CVPR 2023 Highlight

Language-Guided Music Recommendation for Video via Prompt Analogies

Daniel McKee[1*] Justin Salamon[2] Josef Sivic[2,3] Bryan Russell[2]

[1]University of Illinois at Urbana-Champaign [2]Adobe Research

[3]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University

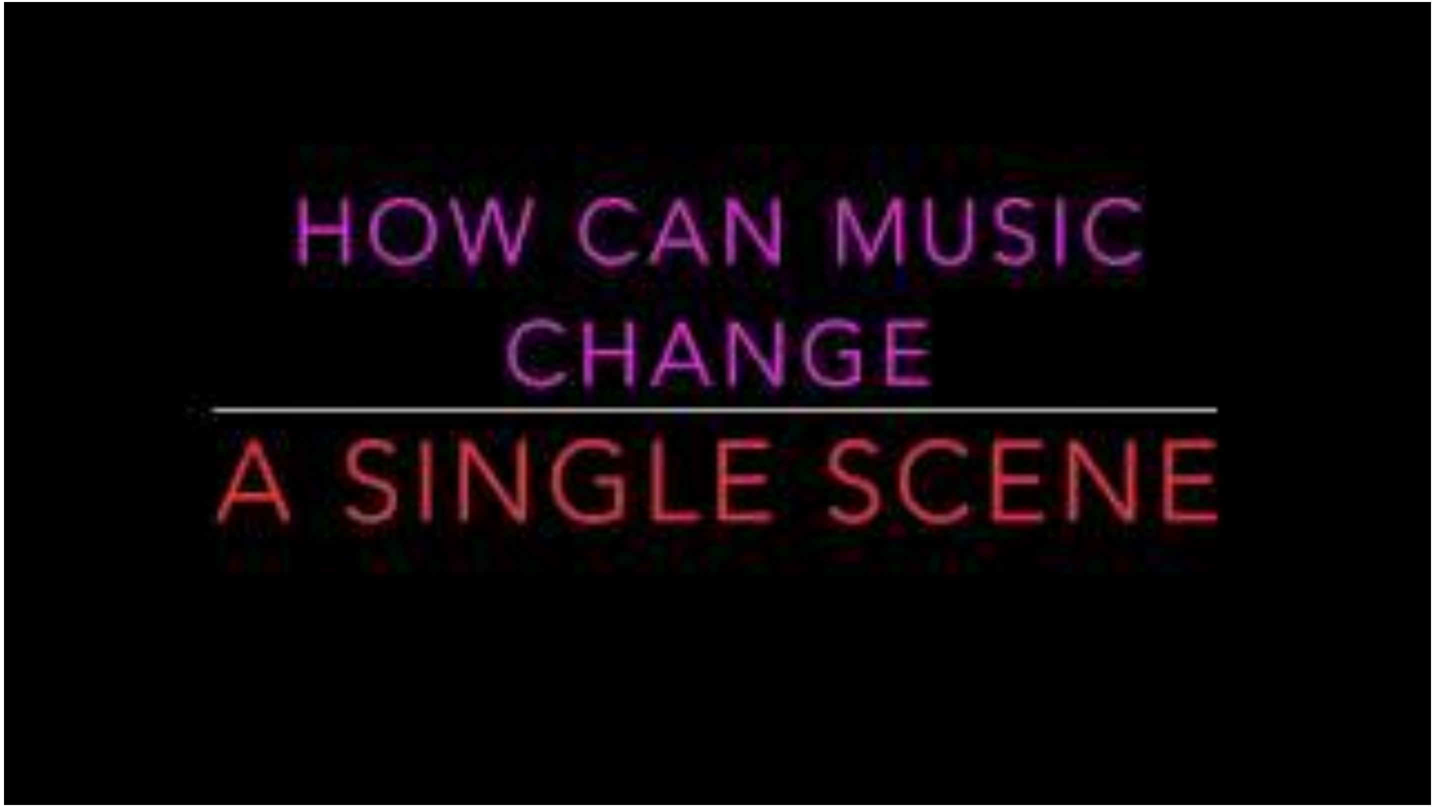
dbmckee2@illinois.edu salamon@adobe.com josef.sivic@cvut.cz brussell@adobe.com

