공동 임베딩을 활용한 음악 자동 태깅

Music Auto-tagging with Joint Embeddings

요 약

공동 임베딩(joint embeddings)이란, 서로 다른 특성을 가지는 도메인을 하나의 공유된 공간에 임베딩하는 기법을 말한다. 음악 자동 태깅의 경우, 음원 소스 도메인과 음원 속성을 나타내는 태그 도메인이 가지는 특성이 크게 다르다. 이에 우리는 음원 데이터의 특징 공간과 태그의 특징 공간을 각각 학습할 수 있는 모델을 구성하고, 두 개의 특징 공간을 하나의 공유된 공간으로 매핑시킬 수 있도록 공동 임베딩 기법을 적용하였다. 기존에 알려진 다양한 딥러닝 음악 자동 태깅 모델들에 해당 기법을 적용하여 실험을 진행한 결과, 기존 SOTA 모델을 포함한 모든 모델에서 눈에 띄는 성능 향상을 이뤄냈다.*

1. 서 론

태그(tag)란 어떠한 물체에 연관된 정보를 담고 있는 메타 데이터의 한 종류이다. 음악의 경우, 태그 정보(예: 장르, 분위기, 연주 악기 등)를 통해 음악을 듣지 않더라도 음악의 특성을 대략 파악할 수 있다. 이에 태그는 개인 맞춤형 음악 추천, 음악 검색 등을 위한 중요한 입력 중 하나로 활용되고 있으며, 최근에는 인공지능의 발전으로 딥러닝을 활용해 음원 데이터로부터 태그를 자동 추출, 분류하는 연구가 활발히 진행되고 있다.

기존 딥러닝 음악 태깅 모델의 경우, 딥러닝을 활용하여 음원 데이터로부터 특징 추출한 결과를 태그와 비교한다. 일반적으로 비교를 위해 태그는 변환이 간단하고 준수한 성능을 내는 멀티-핫 벡터(multi-hot vector) 형태로 전처리한다. 멀티-핫 벡터는 전체 태그 개수만큼의 빈칸을 만들어 태그가 포함되는 자리에는 1, 그렇지 않은 자리에는 0을 넣는 전처리 기법이다. 이때 멀티-핫 벡터의 각 자리에 위치한 태그는 서로 독립적으로 존재하므로 태그 간의 연관성을 포함한 추가적인 내재적 정보(예: 바이올린-클래식 태그 사이의 관계)는 활용되지 않는다.

영상 인식 분야에서는 딥러닝을 통해 이미지의 특징을 추출할 뿐만 아니라 태그에 대한 내재적 정보를 추출하는 딥러닝 모델을 추가로 활용하여 성능 향상을 이뤄냈다[1]. 따라서 본 논문에서는 음원의 특징 공간과 태그의 특징 공간을 개별 학습할 수 있는 모델을 구성하고, 두 개의 특징 공간을 하나의 공유된 공간으로 매핑할 수 있는 공동 임베딩 활용 모델을 제안한다. 이를 통해 기존 딥러닝 음악 태깅 모델이 학습 시 사용하지 않는 추가적인 정보를 획득하여 학습에 활용할 수 있다는 장점이 있다. 또한 해당 기법을 기존 연구된 다양한 태깅 모델에 적용하여 음악 태깅에 있어서 태그에 포함된 내재적 정보의 중요함을 살펴보고자 한다.

2. 관련 연구

우리는 음악과 태그 도메인 간의 공동 임베딩 연구를 찾고자 하였으나, 음원 데이터만을 사용한 기존 연구는 찾을 수 없었다. 이에 본 장에서는 태그 데이터로부터 추출한 내재적 정보를 음악을 포함한 소리 데이터와 공동 임베딩한 기존 연구에 대해 살펴보고자 한다.

