Emotion-Based Speech-to-Music Retrieval: A VAD-Guided Contrastive Learning Approach

조효원(전자전기공학부), 최규원(예술공학부), 최형용(국문학과)

2024 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
Proceeding of 2024 Chung-Ang University Artificial Intelligence Winter Conference

Abstract

본 연구는 감정 기반 음성-음악 매칭 성능을 향상시키기 위해, 차원적 감정 모형을 활용하는 방안을 제안한다.
감정관계를 효과적으로 학습하기 위해 contrastive learning과 kl-divergence 기반 감정 연속성 학습을 적용하였으며, momentum modeling을 도입하여 데이터 제한 환경에서도 안정적인 학습이 가능하도록 설계하였다.

Introduction

인간과 기술의 상호작용은 계속해서 변화하며, 가상현실(VR), 게임, 인공지능(AI), 심리 치료 등 다양한 분야에서 실시간 감정 반응 시스템에 대한 필요성이 점차 커지고 있다.
감정은 기존의 방식의 단순한 태그로 정의하기 어려운 복합적이고 연속적인 개념으로, 이런 시스템이 효과적으로 작동하여 사용자의 정서적 경험을 개선하는 핵심 요소로 자리 잡기 위해서는 속도와 성능에 더불어, 감정에 대한 깊이 있는 이해가 선행되어야 한다.

Aim

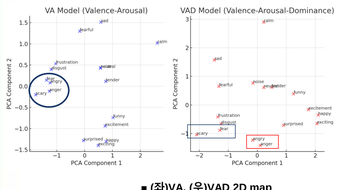
본 연구는 감정을 기반으로 speech를 통해 music을 추천하는 ESMR의 새로운 framework를 소개한다. S. Doh et al.의 Valence-Arousal (VA) 유사도를 활용한 프레임워크를 baseline으로 삼아 연구를 진행하였다. 진행한 실험은 다음과 같다.
1) VA to VAD: 감정 간 관계를 잘 반영 가능한 label matrix 사용 2) contrastive learning 적용: infoNCE를 활용하여 학습을 유도하고, hard negative와의 triplet loss를 통해 어려운 문제 학습 강화 3) similarity loss 개선: 기존 논문의 EmoSim의 오류 해결을 위해 KL_loss로 대체한다.

4) momentum model: 학습의 안정성과 modality 간 gradual alignment를 위해 momentum model을 추가한다.
본 연구는 이를 통해 안정적인 학습 환경에서 **speech-to-music 매칭 성능을 극대화하고, 감정 표현의 정밀도 향상을 이루고자 하였다.**

Methods

VA to VAD

VA based label similarity의 감정 이해 부족
ex.angry/anger vs fear
→ 심리학 모형 PAD를 본 딴 VAD 적용

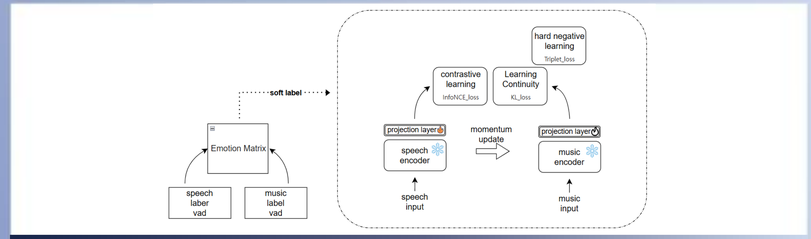


■ (좌)VA, (우)VAD 2D map

■ (좌)VAD, (우)VA soft label

Architecture

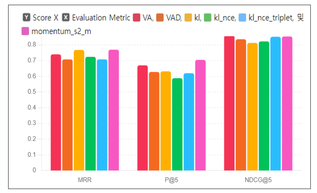
■ InfoNCE + Triplet Loss
Triplet Loss :
1) Hard negative 샘플 하나만 고려하여 전체적 감정 유사도 반영 한계 2)거리기반 손실함수
→InfoNCE: 모든 샘플 이해+Triplet: Hard negative 이해
■ KL_LOSS
EMOSIM Loss:
1)MSE 기반: 감정간 상대적 관계 반영의 부재
2)코사인 유사도와 유클리드 거리 사이 개념적 차이와 범위 차이 반영 X
→softmax + KL-divergence: 전체적인 감정 분포 이해
■ Momentum model:
Speech encoder+projection layer와 같은 구조를 한 Music Momentum model을 사용



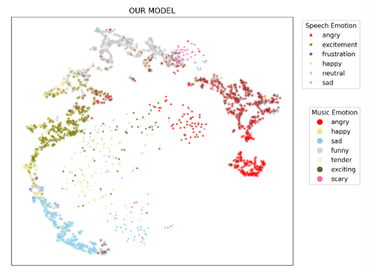
Results

■ Quantitative Results

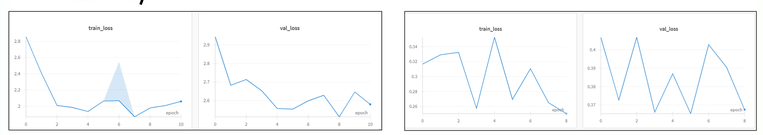
	VAD	momentum _s2_m
rank_eval_results.mrr	0.7077	0.7698
rank_eval_results.P@5	0.6279	0.7049
rank_eval_results.NDCG@5	0.8363	0.8539



