NYU TANDON SCHOOL OF ENGINEERING

The Pinkerton Foundation

Visualizing audiovisual relationships using AI

Nalicha Antoine[1]. Iran Roman[2]. Juan P. Bello[3].

[1] Achievement First Brooklyn High School [2, 3] Music and Audio Research Lab (MARL).

Center for K12 STEM Education

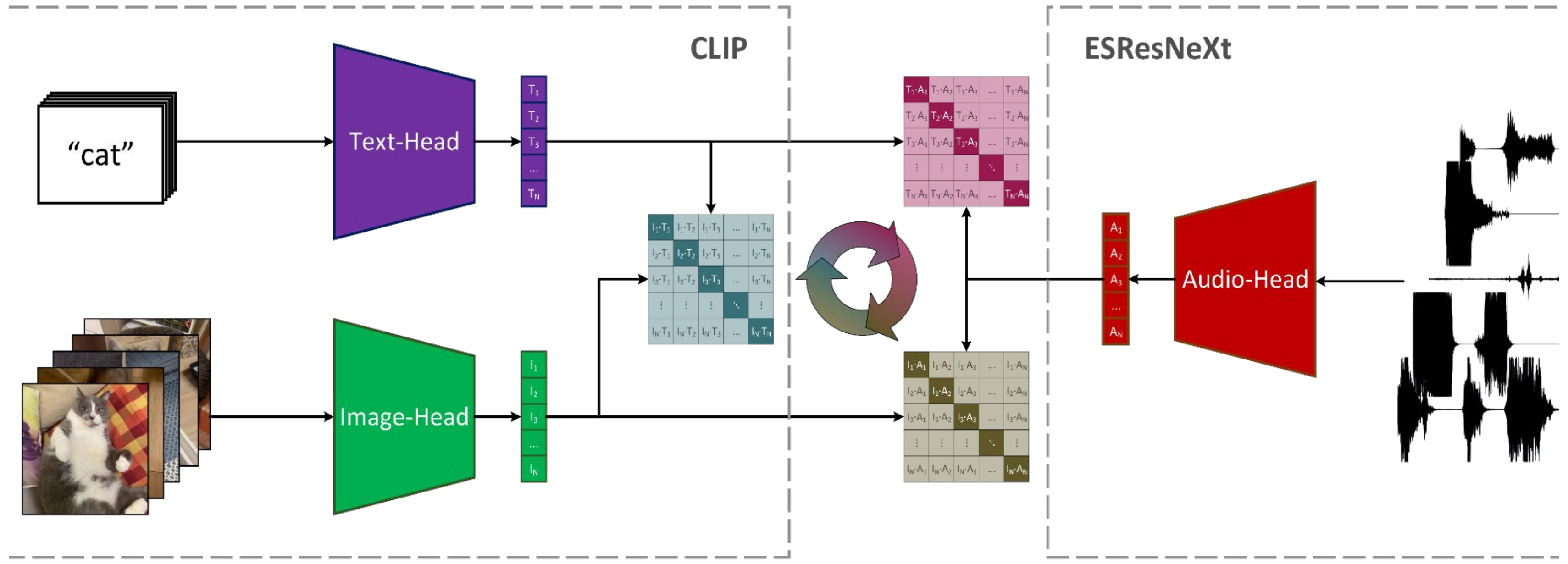
NYU TANDON SCHOOL OF ENGINEERING

INTRODUCTION

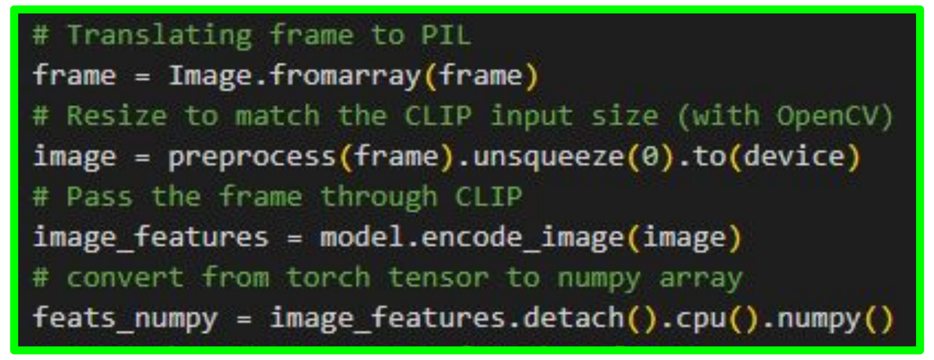
- Epic Kitchens dataset: kitchen videos with head-mounted camera recordings
- Visualize audiovisual relationships using the Epic Kitchens dataset and the AI models CLIP and AudioClip

