딥 러닝을 이용한 오디오 장르 분류

*신성현 장우진 윤호원 박호종
광운대학교
*shinsh1932@kw.ac.kr

Audio genre classification using deep learning

*Shin, Seong-Hyeon Jang, Woo-Jin Yun, Ho-won Park, Ho-Chong
Kwangwoon University

요약

본 논문에서는 딥 러닝을 이용한 오디오 장르 분류 기술을 제안한다. 장르는 music, speech, effect 3가지로 정의하여 분류한다. 기존의 GMM을 이용한 장르 분류 기술은 speech의 인식률에 비해 music과 effect에 대한 인식률이 낮아 각 장르에 대한 인식률의 차이를 보인다. 이러한 문제를 해결하기 위해 본 논문에서는 딥 러닝을 이용해 높은 수준의 추상화 과정을 거쳐 더 세분된 학습을 진행한다. 제안한 방법을 사용하면 미세한 차이의 특성까지 학습해 장르에 대한 인식률의 차이를 줄일 수 있으며, 각 장르에 대해 GMM을 이용한 오디오 장르 분류보다 높은 인식률을 얻을 수 있다.

1. 서론

최근 하드웨어의 발달과 빅 데이터의 등장으로 인해 딥 러닝 (deep learning)에 대한 다양한 연구가 진행되고 있다. 특히, 딥 러닝의 성능에 큰 영향을 미치는 과적응 (overfitting) 문제를 해결할 수 있는 다양한 기술들이 개발되면서, 다른 기계 학습 방법보다 더 우수한 성능을 나타내고 있다[1]. 이와 같은 딥 러닝은 다양한 분야에 적용 가능하며, 주로 GMM (Gaussian Mixture Models)을 이용하였던 오디오 장르 분류에도 사용할 수 있다.

GMM을 이용한 오디오 장르 분류는 학습 과정에서 발생하는 수학적 복잡성을 해결하기 위한 조건이 필요하다[2]. GMM의 좋은 성능을 얻기 위해 조건에 적합한 장르 특성 매개 변수를 입력 데이터로 사용한다. 이러한 방법의 GMM을 이용한 오디오 장르 분류는 장르에 따른 인식률의 차이가 있다. 다른 장르의 특성과 다르게 명확한 특성을 갖는 speech의 인식률은 높게 나오지만, 특성이 유사한 music과 effect는 서로 혼동해 인식률이 speech의 인식률에 비해 낮게 나온다. 즉, 기존의 GMM을 이용한 오디오 장르 분류는 특성이 유사한 장르 간의 분류에 적합하지 않은 구조이다.

본 논문에서는 기존의 GMM을 이용한 오디오 장르 분류의 문제점을 보완하기 위해 딥 러닝을 이용한 오디오 장르 분류를 제안한다. 제안한 방법에서는 사람의 신경망 구조로 디자인된 딥 러닝을 이용해 높은 수준의 추상화 과정을 거쳐 더 세분된 학습을 진행한다. 또한, RBM (Restricted Boltzmann Machine)을 사용해 부족한 학습 데이터 수를 보완해주고, L2-정규화와 dropout을 사용해 특성이 강한 장르에 과적응되는 것을 최소화시켜 딥 러닝의 학습 효과를 높인다[3]. 이와 같은 방법을 사용하는 딥 러닝은 GMM에서는 구분할 수 없는 유사한 특성 차이까지 구분할 수 있다. 따라서 장르에 따라 인식률의 차이를 나타내는 GMM과 달리, 제안한 방법에서는 장르와 관계없이 유사한 인식률을 나타낸다. 또한, GMM보다 높은 인식률을 나타낸다.

2. 제안하는 딥 러닝의 네트워크 설계 방법

제안하는 방법은 조직 (texture) 단위로 장르를 결정하는 on-line 분류이다. 제안하는 방법에 따라 입력 데이터는 각 장르의 특성을 표현하기 위한 최소의 길이를 1초로 잡아 조직 단위로 구성한다. 입력 데이터로 사용하기 위해 특성을 뽑는 과정은 프레임 단위로 진행되며 특성의 구성 요소는 첫 5개의 MFCC, spectral centroid, spectral roll-off, spectral flux, zero-crossing rate이다. 이처럼 구한 프레임 단위 특성의 평균과 분산 값 18개와 조직 단위의 low-energy를 포함해 총 19차 벡터로 입력 데이터를 구성한다[4].

딥 러닝 네트워크 설계에 필요한 핵심 매개 변수는 네트워크의 크기를 결정하는 매개 변수와 학습 과정에서 필요한 매개 변수로 나누어진다.

네트워크의 크기를 결정하는 매개 변수는 은닉층 (hidden layer)의 수와 각 은닉층의 뉴런 수로 구성되며 본 논문에서는 3개의 은닉층과 각 120, 45, 30개의 뉴런을 사용한다. 즉, 총 9,234,000 (19 x 120 x 45 x 30 x 3) 개의 가중치 (weight)를 갖는다. 가중치의 과적응을 막기 위해 최소 가중치 개수 이상의 학습데이터가 있어야 한다[5]. 그러나 9,234,000개의 학습 데이터를 얻기는 쉽지 않아 RBM을 사용해 가중치를 초기화해준다.

