GAN을 이용한 음성 감정 인식 모델의 성능 개선

Performance Improvement of Speech Emotion Recognition Model Using Generative Adversarial Networks

저자
(Authors)

고유정, 김윤중
You-Jung Ko, Yoon-Joong Kim

출처
(Source)

한국정보기술학회논문지 17(11), 2019.11, 77-85(9 pages)
The Journal of Korean Institute of Information Technology 17(11), 2019.11, 77-85(9 pages)

발행처
(Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09263040

APA Style

고유정, 김윤중 (2019). GAN을 이용한 음성 감정 인식 모델의 성능 개선. 한국정보기술학회논문지, 17(11), 77-85

Journal of KIIT. Vol. 17, No. 11, pp. 77-85, Nov. 30, 2019. pISSN 1598-8619, eISSN 2093-7571 77
http://dx.doi.org/10.14801/jkiit.2019.17.11.77

GAN을 이용한 음성 감정 인식 모델의 성능 개선

고유정*, 김윤중**

Performance Improvement of Speech Emotion Recognition Model Using Generative Adversarial Networks

You-Jung Ko*, Yoon-Joong Kim**

요 약

최근 딥 러닝 모델의 발전에 따라 음성 감정 인식 모델의 성능 개선이 이루어지고 있으나 충분한 학습 데이터 확보의 어려움은 여전히 성능 개선의 저하 요인이다. 본 논문은 Generative Adversarial Network(GAN)으로 부정 감정 데이터를 생성하여 부정 데이터 학습을 추가함으로써 음성 감정 인식 모델의 성능을 개선하는 방법을 제안한다. 제안된 시스템은 감정 인식 판별기, 감정 신호 생성기로 구성되어 있다. 생성기는 부정 감정 신호와 부정 레이블을 만들어 학습 데이터 셋을 보완하고, 판별기는 실제 감정 신호와 부정 감정 신호가 포함된 학습 데이터 셋으로 훈련된다. 실험은 IEMOCAP 데이터 셋을 사용하였고 다양한 인식 모델을 구성하여 인식률을 비교한 결과 GAN을 추가한 감정 인식기가 BLSTM과 Attention을 이용한 감정 인식 모델에 비해 1.86% 더 정확한 예측을 제공하는 것으로 나타났다.

Abstract

Recently, with the development of the deep learning model, the performance of the voice emotion recognition system has been improved. However, the difficulty of obtaining sufficient training data is still a deterioration factor of the performance improvement. In this paper, we propose a method to improve the performance of speech emotion recognition model by generating negative emotion data and adding negative data learning using Generative Adversarial Network (GAN). The proposed system consists of emotion recognition discriminator and emotion signal generator. The generator complements the learning dataset by creating negative emotion signals and negative labels. The discriminator is trained with a learning dataset that includes real and negative emotion signals. In the experiment, the IEMOCAP data set was used, and the recognition rate was compared by constructing various recognition models and it was shown that the emotion recognizer with GAN provides 1.86% more accurate prediction than the emotion recognition model using BLSTM and Attention.

Keywords
generative adversarial network, speech emotion recognition, attention mechanism, BLSTM

* 한밭대학교 컴퓨터공학과 강사
- ORCID: https://orcid.org/0000-0003-4882-5784
** 한밭대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-5451-5558

· Received: Oct. 06, 2019, Revised: Nov. 01, 2019, Accepted: Nov. 04, 2019
· Corresponding Author: Yoon-Joong Kim
Dept. of Computer Engineering, Hanbat University, 125, Dongseo-daero, Yuseong-gu, Daejeon, Republic of Korea
Tel.: +82-42-821-1143, Email: yjkim@hanbat.ac.kr

Ⅰ. 서 론

인간의 감정 상태는 인간관계에 있어서 중요한 요소이며 얼굴 표정, 목소리의 특색, 대화 내용 등에 영향을 미친다. 음성은 감정을 표현하는 통로로 인간과 기계간의 상호 작용 시 인간의 음성에 담긴 감정을 인식하고, 번역하여 반응하는 것이 중요하다. 이에 따라 음성 감정 인식 시스템(Speech emotion recognition)에 관한 많은 연구가 이루어지고 있다.

