2 차원 특징점 활용 학습 모델 기반 클래식 악기 자동 분류 기법 비교 분석

조혜원, 김준영
성신여자대학교

20221434@sungshin.ac.kr, jkim@sungshin.ac.kr

Comparative Analysis of Automatic Classification of Classical Music Instruments using 2D-Feature based Learning Model

Hye Won Cho, Joon Young Kim*
Sungshin Women's Univ.

요 약

본 논문은 클래식 악기 오디오 데이터를 분류하기 위해 학습 알고리즘 기반 이미지 분류 기법을 고찰하고자 한다. 기존 악기 데이터셋의 오디오 샘플 일정 시간 단위 분할 및 변환 기법등을 통해 2D 이미지 특징점 추출을 통해 분류를 진행한다. 분류 성능 비교 분석을 위해 4 개 학습 모델을 적용하였으며 이를 통한 고려사항 등도 병행 도출한다.

Ⅰ. 서 론

음향 데이터 분석은 음악 정보 검색, 악기 인식, 자동 채보 등 다양한 응용 분야에서 중요한 역할을 한다. 기존에는 오디오 신호의 주파수 특성을 직접 분석하거나 특정 음향 특징을 추출하는 방식이 주로 사용되었으며, 이러한 접근법은 비교적 직관적이지만 복잡한 신호 처리 기법을 필요로 한다. 그러나 최근에는 오디오 데이터를 이미지 형태로 변환하고, 이미지 분류 모델을 적용하는 연구가 활발히 진행되고 있다. 이미지 변환을 통해 오디오 신호의 시간-주파수 정보를 보다 구조적으로 분석할 수 있으며, CNN 기반 모델 적용을 통한 성능 고도화가 가능하다 [1]. 이를 기반으로 하여 다양한 학습 모델 적용 기반 분류 정확률 비교 분석을 통해서 특정 환경 기반 모델 적용 및 복합 모델 개발 등에도 적용이 가능하다. 본 논문에서는 악기 오디오 데이터를 2 차원 특징점 형태로 변환한 악기 자동 분류 기법을 제안하고자 하며 다양한 모델 기반한 성능 비교 분석도 병행하고자 한다.

Ⅱ. 분류 시스템 모델

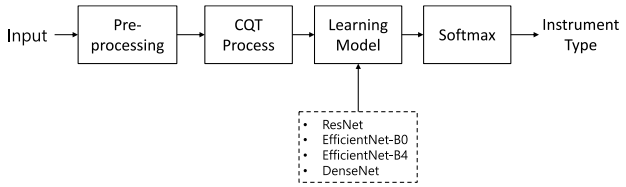

그림 1. 분류 시스템 모델 구조도

본 연구에서는 단일 악기 오디오 샘플을 추출하여 실험에 활용하였다. 수집된 오디오 데이터는 10 초 단위로 분할하여 데이터의 균형성과 다양성을 고려하였으며 길이가 10 초 미만인 파일은 제외하였다. 균일한 클립 분할을 통해 모델 입력에 일정한 데이터셋을 구성하였다. 이후 각 오디오 클립은 주파수 분석 기반의 이미지로 변환되었다. 오디오의 시간-주파수 특성을 효과적으로 반영하기 위해 그림 1 에서와 같이 CQT (Constant-Q Transform) 기법을 사용하였다. CQT 는 모든 주파수에서 시간-주파수 해상도 간 타협을 동일하게 적용하는 STFT 와 달리 낮은 음역은 주파수 정확도, 높은 음역은 시간 정확도에 집중함으로써 음악적 특성을 반영한다[2]. 생성된 CQT 스펙토그램은 각 악기의 음색 및 연주 특성을 이미지 형태로 나타내며, CNN 기반 분류 모델의 입력으로 효과적이다. 데이터 수집 결과는 표 1 을 참고한다. 추가로 본 연구에서는 공개 데이터셋 중 하나인 URMP (University of Rochester Music Performance) 데이터셋을 사용하였다[3]. URMP 는 음악 퍼포먼스 기반의 멀티모달 데이터셋으로, 바이올린, 첼로, 플루트, 트럼펫 등 13 종의 클래식 악기로 구성된 다양한 독주 및 앙상블 연주 데이터를 포함하고 있다.