Elizalde[2]은 소리 데이터의 특징을 추출하기 위해 사전 학습된 모델을 활용하였다. 태그에 포함된 내재적 정보를 추출하기 위해서는 별도의 사전 학습된 단어 임베딩 모델을 활용하였고, 동일한 가중치를 갖는 샴 네트워크(siamese network)를 거쳐 해당 특징들을 공동 임베딩하였다. 이를 통해 유사한 소리를 찾는 소리 검색의 성능을 높였을 뿐 아니라 소리-태그의 양방향 검색을 최초로 수행하였다.

Favory[3, 4]은 소리 데이터의 특징을 추출하기 위해 오토인코더를 활용하였다. [3]에서는 태그에 포함된 내재적 정보를 추출하기 위해 별도의 오토인코더를 활용하였다. 하나의 오토인코더는 소리의 스펙트로그램을 복원하고, 다른 하나는 멀티-핫 벡터 형태의 태그를 복원한다. 한편 [4]에서는 태그의 내재적 정보 추출을 위해 사전 학습된 단어 임베딩 모델을 활용하였다. 두 연구 모두 소리와 태그에서 추출한 특징을 공동 임베딩하여 소리, 음악 장르, 연주 악기 등 다른 분류 작업에 사용할 수 있는 의미론적으로 풍부한 소리 표현을 학습하였다.

우리는 기존에 연구된 소리 데이터와 태그의 공동 임베딩 결과를 통해 태그의 내재적 정보가 소리 표현 학습에 도움을 준다는 사실을 알 수 있었다. 본 논문에서는 기존 연구에서 영감을 받아 서로 다른 전처리 기법(원시 음원 파형[5, 6, 7], 스펙트로그램[8])을 사용하는 딥러닝 음악 태깅 모델에 적용 가능한 일반적인 공동 임베딩 기법을 제안한다.

이후 본 논문의 구성은 다음과 같다. 3장에서는 우리가 제안한 모델과 학습 알고리즘, 손실 함수를 설명하고, 4장에서는 학습에 사용한 데이터 및 실험 결과에 대하여 설명한다. 마지막으로 5장에서는 본 논문의 내용을 정리하고자 한다.

3. 음악-태그 공동 임베딩 활용 딥러닝 모델

본 장에서는 우리가 제안한 모델의 구성 요소, 학습 알고리즘, 그리고 공동 임베딩을 위해 정의한 손실 함수에 관해서 설명한다.

---

* 소스 코드 : https://github.com/jaehwlee/jetatag

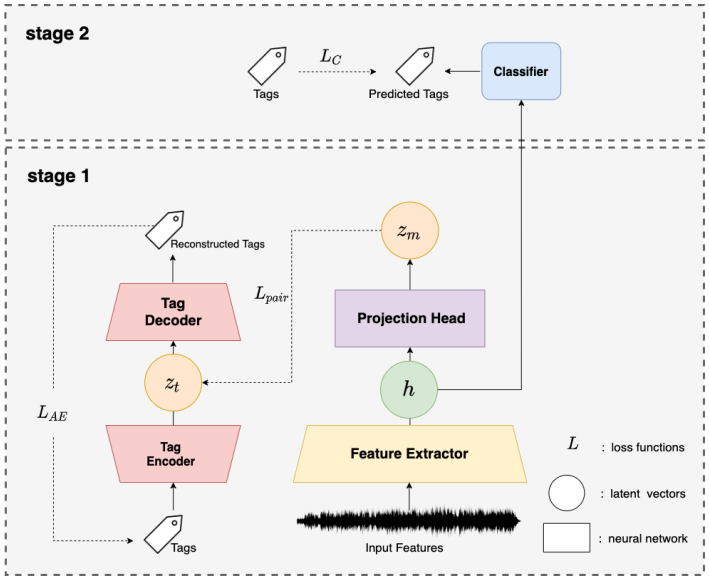


그림 1 음악-태그 공동 임베딩 학습 알고리즘 과정

3.1. 음악-태그 공동 임베딩 활용 모델 구성 요소

우리가 제안하는 모델의 구성 요소는 다음과 같다:

태그 오토인코더(Tag Autoencoder) : 태그로부터 태그 도메인 특징($z_t \in \mathbb{R}^{128}$)을 추출하는 모듈로, 우리는 [3]와 마찬가지로 오토인코더를 활용하였다. 우선 인코더(tag encoder)는 128차원의 완전 연결 층 2개로 구성하여 태그 도메인 특징을 추출한다. 이후 디코더(tag decoder)는 128차원의 완전 연결 층 1개와 50차원의 완전 연결 층 1개로 구성하여 50개의 정답 태그를 복원한 태그(reconstructed tags)를 얻는다. 추가로 같은 층의 노드 간 정규화를 위해 Layer Normalization 기법을 적용하였다.

