2023 경영과학연구실 하계 워크샵

- A Hybrid CNN and RNN Variant Model for Music Classification
- Music Boundary Detection using Convolutional Neural Networks : A Comparative Analysis of Combined Input Features
- A Novel multi-task learning method for symbolic music emotion recognition

경영과학연구실 이태헌
2023.08.18

A Hybrid CNN and RNN Variant Model for Music Classification

Ashraf, Mohsin, et al. Applied Sciences 13.3 (2023)

Introduction

- 기존 hand crafted feature 이용하여 음악이 가진 유니크한 패턴 찾는 연구 시도 많이 진행 되었지만, 음악의 고유한 특성 찾는 것은 어려움

- Rhythm, pitch, tonality, intensity, timbre 등 audio signal acoustics를 이용한 음악 분류 기술도 많이 사용되고 있지만 이것 만으로는 음악의 고유한 특성 찾기 어려움

- 딥러닝 방법은 상대적으로 적은 편향을 가지며 Automatically extract features가 가능함. 또한 기존 방법들에 비해 뛰어난 성능을 입증

음악 분류 모델들의 낮은 정확도

- 하이브리드 모델 결과 비교

Method	Accuracy
George Tzanetakis [9]	61.00%
G. Sun et al. [20]	66.40%
A Heakl et al. [18]	70.60%
Nilesh M. et al. [13]	77.78%
Praseneet Fulzeele et al. [16]	89.00%
N. Farajzadeh [19]	86.00%
Pradeep Kumar D et al. [15]	86.00%
Jan Jakubik [21]	87.70%
Proposed Work	89.30%

Related works

Machine learning method based : svm, knn

- *Patil, N.M, Nemade, M.U. "Music Genre Classification Using MFCC, K-NN and SVM Classifier" Int. J. Comput. Eng. Res, 2017*

Feature extraction with statistical description

- *Elbir, A.; Cam, H.B.; Iyican, M.E.; Ozturk, B.; Aydin, N. "Music Genre Classification and Recommendation by Using Machine Learning Techniques" In Proceedings of the 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 2018*

RNN based model

- *Jakubik, J. "Evaluation of Gated Recurrent Neural Networks in Music Classification Tasks", In Proceedings of the 38th International Conference on Information Systems Architecture and Technology, 2017*

- *Ashraf, Geng G, Wang X, Ahmad, F. Abid, F. A "Globally Regularized Joint Neural Architecture for Music Classification", IEEE Access, 2022*

Problem statement

- End to End 방법으로 음악 장르 분류 문제 접근
- CNN – Variant Model에 적합한 input, parameter 찾고자 함

Key idea

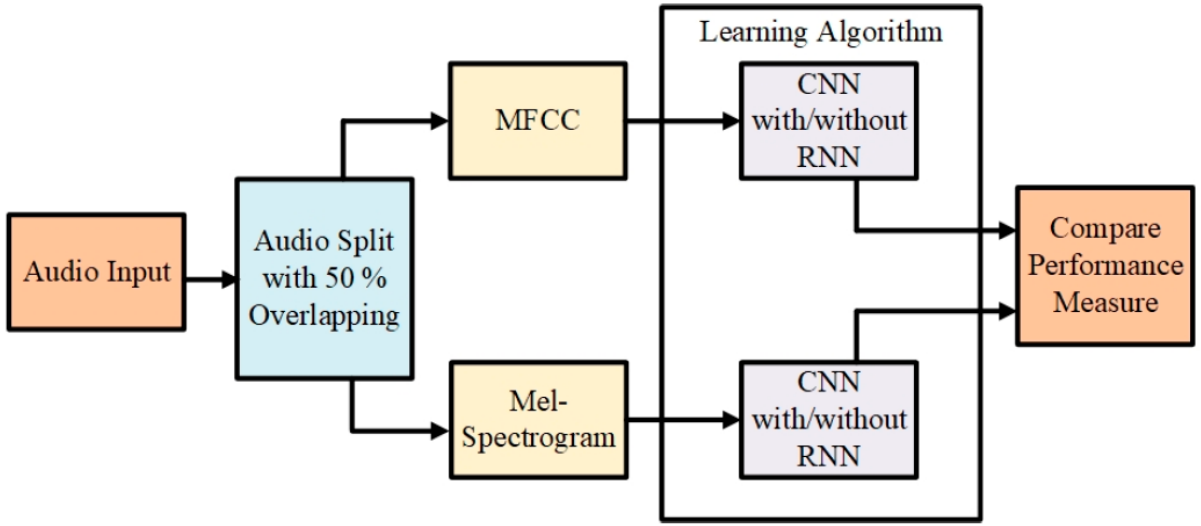
- **CNN + variant RNN(LSTM, GRU, Bi-LSTM, Bi-GRU) 하이브리드 모델 사용**