음성 감정 인식기의 성능은 효과적인 특징 추출과 정확한 분류기를 설계하는 것이 중요하다. 특징 추출은 음성 감정 인식 시스템의 주요 문제이다. 연구자들은 에너지, 피치, 포먼트 주파수, LPCC (Linear Prediction Cepstrum Coefficients), MFCC (Mel-Frequency Cepstrum Coefficients) 및 MSF (Modulation Spectral Feature)와 같은 감정 정보가 포함 된 특징을 추출하는 기능을 제안하였다[1]. 특징은 인식기를 훈련시키는데 사용된다. 음성 감정 인식의 분류 알고리즘으로 GMM, HMM, SVM 등이 제안되어 왔으나 최근에는 심층 신경회로망(Deep neural network)의 발전에 따라 딥 러닝 기술을 적용하여 인식기의 성능을 개선하는 연구가 이루어지고 있다[2]. Wollmer는 각 발화에 대해 4843개의 특징을 추출하고 LSTM-RNN(Long Short Term Memory-Recurrent Neural Network)을 이용하여 감정을 분류하였다[3]. Mirsamadi는 네트워크가 감정적으로 중요한 부분에 집중할 수 있는 Attention Mechanism을 사용하여 양방향 LSTM과 Pooling을 결합하는 방법을 제안하였다[4]. XIE는 심층 신경망의 상위 계층에서 정보를 효율적으로 활용하고 성능이 저하되는 것을 해결하기 위해 음성 감정 인식을 위한 Attention 기반의 LSTM을 제안하였다[5]. Satt는 스펙트럼에 직접 딥 러닝 기술을 적용하고 분류시 RNN과 CNN을 이용하였다[6].

이와 같이 심층 학습 방법으로 음성 감정을 인식하는 연구가 진행되어 인식 시스템의 성능 개선이 이루어지고 있으나 충분한 학습 데이터 확보의 어려움으로 인하여 여전히 성능 개선의 저하 요인이 되고 있다. 딥 러닝의 학습은 충분한 양의 데이터를 학습시킬 때 기계학습 알고리즘보다 좋은 성능을 보인다. 하지만 자연스러운 음성 감정 데이터 셋은 부족하며 음성 데이터는 법적 및 도덕적 문제로 공개적으로 이용할 수 없는 경우가 많다. 대부분의 공공 데이터베이스는 연기자에 의해 생성된 데이터로 실제 상황에 비해 감정 표현이 편향될 수 있다. 또한 표현된 감정과 느낌이 다르기 때문에 연기된 감정에 대해서 레이블이 필요하지만 레이블 작업은 많은 인적자원과 시간이 소요된다[7].

따라서 본 연구에서는 최근 이미지 생성 모델에서 좋은 성능을 보이고 있는 생성적 적대 신경망 중 하나인 GAN(Generative Adversarial Network)[8]을 이용하여 부정 감정 데이터를 생성하여 부정 데이터 학습을 추가함으로써 음성 감정 인식 모델의 성능을 개선하는 방법을 제안한다. 제안된 시스템은 감정 신호 생성기와 감정 인식 판별기로 구성되어 있다. 생성기는 부정 감정 신호와 부정 레이블을 만들어 학습 데이터 셋을 보완하고, 판별기는 BLSTM (Bi-directional Long Short Term Memory)과 Attention Mechanism 신경망 모델을 구성하여 실제 감정 신호와 부정 감정 신호를 학습한다.

실험은 IEMOCAP 데이터 셋을 사용하였고, 다양한 인식 모델을 구성하여 인식률을 비교하였다. 평가 결과 BLSTM과 Attention을 이용한 감정 인식모델은 59.59%, GAN 생성모델을 추가한 음성 감정 인식 모델은 61.45%로 GAN을 적용하지 않은 모델보다 1.86% 개선됨을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 소개하고, 3장에서는 GAN을 이용한 음성 감정 인식 시스템에 대해 기술한다. 4장에서는 네 개의 감정 인식 모델과 비교 실험하여 성능 평가에 대해 논한다. 마지막으로 5장에서는 결론 및 향후 연구에 대하여 기술한다.

Ⅱ. 관련 연구

2.1 GAN

GAN은 생성기 모델과 판별기 모델이 서로 적대적으로 경쟁하며 각 모델의 성능을 향상시키는 비

지도 학습 기반의 모델이다. 생성기 모델은 잠재 벡터(z)를 입력받아 실제 데이터와 유사한 데이터를 만들어낸다. 판별기 모델은 입력된 데이터가 실제 데이터인지 생성기가 만든 가짜 데이터인지 확률값을 계산한다[9]. 성능을 향상시키기 위하여 다음의 목적함수를 최대가 되도록 학습한다. GAN의 목적함수 $V(D, G)$는 식 (1)과 같이 입력 샘플 x, z에 대하여 판별기 D와 생성기 G로 확률을 계산한다.

$$\min_{G}\max_{D} V(D, G) = \tag{1}$$
$$E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

x는 실제 데이터의 확률 분포 $p_{data}(x)$에서 샘플링한 데이터이고, z는 가우시안 분포 $p_z(z)$로부터 샘플링된 잡음이다.

