딥러닝 감정 인식 기반 배경음악 매칭 설계

정문식, 문남미
호서대학교 벤처대학원 융합공학과
moonsik.jung@gmail.com, nammee.moon@gmail.com

Design for Mood-Matched Music Based on Deep Learning Emotion Recognition

Moonsik Chung, Nammee Moon
Dept. of Convergence Engineering, Hoseo Graduate School of Venture

요 약

멀티모달 감정인식을 통해 사람의 감정을 정확하게 분류하고, 사람의 감정에 어울리는 음악을 매칭하는 시스템을 설계한다. 멀티모달 감정 인식 방법으로는 IEMOCAP(Interactive Emotional Dyadic Motion Capture) 데이터셋을 활용해 감정을 분류하고, 분류된 감정의 분위기에 맞는 음악을 매칭시키는 시스템을 구축하고자 한다. 유니모달 대비 멀티모달 감정인식의 정확도를 개선한 시스템을 통해 텍스트, 음성, 표정을 포함하고 있는 동영상의 감성 분위기에 적합한 음악 매칭 시스템을 연구한다.

1. 서론

최근 딥러닝이 다양한 분야에서 활용되면서 사람의 감정을 분석하는 연구가 활발하게 진행되고 있다. 딥러닝의 감정 인식은 심리학자들이 연구한 감성 모델에 기반하여 인간 감정을 분류하는 방식으로 실험되고 있다. 대다수 멀티모달 인간 감정 인식 연구는 텍스트 기반, 음성 기반, 얼굴 표정 기반의 분류 연구가 진행되어 왔으며[1], 각각의 유니모달 방식 감정 분류를 비롯하여 혼합 형태의 멀티모달 방식 감정 분류까지 다양한 시도를 통해 인간 감정 인식의 정확도를 높이는 연구가 진행되어 왔다.

감정 인식은 음악 분야에서도 사용되고 있는데, 사용자 감정에 따른 음악 추천 서비스가 대표적이라고 할 수 있다. 유튜브 뮤직과 같은 음악 서비스는 현재 분위기나 감정의 키워드를 입력함으로써 연관된 음악 콘텐츠를 추천해주는 기능을 갖고 있다. 이는 수많은 음악 콘텐츠의 발생에 따라 거대해진 음악 데이터베이스에서 사용자의 음악 선택 고민을 돕는 유용한 기능으로 모바일 환경의 보편화 이후 대부분 음악 제공 업체들이 채택하고 있는 기능으로 자리 잡고 있다.

음성 기반, 얼굴 표정 기반의 감정 분류 기술 또한 인간 감정을 인식하는 방법으로 활용되고 있는데, IEMOCAP 의 멀티모달 방법의 감정인식은 유니모달 방법과 비교했을 때, 감정인식 성능을 향상시킨 연구 결과[2]가 있으며, 이는 영화나 드라마와 같은 동영상 콘텐츠 제작 시, 배우들의 대화내용이나 말투, 표정을 인식하여 그 감정과 분위기에 맞는 음악 추천도 가능할 것이다.

본 논문에서는 텍스트, 음성, 표정 기반을 모두 포함하고 있는 동영상에 대해 IEMOCAP 을 활용하여 등장인물의 감정을 인식하여 해당 분위기에 맞는 음악추천 시스템을 설계해보고자 하며, 유니모달 방식과 멀티모달 방식의 감정인식 성능을 비교하여 기존의 유니모달 방식 대비 감정인식 성능의 효용성을 검증하고자 한다.

2. 관련연구 및 기존연구 문제점

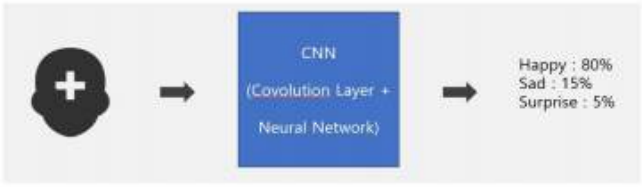
감정에 따른 음악 인식 방법 연구는 다양하게 진행이되어 왔다. 전반적인 연구 절차는 사람의 감정을 어떤 방법을 사용하여 인식하여 분류하고, 분류된 감정에 사전에 준비된 음악 리스트와 매칭하는 방식으로 연구되었다[3].

