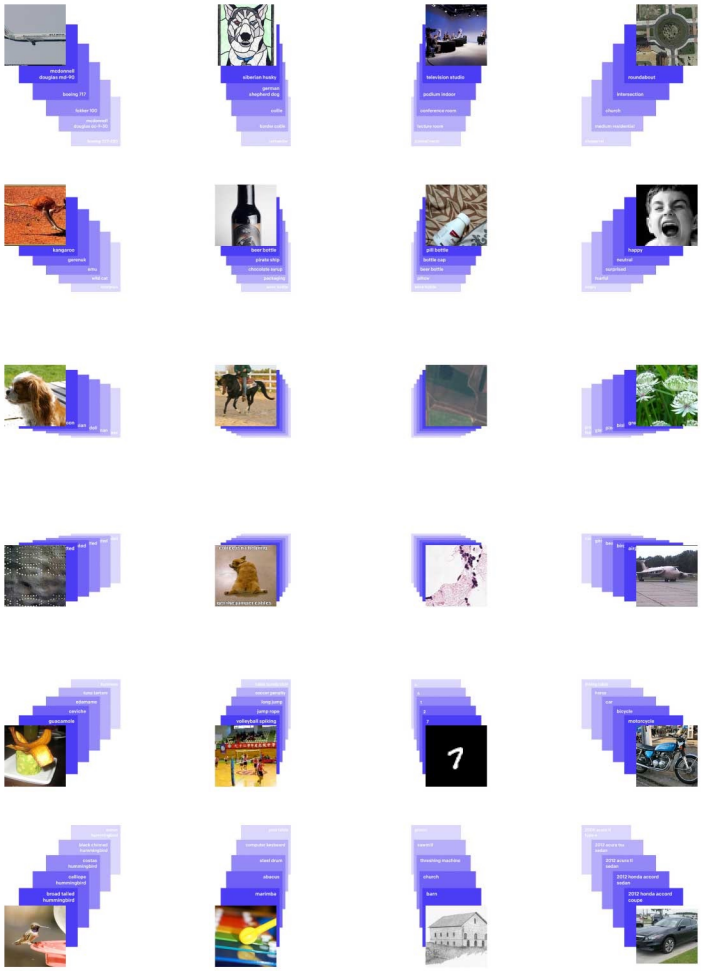
INTRODUCTION

Presentation Overview

1. CLIP Background & Overview
2. Frame Consolidation
3. Applications & Examples
4. Common Issues
5. Future Work & Recommendations

BACKGROUND

Contrastive Language-Image Pre-Training (CLIP)



OpenAI developed CLIP in 2021.

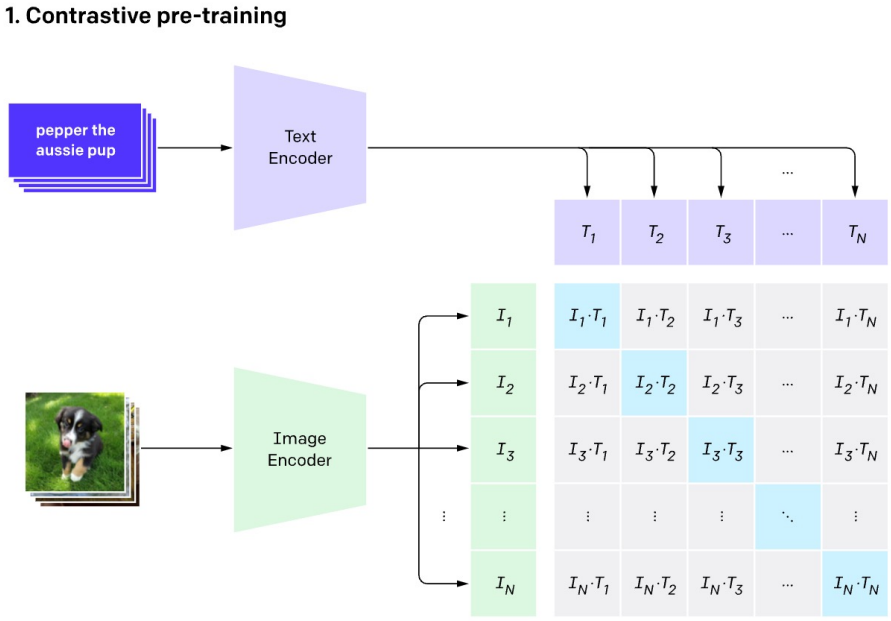
CLIP is a model that connects language and images. OpenAI trained the model on over 400 million image and text pairs.

CLIP can find the closest match of an image to a list of texts, and it can also find the closest match of text to a list of images.

'Zero shot' capabilities mean that CLIP can predict classes that it has never seen before.

BACKGROUND

CLIP Architecture



CLIP uses a text transformer trained from scratch. The transformer has 63M parameters, 12 layers and 512 wide, and it uses 8 attention heads. It takes the text input and outputs a vector embedding that represents the text phrase.

CLIP utilizes Vision Transformer (2020) for the images with a slight modification. It adds a normalization layer. It can then encode images and produce a vector encoding with the same length as the text encoding.

Finally, CLIP compares the text and image encodings using cosine similarity, with the highest similarities being the model's output.