학습 과정에 필요한 매개 변수는 학습률 (learning rate), mini-batch 크기, 정규화 매개 변수 (regularization parameter), dropout 크기, 학습 반복 횟수 (epoch)로 구성된다. 여기서 mini-batch 크기는 학습량을 감소시켜 연산량을 줄이며, 정규화 매개 변수와 dropout 크기는 과적응을 막는다. 본 논문에서는 학습률을 0.1로 설정하며, mini-batch 크기는 1로 설정한다. 또한, 정규화 매개 변수와 dropout

크기는 0.01과 3번째 은닉층만 25로 설정하며, 학습 반복 횟수는 100으로 설정한다.

3. 성능 평가

성능 평가에서 사용한 데이터는 TV 방송에서 사용된 각 32분 길이의 music, speech, effect 음원이다. 이 중 90%를 학습 데이터로 사용하며, 남은 10%를 실험 데이터 (test data)로 사용한다. 기계 학습에 있어 실험 데이터에 hyper-parameter가 과적응 되는 것을 방지해 주는 확인 데이터 (validation data)를 사용해야 한다. 그러나 제안하는 방법은 네트워크에 따른 성능의 차이를 제안하여, 공통으로 적용되는 확인 데이터의 효과는 동일하게 제외하였다.

본 논문의 비교 대상인 GMM은 딥 러닝과 동일한 데이터 집합을 사용하며 k-means 알고리즘은 10개의 무리 (cluster)를 사용한다. EM 알고리즘의 학습 반복 횟수는 200번이며 δ (델타 : 오차 허용 범위) 값은 0.001로 설정한다. 이러한 방법으로 계산된 10개의 가우시안 확률 분포를 혼합하여 사용한다.

위와 같이 설정된 GMM을 이용한 장르 분류 인식률은 표 1과 같다. 전체 인식률은 87.16%이며, speech의 인식률에 비해 music, effect의 인식률이 낮은 것을 확인할 수 있다.

표 1. GMM을 이용한 장르 분류 인식률 (%)
Table 1. Audio genre classification accuracy using GMM. (%)

True \ Estimate	Music	Speech	Effect
Music	**85.41**	4.17	10.42
Speech	5.21	**90.62**	4.17
Effect	11.98	2.60	**85.42**

정규화 과정을 제외한 딥 러닝의 장르 분류 인식률은 표 2와 같다. 전체 인식률은 93.06%로, GMM의 전체 인식률보다 5.9%p 향상했으며, 다른 두 장르의 인식률에 비해 Speech의 인식률이 높지만, 차이가 감소한 것을 볼 수 있다.

표 2. 정규화 과정을 제외한 딥 러닝을 이용한 장르 분류 인식률 (%)
Table 2. Audio genre classification accuracy using deep learning without regularization. (%)

True \ Estimate	Music	Speech	Effect
Music	**93.74**	3.13	3.13
Speech	2.08	**94.27**	3.65
Effect	6.25	2.60	**91.15**

정규화 과정을 적용한 딥 러닝의 장르 분류 인식률은 표 3과 같다. 최종 모델의 전체 인식률은 93.23%로, 정규화 과정을 하지 않은 딥 러닝보다 전체 인식률이 0.17%p 향상했으며, 각 장르의 인식률이 더 평활화된 것을 볼 수 있다.

표 3. 정규화 과정을 적용한 딥 러닝을 이용한 장르 분류 인식률 (%)
Table 3. Audio genre classification accuracy using deep learning applied regularization. (%)

True \ Estimate	Music	Speech	Effect
Music	**92.19**	2.08	5.73
Speech	2.60	**93.23**	4.17
Effect	4.17	1.56	**94.27**

장르별 GMM과의 최종 성능 차이는 표 4와 같으며 GMM에서 성능 떨어졌던 music과 effect의 인식률이 크게 향상되고, 전체 인식률이 향상된 것을 확인할 수 있다.

표 4. 제안하는 딥 러닝과 기존 GMM의 장르 분류 인식률 차이 (%p)
Table 4. The accuracy difference between deep learning and GMM. (%p)

	Music	Speech	Effect	Average
Difference	6.78	2.61	8.85	6.07

4. 결론

본 논문에서는 딥 러닝을 이용한 오디오 장르 분류 방법을 제안하였다. 제안한 방법은 RBM을 이용해 가중치를 초기화시키고 L2-정규화와 dropout을 이용해 딥 러닝 네트워크를 만든다. 이러한 방법으로 만들어진 딥 러닝을 이용해 기존 GMM에서 분류하지 못하던 장르 특성의 미세한 차이까지 분류한다.

제안한 방법으로 기존 방법인 GMM에서 낮은 인식률을 보인 music과 effect의 인식률이 높아지며, 전체 인식률도 GMM을 이용한 장르 분류보다 높은 인식률을 나타낸다.

참고문헌

[1] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning", *Nature*. pp. 436-444, May 2015.

[2] D. Reynolds, "Gaussian mixture models", *Encyclopedia of Biometrics*. pp. 827-832, July 2015.

[3] G. E. Hinton, R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", *Science*. vol. 313, pp. 504-507, July 2010.

[4] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals", *IEEE Transactions on Speech and Audio Processing*. vol. 10, pp. 293-302, July 2002.

[5] S. Lawrence, C. L. Giles, A. C. Tsoi, "Lessons in Neural Network Training: Overfitting May be Harder than Expected", *AAAI*. pp. 540-545, 1997.