KICSP
한국 융합신호처리학회
융합신호처리학회논문지 Vol. 24, No. 4 : 221~228 December 2023
DOI : 10.23087/jkicsp.2023.24.4.007

음성-영상 특징 추출 멀티모달 모델을 이용한 감정 인식 모델 개발

김종구[1], 권장우[2*]
[1]인하대학교 전기컴퓨터공학과, [2]인하대학교 컴퓨터공학과

Development of Emotion Recognition Model Using Audio-video Feature Extraction Multimodal Model

Jong-Gu Kim[1], Jang-Woo Kwon[2*]
[1]Department of Electrical and Computing Science, Inha University
[2]Department of Computer Science, Inha University

요 약 감정으로 인해 생기는 신체적 정신적인 변화는 운전이나 학습 행동 등 다양한 행동에 영향을 미칠 수 있다. 따라서 이러한 감정을 인식하는 것은 운전 중 위험한 감정 인식 및 제어 등 다양한 산업에서 이용될 수 있기 때문에 매우 중요한 과업이다. 본 논문에는 서로 도메인이 다른 음성과 영상 데이터를 모두 이용하여 감정을 인식하는 멀티모달 모델을 구현하여 감정 인식 연구를 진행했다. 본 연구에서는 RAVDESS 데이터를 이용하여 영상 데이터에 음성을 추출한 뒤 2D-CNN을 이용한 모델을 통해 음성 데이터 특징을 추출하였으며 영상 데이터는 Slowfast feature extractor를 통해 영상 데이터 특징을 추출하였다. 감정 인식을 위한 제안된 멀티모달 모델에서 음성 데이터와 영상 데이터의 특징 벡터를 통합하여 감정 인식을 시도하였다. 또한 멀티모달 모델을 구현할 때 많이 쓰인 방법론인 각 모델의 결과 스코어를 합치는 방법, 투표하는 방법을 이용하여 멀티모달 모델을 구현하고 본 논문에서 제안하는 방법과 비교하여 각 모델의 성능을 확인하였다.

• 주제어 : 음성 인식, 비디오 인식, 특징 추출, 멀티모달 모델, 감정 인식

Abstract Physical and mental changes caused by emotions can affect various behaviors, such as driving or learning behavior. Therefore, recognizing these emotions is a very important task because it can be used in various industries, such as recognizing and controlling dangerous emotions while driving. In this paper, we attempted to solve the emotion recognition task by implementing a multimodal model that recognizes emotions using both audio and video data from different domains. After extracting voice from video data using RAVDESS data, features of voice data are extracted through a model using 2D-CNN. In addition, the video data features are extracted using a slowfast feature extractor. And the information contained in the audio and video data, which have different domains, are combined into one feature that contains all the information. Afterwards, emotion recognition is performed using the combined features. Lastly, we evaluate the conventional methods that how to combine results from models and how to vote two model's results and a method of unifying the domain through feature extraction, then combining the features and performing classification using a classifier.

• Key Words : Audio recognition, Video recognition, Feature extraction, Multimodal model, Emotion recognition

Received 10 October 2023 Revised 25 December 2023, Accepted 28 December 2023

*** Corresponding Author** Jang Woo, Kwon, Computer Engineering, Inha University, Hightech building, 100, inha-ro, Inchon, Korea.
E-mail: jwkwon@inha.ac.kr

Ⅰ. 서론

감정의 사전적 정의는 어떤 현상이나 일에 대하여 일어나는 마음이나 느끼는 기분을 의미한다. 감정은 사람에게 신체적 정서적인 변화를 일으켜 사람이 행하는 행동의 결과에도 큰 영향을 미친다. 최근 연구에 따르면 부정적인 감정이 운전자의 운전 양상에 큰 영향을 미칠 뿐만 아니라 실제 외부 자극에 반응하는 시간을 늦춤으로써 유의미한 변화를 보인다는 것을 알 수 있다.