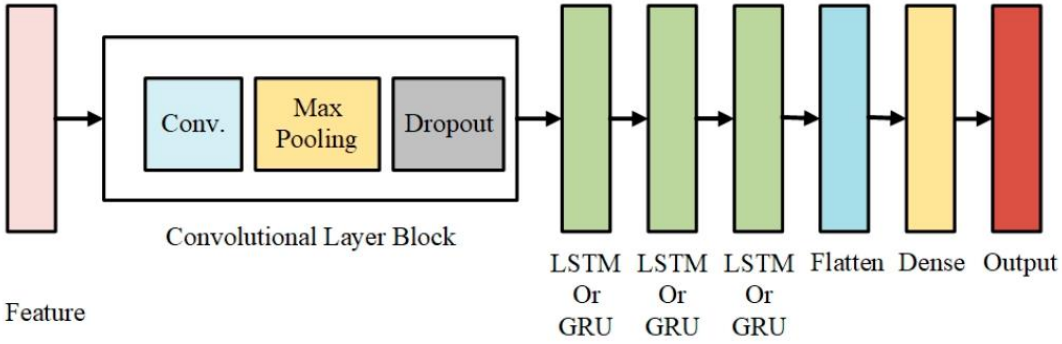
  - 음악 분류 연구에서 CNN과 RNN은 효과적인 딥러닝 모델로 알려져 있음
  - CNN은 공간 정보 Feature 추출
  - 음악은 연속적이므로 시간 정보 Feature 처리 위해 RNN을 합쳐서 사용

- **MFCC, Mel-spectrogram 인풋 사용하여 결과 비교**

Proposed hybrid architecture with CNN and variants of RNN

- CNN + (LSTM, GRU, BI-LSTM, BI-GRU)

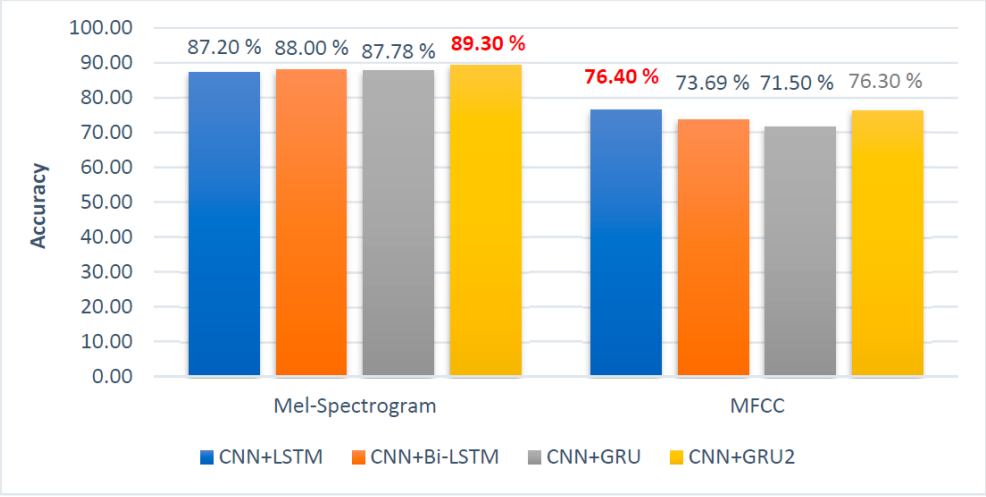




Results of extracted features with proposed hybrid architecture

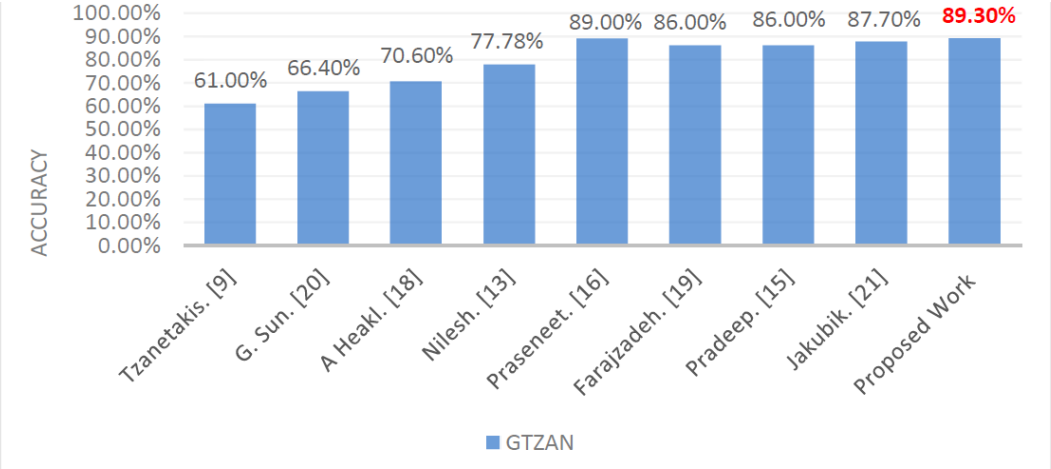
- Mel-Spectrogram + CNN + Bi-GRU 사용하였을 때 Accuracy 89.30%로 가장 좋은 성능을 보임