사람의 감정 인식 방법에는 주로 다음 3 가지 방법을 사용하는데, 텍스트 기반, 음성 기반 분석과 표정을 인식하여 분석하는 방법이 있다. 텍스트 기반 감정 인식은 전형적인 분석 방법 연구는 오랫동안 진행되어 왔으며, 오피니언마이닝(Opinion Mining), LIWC(Lingustic Inquiry and Word Count), Social Metrics, Textat 등과 같은 프로그램 및 서비스들이 개발되어 사람이 입력한 문장 및 단어를 감정과 매칭시켜 분석

및 분류할 수 있다. 딥러닝을 활용한 텍스트 감정 인식으로는 GloVe, BERT 등의 방법이 있으며, GloVe 는 단어의 동시 등장 빈도수를 기반으로 각 단어 간의 연관성을 분석하고, BERT 는 Transformer 의 양방향 인코더를 기반으로 한 언어 표현 모델로서, 단어 간의 관계 뿐만 아니라, 문장 내에서 단어 위치를 통해 문맥의 의미를 파악하여 문장의 이해도를 높일 수 있다[4].

음성을 통한 감정 인식은 사람의 음성신호를 분석하여 감정을 인식하는 것으로 통화 음성 데이터 수집을 통한 연구가 활발히 수행되고 있다. 일반적으로 통화 시 여러 가지 감정이 혼재될 수 있는데, 사용자의 통화 녹음 음원을 일정 시간 단위로 분할하여 MFCC (Mel Frequent Coefficient Cepstral) Filter Bank 알고리즘을 이용한 특징 벡터 값을 추출한다. 다음 단계로 LSTM layer 를 거쳐 학습이 진행되고, Softmax 활성화 함수를 통해 각 감정을 산출하는 방법이 있다[5].

사람의 감정은 얼굴 표정으로 추측이 가능한데, 특정 감정을 나타내는 얼굴 표정의 데이터셋을 수집하여 (그림 1)에서와 같이 CNN 을 활용하여 얼굴 표정 분류를 학습시킬 수 있다[6].



(그림 1) CNN 얼굴 표정 분류 모델

최근에는 위의 3 가지 감정 인식 방법(텍스트, 음성, 얼굴인식)을 멀티모달로 융합하여 감정 인식의 정확성을 높이는 연구가 진행되고 있는데, IEMOCAP 데이터셋을 활용한 감정 인식 연구가 있다. IEMOCAP 이란 Interactive Emotional Dyadic Motion Capture 의 약자로 USC(University of Southern California) 의 SAIL(Speech Analysis and Interpretation Laboratory) 연구소에서 수집된 멀티모달 및 멀티스피커 데이터베이스이다. IEMOCAP 은 비디오, 음성, 얼굴모션캡쳐, 텍스트를 포함하여 약 12 시간 분량의 시청각데이터를 포함하고 있다.[7]

기존 연구에서는 얼굴 표정을 읽고 감정을 분류 후에 음악 매칭하는 연구에서는 행복, 슬픔, 놀라움 3 가지 감정으로 분류하여 각 감정에 맞는 음악을 매칭하는 것으로 각 감정 항목에 맞는 음악 리스트를 사전에 선곡을 하여 분류된 감정에 따라 해당 음악을 재생하는 방식이다.

이는 두가지 취약점을 갖고 있는데, 첫째로 감정에 따라 분류한 임의의 음악 라이브러리와 해당 감정을 매칭시키는 방법에 있어서, 분류된 감정과 매칭되는 음악 추천의 정확성은 보장될 수 있으나, 감정 분류는 3 가지로만 분류되어 그 범위가 매우 제한되어 있다.

둘째로 얼굴 표정 인식에 사용된 방법은 사전에 주어진 감정 키워드에 따라 직접적이고 과장되게 표현되어 실제 감정 표현과 매칭이 안되는 부분이다.