Yuji Byun

EffL@POSTECH

9. 21. 2023

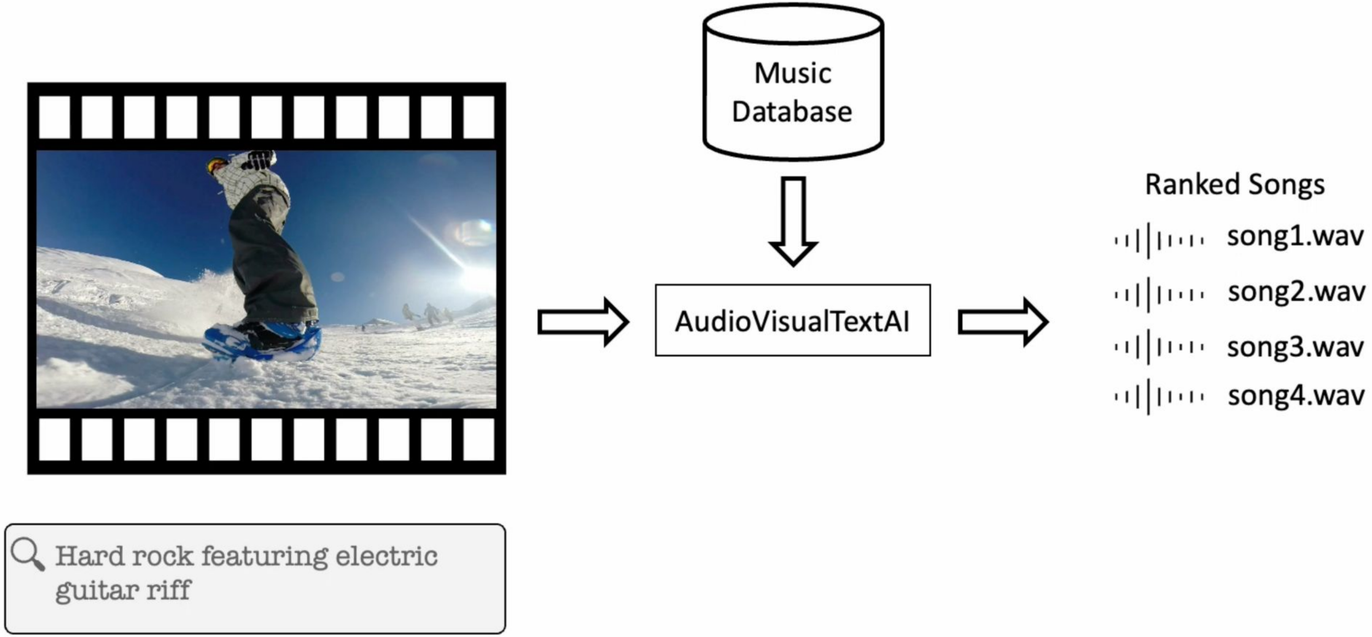
Same scene different music



Introduction

- Music selection can transform a scene into one that is perceived as urgent, sad …
- In previous work, music is retrieved based solely on the visual content from video.
- This paper propose a user guided music-for-video recommendation approach.

Overview


Music Database
AudioVisualTextAI
Ranked Songs
song1.wav
song2.wav
song3.wav
song4.wav
Hard rock featuring electric guitar riff