Features	Model	Accuracy
Mel-Spectrogram	CNN+LSTM	87.20%
Mel-Spectrogram	CNN+Bi-LSTM	88.00%
Mel-Spectrogram	CNN+GRU	87.78%
Mel-Spectrogram	CNN+Bi-GRU	89.30%
MFCC	CNN+LSTM	76.40%
MFCC	CNN+Bi-LSTM	73.69%
MFCC	CNN+GRU	71.50%
MFCC	CNN+Bi-GRU	76.30%



GTZAN 데이터셋을 사용한 다른 분류 모델들과 결과 비교

Method	Accuracy
George Tzanetakis [9]	61.00%
G. Sun et al. [20]	66.40%
A Heakl et al. [18]	70.60%
Nilesh M. et al. [13]	77.78%
Praseneet Fulzeele et al. [16]	89.00%
N. Farajzadeh [19]	86.00%
Pradeep Kumar D et al. [15]	86.00%
Jan Jakubik [21]	87.70%
Proposed Work	89.30%



- GTZAN 데이터셋 : Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock 10개 장르로 구성

Conclusion

- CNN + variants of RNN 하이브리드 모델을 사용했을 때 , CNN 단일 모델에 비해 전반적으로 좋은 성능을 보임

- Input으로 Mel-Spectrogram 사용하였을 때, MFCC를 인풋으로 사용했을 때에 비해 전반적인 성능이 좋았음

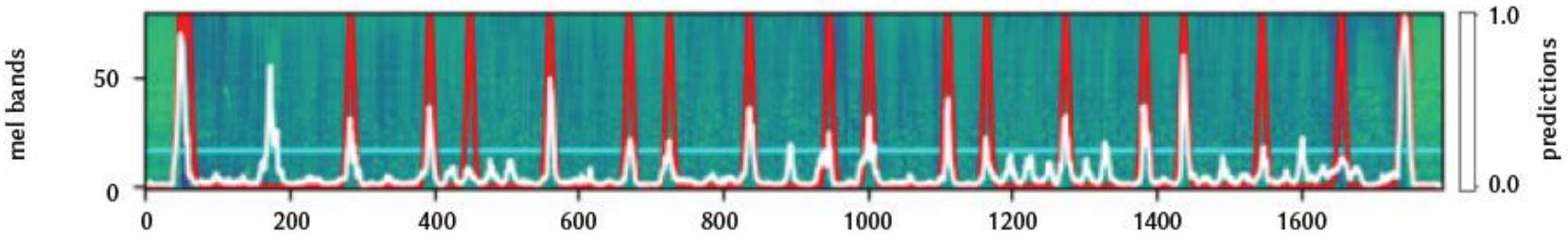
- Mel-Spectrogram input + CNN + Bi-GRU 모델 조합이 89.30%의 Accuracy를 보이며, 이는 모든 조합 중 가장 뛰어난 결과를 보임

Music Boundary Detection using Convolutional Neural Networks : A Comparative Analysis of Combined Input Features

Hernandez-Olivan, Carlos, Jose R. Beltran, and David Diaz-Guerra., arXiv, 2020

Boundary detection

- 음악 구조를 파악하기 위한 프로세스로, 일반적으로 음악 트랙 내에서 다른 섹션 또는 부분(예: 서론, 구절, 후렴 등)으로 전환되는 지점을 자동으로 식별하는 것을 목표로 함
- 일반적인 음악 장르인 pop 장르에 대입하면 해당 노래의 후렴구, 구절 또는 시작 부분을 탐지하고 추출해내는 작업이 됨
- 음악 Boundary detection은 MIDI, Music XML, Audio Representation 방법으로 전처리 된 Input 사용



(a) CNN predictions on MLS.

- Red line : Ground truth
- White line : Predict Value

Introduction

- MIR은 automatic music analysis, pitch tracking, chord estimation, score alignment, music structure detection 등 여러 분야로 이루어짐

- Automatic structural analysis, music structure analysis 는 최근까지도 해결하기에 복잡한 문제이며, 도전적인 문제

- 현재까지도 음악가, 전문가가 수행하는 구조 라벨링 및 분석을 능가하는 충분한 정확도를 달성하지 못함

- 여러 연구에서 다양한 전처리 방법을 이용하므로 일반화된 입력 전처리 방법이 없음

-> 원하는 Audio feature를 추출하기 위한 Audio input에 대한 연구 필요

Related works

Unsupervised Methods

- J. Paulus, M. Müller, A. Klapuri, “State of the art report: Audio-based music structure analysis,” in Proceedings of the 11th ISMIR, 2010

- Unsupervised 접근 방식은 Novelty-based, Homogeneity-based, Repetition-based로 나뉨

Novelty-based : 대비되는 부분 간 음악의 변화 감지

Homogeneity-based : 음악적 특성에 대해 일관성 있는 구간 식별

- Hidden Markov Models

Repetition-based : 반복되는 패턴 감지

- SSM, SSLM

Related works

Supervised Neural Networks

- *K. Ullrich, J. Schlüter, T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR, 2014*

- *T. Grill, J. Schlüter, "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices," in 23rd European Signal Processing Conference, EUSIPCO 2015*

