**노래 장르 분류:**

**1D Conv와 스펙트로그램 CNN의 병렬 적용**

****

**IIE4123 딥러닝과 응용**

**산업공학과**

**Team 13**

**2019147001 김승유 2019147022 정이현**

**2019147027 김정환 2020147049 김현동**

**목차**

1. **Introduction**
2. **음성 데이터와 스펙트로그램**
   1. Audio Data
   2. Spectrogram
   3. Wavegram
   4. Sampling Rate
3. **데이터셋**
   1. GTZAN
   2. FMA(Free Music Archive)
4. **데이터 전처리**
   1. Data Pipeline
   2. Data Augmentation
5. **모델**
   1. Limitation of 1D Convolution
   2. PANNs
   3. Wavegram-Logmel-CNN
6. **결과**
   1. Wavegram-Logmel-CNN Result
   2. Comparison with Baseline Model(ResNet50)
7. **결론**
8. **한계점과 개선 방안**
9. **참고자료**
10. **Introduction**

음악 장르는 청취자가 음악을 선택하는 데에 있어 매우 중요한 요소 중 하나이다. 반면, 대형 음악 스트리밍 사이트 등에서 음악 장르 구분에 오류가 있는 경우가 많다. 노래의 장르가 명확히 구분되지 않음에 따라 장르별로 노래를 청취하고자 하는 이가 원하지 않는 노래가 재생돼 불편함을 초래할 수 있다. 이에 높은 정확도로 음악의 장르 구분을 하는 모델을 구현하는 것을 목표로 설정했다. GTZAN 데이터셋과 FTA 데이터셋을 활용해 음악 장르 구분을 하는 딥러닝 모델 구현을 진행하고자 하였다.

1. **음성 데이터와 스펙트로그램**
   1. **음성데이터**

음성 데이터는 보통 매우 고차원이며, 긴 시간 동안의 신호를 다루기 때문에 처리가 복잡하다. 또한, 환경 소음, 마이크 품질의 차이, 각종 잡음 등 외부 요인으로 인한 노이즈가 존재하며, 이것이 모델의 성능에 영향을 미칠 수 있다.​ 이 외에도 음성 데이터에는 매우 다양한 억양, 악센트, 발음 등이 포함되어 이에 대한 고려가 필수적이다. 이로 인해 모델이 음성 데이터를 학습하는 데에는 어려움이 존재한다.​

* 1. **Spectrogram**

스펙트로그램은 시간에 따른 신호의 주파수 스펙트럼을 시각적으로 나타내는 도구로, ​x축은 시간, y축은 주파수, 색상/밝기는 신호 강도를 나타낸다.

스펙트로그램은 시간과 주파수 두 축에 걸쳐 음성 신호의 변화를 보여줌으로써 신호의 자음과 같은 짧은 시간 변화와 모음과 같은 긴 시간 변화 모두를 포착할 수 있게 해준다. 또한, 음성 신호의 주파수 성분과 그 강도를 시각적으로 표현함으로써 이미지 처리에 사용되는 CNN을 활용해 음성 신호 처리를 가능하게 한다.

* 1. **Wavegram**

Wavegram은 논문에서 제안하는 신경망을 통한 새로운 음악 파일 변환 방법​이다. Wavegram을 생성하기 위해 여러 개의 1D-Convolution Filter를 통해 음악 파일에서 시간축에 따른 진폭에 대한 특징을 추출한다. 이렇게 나온 출력을 T(Number of Frame) × C(Channel) 형태의 텐서로 변환한다. 이후 각 채널마다 F개 그룹으로 분할하여 이 출력을 T × F × C/F 크기의 텐서로 재구성한다. F는 주파수 영역대를 의미하며 이렇게 나온 텐서를 Wavegram이라고 한다. 최종 출력은 N개의 Wavegram이 결합되어 N × T × F × C/F의 형태를 갖게 된다. Wavegram은 주파수 구간을 의미하는 F라는 차원을 새롭게 도입해 주파수 관련 정보로부터 특징을 추출할 수 있게 된다.

* 1. **Sampling Rate**

음성 파일들을 스펙트로그램으로 변환할 때, 높은 Sampling Rate을 가지는 것이 중요하다. Sampling Rate는 초당 샘플의 수를 의미한다. ​이 수치가 높으면 원본 음성 신호에서 더 많은 정보를 포착할 수 있기에 더 세밀한 사운드의 재현을 가능하게 하며, 결과적으로 더 높은 음질을 제공한다.​ 스펙트로그램 변환 시 높은 샘플링 레이트를 사용함으로써 더 정확하고 상세한 음향 정보를 얻고자 하였다.