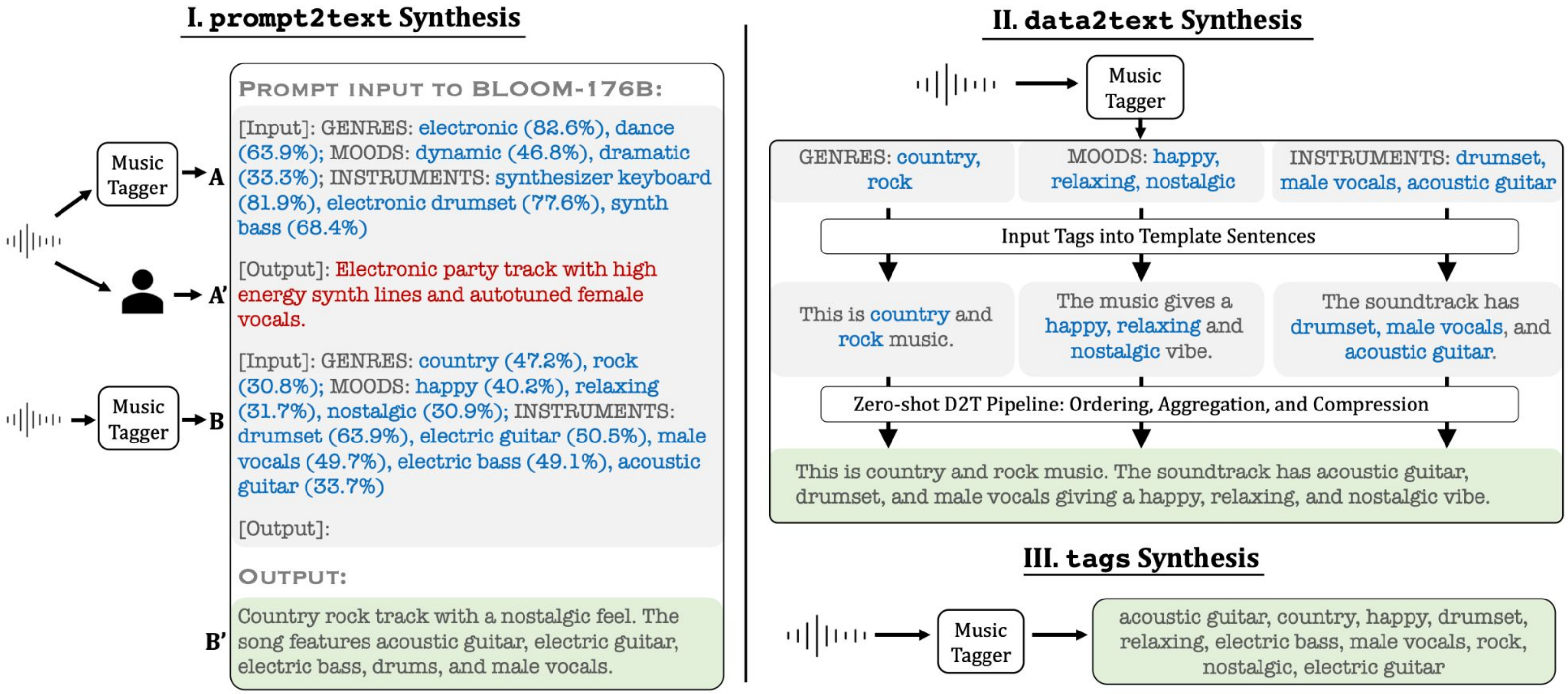
Challenge#1

- There are no available datasets which include music, video, and text together

- The existing datasets that do include text and music focus on a limited vocabulary of tags rather than free-form text. (e.g. electronic, acoustic guitar..)

- It is expensive to obtain high-quality human descriptions.

Challenge#1



I. prompt2text Synthesis
Music Tagger
A
A'
Music Tagger
B
B'
PROMPT INPUT TO BLOOM-176B:
[Input]: GENRES: electronic (82.6%), dance (63.9%); MOODS: dynamic (46.8%), dramatic (33.3%); INSTRUMENTS: synthesizer keyboard (81.9%), electronic drumset (77.6%), synth bass (68.4%)
[Output]: Electronic party track with high energy synth lines and autotuned female vocals.
[Input]: GENRES: country (47.2%), rock (30.8%); MOODS: happy (40.2%), relaxing (31.7%), nostalgic (30.9%); INSTRUMENTS: drumset (63.9%), electric guitar (50.5%), male vocals (49.7%), electric bass (49.1%), acoustic guitar (33.7%)
[Output]:
OUTPUT:
Country rock track with a nostalgic feel. The song features acoustic guitar, electric guitar, electric bass, drums, and male vocals.
II. data2text Synthesis
Music Tagger
GENRES: country, rock
MOODS: happy, relaxing, nostalgic
INSTRUMENTS: drumset, male vocals, acoustic guitar
Input Tags into Template Sentences
This is country and rock music.
The music gives a happy, relaxing and nostalgic vibe.
The soundtrack has drumset, male vocals, and acoustic guitar.
Zero-shot D2T Pipeline: Ordering, Aggregation, and Compression
This is country and rock music. The soundtrack has acoustic guitar, drumset, and male vocals giving a happy, relaxing, and nostalgic vibe.
III. tags Synthesis
Music Tagger
acoustic guitar, country, happy, drumset, relaxing, electric bass, male vocals, rock, nostalgic, electric guitar

