딥러닝을 사용하여 Russell Model 기반 감정 벡터에서 감성적 음악 생성에 대하 연구

윤현세, 서상우, 이정행, 김성진, 동호석, 이상훈

연세대학교

hsyoon97@yonsei.ac.kr, ssw9557@gmail.com, jin.k@yonsei.ac.kr,
hstong09@yonsei.ac.kr, leedoright@yonsei.ac.kr, slee@yonsei.ac.kr

A study on Embedding Russell Model's Emotion Vector to Music Generation Systems

Yoon Hyunse, Seo Sangwoo, Lee Jeonghaeng, Kim Seongjin, Tong Hoseok, Lee Sanghoon

Yonsei Univ.

요 약

본 논문은 얼굴의 감정을 기반하여 감정에 알맞은 음악 생성하는 Framework를 개발하는 것을 목표로 하였다. 음악을 생성하기 앞서 FER2013 데이터셋을 사용하여 얼굴 감정 분류에 중점을 두었고 다양한 감정 불균형을 해결하기 위해 이미지 뒤집기와 회전과 같은 데이터 증강 기술을 적용하였다. 분류된 감정을 Russell 모델의 Valence-Arousal 평면상의 감정 벡터로 표현하였다. 결과적으로 "러셀 모델"을 기반으로 하는 행복, 놀라움, 분노, 혐오, 두려움, 슬픔 및 중립 감정에 대한 '감정 벡터'를 Valence-Arousal 평면에서 얻었다 주어진 감정 벡터를 기반으로 MIDI 데이터로 학습된 Music Transformer로 음악을 생성 하였다.

Ⅰ. 서 론

얼굴 표정은 사람들의 감정을 잘 표현하는 매우 강력한 수단이며 사람들은 감정에 따라 듣는 음악이 다르다. 사람이 어느 특정한 노래를 듣고 싶을 때는 그 노래가 현재 그 사람의 감정과 일치할 경우가 많기 때문이다. 본 논문은 사람들의 표정을 기반으로 사람의 감정을 잘 표현하는 음악을 생성하는 기술을 제안하고자 한다.

FER2013 데이터셋을 활용하였고, 이 데이터셋은 일곱 가지 감정으로 분류된 다양한 얼굴 표정을 포함하고 있다. 그러나 이러한 데이터셋은 종종 다른 감정 범주 간의 데이터 불균형으로 인해 고충을 겪는다. 따라서 다양한 감정의 표현을 균형 있게 만들고 데이터셋의 다양성을 향상시키기 위해 데이터 증강 기술을 활용하여 데이터셋의 퀄리티를 높힌다.

본 연구는 얼굴 표정을 기반으로 감정을 분류하기 위해 합성곱 신경망인 VGG16 [1] 모델을 활용하여 분류된 감정을 기반으로 Music Transformer [2] 모델을 사용하여 MIDI 형식의 해당 음악을 생성한다. 본 논문에서는 사람들의 얼굴 감정 분류 및 이러한 감정을 기반으로 한 음악 생성을 위한 방법론을 명확히 하며, 또한 데이터 불균형을 처리하는 전략과 모델의 정확도 및 Valence-Arousal 평면의 독특한 '감정 벡터'를 포함한 얻은 결과를 탐구한다.

Ⅱ. 본론

얼굴 표정 이미지에서 음악을 생성하기 위해 먼저 이미지에서 감정 벡터를 추출이 필요하다. VGG16 네트워크를 구현하고 FER2013 데이터셋으로 훈련하였다. FER2013 데이터셋은 32,298개의 흑백 이미지로 구성되어 있으며 각 이미지는 해당 감정과 레이블이 지정되어 있다. 이 이미지를 레

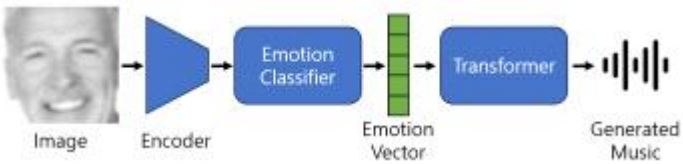

그림 2. 얼굴-음악 생성 프레임워크

이블링하는 데 사용되는 7가지 레이블은 Anger, Disgust, Fear, Happiness, Sadness, Surprise 및 Neutral이다. FER2013 데이터셋은 충분한 데이터를 가지고 있는 것처럼 보이지만 레이블 간 데이터 수에는 불균형이 있으며, 예를 들어 "Disgust"로 레이블이 지정된 이미지는 503개 밖에 없는 반면 "Happiness"으로 레이블이 지정된 데이터는 8,109개가 있습니다. 이 문제에 대처하기 위해 수가 적은 레이블에 대한 데이터를 증가시켰다. 증강된 FER2013 데이터셋으로 VGG16 네트워크를 훈련한 후, VGG16의 출력은 Russell Model을 기반으로 embedding 된 감정 벡터를 출력하였다. Russell Model은 감정을 충동과 가치에 기반하여 모델링합니다. FER2013 데이터셋의 각 레이블에 대한 감정 벡터는 표 1에 정리하였다.

