Evaluation: IMSM

10/11/2024, Chiho, Thanmaya, Mus2Vid

Problem statement

- Find a metric to see how similar image and audio pairs are
- In order to see how our model is performing

IMSM implementation

- Utilizes CLIP and CLAP scores to get a music and image similarity
- Added a softmax to normalize the scores

$$\mathcal{A}_{\text{IMSM}} = \mathcal{A}_{\text{CLIP}}\, \mathcal{A}^{T}_{\text{CLAP}}$$

Progress

Last week's Issues:

- Code was producing scores that were way lower than what was expected
  - Melfusion paper mentioned scores of 0.86 for MelBench dataset
- Issue was I wasn't normalizing the scores with a softmax

This week's progress:

- Updated IMSM code to produce reasonable values with the MelFusion code
- Took a step back and analyzed both the MelFusion paper and code

River Data

River Text

Soothing ambience of flowing water and a forest creek, making it ideal for relaxation, focus, meditation, or sleep.

+

River Audio

+

River Image

=

IMSM Score: 0.7197
CLIP Score: 0.9990
CLAP Score: 0.9457

Piano Data

Piano Text

a soft, flowing piano composition with a gentle, romantic feel. The melody is simple yet deeply emotive, creating a tranquil and introspective atmosphere.

+

Piano Audio

+

Piano Image

=

IMSM Score: 0.7193
CLIP Score: 0.9981
CLAP Score: 0.9989

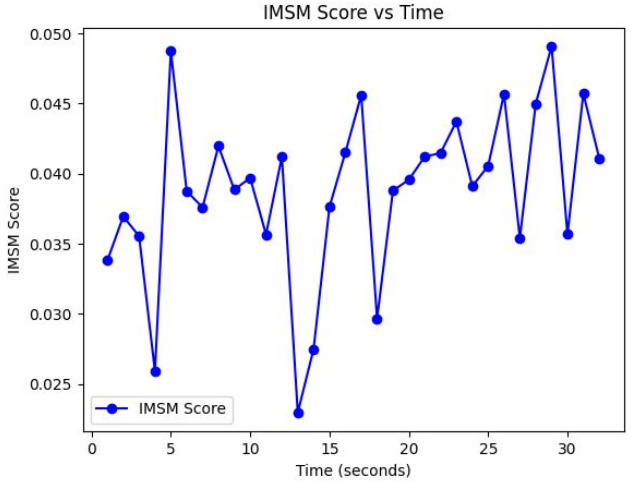
Working on video implication

```python
def process_imsm_melfusion(image_files, text_list, audio_files):
    for i in range(0, len(image_files) - 1, 2):
        image1, image2 = image_files[i], image_files[i+1]
        text1, text2 = text_list[i], text_list[i+1]
        audio1, audio2 = audio_files[i], audio_files[i+1]

        compute_imsm(image1, image2, text1, text2, audio1, audio2)


def extract_frame_audio_and_text1(video_path, interval, audio_length, output_dir):
    video = VideoFileClip(video_path)

    video_duration = video.duration

    image_files = []
    audio_files = []
    hello_array = []

    # Loop through the video at intervals to extract frames and corresponding audio
    current_time = 0
    count = 0
    while current_time < video_duration:
        # Extract the frame at the current time
        frame = video.get_frame(current_time)

        # Save the frame as an image
        frame_path = f"{output_dir}/frame_{count}.png"
        with open(frame_path, "wb") as f:
            img = Image.fromarray(frame)
            img.save(f)
        image_files.append(frame_path)

        # Extract the corresponding audio segment of audio_length duration
        audio_start = current_time
        audio_end = min(current_time + audio_length, video_duration)
        audio = video.audio.subclip(audio_start, audio_end)

        # Save the audio as a file
        audio_path = f"{output_dir}/audio_{count}.mp3"
        audio.write_audiofile(audio_path)
        audio_files.append(audio_path)

        # Append "hello" to the text array
        hello_array.append("hello")

        # Update time and count
        current_time += interval
        count += 1

    print("Extraction complete.")
    return image_files, hello_array, audio_files
```