1. **데이터셋**
   1. **GTZAN**

GTZAN 데이터셋은 총 1,000개의 오디오 트랙으로 이루어져 있으며, 각 트랙은 30초 길이로 구성되어 있다. 이러한 트랙은 10가지 다양한 음악 장르에 속하며, 각 장르는 100개의 트랙으로 세분화되어 있다. GTZAN 데이터셋에 포함된 장르들에는 블루스, 클래식, 컨트리, 디스코, 힙합, 재즈, 메탈, 팝, 레게, 록 등이 있다. 각 트랙은 WAV 형식으로 제공되며, 22,050Hz의 샘플링 레이트와 16비트의 모노 오디오 파일로 이루어져 있다.

* 1. **Free Music Archive(FMA)**

FMA는 MP3로 인코딩된 4가지의 다양한 크기의 오디오 데이터를 제공한다. 이들은 fma\_small, fma\_medium, fma\_large, fma\_full이다. 각 데이터별로 우리는 제공된 것 중에 가장 작지만 우리의 프로젝트를 위해 충분한 fma\_small.zip과 fma\_medium.zip 데이터를 선택했다. fma\_small.zip은 30초 분량의 8,000개 트랙으로 구성되어 있으며, 8개의 장르별로 1,000개의 음성 파일이 들어있는 7.2GiB의 균형 데이터 셋이다. fma\_medium.zip은 30초 분량의 25,000개 트랙으로 구성되어있으며 16개의 장르별로 불균형하게 음성 파일이 들어있는 22GiB의 데이터셋이다.

1. **데이터**
   1. **Data pipeline**

오디오 파일을 이미지로 변환하기 위해 torchaudio 이용해 오디오 파일을 로드하였다.​ torchaudio의 Resample() 함수를 활용하여, Sampling Rate을 동일하게 하여 모든 파일들의 Wavegram에서 시간 축에서 동일한 정보량을 갖도록 하였다.

또한 오디오의 길이를 동일하게 하기 위해서 22khz 30초 기준 길이 661200으로 Wavegram의 길이를 동일하게 패딩 작업을 진행했다. 모델이 1차원 1D 및 2D Convolution을 입력으로 받기 때문에 차원 통일을 위해 스테레오 오디오의 경우 torch의 mean() 함수를 통해 모노 오디오로 변환하여 차원을 낮춰 주었다.

* 1. **Data Augmentation**

과적합 방지와 데이터 다양성을 위해 스펙트로그램에서 시간과 주파수를 랜덤으로 drop하여 Data Augmentation을 추가적으로 진행했다. 다만 음향적인 변형의 경우, 높은 연산 비용으로 인하여 현실적으로 고려하기 어려웠다.

1. **모델**
   1. **1D Convolution의 한계**

1D Convolution의 경우 시간 축에 따라 진폭의 정보로부터 특징을 추출할 수는 있는 반면, 1D Convolution Filter의 차원 한계로 음악의 가장 중요한 요소라고 할 수 있는 주파수를 추가적으로 추출하는 것이 불가하다. 즉, 음악을 신호로만 파악해 분류하면 시간 축에 따른 진폭만을 고려하게 되어 음악의 주요 특징인 음의 높낮이를 의미하는 주파수를 고려할 수 없게 된다. 따라서 우리는 1D Convolution Filter가 아닌 2D Convolution Filter를 사용하는 모델을 우리의 프로젝트에 활용하기로 결정했다.

* 1. **PANNs (Pretrained Audio Neural Networks) ​:**

PANNs는 오디오 신호 처리와 관련된 다양한 작업에 적용되는 사전 훈련된 딥 러닝 모델들로 특히 오디오 패턴 인식 분야에서 주목받고 있다. 이들은 수천 시간에 달하는 대규모 오디오 데이터셋에서 사전 훈련되는데, 이들 중 가장 유명한 예는 AudioSet이다. AudioSet은 다양한 환경과 상황에서 촬영된 오디오 클립들로 구성되어 있으며, 527개의 서로 다른 사운드 클래스를 포함하고 있다. PANNs는 대규모 데이터셋에서 사전 훈련되기 때문에 다양한 오디오 관련 작업에서 높은 성능을 보인다. 사전 훈련된 PANNs는 다양한 오디오 인식 작업에 적용될 수 있으며 음악 분류, 사운드 이벤트 탐지, 음성 인식, 감정 분석 등이 포함된다. PANNs에 대해 특정 음악 장르나 오디오 유형에 대해 추가적으로 훈련을 진행해 모델의 정확도와 효율성을 더욱 개선할 수 있다.