판별기 D의 학습과정에서, $D(x)$는 실제 데이터 x에 대하여 계산되는 판별기 확률 $D(x)$는 1, 잡음 z으로 생성된 가짜(Fake) 데이터 $G(z)$에 대하여 계산되는 판별기 확률 $D(G(z))$는 0이 되어 목적함수가 최대가 되도록 학습하여 판별기의 판별능력을 향상시킨다.

생성기 G의 학습과정에서, 판별기 $D(x)$는 관계가 없고 잡음 z으로 생성된 가짜 데이터에 대한 판별기 확률 $D(G(z))$이 1이 되어 목적함수가 최소가 되도록 학습하여 가짜 데이터의 생성능력을 향상시킨다.

2.2 LSTM

최신 음성인식 모델들은 LSTM RNN을 기반으로 개발되고 있다. LSTM은 FNN(Feedforward Neural Network)에 비하여 재귀적으로 처리하는 기능이 포함되어 있으므로 음성과 같이 시계열 형태의 데이터 처리에 있어서 괄목할 만한 성능을 가지고 있다. LSTM의 처리 과정은 다음과 같다. 순차적으로 입력되는 데이터에 대하여 상태 값을 동적으로 업데이트 한다. 현 상태의 크기는 이전 상태의 크기와 현 입력의 크기가 비선형으로 인터폴레이션(Interpolation)되어 임시의 현재 상태 값을 계산해 내고 그 값을 출력 게이트로 제어하여 최종적인 현 상태 값을 계산한다. 비선형 인터폴레이션은 망각 게이트와 입력 게이트에 의해서 제어되는 것을 의미한다. 모든 게이트는 신경망으로 구성된다. LSTM은 이와 같은 구조를 가짐으로써 이전 정보를 선택적으로 기억할 수 있어 음성신호에 산재되어 있는 감정의 정보를 기억하는데 유효하다.

따라서 LSTM은 음성 특징열의 의존성을 학습할 수 있고 의존성의 패턴은 감정에 따라 고유하게 대응될 수 있으므로 시계열 특징 패턴으로부터 감정을 학습하고 인식하는 모델에서 중요한 역할을 수행한다. BLSTM은 순방향 의존성 뿐만 아니라 역방향 의존성까지 기억할 수 있어 감정학습에 효율적이다.

2.3 Attention Mechanism

Attention Mechanism은 기계 번역을 위하여 개발된 seq2seq(S2S)모델[10]에서 발생하는 장기 의존성 문제를 해결하기 위해 제안되었다. S2S모델은 인코더와 디코더라는 두 개의 LSTM으로 구성된다. 인코더는 LSTM을 사용하여 다양한 길이의 입력 시퀀스를 하나의 문맥 벡터로 변환하고 디코더는 문맥 벡터를 사용하여 출력 시퀀스를 생성한다. 하지만 입력 시퀀스 길이가 길어지면 인코더가 입력 시퀀스의 프론트 엔드 정보를 디코더에 적절하게 전송하지 못하기 때문에 S2S 모델은 장기 의존성 문제가 발생한다.

위에 언급한 바와 같이 장기 의존성 문제를 극복하기 위해 Attention Mechanism[11]이 제안되었다. Attention Mechanism은 디코더가 출력을 계산할 때 중요한 인코더 LSTM 출력에 초점을 맞춘다. 따라서 Attention Mechanism은 소스와 목표 시퀀스 간의 길이에 따른 의존성에 제안되지 않으며 감정 음성인식을 포함한 다양한 분야에서 시도되고 있다.

III. GAN을 이용한 음성 감정 인식 모델

본 연구에서 제안한 GAN을 이용하여 음성 감정 인식 모델의 개념도는 그림 1과 같이 특징 추출 모듈, 감정 인식 판별기, 감정 신호 생성기로 구성된다.