- *T. Grill, J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*

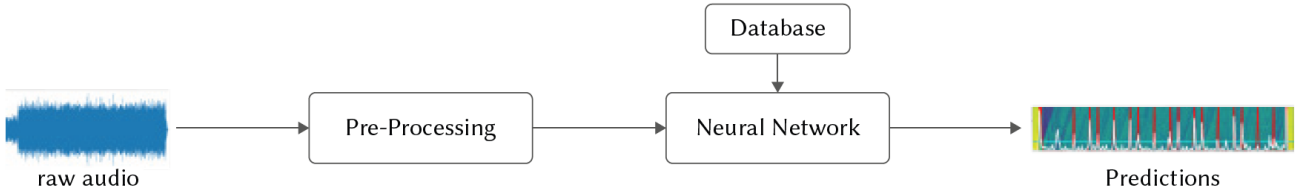


Fig. 1. General scheme of supervised neural networks.

Unsupervised methods are not as high as the results obtained with supervised neural networks

- 비지도 학습 방법의 주요 목표는 Boundary detection (세분화)이 아닌 전체 구조의 식별 (라벨링)
- 비지도 학습 알고리즘이 라벨링 (클러스터링) 부분에서 효율적이지만, boundary detection에서는 그렇지 않음. Boundary detection task는 Supervised Neural Network 모델이 더 좋은 성능을 보임

Unsupervised Methods								
Year[5]	Autors [Ref.]	Algorithm	Input	Method	F-measure (F_1) for Testing Databases			
					MIREX09	RCW-A	RCW-B	SALAMI
2009	Paulus & Klapuri [24]	PK	MFCCs, chromas	*Fitness function*	0.27	-	-	-
2010	Mauch et al. [25]	MND1	MFCCs, Discrete Cepstrum	*HMM*	0.325	0.359	-	-
2011	Sargent et al. [26]	SB-VRS1	Chords estimation	*Viterbi*	0.231	0.324	-	-
2012	Kaiser et al. [27]	KSP2	SSM	*Novelty measure*	0.280	0.366	0.289	0.286
2013	McFee & Ellis [20]	MP2	MLS	*Fisher's Linear Discriminant*	0.281	0.355	0.278	0.317
2014	Nieto & Bello [28]	NB1	MFCCs + chromas	*Checkerboard-like kernel*	0.289	0.352	0.269	0.299
2015	Cannam et al. [29]	CC1	Timbre-type histograms	*HMM*	0.197	0.224	0.203	0.213
2016	Nieto [30]	ON2	Constant-Q Transform Spectrogram	*Linear Discriminant Analysis*	0.259	0.381	0.255	0.299
2017	Cannam et al. [29]	CC1	Timbre-type histograms	*HMM*	0.201	0.228	0.192	0.212
Supervised Neural Networks								
2014	Schlüter et al. [31]	SUG1	MLS	*CNN*	0.434	0.546	0.438	0.529
2015	Grill & Schlüter [32]	GS1	MLS + SSLMs	*CNN*	0.523	0.697	0.506	0.541

Problem statement

- 음악 Boundary detection 문제를 다룸

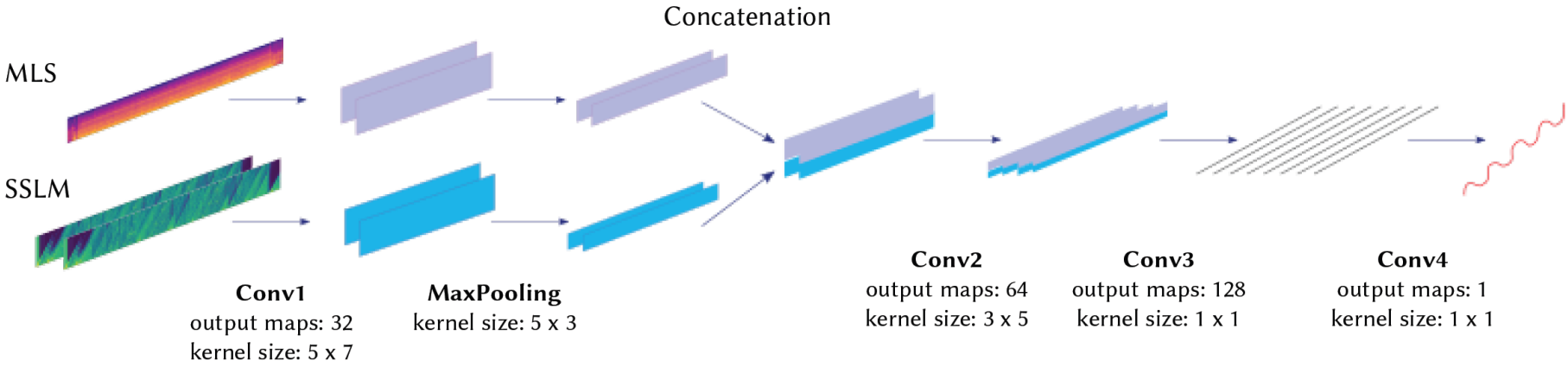
- 다양한 방법 비교 통해 Boundary detection 분야에서 좋은 성능을 보이는 input feature 조합 찾고자 함