표 1. URMP 데이터 정제 결과

악기라벨	악기	개수	악기라벨	악기	개수
Vn.	바이올린	344	Sax.	섹소폰	88
Va.	비올라	147	Bn.	바순	23
Vc.	첼로	129	Tpt.	트럼펫	236
Db.	더블베이스	34	Hn.	호른	59
Fl.	플룻	187	Tbn.	트럼본	99
Ob.	오보에	67	Tba.	튜바	51
Cl.	클라리넷	103			

또한 본 연구에서는 CNN 아키텍처 위주로 적용하였으며 주요 모델인 ResNet, EfficientNet-B0, EfficientNet-B4, DenseNet 들을 선택하여 악기 분류 기법으로 활용하였다. ResNet 은 층이 깊어질수록 발생하는 학습의 어려움을 해결하기 위해 잔차 연결(residual connection)을 도입한 모델로, 깊은 신경망에서도 효과적인 학습이 가능하다[4]. EfficientNet 은 모델의 깊이, 너비, 해상도를 균형 있게 확장하는 컴파운드 스케일링 기법을 적용한 모델로, 효율적인 연산과 높은 정확도를 동시에 보여준다. B4 는 B0 보다 더 깊고 넓은 구조를 가지며 성능이 향상된 모델이다[5]. DenseNet 은 각 층이 이전 모든 층의 출력을 입력으로 받아 연결되는 구조로, 정보 흐름과 그레디언트 전파가 원활하여 학습 효율과 표현 능력이 뛰어난 모델이다[6]. 선택한 모델은 각각 잔차 연결, 효율적인 네트워크 설계, 촘촘한 연결 구조 등의 특성을 지니고 있어 다양한 관점에서 성능을 비교할 수 있다.

모델 학습 시 배치 크기는 16 과 32 두 가지로 설정하였으며, 학습의 안정성과 효율성을 평가하기 위해 학습 에폭은 10, 20, 50, 100, 200, 500 의 단계별로

늘려가며 실험을 진행하였다. 학습률은 Adam 옵티마이저를 사용하여 0.001 로 고정하였으며, 이는 실험 초기 단계에서 과도한 발산을 방지하기 위함이다.

평가지표로는 테스트 데이터셋에 대한 정확도와 Macro F1-score 를 활용하였다. 정확도는 전체 샘플 중 정확히 분류된 비율을 의미하며, Macro F1-score 는 클래스 불균형 문제를 완화하고 모델의 균형 잡힌 분류 성능을 평가하기에 적합하다. 상기 언급된 실험 환경들의 경우 표 2 를 참고한다.

표 2. 분류 실험 위한 주요 환경 세팅 및 파라미터

GPU Type	NVIDIA H100
GPU Core	16 코어
GPU RAM	30GB
RAM	128.00GiB
Optimizer	Adam
Learning Rate	0.001
Batch Size	16, 32
Epochs	10-500

Ⅲ. 실험 결과

모델별 분류 정확률 도출 및 비교를 보여주는 그림 2 내 ResNet 모델의 경우 전반적으로 높은 성능을 나타냈다. 특히 batch size 16 설정에서 500 epoch 학습한 경우, 정확도 97.26%, F1-score 0.9716 으로 가장 우수한 성능을 기록하였다. 이는 같은 조건에서 정확도 92.69%, F1-score 0.9281 의 성능을 보인 EfficientNet-B0 나 정확도 96.35%, F1-score 0.9638 의 성능을 보인 DenseNet 대비 우수한 결과이다.