특징 추출기(Feature Extractor) : 음원 데이터로부터 음악 도메인 특징($h \in \mathbb{R}^{1024}$)을 추출하는 모듈이다. 우리의 공동 임베딩 기법은 이미 존재하는 다른 모델에 적용 가능한 일반적인 접근 기법으로 기존의 태깅 모델[5, 6, 7, 8]에서 사용한 특징 추출기를 활용하였다. 단, 동일한 기법 적용을 위해서는 같은 차원의 음악 도메인 특징이 요구되어 각 특징 추출기로부터 1,024차원의 벡터가 추출되도록 변형하였다.

투영 모듈(Projection Head) : 음악 도메인의 특징(h)을 태그 도메인으로 투영한 내재 벡터($z_m \in \mathbb{R}^{128}$)로 매핑하는 모듈이다. 투영 모듈은 128차원의 완전 연결 층 2개로 구성하여 태그 도메인 특징(z_t)과 공동 임베딩하는 데 사용된다. 참고로 투영 모듈 없이 128차원으로 변환한 h와 z_t를 공동 임베딩하는 것도 가능하지만, 이는 공동 임베딩 적용 전보다 낮은 성능을 보였다.

분류기(Classifier) : stage 1에서 학습된 특징 추출기를 활용하여 추출된 음악 도메인의 특징(h)을 태그로 분류하는 모듈이다. 분류기는 256차원의 완전 연결 층 2개와 태깅을 위한 50차원의 완전 연결 층 1개로 구성하여 멀티-핫 벡터로 표현된 50개의 태그에 대해 각각의 포함 여부를 예측한다.

3.2. 학습 알고리즘

stage 1은 2개의 큰 모듈(태그 오토인코더, 특징 추출기+투영 모듈)로 구성된다. 태그 오토인코더(그림 1 좌측)는 태그를 입력값으로 넣어 태그 도메인의 특징 z_t와 복원 태그를 추출한다. 동시에 특징 추출기+투영 모듈(그림 1 우측)은 음원 데이터를 입력값으로 하여 음악 도메인의 특징(h)과 이를 태그 도메인으로 투영한 z_m을 추출한다. 이후 추출된 값을 바탕으로 태그 오토인코더의 손실 L_{AE}와 음악-태그의 공동 임베딩 손실 L_{pair}을 계산한다. L_{AE}은 정답 태그와 복원 태그의 비교를 위해 평균 제곱 오차(mean squared error)를 사용하였고, L_{pair}은 z_m과 z_t의 거리를 좁히기 위해 코사인 임베딩 손실 기반의 함수를 다음과 같이 제안하였다:

$$L_{pair} = \cos_{loss}(z_m, z_t) + \cos_{loss}(1-z_m, 1-z_t)$$

$$\text{단, } \cos_{loss}(z_m, z_t) = 1 - \frac{z_m \cdot z_t}{\|z_m\| \|z_t\|}$$

z_m과 z_t의 값이 너무 작을 때, 코사인 임베딩 손실이 너무 작아 학습이 안 되는 경우를 방지하기 위해 $1-z_m$, $1-z_t$에 대한 코사인 임베딩 손실을 계산하는 항을 추가하였다. stage 1에서는 이렇게 정의된 L_{AE}와 L_{pair}를 더한 값을 최종 손실로 태그 오토인코더, 특징 추출기, 투영 모듈을 동시에 학습한다.

stage 2는 stage 1에서 학습된 특징 추출기를 이용하여 음원 데이터의 특징을 추출하고, 해당 특징으로 분류기에서 멀티-핫 벡터 형태의 태그를 예측해 최종적인 태그 분류를 수행하였다. 분류기의 손실 L_C는 정답 태그와 예측 태그의 비교를 위해 이진 교차 엔트로피(binary cross-entropy)를 사용하였다.