Key idea

- CNN + variants of RNN(LSTM, GRU, Bi-LSTM, Bi-GRU) 하이브리드 모델 사용

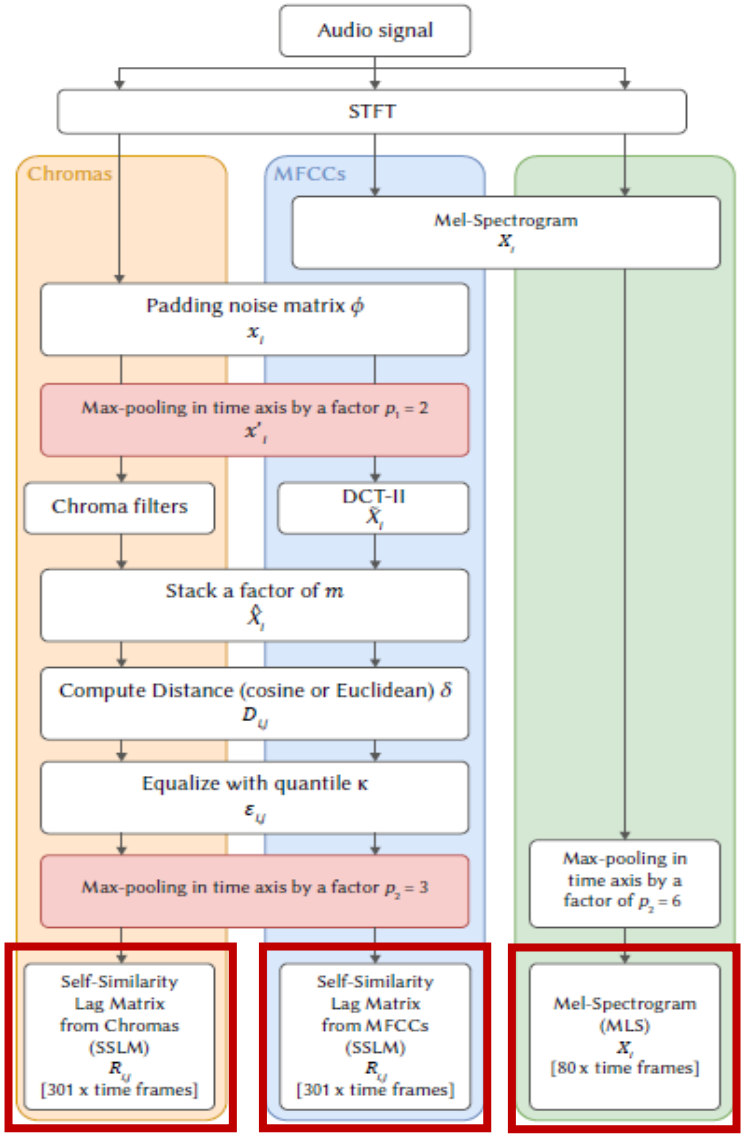
- Mel-spectrogram + Chromas, MFCC 활용한 4가지 SSM 방법을 Input으로 사용

Overall model structure



- Input data : Segmented Audio Labelling And Music Analysis (SALAMI) 데이터 사용 (서론, 구절, 후렴 등 레이블)
- Output : Boundary 확률 분포

Audio Representation method



Results of Boundary Estimation

Boundary Estimation With Tolerance ± 0:5S and Optimum Threshold in Terms of F-score, Precision and Recall

Input	Train Database	Epochs	Thresh.	P	R	F_1(std)
$MLS + SSLM^{MFCCs}_{euclidean}$	SALAMI	140	0.24	0.441	0.415	0.402 (0.163)
$MLS + SSLM^{MFCCs}_{cosine}$	SALAMI	140	0.24	0.428	0.407	0.396 (0.158)
$MLS + (SSLM^{MFCCs}_{euclidean} + SSLM^{chromas}_{euclidean})$	SALAMI	100	0.24	0.465	0.400	0.407 (0.160)
$MLS + (SSLM^{MFCCs}_{cosine} + SSLM^{chromas}_{cosine})$	SALAMI	100	0.24	0.444	0.416	0.404 (0.166)
$MLS + (SSLM^{MFCCs}_{euclidean} + SSLM^{MFCCs}_{cosine})$	SALAMI	100	0.24	0.445	0.421	0.409 (0.173)
$MLS + (SSLM^{chromas}_{euclidean} + SSLM^{chromas}_{cosine})$	SALAMI	100	0.24	0.457	0.396	0.400 (0.157)
$MLS + (SSLM^{chromas}_{euclidean} + SSLM^{chromas}_{cosine} + SSLM^{MFCCs}_{euclidean} + SSLM^{MFCCs}_{cosine})$	SALAMI	100	0.26	0.526	0.374	**0.411 (0.169)**
End-to-end previous works						
$MLS + SSLM^{MFCCs}_{cosine}$ [4] (2015)	Private	-		0.646	0.484	0.523
$MLS + SSLM^{MFCCs}_{cosine}$ [35] (2017)	SALAMI	-		0.279	0.300	0.273 (0.132)
$MLS + (SSLM^{MFCCs}_{cosine} + SSLM^{chromas}_{cosine})$ [35] (2017)	SALAMI	-		0.470	0.225	0.291 (0.120)