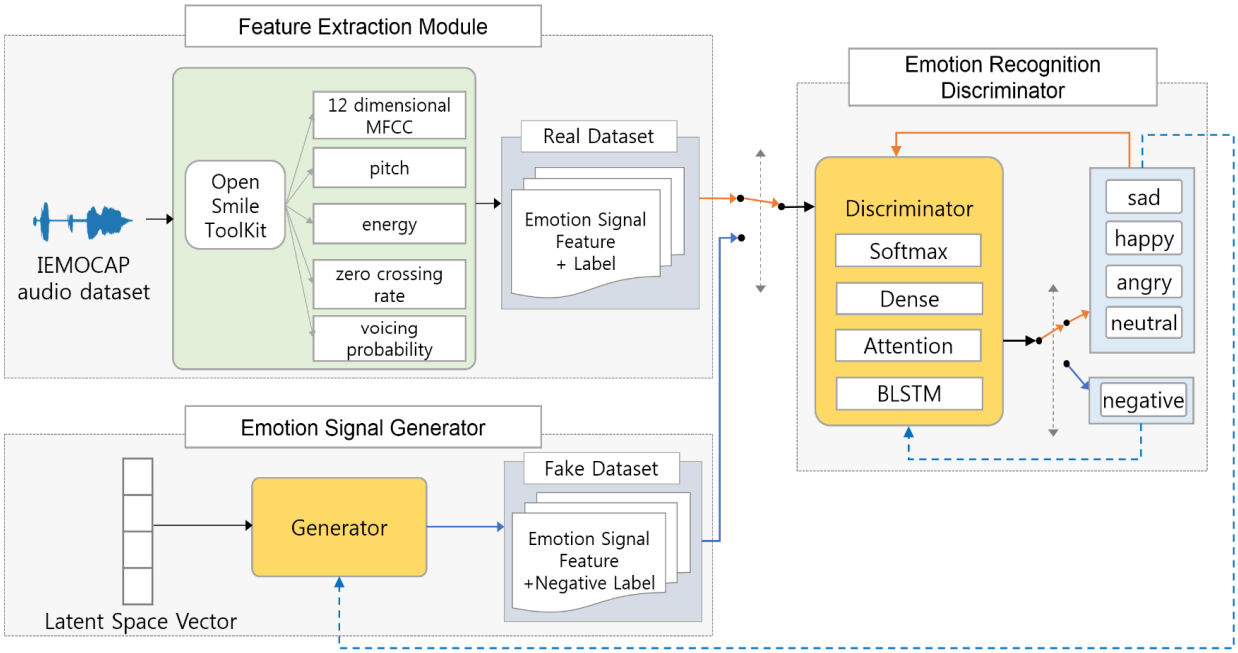

그림 1. 제안한 GAN을 이용한 음성 감정 인식 시스템 구조
Fig. 1. Proposed speech emotion recognition system architecture using GAN

특징 추출 모듈(Feature extraction module)은 IEMOCAP 음성 데이터를 입력받아 특징을 추출하고 추출된 실제 감정 신호와 레이블 모음으로 Real Dataset(실제 데이터 셋)을 생성한다.

감정 신호 생성기(Emotion signal generator)는 잠재 공간 벡터로부터 부정 감정 신호를 생성하고 부정 레이블을 추가하여 가짜 데이터 셋을 생성한다.

감정 인식 판별기(Emotion recognition discriminator)는 실제 데이터 셋의 감정 신호의 특징과 레이블로 학습(주황색 흐름)하고, 가짜 데이터 셋의 생성된 특징과 부정(Negative) 레이블로 학습(청색흐름)한다.

3.1 특징 추출 모듈

특징 추출 모듈은 그림 1과 같이 IEMOCAP (Interactive Emotional Dyadic Motion Capture)[12]로부터 실제 데이터 셋을 생성한다. IEMOCAP 음성 데이터는 5개의 세션으로 구성되어 있으며, 각 세션에서는 남녀 배우가 특정한 감정을 이끌어내기 위해 대본 시나리오와 즉흥 연기를 녹음 하였다. 녹음된 데이터는 여러 명의 평가자가 평가하여 최소 3개의 감정 레이블이 지정되어 있다. 원본 데이터베이스는 9가지 감정 유형인 분노, 흥분, 행복, 슬픔, 평상, 좌절, 두려움, 놀라움, 혐오감이 있다.

본 연구의 실험범주는 이전 연구[13]와 같이 화남, 행복, 평상, 슬픔 4개의 감정을 선정하고, 표 1과 같이 4,490개의 음성파일을 감정별로 7:3으로 분리하여 학습(Train) 및 검증(Test)을 구성하였다.