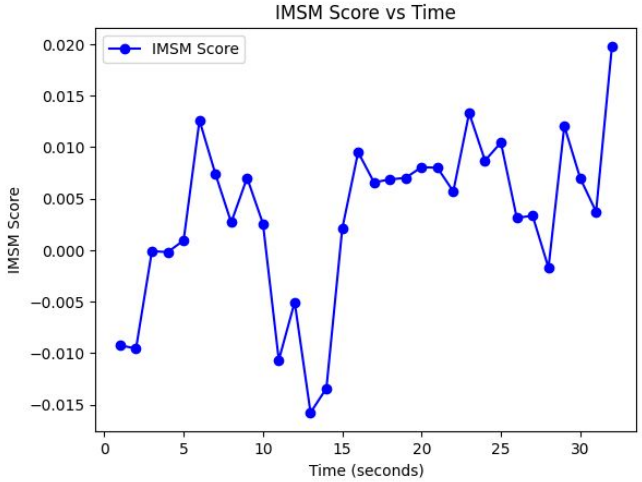
IMSM for video

Still requires text that works as an bridge between audio and image but its difficult for text to match both image and text

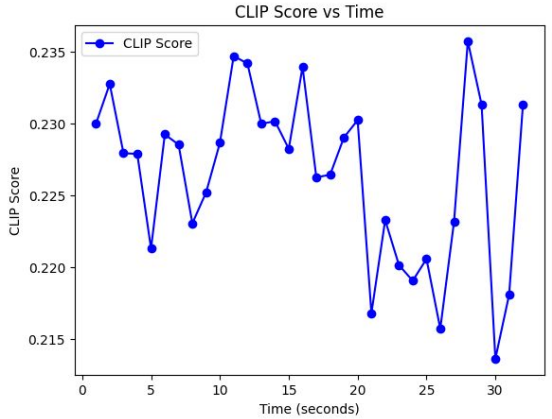
“Hello”

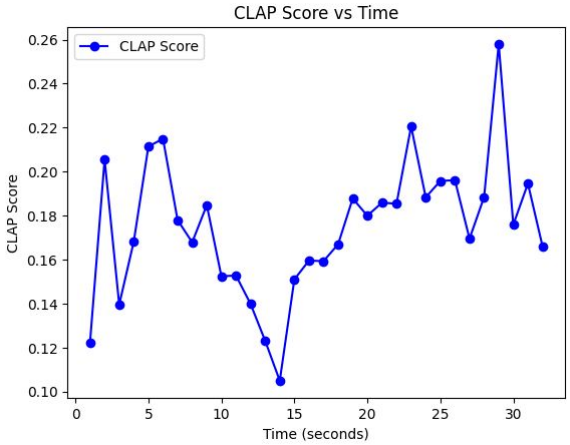


“man playing with dog”