- Tolerance : 예측된 경계와 실제 경계 사이의 허용되는 최대 시간 차이
- Threshold : 경계를 감지할 때의 결정 임계값

Conclusion

- Music Boundary detection 위해 CNN 입력 결정하는 비교 연구 수행
- MLS + 4 SSLM 방법을 input으로 사용하였을 때 가장 높은 성능을 보임

Limitation

- 기존 방법 보다 높은 성능을 보이지만 제안한 모델 또한 높은 성능을 보이지는 못함
- Music boundary detection은 성능이 높지 않은 연구 분야로 꾸준한 후속 연구 필요

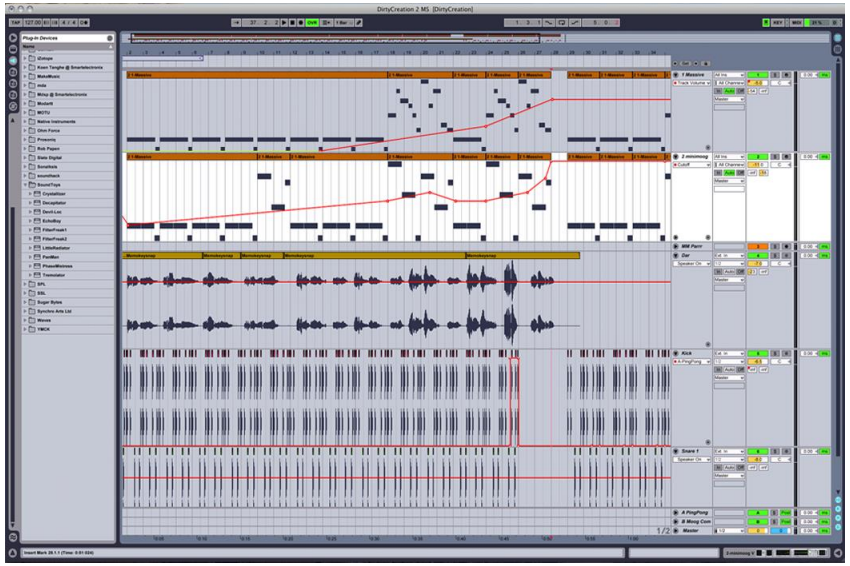
A Novel multi-task learning method for symbolic music emotion recognition

Qiu, Jibao, C. L. Chen, and Tong Zhang, arXiv, 2022

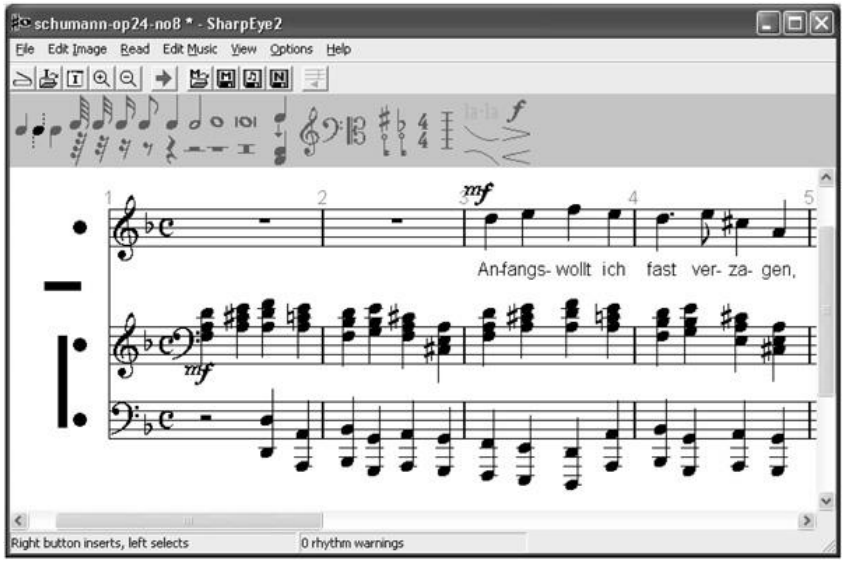
MIDI, Music XML

- Music Symbolic data

MIDI



Music XML



Symbolic Music Emotion Recognition (SMER)

- Symbolic Music Emotion Recognition(SMER) is to predict music emotion from symbolic data, such as MIDI and Music XML
- 최근 Symbolic Music 생성 기술의 발전으로 Symbolic music을 이해하기 위한 연구가 진행 되고 있음
- Emotion Recognition 에서는 symbolic data 이용한 연구가 많이 진행되지 않았음

- 기존 SMER 연구는 NLP 모델 적용하여 연구가 진행 됨
- 단순히 NLP 모델을 이용한 방법을 적용하는 것은 Symbolic data 부족, SMER에 중요한 음악 구조에 대한 부족한 이해로 이어질 수 있음