표 1. 실험을 위한 실제 데이터 셋
Table 1. Real dataset for experiment

	Angry	Happy	Neutral	Sad	All
Train	772	416	1,195	758	3,141
Test	331	179	513	326	1,349
All	1,103	595	1,708	1,084	4,490

Real dataset의 모든 음성신호(파일)는 양자화 주파수는 1600Hz로 녹음되어 있으며 openSMILE[14] 툴킷을 이용하여 특징을 추출 하였다. 음성신호의 특징 벡터(LLD) 열은 25ms 윈도우를 10ms씩 이동하면서 추출된다. 특징 벡터는 피치(Pitch), 에너지(Energy), 제로 크로싱 속도(Zero-crossing rate), 발음확률(Voicing probability), 12차 MFCC와 이 값들은 1차 미분으로 구성되는 32차 벡터이다.

3.2 감정 신호 생성기

생성기는 크기 100의 1차원 잠재 공간 벡터(Latent space vector) z를 입력으로 받아 512×32 부정 감정 신호의 특징 벡터열을 생성하고 부정라벨을 추가하여 가짜 데이터 셋을 생성한다.

생성기는 4개 층으로 구성되어 있고 각 층은 다시 Dense 레이어, LeakeyReLU활성화 함수 및 배치 정규화(Batch normalization) 레이어로 구성된다. Dense 레이어의 크기는 256, 512, 1024로 점점 확장 시키고 마지막 레이어는 특징 벡터 크기로 맞추어 512x32로 재구성하였다. 512는 타임 스텝이고 32는 특징벡터의 크기이다. 모든 활성 함수의 alpha 값은 0.2이고, 배치 정규화 레이어의 momentum값은 0.8을 사용하였다.

3.3 감정 인식 판별기

감정 인식 판별기는 입력되는 실제 데이터 셋 또는 가짜 데이터 셋의 특징 벡터 열에 대하여 확률값을 계산한다. 이 값은 라벨과 entropy가 계산되어 학습에 이용된다. 판별기의 모델의 구성요소 BLSTM, Attention, Dense, Output 레이어와 요소별 파라메터는 그림 2와 같다.

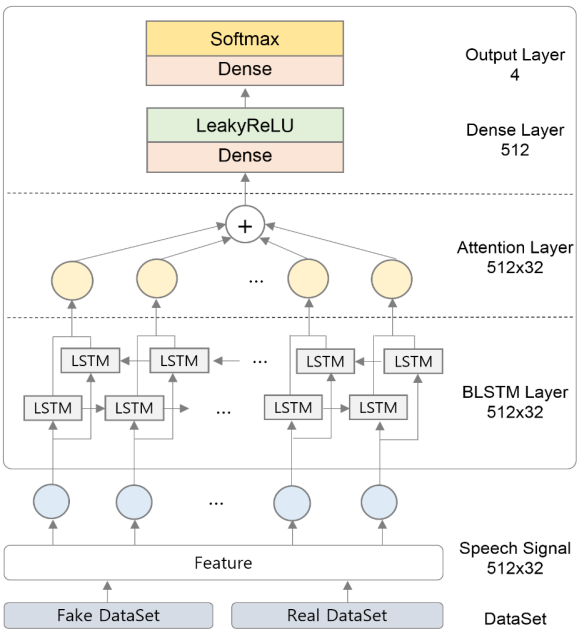

그림 2. Attention과 BLSTM을 이용한 판별기
Fig. 2. Discriminator using attention and BLSTM

하나의 T개의 프레임으로 구성되는 특징벡터열을 $\{x_t|t=[1,T]\}$, BLSTM의 출력을 $\{h_t|t=[1,T]\}$이라고 한다.

BLSTM레이어는 데이터 셋으로부터 입력된 32차원의 512 특징 벡터열 x_t로부터 순방향 및 역방향 프레임간 의존성 정보를 계산하여 감정벡터 열 h_t을 출력한다. 즉 이 감정벡터의 수치는 음성신호에 산재되어 있는 감정의 정보를 의미한다.

Attention 레이어는 감정벡터 열로부터 목적 감정에 관계가 있는 감정벡터들을 가중하여 하나의 감정벡터 c를 출력한다. Attention Mechanism은 연구[10]을 기반으로 다음과 같이 계산된다. 목적감정에 대한 감정벡터 h_t의 가중치는 어텐션벡터 α_t는 정렬 함수 $s(h_t)$로 계산된다.

$s(h_t)$는 식 (2)와 같이 히든 파라메터 벡터 W_a와 tanh 활성화 함수로 구성되는 FCN(Fully Connected Neural Network)으로 변환되고 질의벡터 v_a와 내적으로 연산된다.

$$s(h_t) \equiv v_a^{\mathrm{T}} \tanh(W_a h_t) \tag{2}$$

어텐션 벡터 α_t는 식 (3)과 같이 $s(h_t)$를 softmax 함수로 정규화하여 계산한다.

$$\alpha_t = \frac{\exp(s(h_t))}{\sum_{i=1}^{T}\exp(s(h_i))} \tag{3}$$

감정문맥벡터 c는 식 (4)와 같이 어텐션 벡터 α_t를 감정벡터 h_t에 가중합하여 계산한다.