이러한 사람의 행동에 대한 감정의 영향력으로 인해 감정을 자동으로 인식하고 어떤 감정을 느끼고 있는지 판단하는 과업은 상당히 중요히 다뤄졌다.

특히 최근 딥러닝을 이용하여 객체의 행동을 분류하는 과업의 관심이 많아짐에 따라[4, 16] 서로 도메인이 다른 음성과 영상 데이터를 동시에 이용하기 위해서 각기 다른 모델을 학습하여 붙이는 멀티모달 모델(Multimodal model)을 구현하여 감정을 인식하는 과업을 해결하고자 하는 연구가 많았다. 하지만 멀티모달 모델을 구현하는 방법으로 각기 결과를 합쳐서 분류하는 방법과 투표 방법(Voting)을 이용하는 경우가 많았다[6-7]. 이는 도메인이 다른 데이터를 처리하고 분석하기 위해 사용되는 가장 많이 사용되는 방법이었다. 본 논문에서는 도메인이 다른 두 데이터를 특징 추출법을 이용하여 도메인을 축소하고 하나의 특징으로 합쳐 분류를 진행하는 방법론을 이용하여 멀티모달 모델을 구현하였다. 또한 각 세 방법을 이용한 멀티모달 모델과 음성 데이터를 처리하는 모델 그리고 영상 데이터를 처리하는 모델의 결과를 비교하였다[2-3].

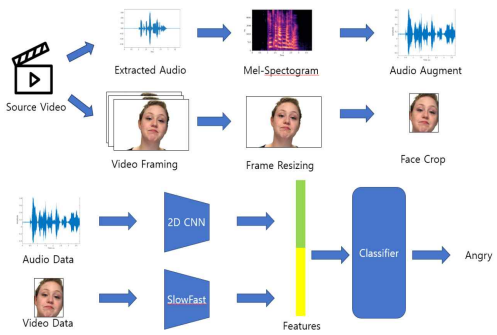

Fig. 1. Mutimodal Model Diagram

Ⅱ. 관련연구

2.1 음성 감정 인식

음성을 이용한 감정 인식은 다양한 방법을 이용하였다. 음성 데이터는 기본적으로 시간에 따라 데이터가 변화하는 시계열 데이터로, 이를 딥러닝을 이용하여 학습 및 추론을 통해 분석하기 위해서는 적절한 전처리가 필요하다. 음성 데이터를 전처리하기 위해서 프레임화, 푸리에 변환, 켑스트럼 분석, MFCC 등이 있다[1, 10].

음성 데이터 전처리 이후 특징을 추출한 뒤 다양한 방법으로 음성 데이터를 분석 및 분류를 진행한다. 2D CNN(Convolution Neural Network)을 이용하여 추출한 특징을 분석하는 방법뿐만 아니라 최근에는 3D CNN을 이용하여 음성 데이터를 직접 입력으로 사용하여 추론을 진행하는 모델 그리고 RNN, LSTM과 같은 회귀 모델(Recursive Model)을 이용하여 음성 데이터가 가지고 있는 시간적 변화 및 흐름에 대한 특징을 학습하는 방법 또한 많은 연구가 진행되고 있다[4].

2.2 비디오 감정 인식

비디오를 이용한 감정 인식 또한 음성을 이용한 감정 인식 못지않게 많은 관심을 받는 과업이다. 비디오를 이용한 감정 인식을 위해서는 다양한 비디오 전처리 기법을 이용한다. 우선 비디오를 이루고 있는 프레임을 프레임 단위로 자르는 작업인 프레임화(framing)를 진행한 뒤 각 프레임에 대한 분류를 진행하는 방법이 있다. 또한 하나의 영상에 하나의 표정 혹은 레이블을 지정하여 자른 뒤 영상을 이용하는 방법이 있다. 이는 트랜스포머나 3D CNN 등 영상을 분석하는 성능이 뛰어난 모델을 이용하지만, 학습과 추론의 속도가 느리고 강한 컴퓨터 파워 환경을 요구하는 단점이 있다[12].

