IMPROVING MUSIC GENRE CLASSIFICATION FROM MULTI-MODAL PROPERTIES OF MUSIC AND GENRE CORRELATIONS PERSPECTIVE

平安科技 PING AN TECHNOLOGY

Ganghui Ru, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd.

Introduction

Music genre classification (MGC) is one of the oldest and most important tasks in music information retrieval (MIR), and has become a research hotspot in both the academic and industrial communities due to its wide application prospects.

Problem in previous researches:

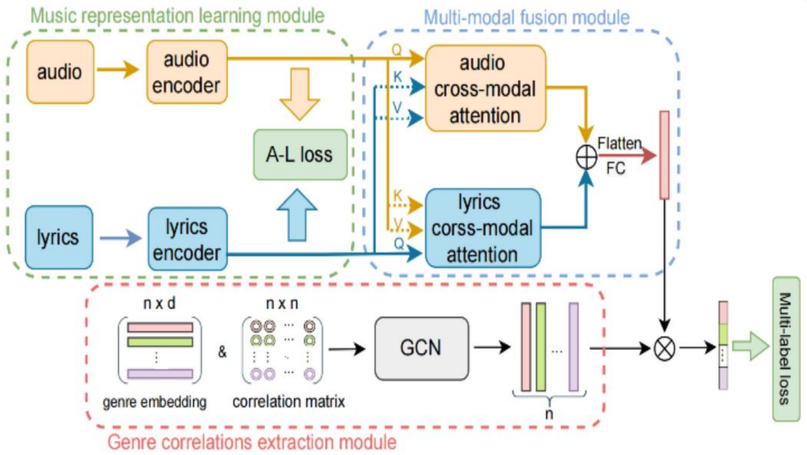
- The uni-modal methods cannot fully utilize the multimodal properties of music, thus leading to prone to performance bottlenecks.
- The existing multi-modal methods ignore that features from different modalities reside in different embedding spaces that are inherently heterogeneous and misaligned.
- Focusing only on single label classification task and lacking research on multi-label music genre classification.

Our contributions:

- We propose a novel multi-modal approach leveraging audio-lyrics contrastive loss and two symmetric cross-modal attention, to align and fuse the features from audio and lyrics.
- Based on the nature of the multi-label classification problem, we design a genre correlations extraction module which can capture and model the potential genre correlations.

Proposed Method

Our proposed model consists of three modules: music representation learning module, multi-modal fusion module, and genre correlations extraction module



Music representation learning module:

The music representation learning module consists of an audio encoder and a lyrics encoder, which takes audio and lyrics as input and outputs corresponding features respectively. In addition, we present an audio-lyrics contrastive loss to align features obtained from different modalities before sending them intothe multi-modal fusion module.

Multi-modal fusion module:

We design a novel multi-modal fusion module which consists of two symmetric cross-modal attention to effectively fuse multi-modal features.

Genre correlations extraction module:

- Considering that the correlations between genres are essentially a topological structure, we design a graph convolution network to model the genre correlations.
- We use a pre-trained BERT model to encode genre names as semantic embedding and treat them as genre node features.
- The correlation matrix is obtained by calculating the co-occurrence conditional probability and the cosine similarity between node features.

Experiments

Dataset: Music4All

- Filtered out music tracks with missing information or non-English lyrics.
- 43650 samples with 87 genre categories are finally selected.
- Approximately 70%, 10%, and 20% samples respectively for the training set, the validation set, and the test set.

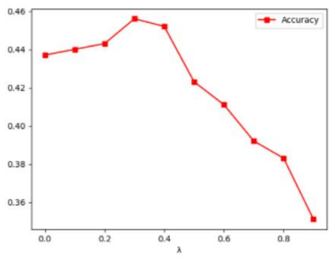
Evaluation metrics: Accuracy and F-measure

Objective evaluation

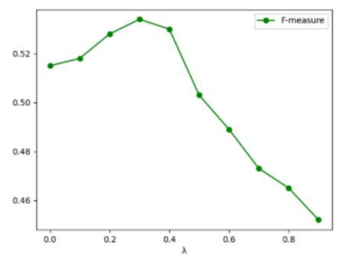
Method	Accuracy	F-measure
Santana's model [21]	0.354	0.419
Pandeya's model [13]	0.387	0.442
Our proposed model	0.456	0.534

Ablation study

A-L loss	SCMA	GCEM	Accuracy	F-measure
			0.372	0.438
✓			0.396	0.475
	✓		0.408	0.491
		✓	0.403	0.487
✓	✓		0.425	0.513
✓		✓	0.419	0.508
	✓	✓	0.437	0.515
✓	✓	✓	0.456	0.534



(a) accuracy



(b) F-measure