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{4}$$

Attention 레이어에서 출력된 512차원의 감정문맥벡터 c가 두 개의 Dense 레이어에서 처리되어 확률을 계산한다. 즉 입력특징벡터 열 $\{x_t|t=[1,T]\}$의 확률을 로짓(logit)으로 변환한다. 첫 번째 Dense 레이어는 유닛 512이고 LeakyReLU로 활성화된다. 두 번째 Dense 레이어는 유닛 4로 Softmax로 활성된다.

3.4 감정 인식 시스템 학습과정

감정 인식 시스템의 학습과정은 감정 인식 판별기 모델 학습과 감정 신호 생성기 모델 학습으로 구성된다. 학습과정은 판별기가 각 훈련데이터 샘플의 확률값을 계산하고 이 값과 레이블의 크로스 엔트로피 연산으로 손실값(Loss)을 산출한다. 이 손실값이 최소가 되도록 아담(Adam) 최적화기로 네트워크의 파라메터를 학습한다.

판별기 모델 학습은 RealDataset을 입력 받아 감정을 분류하는 실제 데이터 학습과정과 FakeDataset을 입력 받아 부정 감정으로 판별하는 부정 데이터 학습과정으로 구성된다. 실제 데이터 학습과정의 훈련 데이터는 실제 데이터 셋을 사용하고, 실제 데이터의 레이블은 원핫 인코딩(One-hot encoding)형태로 바꾸어 화남은 [1,0,0,0], 평상은 [0,1,0,0], 행복은 [0,0,1,0], 슬픔은 [0,0,0,1]로 구성하였다. 부정 데이터 학습과정의 훈련은 생성기가 만들어낸 가짜 데이터 셋이고, 레이블은 훈련데이터가 어떤 감정 분류에도 속하지 않음을 나타나기 위해 [0,0,0,0]으로 설계하였다.

감정 신호 생성기 모델 학습과정은 훈련 데이터는 가짜 데이터 셋이고, 레이블은 실제 데이터 레이블을 사용한다. 학습모델은 생성기와 판별기의 결합된 모델 $D(G(z))$로 구성하였고, 판별기 $D(x)$의 학습기능은 제안된 상태에서 이루어진다.

GAN의 기본모델은 식 (1)의 목적함수에서와 같이 이진분류를 목적으로 하고 있으나 본 연구에서는 4개 클래스의 분류를 목적으로 하고 있으므로 판별기의 확률 계산 및 레이블을 4차원으로 구성하였다.

IV. 실험 및 분석

제안 모델의 성능을 비교평가하기 그림 1의 아키텍쳐에서 판별기 모델을 다양하게 구성하여 4개 실험을 하였다.

4.1 평가 실험과 학습

평가실험은 다음과 같이 구성하였다. 표 2와 같이 Attention, BLSTM, Dense의 조합으로 구성되는 판별기모델, GAN의 유무 및 RealDataset(3.1절) 및 FakeDataset(3.2절)의 조합으로 4개의 비교실험을 구성하였다. 판별기 모든 모델에는 Dense레이어가 포함되어 표 2에는 생략한다. 본 연구에서 제안한 모델의 실험은 BAGM이다. 즉, RealDataset과 FakeDatset이 이용되고 BLSTM, Attention, Dense으로 구성된 판별기를 포함한 GAN의 아키텍쳐이다.

표 2. 데이터 셋, 판별기, GAN에 따른 실험유형
Table 2. Experiment type according to dataset, discriminator and GAN

Experiments	Discriminator	GAN	Dataset
AM	Attention	x	RealDataset
BAM	BLSTM+Attention	x	
AGM	Attention	o	RealDataset FakeDataset
BAGM	BLSTM+Attention	o	

4개의 실험에 사용된 학습용 훈련셋 및 평가셋의 레이블은 3.4절에서 기술한 바와 같고, 학습용 하이퍼 파라미터는 학습률(Learning rates) 0.0001, 에포크(Epoch) 400, 배치 사이즈 1024이다. 사용한 평가지수 WA(Weighted Accuracy)는 검증 샘플 개체수에 대한 정인식 샘플수의 비율이며 UA(Unweighted Accuracy)는 감정별 재현율(Recall)의 평균이다.

