안녕하십니까. 경북대학교 전자공학부 학사과정 4학년 박병준입니다. 오늘 발표드릴 내용은 이제까지 저희가 진행했던 연구에 대해서 발표드릴 건데, 보시는 바와 같이 Convolutional Neural Network 구조를 이용한 음악에서의 악기 소리 분류 방법과 그 결과에 대해 연구했습니다. 우선, 이 연구를 진행한 배경에 대해서 설명을 드리자면, AI 딥러닝 분야에서 sound processing에 관한 다양한 연구가 수행되어왔습니다. 대부분의 연구는 natural language processing과 text-to-speech이지만 음악에서 악기를 분류하는 문제에 대한 연구는 부족한 시점입니다. 따라서 저희는 Convolutional Neural Network로 악기 소리를 분류하는 모델을 설계하고 평가하며 그 접근법에 대해 규명을 해보고자 합니다.

다음으로는 본 연구의 핵심 모델에 통과시킬 데이터셋의 접근법을 간단히 설명을 드리도록 하겠습니다. 저희는 총 7가지의 고전 악기를 선택하여 그 악기들로 이루어진 연주 데이터를 웹에서 수집하였습니다. 그래프에서 보실 수 있듯이, 각 라벨링된 악기별 데이터를 train/validation/test를 하기 위해 각각 6:2:2 비율로 나누었습니다. 이 비율은 각 악기의 solo 연주 뿐 아니라 다른 악기와의 duet 연주도 반영한 것입니다. 또한 이 연주 sound waveform이 CNN 모델을 효과적으로 통과할 수 있도록 constant-Q transform을 사용하여 spectrogram으로 전처리하였습니다.

이제 악기 소리를 학습할 model architecture에 대해 간략히 설명드리도록 하겠습니다. 그림에서 보실 수 있듯이, model에 입력된 spectrogram은 down-sampling layer와 up-sampling layer를 거칩니다. 이렇게 인코딩과 디코딩 과정을 거쳐 특징이 두드러지게 된 데이터는 convolution과 FC레이어를 거쳐 해당 악기의 존재 유무를 binary classification method로 판별합니다. 하나의 연주 spectrogram에 대해 모든 악기를 분류하는 multi-label classification problem을 해결하기 위해 각각의 악기에 대해 동일한 구조의 model을 학습합니다. 결과적으로 각각의 악기의 output을 종합하여 최종 output을 도출해내게 됩니다.

이제 본 연구의 결과에 대해 말씀 드리자면, 그림에서 볼 수 있듯이, test dataset에 대해 15번의 test를 수행하고 각 악기별 accuracy를 box plot으로 나타낸 차트입니다. 또 아래의 표는 각 악기별 best accuracy를 정리한 것입니다. 전체적으로 높은 분류 정확도를 나타내고 있는데, 특징적인 부분은 비록 같은 모델을 통과했지만 악기마다 accuracy가 상이한 모습을 띄고 있습니다. 또한 solo와 duet 연주에 대해서는 상당히 높은 accuracy가 나타난 반면에 3가지 이상의 합주에 대해서는 상대적으로 낮은 accuracy를 보여주었습니다.

저희는 이 연구를 통해 비슷한 주파수 음역대를 가진 악기일수록 accuracy가 낮은 것을 확인했으며 이는 악기의 음역대가 model의 성능에 영향을 주고 있음을 시사했습니다. 데이터셋을 보완한다면 더욱 향상된 성능이 기대될것으로 확인됩니다. 이러한 특성으로 보아, 이러한 소리 분류 메커니즘이 악기 소리 뿐만 아니라 다른 분야에서도 적용될 수 있는 가능성이 있습니다. 제 발표는 여기까지 하겠습니다. 감사합니다.

예상질문

왜 Spectrogram/CNN을 사용하였는가? : Spectrogram은 sound waveform의 frequency 특성을 추출하여 결합한 구조인데, spectrogram으로 sound classification problem이 image classification problem으로 적용할 수 있습니다. 그리고 CNN 구조는 image 데이터를 처리하는데 사용되고 이미지 인식과 분류 문제에 잘 적용되기 때문입니다.

왜 U-net 구조를 사용하였는가? : 스펙트로그램 이미지 전체에 대해 단일 카테고리를 예측하는 이미지 분류와는 달리, 사운드는 각 주파수 픽셀단위로 분류를 수행해야했습니다. 따라서 이러한 이미지 segmentation 문제를 해결하기 적합한 U-net 구조를 사용하였습니다.

구조에 대해 자세히 설명하자면? : down sampling을 통해 spectrogram의 유의미한 특징을 추출하고 다시 up sampling과 skip-connection 과정을 거치면서 원래 spectrogram 크기를 복원하여 위치 파악을 할 수 있게 됩니다.

+인코더와 디코더 사이의 skip connection을 재설계하여 feature map간의 semantic gap을 줄였습니다. 또한 deep supervision은 모든 branch 결과값의 평균을 계산하여 더욱 정확한 segmentation을 수행합니다.