인위적으로 연출된 얼굴 표정이 사람의 감정 상태를 표출하는 것은 일반적이기 보다는 과장된 측면이 있어서, 연기를 구분할 수 있는 장점과 동시에 일상적인 감정의 표현 구분에는 무리가 있다. 이를 보강하는 수단으로 멀티모달 감정인식을 활용한 연구를 진행하고자 한다. 얼굴 표정 및 음성인식을 통한 텍스트 분석과 소리 분석을 통해 감정 분류를 정교화하고, 분류된 감정에 적합한 음악 매칭 시스템을 연구한다.

3. 감정 분류 및 음악 감성 매칭

감정을 분류하는 방법으로 Ekman(1984)의 얼굴 감정 인식의 범주형 이론 분류 방법을 사용한다. Ekman 의 범주형 이론은 6 가지 보편적인 감정으로 행복, 놀람, 혐오, 두려움, 분노, 슬픔으로 분류하고 있다[8]. 이러한 감정은 음악가들의 감정 해석과 음악적 표현 방법을 통해 표현된 다양한 방식의 음악 작품들을 볼 수 있다.

Zentner(2008)는 음악을 들었을 때, 사람이 느끼는 심리적 감정 요소를 도출하여 음악 감성 모델을 만들었다. Zentner 의 음악 감성 모델은 Sublimity(장엄), Vitality(활기), Unease(불안) 의 3 개의 범주로 구성되어 있고, 3 개의 범주는 10 개의 세부 하위 범주와 마지막으로 총 40 개의 세부 감성 요소를 갖고 있다[9]. Zentner 의 모델은 2017 년 ITU(International Telecommunication Union)의 Rep. ITU-R BS.2399 에 음악 감성 속성 체계의 사례로 소개되었으며[10], 통계적 타당성 검증을 거쳐 <표 1>과 같이 음악 감성의 상위 개념부터 하위세부 개념까지 구조적으로 정리되어 있다.

<표 1> Zentner 의 음악 감성 모델 요소

1차 범주	2차 범주	세부감성요소
Sublimity	Wonder	Happy, Amazed, Dazzled, Allured, Moved
	Transcendence	Inspired, Feeling of Trascendence, Feeling of Spiritually, Thrills
	Tenderness	In Love, Affectionate, Sensual, Tender, Softened-Up
	Nostalgia	Sentimental, Dreamy, Nostalgic, Melancholic
	Peacefulness	Calm, Relaxed, Serene, Soothed, Meditative
Vitality	Power	Energetic, Triumphant, Fiery, Strong, Heroic
	Joyful Activation	Stimulated, Joyful, Animated, Dancing, Amused
Unease	Tension	Agitated, Nervous, Impatient, Irritated
	Sadness	Sad, Sorrowful

본 연구에서는 Ekman 의 얼굴 표정 인식으로 분류할 수 있는 6 가지 보편적 감정과 Zentner 의 음악 감성 모델을 매칭시켜 사람의 감정과 그 분위기에 맞는 음악 매칭 연구를 진행하고자 하며, 감정 인식을 정교화 하기 위해 텍스트, 음성 인식을 융합한 멀티모달 감정인식과의 매칭 연구를 진행하고자 한다.

4. 네트워크 구조

감정 인식하는 방법으로 음성과 텍스트, 얼굴 표정의 3 가지 방법을 이용한 네트워크 구조를 사용하며, 각각의 구조는 다음과 같다.

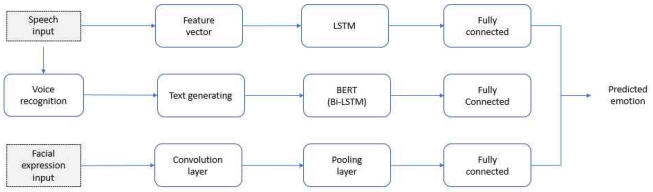
음성을 이용한 네트워크는 음성신호 기반 모델을

거쳐 특징을 추출한 MFCC 데이터로 벡터를 생성하고 생성된 벡터는 LSTM layer 를 거쳐 Softmax 활성화 함수를 통해 감정에 대한 확률값을 산출한다.