학습 패턴을 분석하기 위하여 그림 3과 같이 실험유형별 학습 횟수에 따라 WA값을 시각화하였다.

BAGM은 130회 학습만으로도 인식율의 성능이 좋아지는 것을 확인할 수 있으나 다른 모델과의 더 정확한 비교를 위해 400회를 학습하였다.

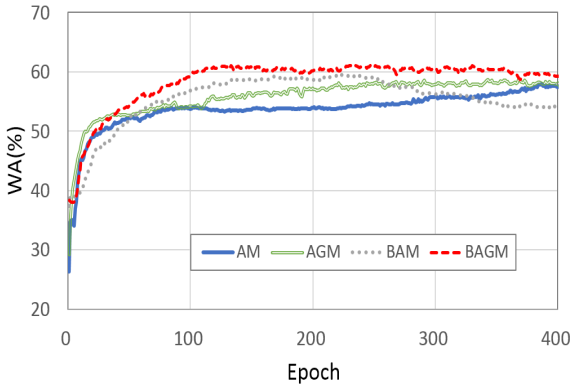

그림 3. 학습 횟수 따라 실험유형별 WA
Fig. 3. WA by experiment type according to epoch

AGM과 AM도 점차 인식율이 높아지고 있음을 알 수 있으나 AM은 300회 이상 학습해야 인식율이 개선되는 것을 알 수 있다. BAM은 250회 이상 학습시 과적합 현상을 보이고 있다. 제안한 BAGM 모델이 학습 속도가 빠르고 정확도가 우수함을 확인할 수 있었다.

4.2 실험결과 및 분석

4개의 실험은 5회 반복하였으며 평균하여 집계한 결과는 표 3과 같다.

표 3. 각 실험유형의 인식 결과
Table 3. Recognition result of each experiment type

Experiments	WA(%)	UA(%)
AM	57.96	50.23
AGM	58.71	53.52
BAM	59.59	54.01
BAGM	61.45	55.19

BLSTM레이어 Dense로 구성된 인식기의 실험 BAM은 59.59%이고 포함하지 않은 AM 57.96%로 1.63% 향상된 결과를 보였다. BAM은 Attention의 감정문맥벡터의 학습특성 뿐만 아니라 BLSTM의 음성 특징열의 의존성(감정정보) 학습기능에 의하여 개선된 것으로 판단된다.

GAN을 포함한 실험 AGM은 AM보다 0.75% 성능이 향상되었고, 실험 BAGM은 BAM보다 1.86% 더 나은 성능을 보임을 알 수 있다.

본 논문에서 제안한 BAGM 모델은 61.45%로 가장 높은 인식률을 나타내었다. BAGM모델은 생성기로 FakeDataset을 만들어 훈련 데이터 셋을 확장하였다. 이는 RealDataset으로만 학습한 모델에 비하여 확장된 데이터를 학습하였기 때문에 나은 성능을 가진 것으로 판단된다.

감정 인식결과의 분포를 분석하기 위하여 표 4와 같이 BAGM 실험의 결과로 혼동행렬을 생성하였다. 행은 평가 감정의 예측 값이고 열은 평가 감정의 이름이다. 예를 들어, 표 1의 평가용 화남 감정의 개체수 331개중 화남(Angry)으로 예측한 개체수는 221개로 재현율은 66.77%이다. 행복(Happy), 평상(Neutral), 슬픔(Sad)도 같은 방법으로 1.81%, 24.17%, 7.25%로 예측됨을 보이고 있다.

표 4. BAGA 실험의 UA의 감정 분포별 인식율
Table 4. Recognition rate by UA emotion distribution of BAGA experiment

BAGA		Predicted emotion			
		Angry	Happy	Neutral	Sad
True Emotion	Angry	66.77	1.81	24.17	7.25
	Happy	15.08	22.35	41.90	20.67
	Neutral	10.14	4.87	64.72	20.27
	Sad	4.60	1.23	27.30	66.87

표 4에서 평가 감정 행복은 재현율은 22.35%로 낮게 예측되었다. 이는 IEMOCAP 데이터 셋의 감정 클래스 분포가 균형화되어 있지 않기 때문이다. 즉 개체수가 적은 행복 감정은 학습 횟수가 상대적으로 작아서 학습율이 저조하기 때문이다.

본 연구에서는 GAN의 기능으로 감정 인식 모델의 성능을 개선할 수 있음을 보이는 것이 목적이고 연구 결과의 객관성을 확보하기 위하여 기본으로 제공되는 데이터 셋을 수정 없이 사용하였다.