다음으로는 영상의 특징을 추출하여 분석하는 방법이 있다. 영상 특징 추출은 slowfast나 X3D 같은 벤치마크를 이용하고 이를 통해 영상이 가지고 있는 시간적 공간적 특징을 추출한다. 이후 추출한 특징을 이용하여 분석 및 분류를 진행하는 방식으로, 특징을 추출하는 추가 단계가 존재하지만, 다른 방식보다 빠르고 성능이 좋다는 장점이 있다.

2.3 멀티모달 모델

멀티모달이란 도메인이 다른 여러 데이터를 이용하여 문제를 해결하는 방법론을 의미한다. 감정 인식에서 멀티모달이란 주로 음성과 영상 그리고 텍스트 데이터를 이용하여 감정 인식하는 모델을 생성하는 것을 의미한다.

연구가 활발한 분야는 음성 데이터와 영상 데이터를 이용하는 멀티모달 모델이다. 이는 아주 성질이 다른 두 데이터를 이용하는 것이기 때문에 대부분 각 데이터의 도메인을 이용하여 다른 모델을 학습한 뒤 각 모델이 결과를 내고 그 결과를 합쳐서 최종 결과를 내는 방법을 이용하고 있다. 이는 도메인이 다른 데이터를 해결하고자 하는 가장 기본적인 방법이다.

음성과 텍스트 데이터를 이용한 멀티모달 모델 또한 상당한 관심을 받고 있다. 두 데이터는 도메인은 다르지만 회귀 모델 학습에 맞는 맥락 성을 가지고 있다는 점을 이용하여 회귀 모델을 이용하여 두 데이터를 동시에 처리하는 멀티 모달 모델이 큰 관심을 받고 있다[8, 9].

Ⅲ. 제안하는 방법

3.1 특징 추출

음성 데이터는 2차원의 정보를 가지고 있는 시계열 데이터이며 영상 데이터는 3차원의 정보를 가지고 있는 이미지 데이터의 집합 데이터이다. 따라서 하나의 모델을 이용하여 동시에 도메인이 다른 영상 데이터와 음성 데이터를 처리할 수 없다.

본 논문에서는 이를 극복하기 위해 데이터 도메인에 맞는 특징 추출 모델을 이용하여 하나의 차원을 가지고 있는 특징으로 통일시켜 음성 데이터가 가지고 있는 정보와 영상 데이터가 가지고 있는 정보를 동시에 가지고 있는 특징을 생성하고자 한다.

이때, 음성 데이터는 2개의 차원을 가지고 있으므로 시계열 데이터의 특징을 효과적으로 추출할 수 있는 2D CNN으로 이루어져 있는 특징 추출 모델을 이용한다. 2D Convolution Neural Network (CNN)은 주로 이미지 처리를 위해 사용되나 시계열 데이터가 가지고 있는 특징을 누적함으로써 이미지화하여 분석하는 방식으로 시계열 데이터를 효과적으로 처리할 수 있다[17]. 이를 위해 음성 데이터 전처리에 가장 많이 쓰이는 Mel-Spectogram[13] 기법을 이용하여 전처리와 특징 추출을 진행하였으며, Gausian Noise, Pitch, Time stretch, shift[14]를 이용하여 데이터 증강을 하였다.

