딥 러닝 기반의 오디오 장르 및 품질의 다중 분류 기술

*신성현 조효진 장 원 박호종

광운대학교

*shinsh1932@kw.ac.kr

Multiple Classification of Audio Genre and Quality based on Deep Learning

*Shin, Seonghyeon Cho, Hyojin Jang, Won Park, Hochong

Kwangwoon University

classificationclassification

요약

본 논문에서는 스펙트로그램을 이용하여 딥 러닝 기반으로 오디오 장르와 품질의 다중 정보를 동시에 분류하는 기술을 제안한다. 기존 딥 러닝 기반의 오디오 정보 인식 기술은 각각의 정보 인식을 목표로 독립 네트워크를 설계하고, 여러 정보를 동시에 인식하기 위하여 각각에 특화된 여러 네트워크를 사용한다. 이러한 문제점을 보완하기 위해 본 논문에서는 디지털 오디오의 대표 특성인 스펙트로그램을 기반으로 범용성이 있는 특성을 추출하고, 단일 네트워크로 학습시켜 장르 및 품질을 동시에 분류하는 다중 분류 기술을 제안한다. 제안하는 방법으로 단일 분류 성능과 유사한 다중 분류 성능을 얻을 수 있다.

1. 서론

인간의 신경망을 모델로 만들어진 딥 러닝을 사용한 오디오 기술은 오디오 장르, 품질 분류에서 높은 성능을 보여주고 있다[1, 2, 4]. 이러한 기술을 사용하여 장르별 최적의 이퀄라이저 적용, 음원에 표시된 품질이 아닌 실제 품질 확인 등의 다양한 서비스를 제공할 수 있다. 하지만, 각 정보 인식 분야마다 최적의 특성 벡터가 달라 다양한 서비스를 제공하기 위해서는 그에 맞는 다수의 시스템이 필요하다. 단일 시스템으로 모든 오디오 정보를 인식하는 인간의 청각 시스템과 대조되는 부분이다[3]. 또한, 기존 장르 분류의 학습과 성능 평가는 유통되는 오디오 음원과 다르게 원본 음원만을 사용하여 실제 기술 적용에 한계가 있다.

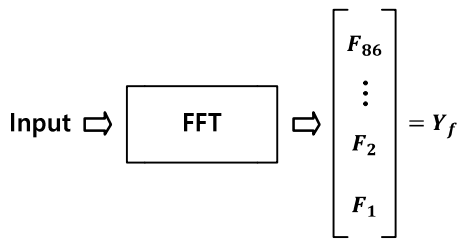
기존의 오디오 장르 및 품질 분류는 수학을 기반으로 신호를 분석하여 각 분야를 잘 수행할 수 있는 특성을 추출한다. 그 예로, 오디오 장르 분류는 장르별 주파수 성분을 확인하기 위해 복잡한 수학을 기반으로 특성을 추출하며, 품질 분류는 noise, bitrate 변화 등을 감지하기 쉬운 MDCT (modified discrete cosine transform)를 기반으로 특성을 추출한다[1, 2]. 이러한 특성은 단일 분류에서 우수한 성능을 보이지만 각각의 정보 인식을 목표로 만들어진 특성과 학습된 네트워크로 다중 분류에는 적합하지 않으며, 인간의 신경망을 더욱 정확히 구현하기 위해서는 범용성이 있는 시스템이 필요하다.

본 논문은 스펙트로그램 (spectrogram)을 기반으로 특성을 추출하여 오디오 장르 분류 및 품질 분류를 동시에 진행하는 다중 분류 방법을 제안한다. 스파이크그램을 기반으로 texture frame 단위의 통계적 특성으로 이루어진 특성 벡터를 추출하고 단일 네트워크를 사용해 장르와 품질을 동시에 분류한다. 오디오의 대표 특성인 스파이크그램을 기반으로 얻은 특성 벡터를 사용하여 범용성을 높이며, on-line 분류가 가능하다. 또한, 단일 분류 성능과 유사한 성능을 얻을 수 있다.

