Evaluation: IMSM

9/27/2024, Thanmaya, Mus2Vid

Overview of progress on tasks

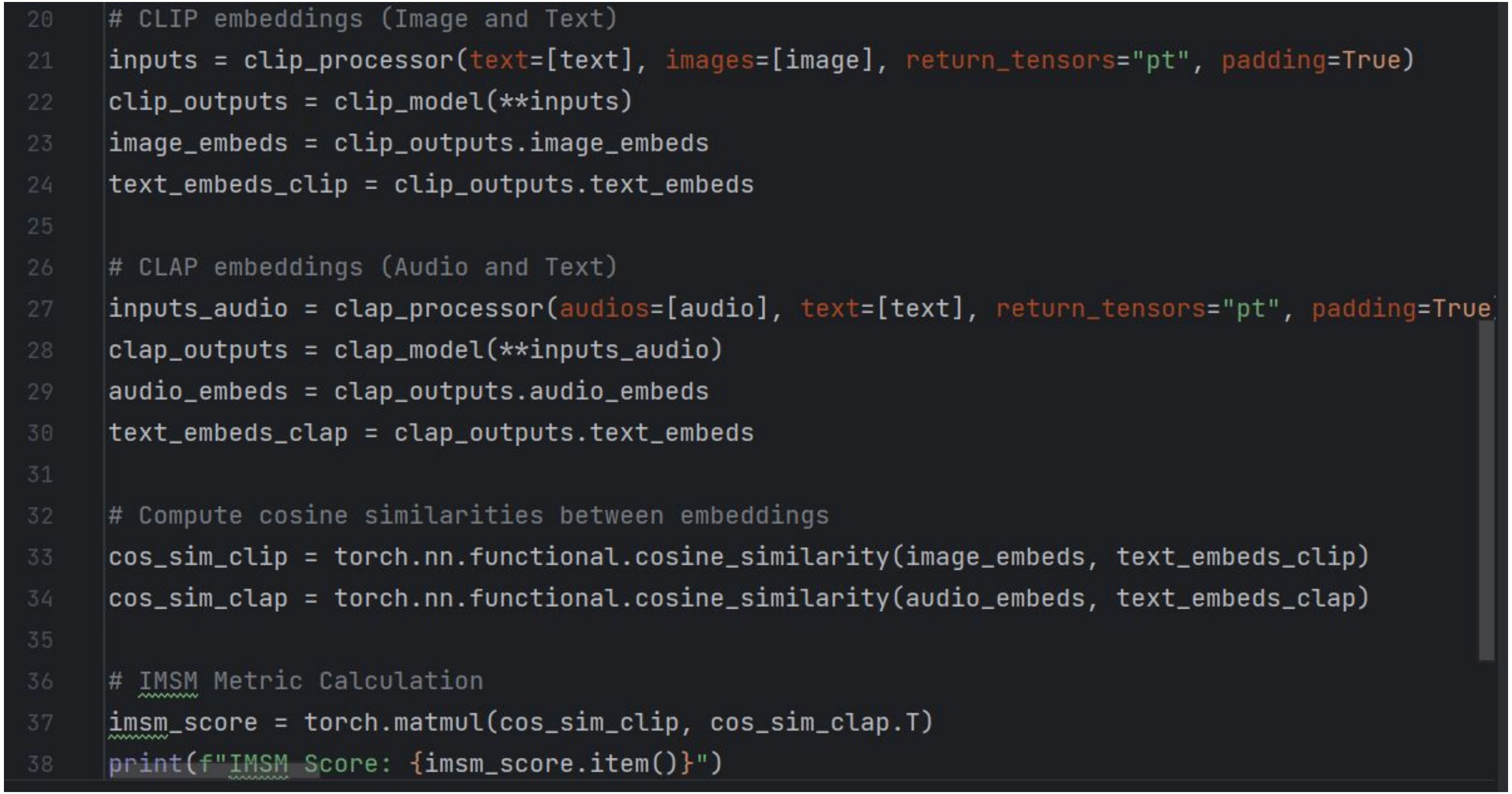
- Read the MeLFusion Paper especially their evaluation metric
- Worked through IMSM implementation basing my code off of the MeLFusion implementation

IMSM (Image Music Similarity Metric)

- MeLFusion discusses a Multimodal approach to generating audio music from text and image
- Utilizes CLIP and CLAP scores to get a music and image similarity

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

IMSM implementation



```python
# CLIP embeddings (Image and Text)
inputs = clip_processor(text=[text], images=[image], return_tensors="pt", padding=True)
clip_outputs = clip_model(**inputs)
image_embeds = clip_outputs.image_embeds
text_embeds_clip = clip_outputs.text_embeds

# CLAP embeddings (Audio and Text)
inputs_audio = clap_processor(audios=[audio], text=[text], return_tensors="pt", padding=True
clap_outputs = clap_model(**inputs_audio)
audio_embeds = clap_outputs.audio_embeds
text_embeds_clap = clap_outputs.text_embeds

# Compute cosine similarities between embeddings
cos_sim_clip = torch.nn.functional.cosine_similarity(image_embeds, text_embeds_clip)
cos_sim_clap = torch.nn.functional.cosine_similarity(audio_embeds, text_embeds_clap)

# IMSM Metric Calculation
imsm_score = torch.matmul(cos_sim_clip, cos_sim_clap.T)
print(f"IMSM Score: {imsm_score.item()}")
```

Next Steps:

- Test IMSM on MeLBench data set to see results
- Try using on CoLab, and Gilbreth to get different results