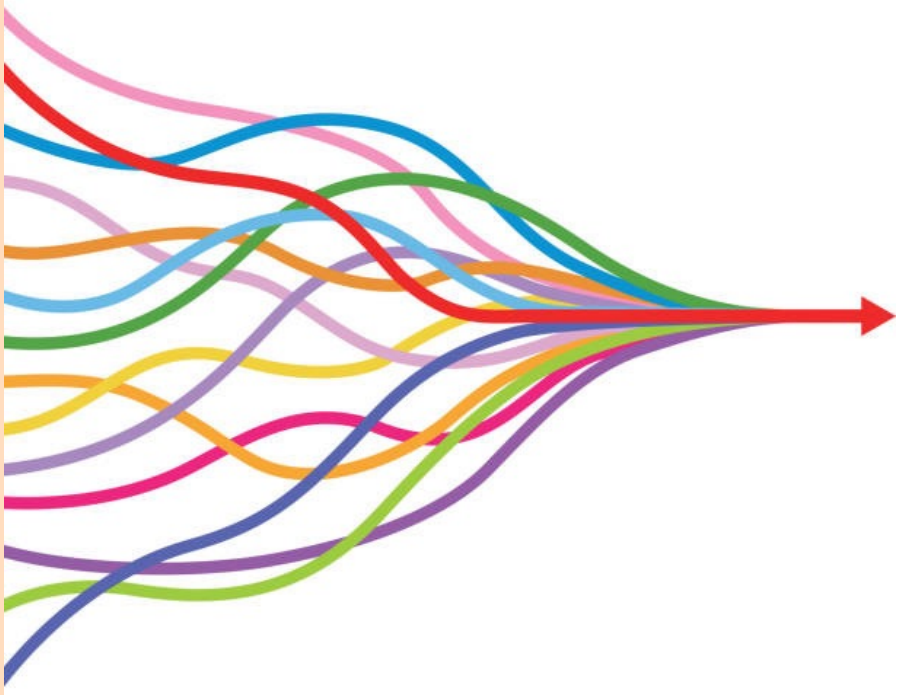
BACKGROUND

Frame Consolidation

Our goal was to increase the computation time of searching videos with CLIP. Initially, CLIP would search through videos frame by frame and output the frames with the highest cosine similarity with the inputted text.

We decided to group the videos into blocks, take the average values of all the blocks, and feed those through the model.

From there, we can find the blocks with the highest similarity and search those more thoroughly. Searching the blocks seems to be somewhat promising, and we may need to find a more robust way to combine the blocks of images to not lose information. However, especially with longer videos, this should significantly reduce computation time.



APPLICATIONS & EXAMPLES

Netflix Episode Search

Found at 0:02:06 (link)

Found at 0:07:46 (link)

Video length: 14:57

Search Query: “wrapping paper” & “man holding bad of seeds”

Execution time: 1 minute

Frames extracted: 269

Applications for users that would like to search where they were in a show or series if they remember specific scenes but not the name of the episode

APPLICATIONS & EXAMPLES

Concussion Diagnosis in Football

Video length: 18:36

Search Query: "Strong Collision"

Execution time: 9 minutes

Frames extracted: 2,232

Able to accurately identify frames that contain big hits. Doctors and team personnel would be interested in identifying the plays that resulted in concussive or sub-concussive hits and pair with player tracking data to determine which players are most at risk

APPLICATIONS & EXAMPLES

Concussion Diagnosis in Football

Video length: 18:36

Search Query: "big hit"

Execution time: 9 minutes

Frames extracted: 2,232

With query "big hit" the parser is not as accurate

APPLICATIONS & EXAMPLES

Traffic Cam Footage

Video length: 1:32

Search Query: “taxi”

Execution time: 30 seconds

Frames extracted: 27

All frames returned included taxis – applications in government (traffic optimization) or criminal justice

Found at 0:00:27 (link)

Found at 0:00:23 (link)

APPLICATIONS & EXAMPLES

Traffic Cam Footage

Video length: 1:32

Search Query: "White Sedan"

Execution time: 1 minute

Frames extracted: 177

Able to accurately identify the only white sedan in the video; longer processing time and more frames extracted compared to "taxi" query

APPLICATIONS & EXAMPLES

Interrogation Footage

Found at 0:47:13 (link)

Found at 0:39:36 (link)

Video length: 2:29:48

Search Query: “nervous man”

Execution time: 8 minutes

Frames extracted: 2,694

Retrieves frames of a suspect that appears nervous – applications in criminal psychology to re-visit moments where suspect could be demonstrating deceptive behavior

APPLICATIONS & EXAMPLES

Interrogation Footage

Video length: 1:12:01

Search Query: “deceptive woman”

Execution time: 7 minutes

Frames extracted: 1,296

Retrieves frames of video where the suspect is acting agitated and defensive. Potential applications in criminal psychology to identify signs of deceptive behavior for further analysis

Found at 0:14:34 (link)

Found at 0:48:13 (link)

CONCLUSION

Future Work & Recommendations

Frame Consolidation

- Groups of frames could be consolidated by averaging their pixel layouts

- Speed processing times and minimize required computing resources

Query Engineering

- Pair domain-specific applications with a search query converter

- Generalize domain-specific queries with a converter to yield more accurate results and ease use of the program

Thank you!