Algorithm 1: 음성 데이터 전처리 알고리즘

Input: Video data $\{V_1, V_2, ..., V_n\}$, n is data size
Output: audio feature data set

V_k = k'th Video file
extract = take Audio file from Video
cut(n, start, end) = crop audio n start to end
stft = Short Time Fourier Transform
mel = Mel Filter Bank Application
noise = Gaussian Noising
pitch = Pitch changing
stretch = Time stretch
shift = Audio shifting
dataset = [ ]

for $0 \le i \le n$ **do**
  A_i = extract(V_i)
  $A_i^{'}$ = cut(A_i, start = 0, end = 126000)
  W_j = cut($A_i^{'}$, start = $j * 400$, end = $j + 400$)
  $W = \{W_1, W_2, ..., W_{315}\}$
  for w ***in*** W **do**
    f = mel(sftf(w))
  F_i = concat(W)
  dataset.append(F_i)
  G = noise(F_i)
  P = pitch(F_i)
  S = stretch(F_i)
  H = shift(F_i)
  dataset.append(G, P, S, H)

return dataset

Fig. 2. Algorithm of audio preprocessing

다음으로 영상 데이터는 영상에서 특징을 추출하는데 자주 사용되는 slowfast feature extractor를 이용하였다. Slowfast는 영상이 가지고 있는 시간적 공간적 정보를 이용하여 하나의 차원을 가지고 있는 특징으로 생성한다.

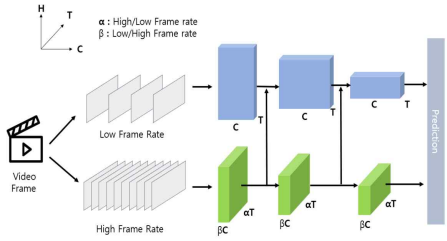

Fig. 3. Diagram of Slowfast Feature Extractor

3.2 멀티모달 모델 구현

멀티모달 모델을 구현하기 위해 가장 중요한 것은 서로 다른 도메인을 가지고 있는 데이터를 처리하는 구조를 설계하는 것이다. 본 논문에서 제안하고자 하는 멀티모달 모델 구조는 서로 다른 도메인을 가지고

있는 데이터를 하나의 차원을 가지는 특징으로 추출하여 분류를 진행할 때 각 데이터가 가지고 있는 모든 정보를 고려할 수 있도록 하였다. 이를 위해 입력으로 들어온 영상의 음성을 추출한 뒤 각 데이터에 맞는 특징을 추출한다. 이후 추출한 특징을 합쳐 만든 통일 도메인 특징을, 분류기를 이용하여 분류를 진행한다. 본 논문에서 제안하는 멀티모달 모델의 모식도는 그림 1과 같다.

Ⅳ. 실험

본 논문에서는 음성을 포함한 비디오 데이터인 RAVDESS 데이터를 이용하였다. 음성 데이터는 영상 데이터가 포함하는 음성을 우선 추출하여 'wav' 파일로 만든다. 이후, 전처리 작업을 진행 후에 가장 많이 사용되는 특징 추출 방법인 Mel-spectrogram 방식을 이용하여 특징을 추출한다.

영상 데이터는 각 영상 데이터를 하나의 특징으로 추출하는 벤치마크인 slowfast를 이용하여 특징 추출을 진행하였다. 이를 통하여 음성과 영상 데이터에서 특징을 추출함으로써 도메인을 통일시키는 역할을 진행한다.

마지막으로 두 모델을 이용하여 추출한 특징을 합쳐서 분류를 진행한 멀티 모달 모델, 각 모델에 분류를 진행하여 그 결과를 합친 모델, 각 모델 중 더 강하게 감정을 인식한 모델의 결과를 따르는 투표 멀티 모달 모델 방식을 구현하여 어떤 방식을 이용한 모델이 더 좋은 결과를 보여주는지를 증명하였다.

본 논문에서 실험은 Linux 20.04 LTS, Geforce RTX 3090, RAM 64 GB, AMD Ryzen 9 5900X, Python 3.8, Tensorflow 2.3.1, Cuda 11.2, and cuDNN 8.7 환경에서 진행하였다.

