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

CUAI 7기 MM팀

조효원(전자전기공학부), 최규원(예술공학부), 최형용(국문학과)

**[요약]**

**본 연구는 기존 연구에서 벗어나 차원적 감정 모형을 기반으로 하여 감정의 본질적 특성에 집중한다. InfoNCE와 triplet loss을 통한 contrastive learning, kl divergence를 활용한 감정 연속성 학습을 통해 감정 간 관계를 반영하는 데 집중했다. 또, momentum encoder로 안정성을 높여 제한된 데이터에 대한 효과적인 감정 학습 방법을 제안한다.**

1. **서 론**

**인간과 기술의 관계는 지속적으로 변화하며 상호작용한다. 이러한 흐름 속에서 VR, 게임, AI 기기, 심리 치료 등 다양한 분야에서는 실시간 감정 반응 시스템이 요구되고 있다. 사용자의 정서적 경험을 개선하는 핵심 요소로 자리 잡기 위해서는 속도와 성능에 더불어, 감정에 대한 깊이 있는 이해가 선행되어야 한다.**

**본 연구는 감정을 기반으로 speech를 통해 music을 추천하는 emotion based speech to music retrieval (ESMR)의 새로운 frame work를 소개한다. [1] S. Doh et al.는 Valence-Arousal (VA) 유사도를 활용한 프레임워크를 제안하며, 우리는 이를 baseline으로 삼아 연구를 진행했다. 실험은 다음과 같다 1) VA를 VAD로 바꿈으로써, 감정 간 관계 더 잘 반영 가능한 label matrix를 사용했다. 2) InfoNce를 통해 contrastive learning을 유도하고, hard negative와의 triplet loss를 통해 어려운 문제에 대한 학습을 추가적으로 시켰다. 3) 기존 논문에서 사용한 EmoSim의 문제점을 발견하고, KL\_loss로 대체했다. 4) 학습의 안정성과 modality 간 gradual alignment를 위해 momentum model을 추가했다. 본 연구는 이러한 방법론을 통해 안정적인 학습 환경에서 speech-to-music 매칭 성능을 극대화하고, 감정 표현의 정밀도를 향상을 이룰 수 있었다**.

**2. 본 론**

**ESMR task는 두개의 데이터 셋을 연결시켜 학습시키는 일반적이지 않은 구조를 가진다. 따라서 단순한 model의 교체보단, 감정의 특성과 학습의 안정성을 반영한 구조적 향상을 위한 실험을 주로 진행했다.**

**1)VA to VAD**

**[1]S. Doh et al.에서 제시된 emotion label matrix는 emotion tags 의 VA 사이 유클리드 거리를 통해 구해졌다. 우리는 더 세밀한 감정간 관계를 포착하기 위해 [2]Mehrabian과 [3]James Russell의 Pleasure Arousal Dominance, PAD 감정 모형을 기반으로 하여, VA에 Dominance 차원이 추가된 VAD를 적용해 emotion label matrix를 만들었다.**

텍스트, 스크린샷, 사각형, 직사각형이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

Fig. 1 emotion label matrix VAD(Left), VA(Right)

유클리드 거리로 구한 IEMOCAP과 AudioSet의 감정 label 사이의 유사도 값이다. 왼쪽 VAD와 오른쪽 VA의 차이가 존재한다.

전반적으로 더 섬세한 차이가 반영되었다. 특히 Dominance를 통해 angry 와 scary 사이 모호함은 비교적 더 뚜렷하게 구분하게 되었고, sad와 neutral 사이 거리가 가까웠던 문제가 완화되었다. VAD로 교체 시 평가 지표의 미세한 저하가 있었지만, 이는 감정 간 미묘한 차이를 더 잘 반영하고, 더 난이도가 상승되었기 때문일 것으로 예측된다. 추가적으로 감정 간 유사도 계산을 위해 cosine similarity, Jsd, Dtw 등을 사용해보았지만, Euclidean distance가 감정 관계를 가장 잘 반영함을 확인했다.