Challenge#1

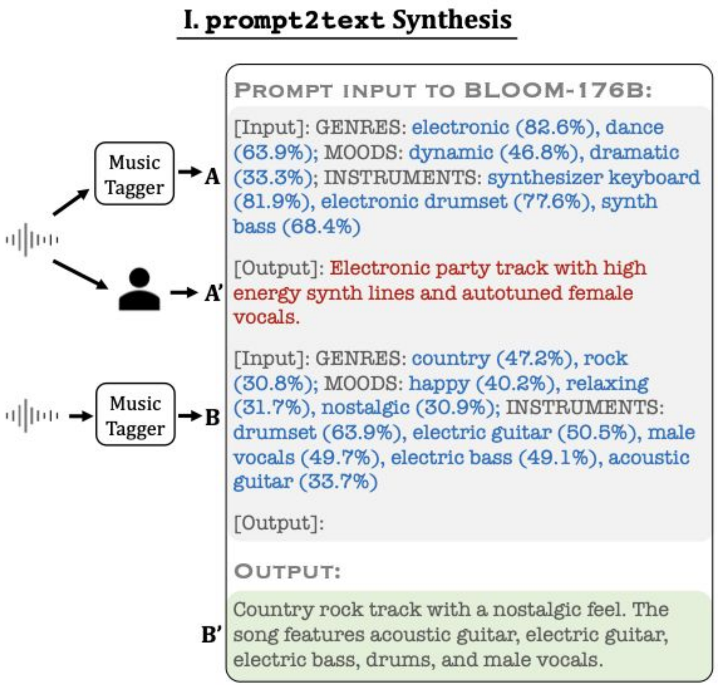
- **I**.prompt2text

  : Combining a pre-trained music tagger

  and human annotation

  with a large-scale language model.

- LLM : Bloom-176B*
- Music Tagger**



* BLOOM (Scao et al., 2022) /https://huggingface.co/bigscience/bloom

** Disentangled multidimensional metric learning for music similarity (Lee et al., 2020)

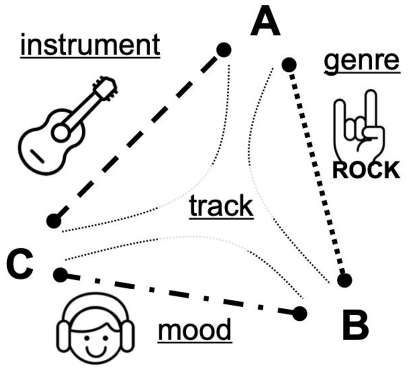
Challenge#1

- **I**.prompt2text

  : Combining a pre-trained music tagger

  and human annotation

  with a large-scale language model.



- LLM : Bloom-176B*
- Music Tagger** : 41 instrument tags, 20 genre tags, 28 mood tags

  keeping only those above a threshold(0.3 in this paper)

* BLOOM (Scao et al., 2022) /https://huggingface.co/bigscience/bloom

** Disentangled multidimensional metric learning for music similarity (Lee et al., 2020)

Challenge#1

- Dataset : YouTube8M - Music Video

- Music sampled 10 seconds audio clips from middle of each music video.
- An annotation describes only the music from YT8M sample, Annotators do not see the corresponding video.

- 3,000 for test, 1,000 for training

Hip-hop track with a dark synth pad with male aggressive rapping along with a chipmunk voice.

Instrumental track featuring an ambient pad and bell-like sounds. Seems to be a film score.