4. 실 험

4.1. 데이터 셋 및 실험 환경 구성

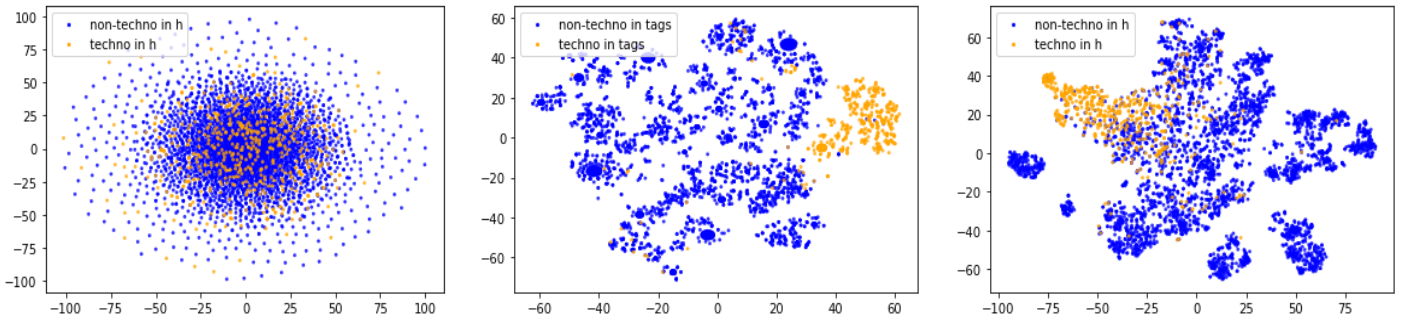
MTAT[9]는 TagATune 게임과 Magnatune 음악 저장소로부터 수집된 데이터 셋으로 음악 태깅 문제에서 가장 많이 사용되는 데이터 셋 중 하나이다. 해당 데이터 셋은 259개의 태그가 속한 25,863개의 음악 파일을 포함하며, 이전 실험[5, 6, 7, 8]과 성능 비교를 위해 사용 빈도가 높은 상위 50개의 태그만을 사용하였다. 모델의 성능 평가 지표는 모든 테스트 세그먼트에 대한 ROC-AUC의 평균을 사용하였고, early stopping 없이 stage 1은 epoch 100, stage 2는 epoch 200 동안 실험하였다.

4.2. 실험 결과

본 논문에서 제안한 공동 임베딩 기법(JE)의 성능을 검증하기 위해 공동 임베딩을 적용하지 않은 모델과 공동 임베딩을 적용한 모델의 ROC-AUC, PR-AUC 성능을 비교하였다. 음원의 특징 추출기는 기존 논문의 모델을 가능한 변형없이 사용하였으나, 공동 임베딩 적용 모델과의 성능 비교를 위해 최종 특징의 차원은 1,024로 모두 통일하였음을 밝힌다.

표 1 ROC-AUC, PR-AUC 비교 실험 결과

Feature Extractor	no JE		with JE (ours)	
	ROC AUC	PR AUC	ROC AUC	PR AUC
Harmonic-CNN[8]	0.9034	0.5137	**0.9185**	**0.5595**
Music-SincNet[7]	0.8616	0.3937	0.8823	0.4551
Sample-CNN[5]	0.8744	0.4214	0.8933	0.4853
+SE[6]	0.8809	0.4456	0.8840	0.4581
+Res[6]	0.8758	0.4271	0.8859	0.4915
+Res+SE[6]	0.8775	0.4358	0.8951	0.4903



(a) h 분포 (JE 미적용) (b) 태그 분포 (c) h 분포 (JE 적용)

그림 2 JE 적용 여부에 따른 h 분포 vs 태그 분포 비교 결과

표 1의 실험 결과를 통해, 모든 실험 결과에서 공동 임베딩 적용 모델(with JE)의 성능이 기존 대비 향상된 것을 확인하였다. 특히, 현재의 SOTA 모델인 Harmonic-CNN에서도 ROC-AUC가 0.9034에서 0.9185, PR-AUC가 0.5137에서 0.5595로 각각 1.5%, 4.5% 이상의 성능 향상을 보여 우리의 공동 임베딩 기법이 음악 태깅에 효과적임을 확인할 수 있었다.