**2) architecture**

텍스트, 도표, 평면도, 기술 도면이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

Fig. 2 Architecture

**아키텍처의 기본 구조이다. Speech label과 music label의 vad 값을 통해 emotion Matrix를 구한다. 이는 training에서 soft label로 사용된다. 학습 시 encoder는 Freeze 시키고, Projection layer만을 학습시킨다. 이때 encoder는 [1]S. Doh et al.에서 사용된 [4]Music Tagging Transformer와 [5]Wav2vec2.0을 사용하였다. Projection layer를 지난 speech embedding은 같은 구조를 한 Momentum Model에서 나온 music embedding과 contrastive learning을 하게 된다. 이때 loss는 InfoNCE, KL divergence, Triplet loss를 통해 계산된다.**

**3) InfoNCE & triplet loss for contrastive learning**

**우리는 Contrastive learning을 위해 InfoNCE를 사용했다. 이 손실 함수는 positive pairs 사이의 유사도를 증가시키고, 모든 negative pairs 사이의 유사도를 감소시키는 방향으로 학습을 유도한다. 여기선 Speech를 anchor로 music과의 유사도를 학습시킨다. 감정 라벨간 유사도가 높은 하나의 positive pair에 대해선 유사도를 증가하고, 나머지 negative pairs와의 유사도는 줄이는 방향으로 학습이 진행된다. [1]S. Doh et al.에서 사용된 triplet loss는 하나의 negative pair만을 고려한다는 점, 거리기반 손실함수라는 점에서 감정 간 관계 학습에 불리하다고 판단해 InfoNCE를 채택했다. 대신 비슷한 유사도를 지닌 hard negative pair에 대한 학습을 추가적으로 시키기 위해 낮은 가중치를 두고 triplet loss를 사용했다. Label matix에서 유사도가 높을수록 negative sample로 뽑힐 확률이 크도록 설계되었다. 이는** **[6]Li et al.에서 사용된 hard negative mining과 유사하다.**

**InfoNCE**

텍스트, 폰트, 라인, 화이트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

**Triplet loss**



**4) KL\_LOSS for emotion similarity regularization**

**[1]S. Doh et al.에선 감정 간 연속적인 관계를 반영하기 위한 emotion similarity regularization으로 emosim loss를 사용한다.**

**Emosim loss는 라벨 유사도(유클리드 거리)와 임베딩 유사도(코사인 유사도) 사이 MSE를 줄여나가는 방식이다. 그러나 이것은 감정 간 상대적 관계를 고려하지 않고, 단순히 각 유사도 간 수치 차이를 줄이는 방식이다. 이러한 방식은 감정 간 유사도 분포의 세부적인 패턴을 놓치기 쉽고, 전체적인 평균 중심적(Mean-seeking) 경향이 강하다는 한계가 있다. 또한 이러한 방식은 유클리드 거리 (라벨 유사도)와 코사인 유사도(임베딩 유사도) 사이 개념적 차이와 범위 차이를 무시한 방식임을 확인했다. 우리의 목적은 감정 분포를 연속적인 구조로 보존하려는 것이다. 각각의 절대적 차이 값을 줄이기 보단, 정답의 분포를 따라가, 감정 간 유사도 관계를 유지하도록 유도해야 한다.**

**우리는 각 유사도에 softmax를 적용시켜 확률 분포로 만들고, kl\_divergence 손실함수를 적용시켜 둘 사이 확률 분포 간의 차이를 계산하는 방식을 제안한다. KL-divergence는 확률 분포의 상대적인 차이를 줄이는 방식으로 동작 하며 단순히 값 자체를 맞추는 것이 아니라, 각 감정 간의 상대적 비율을 유지하도록 강제한다. 이러한 방식을 통해 감정 간 연속적 관계를 확률분포 상에서 세밀히 고려할 수 있도록 했다.**

**텍스트, 도표, 기술 도면, 평면도이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

Fig. 3 MSE는 단순 수치 차이를 계산하면서 기쁨-분노 와 기쁨- 우울, 공포 사이 관계가 어긋난 prediction 1의 loss를 더 작게 계산한다. 반면, KL\_loss는 prediction 2의 loss를 상대적으로 작게 계산한다.