4.1 데이터셋

데이터셋은 음성이 포함된 7가지 감정을 연기한 RAVDESS 연기 데이터셋을 이용하고자 하였다[11]. RAVDESS 데이터셋은 12명의 남성과 12명의 여성으로 이루어져 있는 24명의 배우가 총 7개의 감정(평상시, 행복, 슬픔, 분노, 공포, 놀람, 혐오스러움)을 연기한 데이터셋이다. 놀람과 혐오스러움을 제외한 5개의 레이블은 각각 384개의 데이터를 가지고 있고 놀람과 혐오스러움은 192개의 데이터를 가지고 있어 총 2,304개의 데이터를 가지고 있다. RAVDESS 데이터셋은 하나의 영상당 하나의 레이블에 해당하는 행동을 하고 있도록 구분되어 있는 잘려진 비디오(Trimmed Video)이며, 20초~24초가량으로 이루어져 있다. 또한 30프레임에 1,920 x 1,080의 크기를 가지고 있다. 본 논문에서는 2,304개의 영상 중 1,600개를 학습 데이터, 320개를 검증 데이터, 376개를 테스트 데이터로 이용하였다.

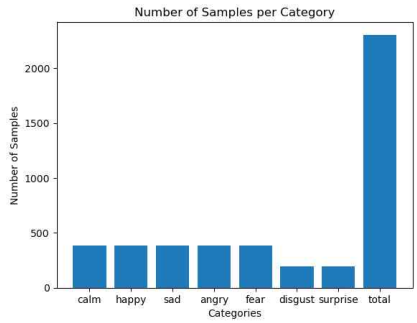

Fig. 4. The number of Label in RAVDESS dataset

4.2 데이터 전처리

데이터 전처리는 모델 학습을 위해서 필수적으로 거쳐야 하는 단계이다. 특히 음성 데이터와 영상 데이터는 그 자체로 학습에 이용하면 상당한 컴퓨팅 파워를 요구할 뿐만 아니라 정확도마저 떨어지는 단점을 가지고 있다. 본 논문에서는 음성 데이터 전처리를 위해 다양한 transform 함수를 사용하고 mel-spectrum 필터링을 통해 특징을 추출하였으며 영상 데이터는 slowfast를 이용하여 특징을 추출하였다.

음성 데이터는 우선 영상 데이터에 담겨있는 음성 데이터를 추출하는 것에서부터 시작한다. 음성 데이터를 추출한 뒤 일정한 크기로 자른다.

이후 Mel-spectrogram 방식으로 특징을 추출한 뒤 데이터 증강(data augmentation)을 진행한다. 데이터 증강은 데이터에 일정 수준 이상의 잡음을 추가하는 변환 기법인 Gaussian Noise 추가를 진행하고 이후에, pitch 즉 소리의 높낮이를 변경하는 피치 변화 증강을 진행한다. 다음은 pitch는 그대로 두되 시간을 늘리거나 줄이는 time stretch 증강을 적용한 뒤 마지막 shift를 적용했다.

또한 RAVDESS 데이터셋은 한 영상에 하나의 레이블이 담겨 있는 trimmed video data set이다. 따라서 각 영상을 잘라서 레이블 화하는 작업은 필요하지 않

다. 영상이 가지고 있는 특징 추출을 위해 영상 특징 추출에 가장 많이 사용되는 slowfast feature extractor를 이용하였다.

영상 특징은 50x56의 shape를 가지고 있으며 총 600프레임으로 이루어져 있어 10개의 윈도우로 나눠 60프레임의 윈도우 중 10프레임을 무작위로 추출하게 된다.

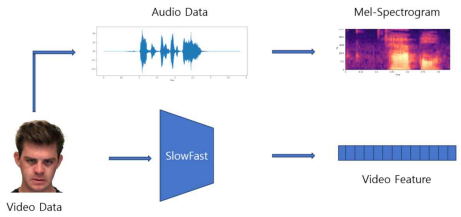

Fig. 5. Diagram of data preprocessing

4.3 음성 인식 모델

멀티모달 모델을 구성하는 음성 인식 모델 및 영상 인식 모델은 우선 각각 학습을 진행한 뒤 멀티모달 모델을 형성할 수 있는 다양한 방법으로 학습된 두 모델을 합치는 방법을 사용하였다. 우선 음성 인식 모델은 2D convolution으로 이루어져 있는 CNN 모델을 이용하였다. 입력값으로는 음성 특징의 shape인 (128,282,1)을 이용하였고, 우선 음성 인식 모델 자체 성능을 평가하기 위해서 마지막에 Fully Connected MLP를 이용하여 분류를 진행하는 모델과 2048의 shape를 가지는 음성 특징을 추출하는 모델로 따로 구분하였다.