텍스트를 이용한 네트워크는 음성인식으로 생성된 텍스트를 BERT 를 이용하여 토큰이 포함된 문장 정보를 인코딩하고, Bi-LSTM layer 와 Attention layer 를 거쳐 Max pooling layer 를 사용하여 Attention 값을 줄이고, Softmax 활성화 함수를 사용하여 결과를 출력한다.

얼굴 표정은 Convolution layer 와 Pooling layer 를 거쳐 Softmax 활성화 함수를 통해 감정에 대한 확률값을 산출한다.

각 감정에 대한 오디오, 텍스트, 얼굴표정의 확률값의 평균을 구해 (그림 2)와 같이 가장 높은 값을 갖는 감정 범주로 분류한다.



(그림 2) 멀티모달 감정인식 모델 구조

4. 실험 계획

본 연구에서는 3 가지 감정인식(음성, 텍스트, 얼굴표정) 방법을 사용한 멀티모달을 사용하여 사람의 감정을 분류한다. 3 가지 감정을 포함하는 데이터셋으로는 IEMOCAP 을 활용하여 실험을 진행하며,

Ekman 의 보편적 6 개 감정과 Zentner 의 음악 감성 모델 2 차 범주의 10 개 감성 키워드를 매칭시키는 작업을 진행한다. 이는 Zentner 의 10 개의 음악 감성에 매칭되는 음악 샘플들을 먼저 선별한 후, 선별된 음악 샘플들을 다시 보편적 6 개의 감정에 대입시키는 방법을 통해 유사 감정으로 범주화한다.

각 음악 감성에 해당하는 테스트 음원을 분류된 감정에 매칭하는 방식으로 테스트를 진행하며, 이 과정에서 각각의 유니모달로 추출된 감정과 멀티모달로 추출된 감정 간의 매칭 결과를 비교하여 멀티모달 감정인식을 통한 음악 추천의 정확도의 개선을 기대할 수 있다.

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2021R1A2C2011966).

참고문헌

[1] 강도희, 김대하, 송병철, 멀티모달 감정인식 모델의 입력 특징에 따른 성능 분석, 대한전자공학회 학술대회, pp.1045-1047, 2021.

[2] 김석민, 조원익, 김형주, 김남수, 멀티모달 접근을 통한 딥러닝 기반 감정인식 알고리즘, 한국통신학회 학술대회논문집, pp.1227-1228, 2019.

[3] 배경아, 인관호, 김웅모, 감정분석을 통한 음악 추천 기법, 한국지능정보시스템학회 학술대회논문집, pp.229-232, 2012.

[4] 박호연, 김경재, BERT 기반 감성분석을 이용한 추천시스템, 지능정보연구, Vol.27, No.2, pp.1-15, 2021.

[5] 김주희, 이석필, 음성 특징과 텍스트 임베딩을 이용한 멀티모달 감정인식, 전기학회논문지, Vol.70, No.1, pp.108-113, 2021.

[6] 윤경섭, 이상원, 얼굴표정을 통한 감정 분류 및 음악재생 프로그램, 한국컴퓨터정보학회 학술대회논문집, Vol.27, No.1, pp.243-246, 2019.

[7] Carlos B. et al., “IEMOCAP :interactive emotional dyadic motion capture database”, Language Resources and Evaluation, 335, p.1, 2007.

[8] Ekman, P., “Approaches to emotion”, NJ, Erlbaum, 1984.

[9] Zentner, M., Grandjean, D., Scherer, K. R., Emotions evoked by the sound of music: Characterization, classification, and measurement, Emotion, Vol.8, No.4, pp.494-521, 2008.

[10] Report ITU-R BS.2399-0, “Methods for selecting and describing attributes and terms, in the preparation of subjective tests”, Geneva, ITU, p.8, 2017.