- Ming liang Zeng, et al., Music bert: Symbolic music understanding with large-scale pre-training, arXiv, 2021

Problem statement

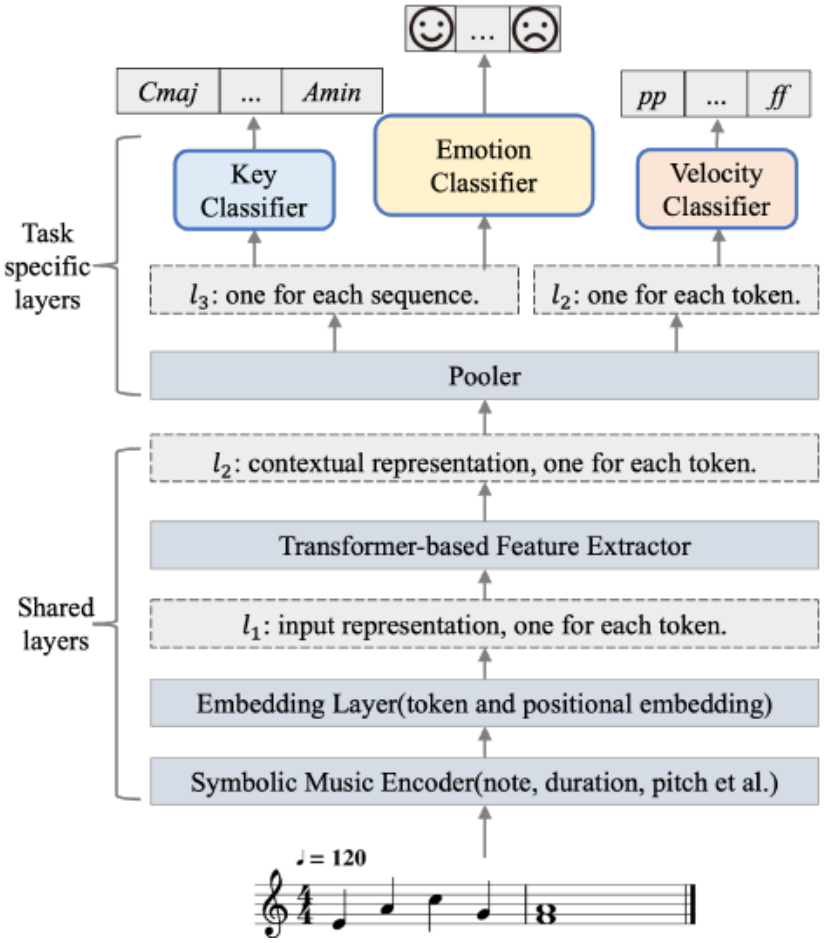
Symbolic Music Emotion Recognition 문제를 다룸

Key idea

- SMER 위해 Key, velocity classifier를 auxiliary multi-task로 사용
- 음악을 sequential token으로 사용하여 Transformer based model에 적용

Multi-Task Symbolic Music Neural Network (MT-SMNN)

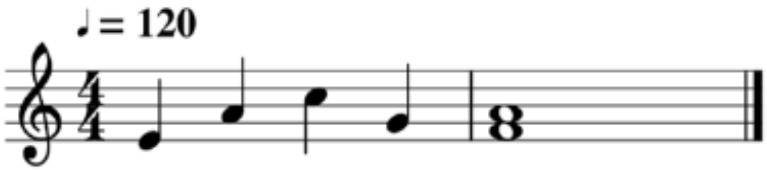
- Multi-task learning은 여러 Task를 동시에 학습함으로써 각 task간 정보를 공유할 수 있음
- Auxiliary task인 Key, Velocity feature 정보가 Emotion classifier 성능 향상



- Sequence level : 전체 음악이나 노트의 긴 시퀀스 분석에 초점. -> l3
  (ex : 음악적 요소의 상호 관계나 전체 구조 등)
- Key classifier : 12개 장조 key, 12개 단조 key 분류 (Sequence level task)
- Note level : 개별인 음표나 음악적 이벤트 사운드 분석에 초점 -> l2
  (ex : 지속 시간, 강도, 피치)
- Velocity classifier : pp (0-31), p (32-47), mp (48-63), mf (64-79), f (80-95), ff (96-127) 6개의 속도 구간 분류 (Note level task)

Symbolic Music Encoder

- CP representation : 음악을 기호적 데이터로 표현할 때 사용되는 인코딩 방식
- 음악 조각은 ' Super token' 시퀀스로 인코딩 됨.
- 하나의 ' Super token' 은 여러 개의 'Sub token' 으로 구성 됨



(a) A piece of symbolic music

Super token

Duration(8)	*Duration*(8)	*Duration*(8)	*Duration*(8)	*Duration*(32)	*Duration*(32)
Pitch(64)	*Pitch*(69)	*Pitch*(72)	*Pitch*(67)	*Pitch*(65)	*Pitch*(69)
Sub-beat(1)	*Sub-beat*(5)	*Sub-beat*(9)	*Sub-beat*(13)	*Sub-beat*(1)	*Sub-beat*(1)
Bar (new)	*Bar* (cont)	*Bar* (cont)	*Bar* (cont)	*Bar* (new)	*Bar* (cont)