분류기를 이용한 음성 인식 자체 모델의 성능을 평가하기 위해서 학습률 0.001을 가지는 Adam 옵티마이저를 이용하였다. Loss는 분류에서 가장 많이 사용하는 Categorical crossEntropy를 이용하였으며, 배치 크기는 32에 100번의 학습을 진행하였다.

4.4 영상 인식 모델

영상 인식 모델은 추출한 feature를 분류하는 분류기로 만들었다. 멀티 모달 모델에서는 이 특징을 그대로 사용할 것이기 때문에 영상 인식 모델의 성능을 구현하기 위해 2800의 shape를 가지는 특징을 분류하는 MLP 모델을 이용하여 영상 인식 모델 성능을 평가하였다. 이후 멀티모달 모델에서는 이 모델과 음성 인식 모델의 결과를 합치는 멀티모달 모델과 함께 특징을 합쳐 분류를 진행하는 모델을 구현하여 성능을 평가할 것이다. 영상 인식 모델의 성능을 평가하기 위해서 학습률 0.01을 가지는 Adam 옵티마이저를 이용하였으며, 또한 Categorical crossEntropy를 이용하였다. 영상 인식 모델의 배치 사이즈는 32에 20번의 학습을 진행하였다. 이후 각 모델은 분류기를 분리한 뒤 특징 추출기로 이용한다.

4.5 멀티모달 모델

멀티모달 모델은 본 논문에서 적용한 도메인을 일치하여 특징을 추출하고 분류를 진행하는 Multi-ⅰ모델과 기존 멀티모달 모델을 구현할 때 자주 사용되었던 방법론들인 각 모델을 각자 학습 및 추론한 뒤 그 결과가 가지고 있는 scoce 값을 합하여 가장 높은 값을 가지는 레이블을 정답으로 도출하는 Multi-ⅱ모델, 마지막으로 두 모델 중 score값이 가장 높은 값을 결과로 하는 Multi-ⅲ모델을 구현하였다. 각 모델에 대한 다이어그램은 아래 <Fig 6>에서 나타난다[15]. 별도의 분류기(Classifier)를 사용하지 않는 Multi-ⅱ, Multi-ⅲ 모델이 아닌 Multi-ⅰ 모델은 분류기를 사용하므로 재학습을 진행한다. 이때 음성 인식 모델과 영상 인식 모델은 미리 학습한 모델을 사용하여 특징을 추출하며 분류기는 3개의 은닉층을 가지는 MLP를 이용하였다. 분류기의 학습은 Categorical crossEntropy 손실함수를 이용했으며 배치 사이즈는 32에 20번의 학습을 진행하였다.

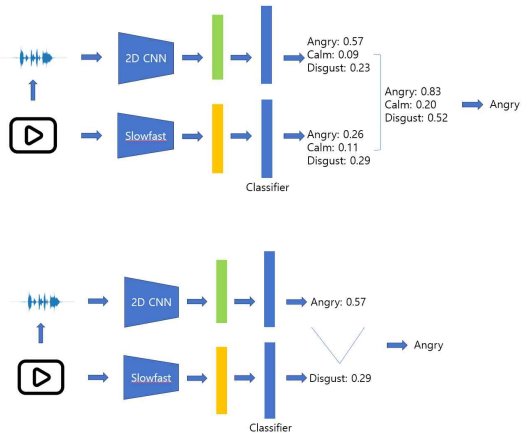

Fig. 6. Diagram of Mutimodal models

4.6 실험 결과

각 모델의 정확도는 분류에서 가장 많이 사용되는 정확도(Accuracy)를 이용했으며 이는 아래 식으로 나타난다.

$$ACC = \frac{TP + TN}{P + N}$$

이때, TP와 TN은 각각 분류를 정확히 한 횟수를 나타내며 T+N은 전체 시행 수를 나타낸다.