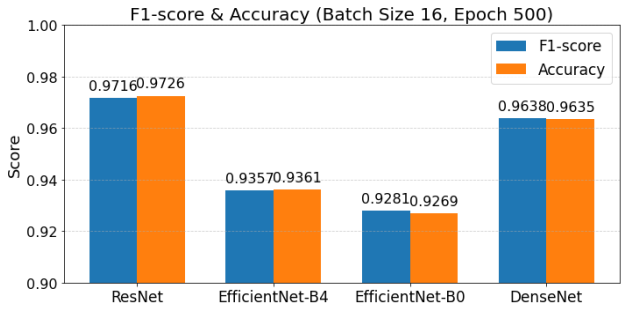

그림 2. 모델별 F1-score & Accuracy

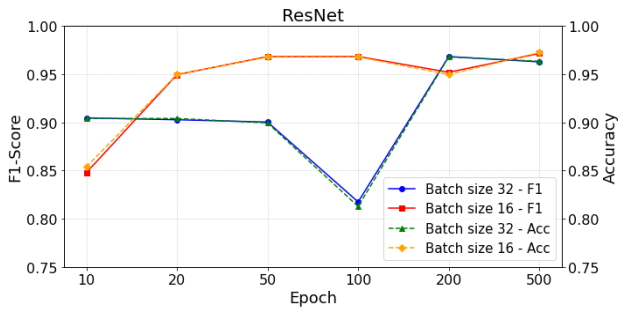

그림 3. Resnet 의 Epoch 별 F1-score & Accuracy

ResNet 경우 그림 3 과 같이 비교적 적은 Epoch 수인 50 에서도 F1-score 0.9684 를 기록하여 빠른 수렴과 높은 안정성을 보여주었다. 이는 ResNet 의 Skip connection 구조가 심층 모델 내 정보 손실 최소화 및 학습 안정성을 높였다. 그림 2 에서 92% 이상 안정적인 성능을 보인 EfficientNet-B0 는 상대적으로 파라미터 수가 적고 연산 효율이 높아, 경량 모델로서의 활용 가능성을 제시한다. 반면 EfficientNet-B4 는 전반적으로 ResNet 및 EfficientNet-B0 보다 낮은 성능을 보였다. 이는 B4 가 더 깊고 복잡한 구조로 인해 과적합 취약 및 학습 데이터양에 의존성이 강한 것으로 고려된다.

또한 그림 4 경우 모델 구조뿐만 아니라 학습 조건의 선택이 모델 성능에 큰 영향을 미친다는 것을 보여준다. 대부분의 모델이 batch size 16 조건에서 더 높은 성능을 보였으며, 특히 200 epoch 이상의 충분한 학습이 이루어질 경우 안정적인 분류 정확도를 보이는 것이 확인되었다. ResNet 과 EfficientNet-B0 모델 경우 전체 실험에서 가장 일관된 고성능을 보였으며, DenseNet 은 장기 학습 시 효과적인 성능 향상을 보여주었다.

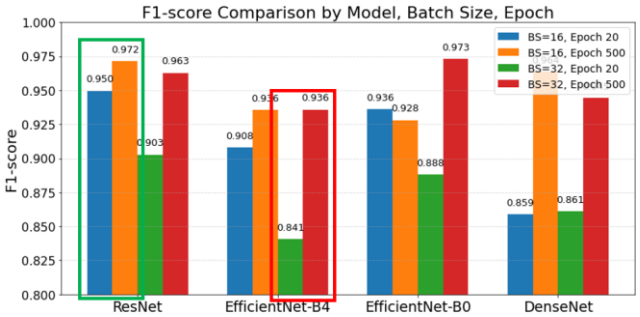

그림 4. batch size 에 따른 모델별 F1-score

Ⅳ. 결론

본 연구는 악기 분류 자동화를 위한 기반으로써, 음악 신호 분석과 인공지능 기술의 융합 가능성을 실증적으로 보여주었다. EfficientNet-B4 는 모델 복잡도 대비 성능 향상에 한계를 보인 바 추가적인 하이퍼파라미터 튜닝이나 정확도 개선 연구가 진행되어야 할 것이다. CQT 이외에도 CWT(continuous Wavelet Transform), Mel-spectrogram 등 다양한 시간-주파수 변환 기법과 모델 구조를 실험하여 최적 조합을 탐색하는 연구가 병행되어야 할 것으로 고려된다.

참 고 문 헌

[1] Han, H. and Jung, Y., “Comparison of audio input representations on piano transcription using neural networks,” Journal of the Korean Data & Information Science Society, Vol.32, no.2, pp.439-452, 2021.

[2] Brown, J. C., “Calculation of a constant Q spectral transform”, Journal of the Acoustical Society of America, vol. 89, no.1, pp. 425-434, Jan. 1991.

[3] University of Rochester, “University of Rochester Music Performance (URMP) Dataset”, 2019, (https://labsites.rochester.edu/air/projects/URMP.html).

[4] Kaiming He, Xiangyu Zhang, Shaoping Ren, and Jian Sun, “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770- 778, 2016.

[5] Mingxing Tan and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 6105- 6114, 2019.

[6] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely Connected Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700- 4708, 2017.