한국음향학회지 제40권 제5호 pp. 515~522 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.515

pISSN : 1225-4428
eISSN : 2287-3775

음성감정인식 성능 향상을 위한 트랜스포머 기반 전이학습 및 다중작업학습

Transformer-based transfer learning and multi-task learning for improving the performance of speech emotion recognition

박순찬,[1] 김형순[1†]

(Sunchan Park[1] and Hyung Soon Kim[1†])

[1]부산대학교 전자공학과

(Received July 16, 2021; accepted August 25, 2021)

초 록: 음성감정인식을 위한 훈련 데이터는 감정 레이블링의 어려움으로 인해 충분히 확보하기 어렵다. 본 논문에서는 음성감정인식의 성능 개선을 위해 트랜스포머 기반 모델에 대규모 음성인식용 훈련 데이터를 통한 전이학습을 적용한다. 또한 음성인식과의 다중작업학습을 통해 별도의 디코딩 없이 문맥 정보를 활용하는 방법을 제안한다. IEMOCAP 데이터 셋을 이용한 음성감정인식 실험을 통해, 가중정확도 70.6 % 및 비가중정확도 71.6 %를 달성하여, 제안된 방법이 음성감정인식 성능 향상에 효과가 있음을 보여준다.
핵심용어: 음성감정인식, 트랜스포머, 전이학습, 다중작업학습

ABSTRACT: It is hard to prepare sufficient training data for speech emotion recognition due to the difficulty of emotion labeling. In this paper, we apply transfer learning with large-scale training data for speech recognition on a transformer-based model to improve the performance of speech emotion recognition. In addition, we propose a method to utilize context information without decoding by multi-task learning with speech recognition. According to the speech emotion recognition experiments using the IEMOCAP dataset, our model achieves a weighted accuracy of 70.6 % and an unweighted accuracy of 71.6 %, which shows that the proposed method is effective in improving the performance of speech emotion recognition.
Keywords: Speech emotion recognition, Transformer, Transfer learning, Multi-task learning

PACS numbers: 43.72.Bs, 43.72.Ne

I. 서 론

음성감정인식은 주어진 사람의 음성으로부터 화자의 감정 상태를 추정하는 기술로, 고객 응대, 인공지능 비서, 헬스케어 등 다양한 분야에서 유용하게 사용될 수 있다. 그러나 음성인식, 화자인식 등 다른 음성신호기반 기술들이 심층신경망의 발전과 함께 비약적인 발전을 이룬 것에 비해 음성감정인식에는 여전히 어려운 문제가 많이 남아 있는데, 훈련을 위한 감정 레이블링된 음성 데이터의 부족이 주요 원인으로 여겨진다. 음성인식을 위한 텍스트 레이블이나 화자인식을 위한 화자 레이블과 달리 감정 레이블은 평가자의 주관적 기준에 따라 큰 차이가 발생할 수 있어 전문가의 판단을 필요로 한다. 또한 자연스러운 감정 음성은 개인적인 내용을 담고 있는 경우가 많아 공개가 어렵고, 전문 성우의 연기를 통해

†**Corresponding author:** Hyung Soon Kim (kimhs@pusan.ac.kr)
Department of Electronics Engineering, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Republic of Korea
(Tel: 82-51-510-2452, Fax: 82-51-515-5190)

수집되는 경우가 많아 데이터를 충분히 확보하기 어렵다.

응용 분야에 따라 여러 형태의 감정인식 모델이 존재하나, 여기서는 입력된 문장 단위 발화를 사전에 정의된 감정 범주 중 한 가지로 분류하는 모델에 대해 살펴보도록 한다. 전통적인 방법은 프레임 에너지, Mel-Frequency Cepstral Coefficient(MFCC), 기본 주파수 등 저수준 특징 통계값을 입력으로 Support Vector Machine(SVM)과 같은 머신러닝 기반 분류 모델을 통해 구현되었다.[1] 그러나 심층 신경망의 발전과 함께 전통적 시스템의 여러 부분들을 신경망 기반 모델로 대체하는 연구들이 진행되었다. 먼저 기존의 전통적 머신러닝 기반 분류 모델을 심층 신경망으로 대체하는 시도가 성공적인 결과를 보여주었다.[2] 또한 여러 저수준 특징의 조합 대신 스펙트럼 특징을 단일 입력으로 하고, Convolutional Neural Networks(CNN), Recurrent Neural Network(RNN) 등의 모델을 통해 신경망 내부에서 특징을 추출하는 방법들이 연구되었다.[3] 마지막으로 발화를 구성하는 프레임 사이 단순 통계값 대신 어텐션 메커니즘을 통해 감정인식에 유용한 프레임에 더 높은 가중치를 부여한 뒤 최종 분류를 수행하는 방법들도 제안되었다.[4]