각 모델의 정확도는 아래와 같이 <Table 1>에 나타난다. 우선 음성 데이터만을 이용했을 경우 테스트 데이터의 정확도가 38.42퍼센트로 낮은 결과를 나타냈다. 영상 모델만을 이용했을 때 또한 40.55%로 낮은 결과를 나타내었다.

Table 1. The result of model accuracy

number	Model	Validation Accuracy (%)	Test Accuracy (%)
01	Audio model (CNN)	43.25	38.42
02	Video model (MLP)	46.39	40.55
03	Multi-ⅰ	61.58	60.11
04	Multi-ⅱ	52.15	45.52
05	Multi-ⅲ	39.14	36.57

멀티모달 모델의 경우 음성-영상 데이터 모델에서 추출한 특징을 붙여 하나의 도메인으로 축소한 뒤 분류를 진행한 Multi-ⅰ모델의 경우에는 60.11%의 준수한 결과를 나타냈다. 이는 각 도메인에 추출한 특징을 합침으로써 하나의 감정을 나타내는 두 데이터가 가지고 있는 정보를 전부 반영할 수 있기 때문이다. 다음으로 Multi-ⅱ모델의 경우 각 모델을 따로 학습 및 추론을 진행한다. 이후 각 모델의 추론 결과를 나타내는 스코어 텐서의 합을 이용하여 최종 예측 결과를 도출하는 모델이다. 이 모델의 경우 45.52%로 기존 모델보다 성능이 좋지만 Multi-ⅰ모델보다는 좋은 성능을 내지 못하는 것을 볼 수 있다. 이는 도메인이 다른 음성과 영상 데이터를 따로 학습하여 결과를 내는 것보다 Multi-ⅰ와 같이 각기 다른 도메인을 가지고 있는 데이터가 가지고 있는 정보를 한 번에 학습하여 추론하는 것이 더 결과가 좋다는 것을 보여준다.

마지막으로 Multi-ⅲ 모델은 음성 인식 모델과 영상 인식 모델의 결과 스코어(각 모델이 분류한 감정의 점수) 중 가장 높은 값을 결과로 도출하는 방법을 이용하였다. 예를 들어 음성 인식 모델이 분노의 감정을 62.5의 스코어로 인식하고, 영상 인식 모델이 역겨움의 감정을 45의 스코어로 인식하였을 경우 본 멀티모달 모델은 더 높은 스코어를 가진 음성 인식 모델에 따라 분노의 감정으로 분류하는 것이다. 이러한 방법을 이용한 Multi-ⅲ 모델은 가장 성능이 낮은 36.57%를 나타냈다. 이는 각 모델이 가지고 있는 평균적인 스코어값이 달라 두 모델 중 하나의 모델의 결과가 총 결과를 선택하는 데 영향을 미치지 못하기 때문으로 나타난다.