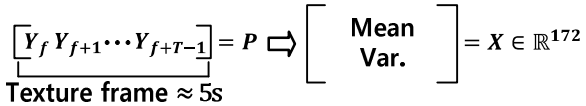
2. 제안하는 방법

2.1 스펙트로그램 기반 특성 추출

본 논문의 특성 벡터는 스펙트로그램을 기반으로 추출한다. 그림 1은 스펙트로그램을 기반으로 특성 벡터의 추출 과정과 구성을 나타낸다. 먼저, 86개의 Mel-scale 밴드로 구성된 스펙트로그램 Y_f를 약 100ms 길이의 frame 단위로 구한다.



(a)



(b)

그림 1. 특성 벡터 X의 추출 과정 및 구성 (a) frame 단위 과정 (b) texture frame 단위 과정

Fig. 1. Extraction process and composition of feature vector X (a) frame-based process (b) texture frame-based process

계산된 Y_f를 약 5초의 길이를 갖는 texture frame으로 묶은 P를 만든다. 마지막으로 P의 밴드별 평균 분산을 구하여 172-D 특성 벡터 X를 만든다.

2.2 네트워크 구조

본 논문에서는 오디오 장르 및 품질 분류기로 deep neural network (DNN)를 사용한다[4]. Fully-connected 구조의 일반 DNN으로 그림 2의 구조를 갖는다.

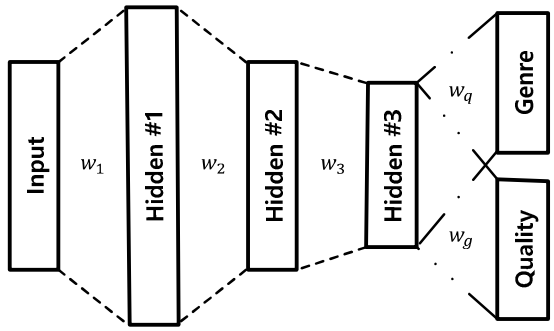


그림 2. DNN과 출력단의 구조
Fig. 2. Architecture of DNN and output layer.

본 논문에서는 3개의 hidden layer를 갖는 DNN을 사용한다. DNN에 필요한 매개 변수 (hyper-parameter)는 실험을 통해 얻었고 각 매개 변수의 설정값은 다음과 같다. 네트워크 구조 파라미터로 3개의 hidden layer는 각 [300, 60, 30] 개의 뉴런과 0.8의 dropout 유지 확률을 갖는다. Hidden layer의 활성화 함수는 ReLU (rectified linear unit)를 사용하였고, 출력단은 softmax 함수를 적용하였다. 네트워크 학습률은 0.007로 설정하고 학습 반복 횟수는 1000번으로 한다.

3. 성능 평가

성능 평가에는 오디오 장르 분류 성능 평가에 대표적으로 사용되는 GTZAN 데이터 세트를 사용한다. GTZAN 데이터 세트는 classical, country, disco, hip hop, jazz, rock, blues, reggae, pop, metal로 총 10개의 장르로 구성되며, 장르별 100개의 파일로 구성되어 있다. 각 파일은 30초이며 5초 단위로 추출된 특성을 30초가 되는 6번의 평가 결과 중 가장 높은 평균 확률을 갖는 장르와 품질을 해당 파일의 장르 및 품질로 판단한다. 예로, 임의의 음원 파일이 입력되면 출력단의 장르 분류단은 장르에 대한 6번의 판정 결과를 출력하고, 품질 분류단은 품질에 대한 6번의 판정 결과를 출력하여 파일에 대한 장르 및 품질을 동시에 판정한다.

표 1는 제안하는 방법의 장르 및 품질의 다중 분류 성능을 보여준다.