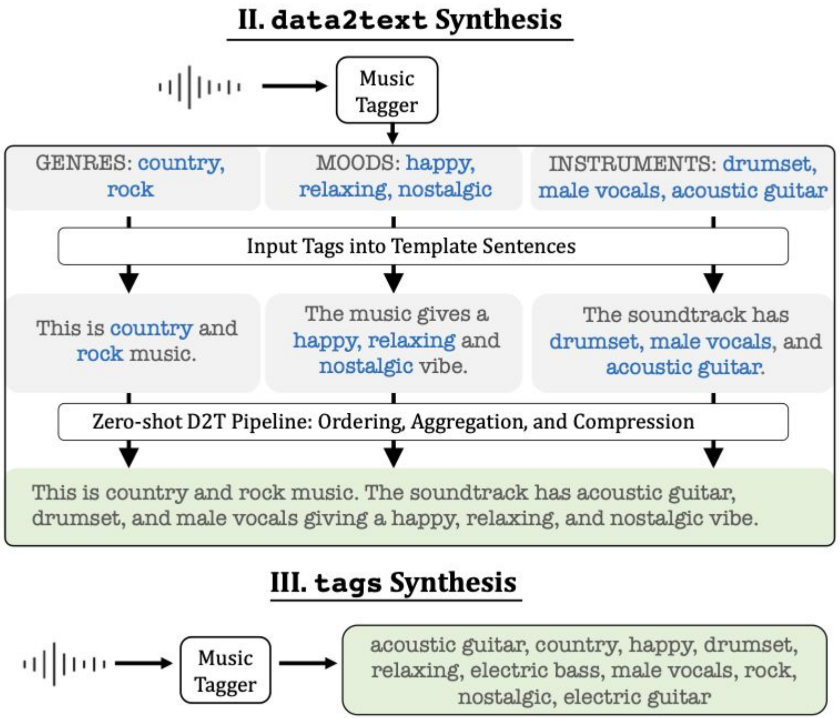
Challenge#1

- **II**:data2text Synthesis

  : Using Zero-shot D2T approach*

- **III**.tags Synthesis

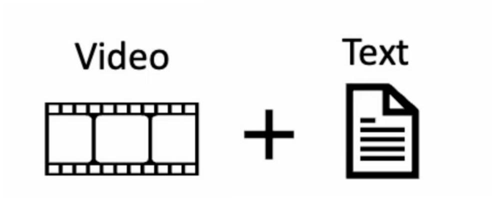
  : Shuffling tags randomly



* Neural pipeline for zero-shot data-to-text generation (Kasner et al., 2022)

Challenge#2

- Balancing Influence of Video/Text Inputs for Audio Retrieval



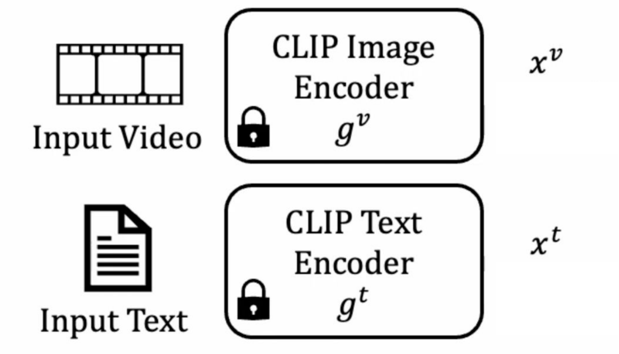
- Propose a video+text fusion model architecture
- Using text dropout to prevent overfitting on text

Method

- Split the video into 10-second segments(6 frames per second) and get features using CLIP ViT-B/32.

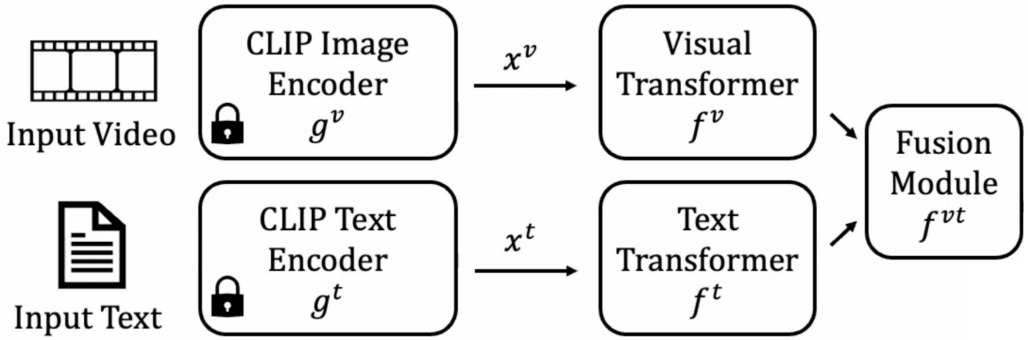
  Compute a feature by averaging CLIP embedding features.