**우리는 수식을 통해 InfoNCE와 KL\_loss를 함께 사용하는 것이 결국 positive pair에 대한 강조의 효과를 준다는 것 또한 확인할 수 있었다. 수식은 다음과 같다.**

**최종 Loss 값; Loss total**

**5) Momentum modelling**

**기존 모델에선 train loss와 validation loss가 변화폭이 크고, 불안정한 모습을 보였다. 이러한 학습의 불안정성으로 finetuning과 학습의 진행 여부 파악에 어려움이 있었다.**

**따라서 우리는 momentum model을 통해 보다 안정적인 contrastive learning을 도모하고자 했다. architecture에서 볼 수 있듯이 speech encoder를 통과한 speech embedding은 같은 구조의 momentum model을 지난 music embedding과 contrastive learning이 진행된다. 보통의 momentum을 이용한 contrastive learning에선 speech to music, music to speech의 contrastive loss를 모두 사용하지만([6]Li et al.의 ITC), ESMR같은 단방향 retrieval task에선 speech to music loss에 관해서만, layer를 최적화하는 방법이 더욱 효과적임을 실험을 통해 확인했다.**

**정리하자면, 기존 모델과 momentum model을 통해 music embedding을 도출한다는 점에서 차이가 존재한다. projection layer는 speech feature를 music feature와 정렬되도록 업데이트하는 역할이다.**

**Momentum model은 [7]Moco v3를 따라서 EMA(Exponential Moving Average) 방식으로 따른다. 이를 통해 성능 향상 뿐만 아니라, 더 안정적인 target embedding으로써 music embedding을 다루고, 서로 다른 modality 간 gradual alignment를 이룰 수 있었다. 특히 Fig. 4에서 보이듯, 학습이 안정적으로 진행됨을 확인할 수 있었다. train loss와 validation loss의 감소도 안정적이고 뚜렷해 졌다. 이외에도 projection layer 위에 prediction layer를 추가해봤지만, 성능에 악영향을 끼쳤다.**

**라인, 그래프, 도표이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

**라인, 도표, 그래프이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

**Fig. 4 기존 모델(위) vs momentum 추가 모델(아래)**

**6) Result**

**우리는 IEMOCAP 데이터 셋과 AudioSet을 사용해 훈련, 평가하였다. 평가는 MRR, P@5, NDCG@5 점수를 사용했다. 표 1에서 보이듯 기존 모델 과 비교했을 때 성능 향상을 이뤄냈다.**

표 1 IEMOCAP 성능 평가

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

**3. 결 론**

**우리는 ESMR task에 보다 효과적인 새로운 frame work를 제안한다. 음악과 음성 사이 감정을 통한 매칭의 본질을 더 잘 파악하고, 이를 위한 학습 방식을 도시하였다. 첫째 감정 간 관계 더 잘 반영한 label matrix를 위해 VAD를 사용했다. 둘째, InfoNCE를 통한 contrastive learning을 메인으로 하여 triplet loss을 추가하여 hard negative에 대한 학습을 추가하였다. 또, KL\_loss를 통해 감정 간 유사도 분포를 이용하여 감정 사이 연속적 특징을 잡아내도록 만들었다. 마지막으로 momentum model을 추가해 학습의 안정성과 modality 간 gradual alignment를 이룰 수 있었다. 이러한 frame work는 성능 평가에서 모두 더 나은 성능을 보였다.**

**참고 문헌**

***[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.***