표 1. 제안하는 방법의 성능 (%)
Table 1. Performance of proposed method (%).

Est. / True	Genre										Quality	
	cl	co	di	hi	ja	ro	bl	re	po	me	ori	dis
cl	**93.5**	2.0	-	-	-	1.5	-	-	-	-		
co	1.5	**70.5**	2.0	-	4.5	12.0	5.0	-	2.0	-		
di	2.0	9.0	**62.0**	3.5	2.0	7.5	2.0	3.5	7.0	1.5		
hi	2.0	-	5.0	**67.5**	1.5	-	-	7.5	10.0	4.0		
ja	4.5	3.5	1.5	-	**85.5**	-	-	-	-	2.5		
ro	1.5	16.0	14.0	-	3.0	**50.0**	2.0	-	1.5	10.5		
bl	-	6.0	3.0	-	3.5	4.5	**77.5**	2.5	-	-		
re	1.5	10.0	6.0	7.0	3.0	-	2.0	**65.0**	5.0	-		
po	2.0	9.0	7.0	5.0	-	3.5	-	-	**71.0**	-		
me	-	-	-	-	1.5	5.0	-	-	-	**91.0**		
ori											**80.6**	19.4
dis											15.5	**84.5**

1% 이하의 오차는 '-'로 표현하였고, 장르 분류는 10개의 장르를 모두 분류하며, 품질 분류는 원본과 다양한 음질 저하를 적용한 것으로 나눈 바이너리 분류로 진행된다. 품질 분류에 사용된 음질 저하 기술은 pink noise와 AAC codec을 사용하였다. 평가 방법은 10-fold cross validation을 사용하며, 품질 분류를 위해 만들어진 음질이 저하된 음원의 장르도 함께 분류한다. 표 1로 장르 분류의 평균 정확도 73.35%와 품질 분류의 평균 정확도 82.6%를 확인할 수 있다.

표 2는 제안하는 방법의 단일 분류 성능과 다중 분류 성능을 비교한 것이다. 비교를 위해 진행된 단일 분류 성능은 장르 분류를 위한 시스템과 품질 분류를 위한 시스템에 다중 분류와 동일한 데이터 세트를 각각 적용하여 얻은 성능이다.

표 2. 제안하는 방법의 다중 분류와 단일 분류 성능 (%)
Table 2. Performance of proposed method comparison between multi task and single task (%).

	Multi Task	Single Task
Genre	73.35	73.75
Quality	82.6	83.15
Average	77.95	78.45

오디오 장르 분류는 0.4%p, 품질 분류는 0.55%p 낮은 성능으로 단일 분류 성능보다 평균 0.475%p 낮은 성능을 얻어, 다중 분류와 단일 분류가 유사한 성능을 갖는 것을 확인할 수 있다.

4. 결론

본 논문에서는 스펙트로그램을 이용한 딥 러닝 기반 오디오 장르 및 품질의 다중 분류 방법을 제안하였다. 오디오의 대표 특성인 스펙트로그램을 이용하여 특성 벡터를 만들고, 단일 네트워크로 2개의 분류를 동시에 분류하여 시스템의 범용성을 높였다. 또한, 단일 분류와 유사한 성능을 얻을 수 있다.

감사의 글

본 연구는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B03930923).

참고문헌

[1] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, "Unsupervised Learning of Sparse Features for Scalable Audio Classification," *ISMIR*, pp.681-686, Sep. 2011.

[2] D. Luo, W. Luo, R. Yang, and J. Huang, "Identifying compression history of wave audio and its applications," *ACM Trans. Multimedia Comput., Commu., Appl.*, Vol. 10, No. 3, pp. 30, 2014.

[3] C. Alain, SR. Arnott, S. Hevenor, S. Graham, and CL. Grady, ""What" and "where" in the human auditory system," *Proc. Natl. Acad. Sci. U.S.A.* 98, 12301 - 12306, 2001.

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