■ Qualitative Results



■ Stability



■ 기존 모델의 Loss

Conclusion

우리는 ESMR task에 보다 효과적인 새로운 framework를 제안한다. 음악과 음성 사이 감정을 통한 매칭 본질을 파악하고, 이를 위한 학습 방식 도시하였다.
첫째 감정 간 관계 더 잘 반영한 label matrix를 위해 VAD를 사용했다. 둘째, InfoNCE를 통한 contrastive learning을 메인으로, triplet loss을 추가하여 hard negative에 대한 학습을 추가하였다. KL loss를 통해 감정 사이 연속적 특징을 더 잘 잡아내도록 만들었다. 마지막으로 momentum model을 추가해 학습의 안정성과 modality 간 gradual alignment를 이룰 수 있었다. 이러한 framework는 성능 평가에서 모두 더 나은 성능을 보였다.

Reference

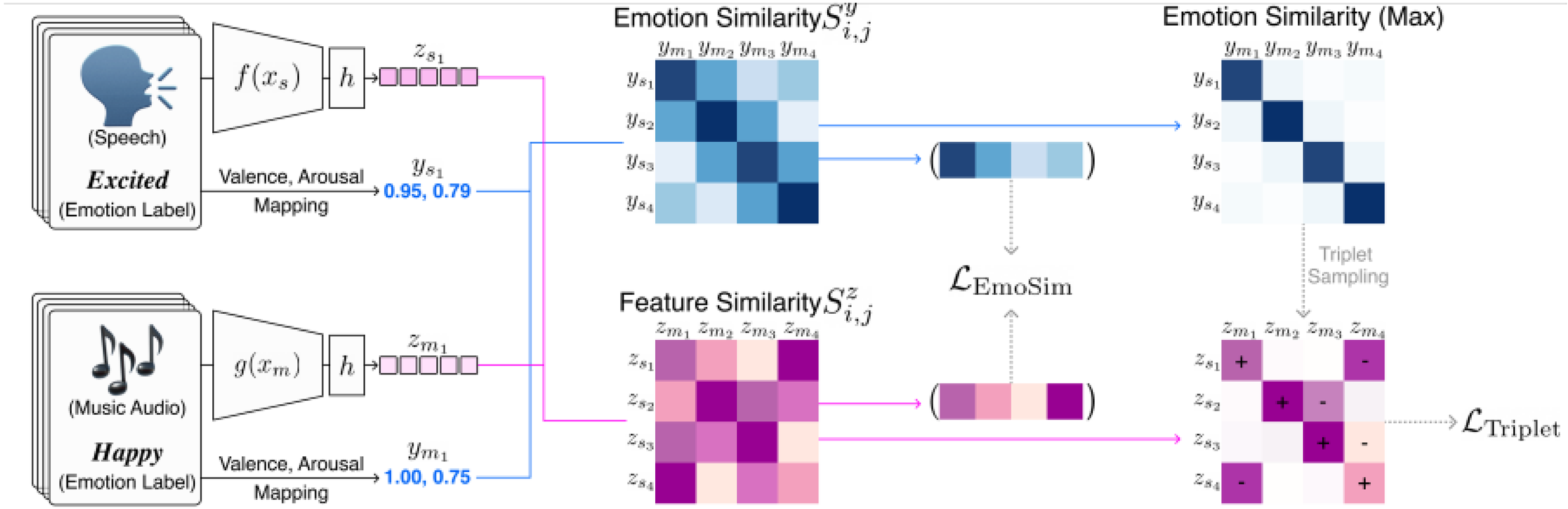
[1] S. Doh, M. Won, K. Choi and J. Nam, "Textless Speech-to-Music Retrieval Using Emotion Similarity," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5
[2] Mehrabian, A. (1996). "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament." Current Psychology, 14(4), 261-292.
[3] Russell, J. A. (1980). "A circumplex model of affect." Journal of Personality and Social Psychology, 39(6), 1161-1178.
[4]Minz Won, Keunwoo Choi, and Xavier Serra, "Semisupervised music tagging transformer," in Proceedings of International Society for Music Information Retrieval, ISMIR,
2021.
[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and
Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems, NeurIPS, 2020.
[6] Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In NeurIPS, 2021.
[7] Chen, Xinlei, Saining Xie, and Kaiming He. "An empirical study of training self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Emotion-Based Speech-to-Music Retrieval: A VAD-Guided Contrastive Learning Approach

조효원(전자전기공학부), 최규원(예술공학부), 최형용(국문학과)

2024 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
Proceeding of 2024 Chung-Ang University Artificial Intelligence Winter Conference

부록



Emotion-Based Speech-to-Music Retrieval: A VAD-Guided Contrastive Learning Approach

조효원(전자전기공학부), 최규원(예술공학부), 최형용(국문학과)

2024 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
Proceeding of 2024 Chung-Ang University Artificial Intelligence Winter Conference

부록