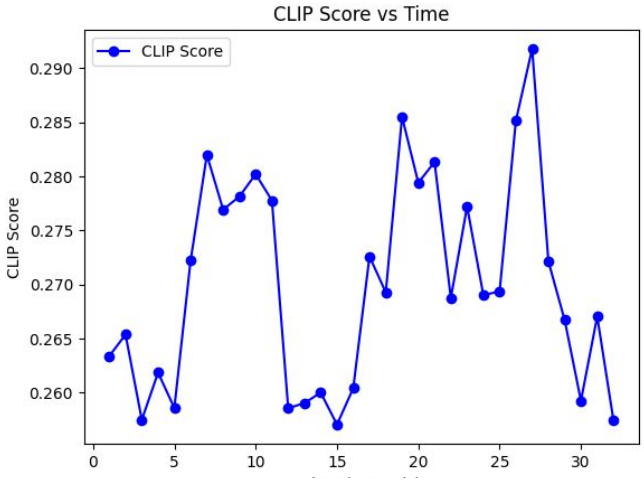


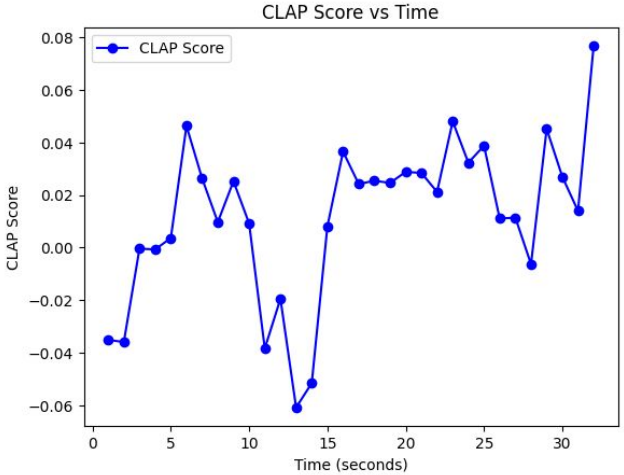
“Hello”





“man playing with dog”





Confusions

- MelFusion paper proposes this metric as if it is a universal metric
- Their code takes in 2 image, audio, text groups
- Produces high scores for piano and river data but not the MelBench dataset

High IMSM score: MelBench

Score: 0.048997

Text: The track belongs to the blues genre, evoking a somber and heart-wrenching emotional landscape, reflecting the deep sadness and pain of lost love. It is characterized by emotive and soulful vocals that convey the profound emotions associated with blues music. The instrumentation typically includes electric guitar, piano, drums, and bass, with the electric guitar often taking a prominent role, delivering expressive and soul-stirring bluesy solos. The central theme revolves around heartbreak and the emotional struggle of losing love, delving into the pain and despair associated with such loss, and longing for what once was.

MelFusion vs River

```
CLIP Scores (Image-Text Similarity):
CLIP Score between image 1 and text 1: 1.0000
CLIP Score between image 1 and text 2: 0.0000
CLIP Score between image 2 and text 1: 0.0000
CLIP Score between image 2 and text 2: 1.0000

CLAP Scores (Audio-Text Similarity):
CLAP Score between audio 1 and text 1: 0.0910
CLAP Score between audio 1 and text 2: 0.9090
CLAP Score between audio 2 and text 1: 0.5605
CLAP Score between audio 2 and text 2: 0.4395

IMSM Scores (Image-Audio Similarity):
IMSM Score between image 1 and audio 1: 0.3847
IMSM Score between image 1 and audio 2: 0.6153
IMSM Score between image 2 and audio 1: 0.6153
IMSM Score between image 2 and audio 2: 0.3847
```

Label 1: MelFusion Data from previous Slide
Label 2: River Data

Piano vs River

```
CLIP Scores (Image-Text Similarity):
CLIP Score between image 1 and text 1: 0.9981
CLIP Score between image 1 and text 2: 0.0019
CLIP Score between image 2 and text 1: 0.0010
CLIP Score between image 2 and text 2: 0.9990

CLAP Scores (Audio-Text Similarity):
CLAP Score between audio 1 and text 1: 0.9748
CLAP Score between audio 1 and text 2: 0.0252
CLAP Score between audio 2 and text 1: 0.0081
CLAP Score between audio 2 and text 2: 0.9919

IMSM Scores (Image-Audio Similarity):
IMSM Score between image 1 and audio 1: 0.7237
IMSM Score between image 1 and audio 2: 0.2763
IMSM Score between image 2 and audio 1: 0.2759
IMSM Score between image 2 and audio 2: 0.7241
```

Label 1: Piano Data
Label 2: River Data

Next Steps:

- Final sanity check run it through more MelBench data
- Any guidance? Suggestions?
- Look through more metrics: Fréchet inception distance (FID score)