CLIP & AudioCLIP

- CLIP: openAI model that was trained to connect image and text
- When CLIP is fed an image and a piece of text it would match them into a pair if they're alike
- AudioCLIP: similar to CLIP but instead can match audio to either image or text

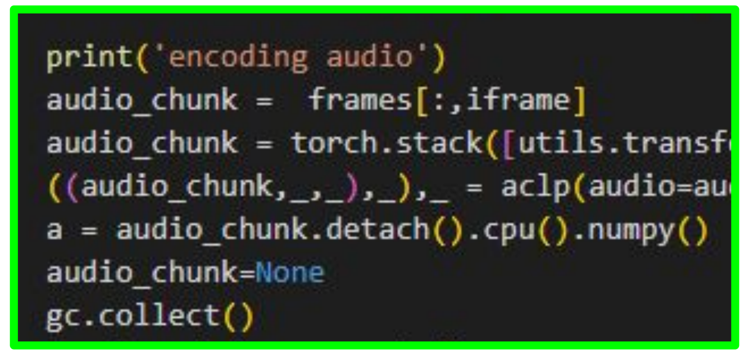

CLIP
ESResNeXt
"cat"
Text-Head
Image-Head
Audio-Head
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ... I_N
A_1 A_2 A_3 ... A_N
$I_1 \cdot T_1$ ... $I_N \cdot T_N$
$T_1 \cdot A_1$ ... $T_N \cdot A_N$
$I_1 \cdot A_1$ ... $I_N \cdot A_N$