- Get text feature from input text using CLIP.



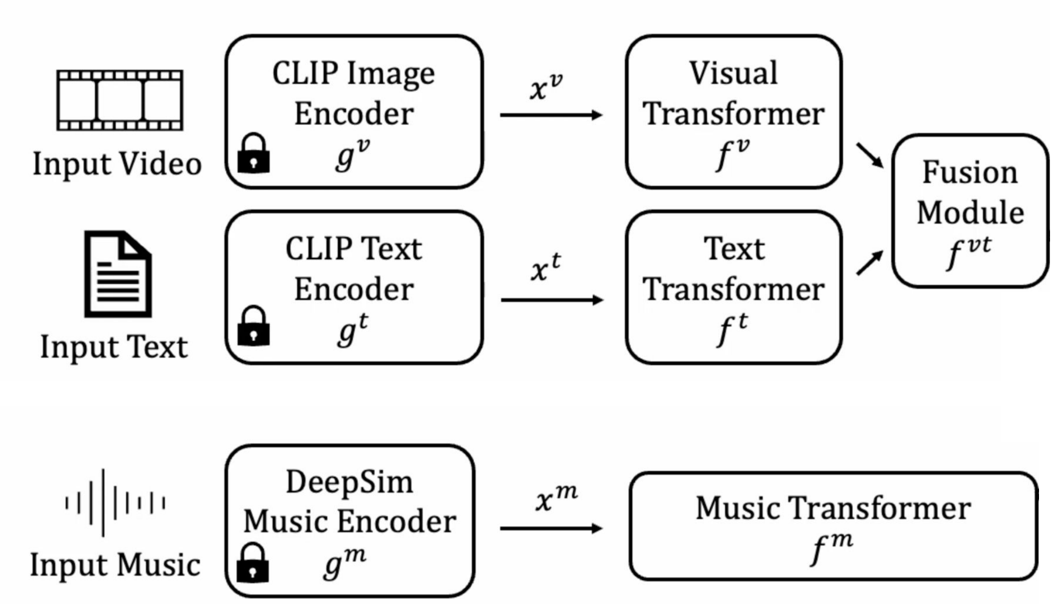
Method

- Input feature to each modality transformer encoders, and combined video and text embedding.
- Get fused embedding from the fusion model.