Ⅴ. 결론

본 논문에서 진행한 연구는 각기 다른 도메인을 가지고 있는 음성-영상 데이터를 이용하여 감정을 인식하는 멀티모달 모델을 구현할 때 어떤 방식이 가장 좋은 결과를 도출하는지를 실험으로 나타냈다. 본 논문의 결과에 따르면 각 모델의 특징을 합친 뒤 분류하는 방법이 60.11%의 결과로 도메인이 다른 데이터를 각자 학습 및 추론하여 도출한 결과를 더하는 것보다 모델이 결과를 도출해 내는 과정 중간에 두 데이터의 특징 및 정보를 합쳐 도메인을 통일시킨 뒤 한 번에 분석하는 것이 더 유리하다는 결과를 도출할 수 있었다. 7개의 감정을 분류하는 과업에서 60.11%의 정확도는 뛰어난 수치는 아니지만 충분히 유의미한 수치이다. 본 논문에서는 감정 인식 과업에 있어서 멀티모달 모델의 성능을 확인했으며 향후 성능 향상 및 데이터셋 도메인 확대를 위해 음성 및 영상 데이터뿐만 아니라 동기화된 생체 데이터(맥동 데이터, 심전도 데이터 등)를 이용함으로써 정확도 향상 효과를 얻을 수 있는 연구를 수행하고자 할 것이다.

ACKNOWLEDGMENTS

이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국에너지기술평가원의 지원을 받아 수행된 연구임(20224B10100060, 회전설비 인공지능형 진동 감시 시스템 개발)

REFERENCES

[1] Bum-Jun kim et al. (2018). Research on data augmentation method for audio marking based on deep neural network. Journal of the Korean Acoustical Society, 37(6), 475-482.

[2] Ju-Hyuk Han et al. (2022). MR-CL: Contrastive Loss-based multimodal convergence learning method for emotion recognition. Korean Society of Information Scientists and Engineers Academic Presentation Papers

[3] Chung-bin Kim et al. (2022). Feature extraction from self-supervised learning model and transformer-based multimodal expression learning for combined voice-text emotion recognition. Korean Society of Information Scientists and Engineers Academic Presentation Papers

[4] Yong-hwa Jo el al. (2020). Implementation of a Classification System for Dog Behaviors using YOLI-based Object Detection and a Node.js Server. Journal of Convergence Signal Processing Society. v.21, no.1, pp.29-37

[5] Lee-sun Mun et al. (2023). Multimodal emotion recognition based on biometric signals and voice data. Proceedings of the Domestic Conference of the Society of Control and Robotics Systems

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017,

[7] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9,

[8] Kahou, S.E., Bouthillier, X., Lamblin, P. et al. EmoNets: Multimodal deep learning approaches for emotion recognition in video. J Multimodal User Interfaces 10, 99-111 (2016).

[9] Xu, Haiyang, et al. "Learning alignment for multimodal emotion recognition from speech."arXiv preprint arXiv:1909.05645 (2019).

[10] Alu, D. A. S. C., Elteto Zoltan, and Ioan Cristian Stoica. "Voice based emotion recognition with convolutional neural networks for companion robots." Science and Technology 20.3 (2017): 222-240.

[11] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13.5 (2018): e0196391.

[12] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[13] Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.

[14] Schlüter, Jan, and Thomas Grill. "Exploring data augmentation for improved singing voice detection with neural networks." ISMIR. 2015.

[15] Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

[16] Jae-Eun Lee et al. (2022) A neck healthy warning algorithm for identifying text neck posture prevention. Journal of Convergence Signal Processing Society. v.23, no.3, pp.115-122

[17] HATAMI, Nima; GAVET, Yann; DEBAYLE, Johan. Classification of time-series images using deep convolutional neural networks. In: Tenth international conference on machine vision (ICMV 2017). SPIE, 2018. p. 242-249.

저자소개

김 종 구 (Jong-Gu Kim)

2022년 2월 : 인하대학교
컴퓨터공학과(공학사)
2023년 3월-현재 : 인하대학교
전기컴퓨터공학과(공학석사)
관심분야 : 인공지능, 컴퓨터 비전,
신호 처리, 대조 학습

권 장 우 (Jang-woo Kwon)

1990년 2월 : 인하대학교
전자공학(공학사)
1992년 2월: 인하대학교
정보공학(공학석사)
1992년 2월: 인하대학교
정보공학(공학박사)
2012년-현재 : 인하대학교 교수
관심분야 : 인간 컴퓨터 상호작용