본 연구에서 GAN으로 부정 감정 데이터를 생성하여 부정데이터의 학습을 추가함으로써 실제데이터만을 이용하는 학습에 비하여 음성 감정 인식 모델의 성능이 개선됨을 확인하였다.

V. 결론 및 향후 과제

본 연구에서는 GAN의 기능으로 감정 인식 모델의 성능을 개선할 수 있음을 보이는 것이 목적이고 연구 결과의 객관성을 확보하기 위하여 IEMOCAP 데이터베이스에서 제공하는 기본 데이터 셋을 사용하였다. 제안된 모델은 BLSTM과 Attention, Dense 레이어로 구성되는 판별기를 포함하는 GAN의 아키텍쳐이다. RealDateSet은 4개의 감정으로 된 IEMOCAP 데이터 셋이고, 레이블은 4개의 원핫인코딩 벡터이다. FakeDataSet은 생성기가 만들어낸 데이터 셋이고 레이블은 부정 레이블을 설계하였다. 데이터 셋의 특징은 피치, 에너지, 제로 크로싱 속도, 발음 확률, 12차 MFCC와 이 값들은 1차 미분으로 구성되는 32차 벡터이다.

실험 결과 제안 모델은 61.45%의 인식율을 성취하였고 GAN을 사용하지 않은 모델에 비하여 1.86% 개선됨을 확인하였다. GAN으로 부정 감정 데이터

를 생성하여 부정데이터의 학습을 추가함으로써 실제 데이터만을 이용하는 학습에 비하여 음성 감정 인식 모델의 성능이 개선됨을 확인하였다.

향후 연구에는 행복의 낮은 인식율을 해결하기 위하여 네트워크상에서 행복에 대한 손실함수를 수정하여 인식기의 성능을 제고할 예정이다. 또한 감정별 데이터의 불균형을 해소하기 위해 데이터 균형을 조절하기 위한 방안을 마련하고자 한다.

References

[1] Automatic Speech Emotion Recognition Using Machine Learning, https://www.intechopen.com/online-first/automatic-speech-emotion-recognition-using-machine-learning. [accessed: Sep. 22, 2019]

[2] I. K. Hwang and H. B. Song, "AI-based Infant State Recognition Using Crying Sound", Journal of KIIT, Vol. 17, No. 7, pp. 13-21, Jul. 2019.

[3] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies", Rubber Chemistry & Technology, Vol. 24, pp. 638-639, 2008.

[4] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention", 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, pp. 2227-2231, Mar. 2017.

[5] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-Based Dense LSTM for Speech Emotion Recognition", IEICE Transactions on Information and Systems, Vol. E102–D, No. 7, pp. 1426-1229, Jul. 2019.

[6] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms", Proc. Interspeech, Stockholm, Sweden, pp. 1089-1093, Aug. 2017.

[7] F.Bao, "Improving Speech Emotion Recognition via Generative Adversarial Networks", University of Stuttgart, May 2019.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", in Advances in neural information processing systems, pp. 2672-2680, Jun. 2014.

[9] C. Y. Park, Y. S. Choi, and K. J. Lee, "Evaluation of Sentimental Texts Automatically Generated by a Generative Adversarial Network", Korea Information Processing Society, Vol. 8, No. 6, pp. 257-264, Jun. 2019.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, Vol. 2, pp. 3104-3112, Dec. 2014.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv:1409.0473, Sep. 2014.

[12] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", Language resources and evaluation, Vol. 42, No. 4, pp. 335-359, Dec. 2008.

[13] C. W. Huang and S. S Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition", INTERSPEECH 2016, San Francisco, USA, pp. 1387-1391. Sep. 2016.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor", Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain, pp. 835-838, May 2013.

저자소개

고 유 정 (You-Jung Ko)

2004년 2월 : 한밭대학교 컴퓨터공학과(공학석사)
2009년 2월 : 한밭대학교 컴퓨터공학과(공학박사)
2018년 3월 ~ 현재 : 한국교원대학교 컴퓨터교육학과 석사과정
2005년 3월 ~ 현재 : 한밭대학교 강사
관심분야 : 딥러닝, 음성감정인식, 컴퓨팅 사고력

김 윤 중 (Yoon-Joong Kim)

1981년 2월 : 충남대학교 전자공학과(공학사)
1983년 2월 : 충남대학교 컴퓨터공학과(공학석사)
1999년 2월 : 충남대학교 퓨터공학과(공학박사)
1984년 3월 ~ 현재 : 한밭대학교 컴퓨터공학과 교수
관심분야 : 딥러닝, 인공지능, 음성 인식