* 1. **Wavegram-Logmel-CNN**

Wavegram-Logmel-CNN은 음성 차원에서 생성된 Wavegram과 이미지 차원의 스펙트로그램을 동시에 활용해 음악 파일을 분류하는 Multi-Modal Model이다. Wavegram-Logmel-CNN은 음성 데이터를 활용해 Wavegram과 스펙트로그램을 생성한 이후, Wavegram과 스펙트로그램을 concat해 Feature Map을 만든다. 이렇게 만들어진 Feature Map은 14개의 Convolution Layer로 구성된 CNN 모델을 학습시키는데 사용된다. 이렇게 학습된 모델은 음성 데이터 분류 작업을 수행하는 데에 사용된다. 아래 그림은 Wavegram-Logmel-CNN의 아키텍처이다.

텍스트, 도표, 스케치, 평면도이(가) 표시된 사진

자동 생성된 설명

1. **결과**
   1. **Wavegram-Logmel-CNN Result**

Wavegram-Logmel-CNN의 성능을 확인해보기 위해 GTZAN 데이터셋을 이용해 학습을 진행했다. A6000 GPU 환경에서 학습을 진행하였으며 Epoch은 100으로 설정하여 가장 성능이 좋은 Checkpoint를 선택하였다. Optimizer는 Adam을 사용하였으며, Scheduler는 ReduceOnPlateau를 사용하였다. 초기 Learning Rate은 5e-5로 설정하여 학습을 진행하였다.

우선, 사전 학습된 파라미터를 활용해보고자 사전 학습된 모델 파라미터를 기준으로 전체 모델에 대해 재학습을 진행해본 결과 Test Data에 대한 정확도는 78.5%였으며 ​모델 출력층만 재학습한 결과 Test Data에 대한 정확도는 21.5%였다. 음성 데이터의 경우 시간 축의 정보를 포함하고 있기 때문에, Sampling Rate 또는 음원의 길이에 따라 정보가 크게 변해 학습 데이터의 차이에 따라 모델의 성능이 크게 달라질 수 있다. 이에 따라 영상 데이터에 비해 전이 학습이 효과적이지 않다고 판단했다. Wavegram-Logmel-CNN에서는 GTZAN 데이터셋과 달리 22.05kHz보다 높은 음질의 데이터와 30s 이외의 데이터를 사용하기 때문에, 사전 학습된 파라미터를 사용하지 않기로 결정했다. 사전 학습된 모델 파라미터를 사용했을 때 모델의 학습데이터에 대한 정확도 곡선과 학습 데이터에 대한 손실함수 곡선은 아래 그림과 같다. 모델 출력층만 재학습한 경우 Train Data에 대한 정확도 곡선이 전혀 수렴하지 못하고 있는 것을 확인할 수 있다.

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

데이터 증강을 위해 학습 대상인 스펙트로그램들에 대해 Mixup을 수행한 뒤 학습을 진행한 경우 Test Data에 대해 78.5%의 정확도를 기록했다. 음성 데이터의 경우 Mixup이 영상 데이터에서의 Mixup과 같이 긍정적인 효과를 주지 못하는 것을 확인할 수 있었다. 영상 데이터에서 사용되는 다른 데이터 증강 기법 또한 Test Data에 대한 정확도 향상에 긍정적인 영향을 주지 못해 Mixup과 같은 다양한 데이터 증강 기법을 사용하지 않기로 했다. Mixup을 사용했을 때 모델의 학습데이터에 대한 정확도 곡선과 학습 데이터에 대한 손실함수 곡선은 아래 그림과 같다.

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명 텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

데이터 증강과 사전 학습된 파라미터를 사용하지 않은 채 Wavegram-Logmel-CNN을 학습시킨 결과 50번째 Epoch에서 최고 성능을 보였으며, 이때 Test Data에 대한 Loss는 0.8504였으며 Test Data에 대한 정확도는 85%였다. 모델의 학습데이터에 대한 정확도 곡선과 학습 데이터에 대한 손실함수 곡선은 아래 그림과 같다.

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

* 1. **Baseline Model(ResNet50)과의 비교​**

PANN 모델과 기존 CNN 모델의 성능 비교를 위해 ResNet50과의 비교를 진행하였다. 비교를 위해 Wavegram-Logmel-CNN과 ResNet50 모두 같은 데이터셋을 활용해 학습하였다. 총 두 가지의 데이터셋을 활용하였으며, GTZAN 데이터셋과 fma\_small 데이터셋을 이용해 학습을 진행하였다.

ResNet50의 학습을 위해 GTZAN의 .wav 파일과 fma\_small의 .mp3 파일을 스펙트로그램으로 변환하였으며, 스펙트로그램에 따라 각 장르별로 분류가 되도록 학습을 진행하였다. GTZAN의 경우 Batch Size는 16, fma\_small의 경우 Batch Size를 4로 설정하여 학습을 진행했으며 두 데이터셋 모두 50회의 Epoch을 반복하여 학습하였다.