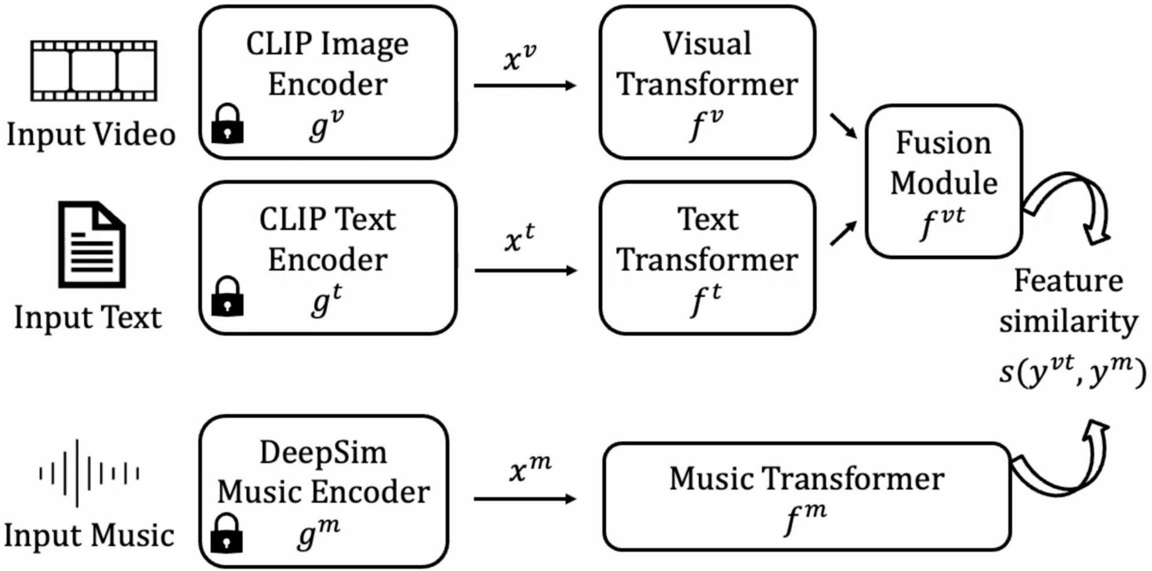
Method

- Extract music feature using Music Encoder and get music embedding using music transformer
- DeepSim → Tagger



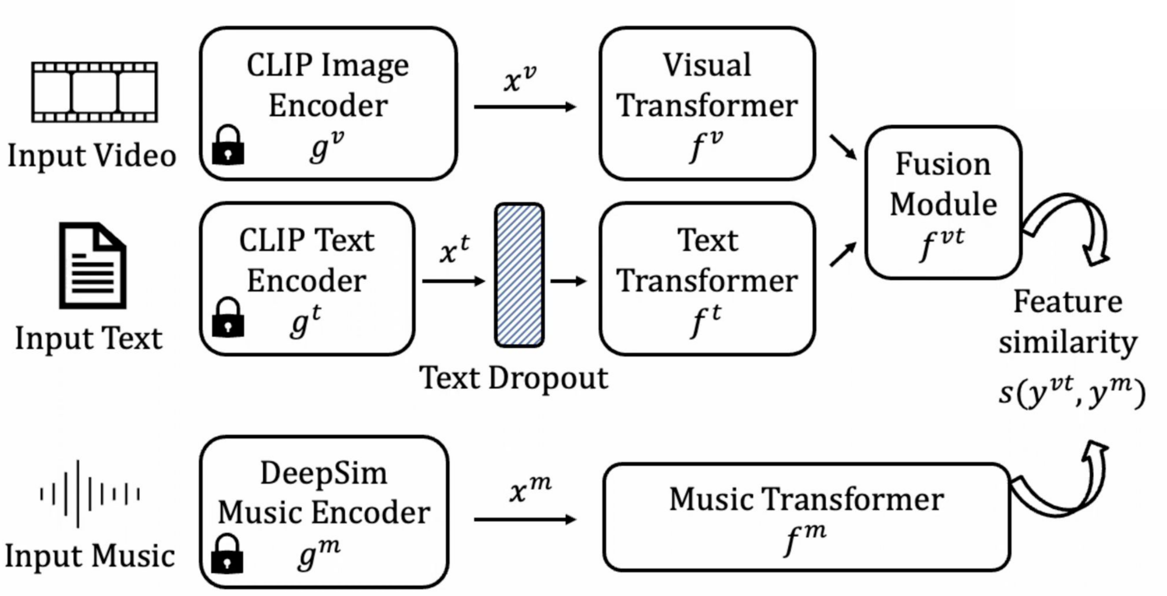
Method

- Get feature similarity between fusion embedding(video-text) and music embedding.
- But the model starts overfitting to the training text input …



Method

- With probability p(0.8), set the input text embedding x^t to a specific value x^{NULL}
- It yields improving the performance



Method

- Use InfoNCE loss between music and fused video-text embeddings.
- τ is hyperparameter (0.03 in this paper), similarity metric is cosine similarity.