Encoding Image



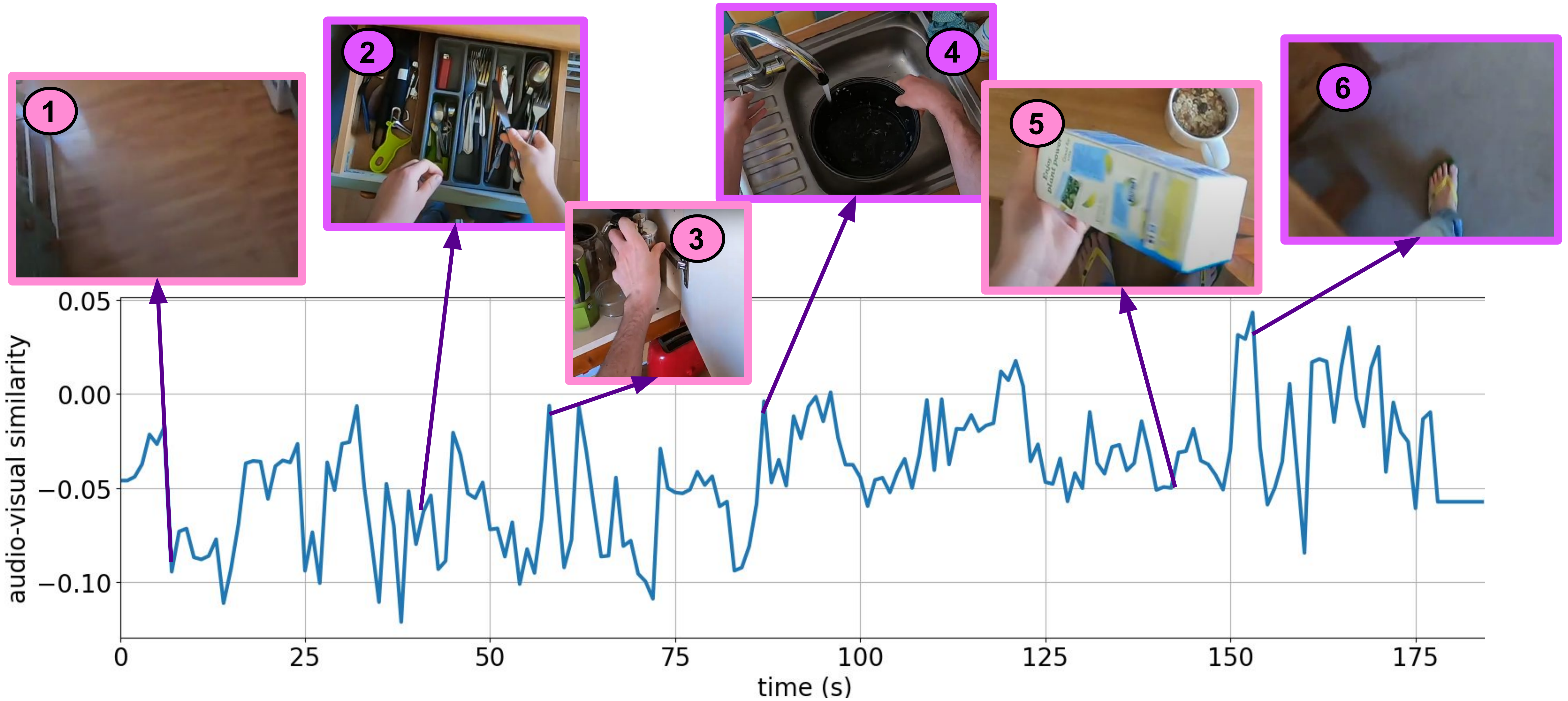
```
# Translating frame to PIL
frame = Image.fromarray(frame)
# Resize to match the CLIP input size (with OpenCV)
image = preprocess(frame).unsqueeze(0).to(device)
# Pass the frame through CLIP
image_features = model.encode_image(image)
# convert from torch tensor to numpy array
feats_numpy = image_features.detach().cpu().numpy()
```

Encoding Audio



```
print('encoding audio')
audio_chunk =  frames[:,iframe]
audio_chunk = torch.stack([utils.transf
((audio_chunk,_,_),_),_ = aclp(audio=au
a = audio_chunk.detach().cpu().numpy()
audio_chunk=None
gc.collect()
```

RESULTS



INTERPRETATION

1. AudioCLIP establishes a weaker relationship between audio and image since it can hear the person walking, but cannot see their feet in the frame
2. Actions in between cause small spikes
3. Impact sounds cause these spikes
4. AudioCLIP establishes a stronger relationship between audio and frame since it can hear and see the running water
5. AudioCLIP can hear the sound of the milk in the carton but can't see the milk, which establishes a weaker relationship
6. Compared to when there was no feet in the frame, when there is feet in the frame AudioCLIP establishes a stronger relationship

CONCLUSION

- CLIP can easily identify similarity and differences
- AudioCLIP relates what it can actually see and hear
- Since CLIP just follows directions it can be misused
- I would like to run CLIP on GPUs with more VRAM

REFERENCES

- McFee, Brian, et al. "librosa: Audio and music signal analysis in python." *Proceedings of the 14th python in science conference*. Vol. 8. 2015.
- Hafner, Markus, et al. "CLIP and complementary methods." *Nature Reviews Methods Primers* 1.1 (2021): 20.
- Guzhov, Andrey, et al. "Audioclip: Extending clip to image, text and audio." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.
- Damen, Dima, et al. "The epic-kitchens dataset: Collection, challenges and baselines." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2020): 4125-4141.