Embedding 된 감정벡터를 기반으로 음악을 생성하기 앞서 음악 생성 모델을 MetaMIDI dataset을 [3] 기반으로 학습시켜야 한다. 하지만 MetaMIDI dataset에 감정이 레이블이 포함되어 있지 않아 각 음악의 감정을 분류해야 한다. 감정에 기반하여 MetaMIDI 데이터셋에 포함되어 있는 MIDI 파

Happiness	[0.60, 0.85]
Surprise	[0.05, 0.50]
Anger	[-0.6, 0.85]
Disgust	[-0.3, 0.45]
Fear	[-0.62, 0.1]
Sad	[-0.7, -0.2]
Neutral	[0.0, 0.0]

표 1. Russell Model 기반 FER2013 데이터셋 레이블의 감정 벡터 값

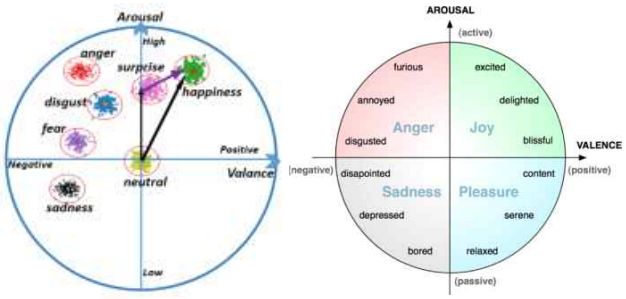

그림 2. Russell Model 기반 Valence-Arousal Plane 상의 감정 벡터

일을 분류하기 위해 Spotify API에서 제공한 Energy 및 Valence 값을 사용하여 해당 파일을 분류하였다. 비록 Energy와 Arousal 동일한 개념은 아니지만, 이들은 유사한 개념으로 사용할 수 있다. Energy와 Arousal 간에는 관련성이 있으며 Arousal은 "The strength and energy of music"으로 언급되고 있습니다 [4]. Arousal은 "an emotional dimension of musically Energy level"으로 설명되었다 [5]. 또한 SVM classifier를 사용하여 감정 분류를 수행했으며 실험 결과에서는 high-energy 음악이 주로 Russell Model의 Arousal-Valence 평면 상 Q1과 Q2에 분포하며, low-energy 음악은 Q3과 Q4에 분포하는 것으로 나타났습니다 [14].. 하지만 Spotify API가 분류한 감정의 레이블은 총 5개이며 그 감정들은 Happiness, Anger, Surprise, Sadness 와 Others 이다. 하지만 Embedding에 사용된 FER2013 데이터셋의 감정 레이블 수와 MIDI 데이터의 레이블 데이터가 일치하지 않아. Disgust, Fear, Neutral를 Others에 해당하는 감정으로 embedding을 하였다.

Energy 값이 0.5 이상이거나 같고 가치 값이 0.5 이상이면 감정은 사분면 Q1에서 Happiness 및 Surpise 범위에 속한다. 이 경우 Energy 값이 가치 값보다 작으면 Happiness으로 분류되며, Energy 값이 Valence 값보다 크거나 같으면 Surprise으로 분류된다. Energy 값이 0.5 이상이고 Valence 값이 0.5 미만이면 감정은 사분면 Q1에서 Anger로 분류된다. Energy 값이 0.5 미만이고 Valence 값도 0.5 미만이면 감정은 Sadness으로 분류된다. 마지막으로, Energy 값이 0.5 미만이고 Valence 값이 0.5 이상이면 감정은 Others에 분류된다.

주어진 motion vector와 MetaMIDI dataset 기반으로 본 프레임워크에 알맞게 수정된 Music Transformer를 학습 시킨다. 네트워크를 사용하여 사람의 표정에 맞는 음악을 생성하여 각 음악의 감정과 주어진 사람의 표정의 유사성의 정확도를 얻기 위해 Subjective Test를 행하였다. 각 감정에 대한 정화도의 결과는 표 2에 정리하였다.

Emotion	Accuracy
Happy	0.851
Anger	0.755
Surprise	0.799
Sad	0.873
Other	0.683

표 2. 생성된 음악의 정확도.

표 2가 나타내는 결과에서 Others의 결과가 제일 낮았다. 그 이유는 Others 분류에 포함되어 있는 감정인 Disgust가 Anger와 유사성이 있어 Subjective Test에 참가한 많은 실험 참가자들은 Others에 분류되는 음악을 Anger에 분류하였다. 그리고 명확한 레이블인 Surprise인 경우 다른 감정과 섞여 표현되는 경우가 많아 Surprise 표정에서 생성된 음악을 Happiness에 분류하는 경우도 있었다.

Ⅲ. 결론

본 논문에서는 VGG16을 사용하여 빠르게 주어진 사람의 얼굴 표정을 Russell Model의 Arousal-Valence 평면 상의 감정 벡터를 구했다. 그리고 MetaMIDI dataset에 감정 레이블을 얻기 위해 Spotify API를 사용하여 5개의 감정으로 음악을 분류하였다. 분류된 음악과 감정 벡터를 통해 Music Transformer를 학습시켜 주어진 얼굴 표정에 알맞은 노래를 생성하였다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C3011697) and the Yonsei Signature Research Cluster Program of 2023 (2023-22-0008).

참 고 문 헌

[1] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[2] Huang, Cheng-Zhi Anna, et al. "Music transformer." arXiv preprint arXiv:1809.04281 (2018).

[3] Ens, Jeffrey, and Philippe Pasquier. "Building the MetaMIDI Dataset: Linking Symbolic and Audio Musical Data." ISMIR. 2021.

[4] Han, Xiao, Fuyang Chen, and Junrong Ban. "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure." Electronics 12.4 (2023): 978.

[5] Rachman, Fika Hastarita, Riyanarto Samo, and Chastine Fatichah. "Song emotion detection based on arousal-valence from audio and lyrics using rule based method." 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). IEEE, 2019.

[6] Panda, Renato, et al. "How does the spotify api compare to the music emotion recognition state-of-the-art?." 18th Sound and Music Computing Conference (SMC 2021). Axea sas/SMC Network, 2021.