Unified Pretraining Target Based Video-Music Retrieval with Music Rhythm and Video Optical Flow Information

Tianjun Mao, Shansong Liu, Yunxuan Zhang, Dian Li, Ying Shan
School of Management Fudan University, ARC Lab Tencent PCG

Background

➢ **Task: Video Background Recommendation**

Select an appropriate BGM for users' uploaded video

➢ **Challenges:**

1. Mismatch of target sets for video/music pretraining

2. Underlying temporal correlation between video and music is ignored

Innovations

Drawback1: **Mismatch of modalities**
Innovation1:
Collect a unified target set to match two modalities
Cross-modality attention is adopted to fuse video and music modalities

Drawback2: **Temporal correlation**
Motivation2:
To adopt clip-level embeddings from pretrained Conformer
To introduce optical flow and rhythm information

Framework

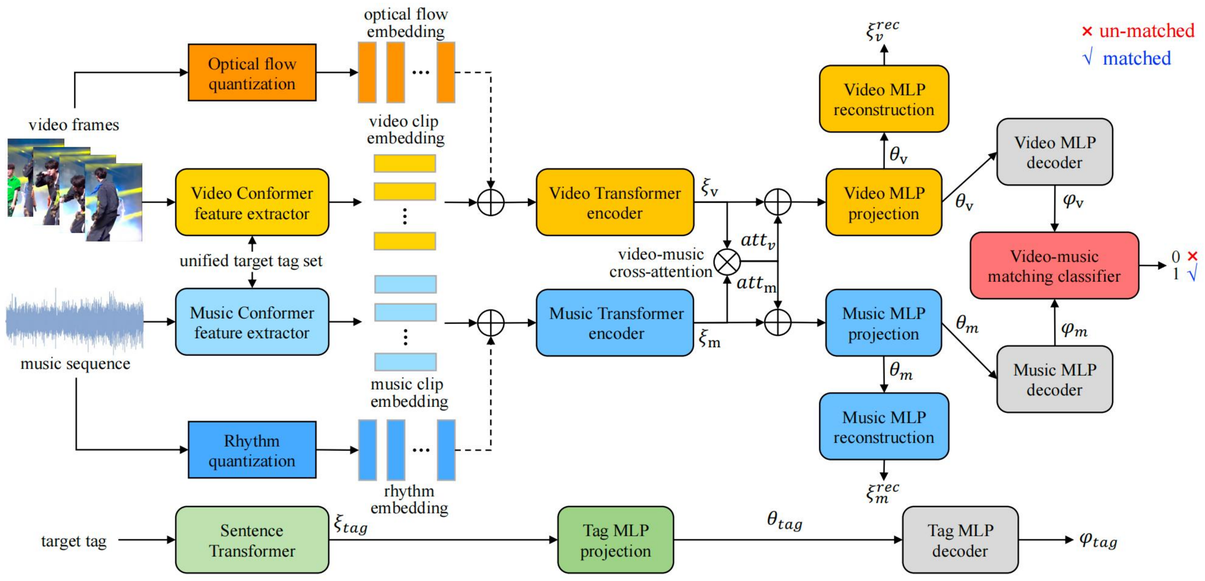

Fig. 1. Illustration of the unified pretraining target based cross-modal video-music retrieval (UT-CMVMR) framework.

➢ **Framework**

➢ Pretrained Conformers on unified target set
➢ Extraction of optimal flow and rhythm information
➢ Two-Branch Structure
➢ Cross-Modalities Attention Module

➢ **Loss Function**

➢ Triplet loss

$$\mathcal{L}_{av} = t(\theta_v^+, \theta_m^+, \theta_m^-) + t(\theta_m^+, \theta_v^+, \theta_v^-)$$
$$\mathcal{L}_{vtag} = t(\theta_v^+, \theta_{tag}^+, \theta_v^-) + t(\theta_{tag}^+, \theta_v^+, \theta_{tag}^-) + t(\varphi_{tag}^+, \varphi_v^+, \varphi_v^-)$$
$$\mathcal{L}_{atag} = t(\theta_m^+, \theta_{tag}^+, \theta_m^-) + t(\theta_{tag}^+, \theta_m^+, \theta_m^-) + t(\varphi_{tag}^+, \varphi_m^+, \varphi_m^-)$$

➢ Regularization loss

$$\mathcal{L}_{regular} = d(\xi_v, \xi_v^{rec}) + d(\xi_m, \xi_m^{rec}) + d(\varphi_v, \xi_{tag}) + d(\varphi_m, \xi_{tag}) + d(\varphi_{tag}, \xi_{tag}) + d(\theta_v, \theta_{tag}) + d(\theta_m, \theta_{tag})$$

➢ Cross-entropy matching loss $\mathcal{L}_{ce}$

Preparations

➢ **Process**

➢ Collect Unified Target Label Set
➢ Pretrain Comformers on video-tag pairs and audio-tag pairs
➢ Length-normalized videos and music are chopped into clips
➢ To get clip-level embeddings from Conformers
➢ Video optical flow information: average pixel displacement between adjacent frames
➢ Audio rhythm: number of beats+ average beat strength+average interval length

Experiments

Question1: Is it works better on our collected data?

Sys	Model	Setting	Recall@K (%)			
			K = 1	K = 5	K = 10	K = 25
1	CBVMR	AE	4.54	15.69	27.71	43.35
2	CMVAE		4.95	18.00	29.70	43.86
3	(UT-) CMVMR	AE	5.13	19.83	30.32	44.80
4		A-SE	5.18	22.38	35.76	45.50
5		SE	5.58	21.02	35.80	46.11
6		SE&R	**8.82**	**22.92**	**36.28**	**53.82**

➢ **Effectiveness of CMVMR**
- **Sys. 3 vs. Sys. 1-2**

➢ **Effectiveness of Unified Tag Set for Conformer Extractors**
- **Sys. 4 vs. Sys. 1-3**

➢ **Effectiveness of Temporal Information and Rhythm Information**
- **Temporal correlation: Sys. 5 vs. Sys. 4**
- **Rhythm information: Sys. 6 vs. Sys. 5**

Question2: How it works on HIMV-200K?

Model	Setting	Recall@K(%)				
		K=1	K=5	K=10	K=25	
CBVMR	AE	3.40	5.20	15.30	22.70	①
	SE&R	5.20	7.10	18.20	29.10	②
CMVAE	AE	4.70	9.10	17.00	41.20	③
	SE&R	6.10	11.80	20.40	44.00	④
CMVMR	AE	9.70	13.90	21.30	45.90	⑤
	SE&R	**10.80**	**28.10**	**36.50**	**51.60**	⑥

- ***With the addition of temporal information and rhythm information, it works better***
- ⑥ **vs.** ⑤
- ***The framework of CMVMR is better than the baselines***
- ① ③**vs.** ⑤

Question3: How it works for human evaluation?

Model	Our UT-CMVMR model	CBVMR	CMVAE
Preferred	50.00%	40.00%	10.00%

➢ **24 participants**
➢ **15 out of 24 people have knowledge of music theory**
➢ **21 out of 24 people maintain the habit of listening to music per week**