(b) CP representation

v_76 d_8 n_64 . v_48 n_69 . v_32 n_72 . v_116 n_67 . v_80 d_30 n_65 n_69

(c) Ferreira's representation

- Ferreira representation : 음악을 sequential token으로 인코딩
- 음표의 특성을 나타내는 token들로 인코딩
- V : Velocity , D : Duration, N : Pitch

Main result of SMER in the EMOPIA and VGMIDI dataset

Model	EMOPIA		VGMIDI	
	Accuracy(%)	macro-F1	Accuracy(%)	macro-F1
SVM([Lin *et al.*, 2013])	47.72	0.4763	45.12	0.3779
SVM([Panda *et al.*, 2013])	39.77	0.3624	36.93	0.2146
MIDIGPT[Ferreira *et al.*, 2020]	58.75±3.13	0.572±0.029	53.88±3.48	0.505±0.041
MIDIBERT-Piano[Chou *et al.*, 2021]	63.41±3.52	0.628±0.033	47.30±2.81	0.432±0.021
MT-MIDIGPT(proposed)	62.50±4.45	0.611±0.047	**55.85±1.97**	**0.509±0.017**
MT-MIDIBERT(proposed)	**67.58±2.39**	**0.664±0.027**	49.81±2.52	0.453±0.019

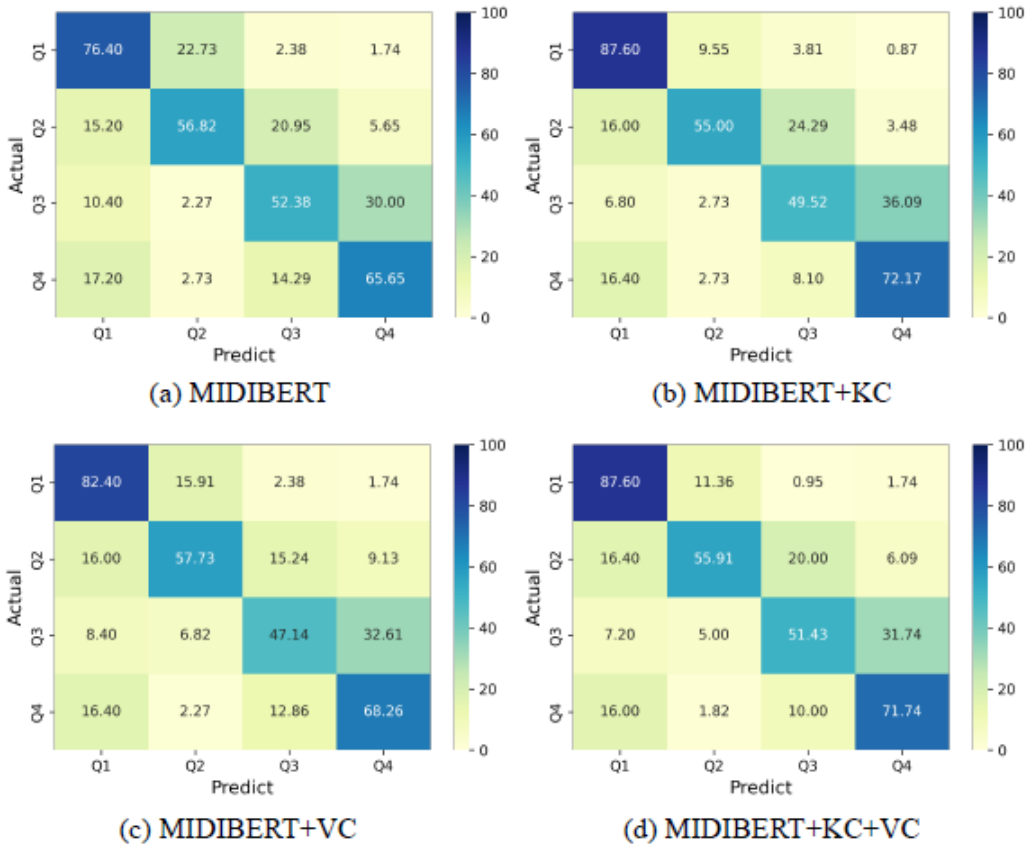
2 auxiliary task result

Performance verification based on the presence or absence of key and velocity classification

Key Classification	Velocity Classificaiton	Accuracy
✗	✗	63.41±3.52
✓	✗	67.03±2.54
✗	✓	64.73±5.47
✓	✓	**67.58±2.39**

Confusion matrix result

- Q1 : Happy, Q2 : Angry, Q3 : Sad, Q4 : Calm



Conclusion

- Symbolic Music Emotion Recognition에 중점을 둔 multi-task framework MT-SMNN 제안
- MT-SMNN framework는 key classification, velocity classification 작업을 함께 수행하여 EMOPIA, VGMIDI 데이터셋에서 성능을 입증
- 2개의 auxiliary task에서도 성능 입증

Q & A