GTZAN과 fma\_small을 이용해 학습이 완료된 Wavegram-Logmel-CNN과 ResNet50의 Test Data에 대한 정확도를 비교한 결과, GTZAN에 대해 Wavegram-Logmel-CNN은 85%의 정확도를 보였으며 ResNet50은 79.5%의 정확도를 보였다. fma\_small에 대해 Wavegram-Logmel-CNN은 60.9%의 정확도를 보였으며 ResNet50은 50.8%의 정확도를 보였다. 이를 통해 PANN이 기존의 CNN 모델에 비해 좋은 성능을 보인다는 것을 확인할 수 있었다.

1. **결론**

모델의 학습 결과 각 장르별 분류 정확도는 아래 왼쪽 그림과 같았으며, 아래 오른쪽에 혼동행렬을 통해 이를 시각화하였다.

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

전반적으로 장르별 분류 성능이 우수한 것을 확인할 수 있으며, 재즈 장르에 대한 정확도와 f1-score가 가장 높은 것을 확인할 수 있다. 반면, 메탈은 락으로, 힙합은 레게로 오분류되는 경우가 가장 많았는데, 이는 장르적 유사도를 고려했을 때 충분히 납득할 만한 결과라고 해석할 수 있다. 새롭게 학습된 모델을 활용해 다양한 노래 장르를 분류해볼 수 있었으며, 아래 그림처럼 외부 데이터를 사용해 실제로 장르 분류가 되는 것을 확인해볼 수 있었다. 해당 외부 데이터들은 전반적으로 장르 분류가 정확히 이루어졌지만, Louis Armstrong의 What a Wonderful World가 클래식으로 오분류되는 것을 확인할 수 있었다.

​ 텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Louis Armstrom의 What a Wonderful World의 장르 예측 확률을 확인한 결과는 아래 그림과 같다. 클래식일 확률을 54.26%, 락일 확률을 14.32%, 디스코일 확률을 7.38%로 예측한 것을 알 수 있다. 이는 mixup을 통한 예측의 confidence 증가를 통해 보완할 수 있을 것으로 예상된다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

추가로 fma\_medium에 대해 학습을 진행해본 결과 fma\_small에 대해 학습을 진행했을 때에 비해 분류 성능이 낮아진 것을 확인할 수 있었다. 이는 fma\_medium에 포함된 장르의 수가 상대적으로 많기 때문이기도 하지만, 장르별로 음성 데이터의 수가 상이한 불균형 데이터셋이기 때문일 것으로 예상된다. 아래 왼쪽 그림은 fma\_medium, 아래 오른쪽 그림은 fma\_small에 대한 장르별 오디오 데이터의 분포를 나타낸 그래프이다.

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

아래 왼쪽 그림은 fma\_medium으로 학습을 진행한 모델의 분류 성능을 시각화한 혼동 행렬이며, 아래 오른쪽 그림은 fma\_small로 학습을 진행한 모델의 분류 성능을 시각화한 혼동 행렬이다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 사각형, 도표이(가) 표시된 사진

자동 생성된 설명

1. **한계점 및 개선 방안**

장르 분류, 즉 데이터 라벨링에 주관성이 포함되어 있을 수 있다는 한계가 존재한다. 이는 모델 성능 저하에 기여할 수 있을 것으로 생각된다. 또한, fma\_medium을 활용해 학습을 시킨 모델에서 노래 장르별 데이터 불균형이 존재함을 확인하였다. 데이터 수가 가장 적은 장르에 맞춰서 데이터 수를 맞춰 주기 위해서는 극히 일부의 데이터만 사용하게 되는 문제가 있었다. 반면, 데이터 수가 많은 장르에 데이터 수를 맞추려면 몇몇 장르를 제외해야 하는 문제가 있었다. 이러한 데이터 불균형을 GAN 등의 생성모델을 통해 해소할 수 있을 것으로 예상된다. 데이터 불균형 해소를 위한 데이터 증강에 사용된 GAN 모델은 노래를 작곡하는 데에 도움이 될 수도 있어 흥미로운 추가과제가 되리라 기대된다.

1. **참고자료**
   1. Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2880-2894
   2. Defferrard, Michaël, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. "FMA: A Dataset for Music Analysis." In 18th International Society for Music Information Retrieval Conference (ISMIR), 2017. [arXiv preprint] (<https://arxiv.org/abs/1612.01840>).
   3. “GTZAN Dataset - Music Genre Classification”, Kaggle, last modified 2019, accessed Dec 1, 2023, <https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification>.
   4. “The Nyquist–Shannon Theorem: Understanding Sampled Systems”, All About Circuits, last modified May 06, 2020, accessed Dec 1, 2023, <https://www.allaboutcircuits.com/technical-articles/nyquist-shannon-theorem-understanding-sampled-systems/>.
   5. “What is a Spectrogram?”, Vibration Research, last modified July 19, 2023, accessed Dec 1, 2023, <https://vibrationresearch.com/blog/what-is-a-spectrogram/>.