공동 임베딩(JE)의 효과를 분석하기 위해, 임베딩 적용 여부에 따른 특징 추출기의 특징(h) 분포와 이와 쌍을 이루는 정답 태그 분포를 2차원에서 시각화하였다. 이때 데이터는 MTAT의 테스트 데이터 셋, 특징 추출기는 Harmonic-CNN, 시각화는 t-SNE 기법을 사용하였다. 그 결과, JE 미적용 h(그림 2-a)는 무작위에 가까운 산개한 분포를 보였다. 반면 정답 태그(그림 2-b)는 동일한 태그('techno')가 포함된 데이터끼리 밀집된 분포를 보였으며, JE 적용 h(그림 2-c)의 경우도 정답 태그와 유사하게 동일한 태그가 포함된 데이터끼리 밀집된 분포를 보이는 것을 확인할 수 있었다. 결론적으로 우리는 공동 임베딩 적용 특징 추출기가 태그 도메인의 내재적 정보를 학습함을 t-SNE 시각화 기법을 통해 간접적으로 확인할 수 있었다.

5. 결 론

본 논문에서는 태그 도메인의 내재적 정보를 학습할 수 있는 공동 임베딩 활용 음악 자동 태깅 모델을 제안하였다. 음악과 태그는 서로 독립된 도메인이므로 특징을 곧바로 비교하는 대신, 투영 모듈을 추가로 두어 음악 도메인 특징을 태그 도메인으로 투영한 뒤 공동 임베딩을 진행하였다. 시각화 기법을 통해 특징 추출기에서는 음원 데이터의 음향학적 특징뿐만 아니라 투영 모듈로부터 전달되는 태그 데이터의 내재적 정보도 학습하는 것을 간접적으로 확인하였다.

우리가 제안한 모델은 기존 공동 임베딩 모델과 달리 추가 데이터 없이 학습 데이터만을 사용하면서 기존 모델에 적용 가능한 일반적인 접근이라는 장점이 있다. 우리의 공동 임베딩 적용 모델은 MTAT 데이터 셋에서 ROC-AUC 0.9185, PR-AUC 0.5595를 기록하여 SOTA를 달성하였다.

참 고 문 헌

[1] A. Karpathy, and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. of the Computer Vision and Pattern Recognition (CVPR), pp.3128-3137*, 2015.

[2] B. Elizalde, S. Zarar, and B. Raj, "Cross Modal Audio Search and Retrieval with Joint Embedding based in Text and Audio," in *Proc. of the International Conference in Acoustics, Speech, and Signal Processing (ICASSP), pp.4095-4099*, 2019.

[3] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations," in *Proc. of the International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech*, 2020.

[4] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Learning Contextual Tag Embeddings for Cross-Modal Alignment of Audio and Tags," in *arXiv preprint arXiv:2010.14171*, 2020.

[5] J. Lee, J. Park, K. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proc. of the 14th Sound and music computing (SMC)*, 2017.

[6] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 366-370*, 2018.

[7] 이재환, 신종호, "필터 학습을 이용한 음악 자동 태깅," *한국정보과학회 학술발표논문집, Vol.2020 No.12 pp.388-390*, 2020.

[8] M. Won, S. Chun, O. Nieto, and X. Serra, "Data-driven harmonic filters for audio representation learning," in *Proc. of the International Conference on Acoustics, Speech and Singal Processing (ICASSP)*, 2020.

[9] E. Law, K. West, M. Mandel, M. Bay, and J. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR), pp. 387-392*, 2009.