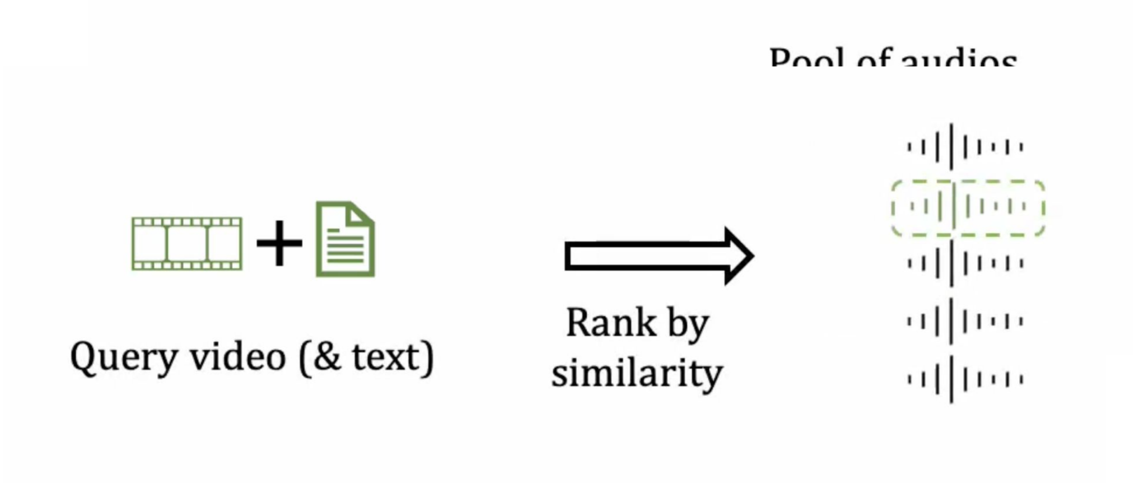
- Loss is not symmetric, so final loss is summed loss.

$$\mathcal{L}_{vt\to m} = -\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \ldots \sum_{j\in\mathcal{D}} \exp(s(y_i^{vt}, y_j^{m})/\tau)$$

$$\mathcal{L}_{m,vt} = \mathcal{L}_{vt\to m} + \mathcal{L}_{m\to vt}$$

Feature similarity $s(y^{vt}, y^{m})$

Evaluation



- The poll of N music, N=2000 in track level and N=500 for evaluation clips
- Recall metric

Evaluation

Method	Train Text	Query Text Input	Median Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. Pretét et al. [31]	-	-	234	0.76	3.42	5.90
b. MVPt [36]	-	-	13	6.09	24.91	41.89
c. MVPt+ [36]	-	-	5	27.93	50.64	60.68
d. ViML (ours)	`tags`	-	3	29.43	62.49	75.40
e. ViML (ours)	`tags`	`tags`	**2**	**49.49**	**81.61**	**89.41**
f. Chance			1000	0.05	0.25	0.50

- Tag-based music retrieval on full YoTube8M-MusicVideo
- MVTPt+ is improved version of MVP (tuned the parameter τ in the InfoNCE loss to 0.03 from 0.3)

* Cross-modal music-video recommendation(Pretet et al., 2021)

** It’s time for artistic correspondence in music and video(Suris et al. 2022)

Evaluation

Method	Train Text	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. MVPt+	-	17	12.20	29.43	40.46
b. ViML	`tags`	15	11.95	30.34	42.62
c. ViML	`data2text`	13	13.61	33.94	46.24
d. ViML	`prompt2text`	**12**	**14.09**	**35.04**	**47.88**
Chance		250	0.20	1.00	2.00

- Music retrieval on YT8M-MusicTextClips
- Video includes only a 30sec clip surrounding the 10sec of audio labeled by human annnotato

Evaluation

Method	Train Text	Dropout	Text Inputs	Median Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. MVPt+	-	-	-	17	12.20	29.43	40.46
b. ViML	`prompt2text`	✗	-	20	9.94	26.42	37.01
c. ViML	`prompt2text`	✗	`human`	15	11.45	30.45	42.77
d. ViML	`prompt2text`	✓	-	16	12.27	30.34	41.51
e. ViML	`prompt2text`	✓	`human`	**12**	**14.09**	**35.04**	**47.88**

- Performance of with and without text dropout
- Dropout technique is most effective in the range of 0.8 - 0.95

Output


JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
CTU
Language-Guided Music Recommendation for Video via Prompt Analogies
Daniel McKee
Justin Salamon
Josef Sivic
Bryan Russell
University of Illinois at Urbana-Champaign
Adobe Research
Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University
Poster: WED-PM-232

Thank you