신경망 기반 모델이 음성감정인식에 성공적으로 적용되면서, 부족한 훈련 데이터를 보완하기 위한 방법들이 함께 연구되었다. 먼저 모델이 감정 범주와 성별, 자연스러움 등의 보조 레이블을 함께 예측하도록 하는 다중작업학습 방법이 시도되었다.[5] 이와 같은 연구에서는 보조 과제를 통해 모델이 보다 다양한 특징을 추출하고, 특정 방향으로 과적합 되는 것을 방지하고자 하였다. 한편 음성 신호 뿐만 아니라 그에 해당하는 텍스트를 함께 입력으로 사용하는 바이모달 접근법도 많이 연구되었다.[6] 별도의 데이터 수집이 필요한 다른 데이터와 달리, 텍스트 데이터는 주어진 음성으로부터 자동 음성 인식(Automatic Speech Recognition, ASR)을 통해 얻을 수 있다. 사람의 감정의 많은 부분이 음성 신호 뿐만 아니라, 어휘나 문맥으로부터 추정할 수 있기 때문에 텍스트 입력을 이용한 시도는 성공적인 결과를 보여주었다. 그러나 음성인식 오류가 발생하는 경우 감정인식 결과에 부정적인 영향을 미치는 한계가 존재한다. 최근 연구에서는 감정인식이 아닌 다른 작업을 통해 모델을 사전훈련하고, 음성감정인식에 대해 미세조정하는 전이학습 방법을 통해 모델 훈련에 사용되는 전체 데이터의 양을 늘리는 접근이 시도되고 있다.[7]

한편 자연어 처리, 음성 인식 등 여러 분야에서 트랜스포머[8] 모델과 전이학습을 활용한 연구가 뛰어난 성과를 보여주었다.[9,10] 특히 자연어 처리 분야에서는 Bidirectional Encoder Representations from Transformers (BERT)[9]로 대표되는 언어 모델을 통한 사전훈련 방식이 널리 활용되고 있다. BERT는 입력 시퀀스에서 마스킹 된 토큰을 추정하는 언어 모델과 두 문장의 연결이 자연스러운지 판단하는 두 가지 과제를 동시에 해결하도록 사전훈련된다. 두 문제를 동시에 풀기 위해 훈련 가능한 파라미터로 구성된 임베딩 벡터를 추가하고 어텐션 메커니즘을 통해 시퀀스 전체의 특성을 모으도록 한다. BERT는 증강된 출력을 각각의 문제에 대해 분리하여 시퀀스 전체를 다루는 문제와 각각의 시점에 관한 문제를 동시에 다룰 수 있게 하였다.

본 논문에서는 트랜스포머 모델을 기반으로 전이학습과 다중작업학습을 함께 적용하여 음성감정인식 성능을 향상시키는 방법을 제안한다. 우선 트랜스포머 모델을 종단간 음성 인식으로 사전훈련하고, 감정인식을 통해 미세조정하는 방법으로 전체 훈련 데이터의 수를 늘린다. 미세조정 시 BERT와 유사하게 임베딩 벡터를 도입하여 감정인식을 수행하도록 하여, 감정인식과 음성인식을 동시에 수행할 수 있는 구조로 만든다. 다중작업학습을 통해 모델을 미세조정하면 별도의 디코딩 없이 모델 내부에서 문맥 정보가 반영되는 것을 기대할 수 있다. 제안한 모델의 감정인식 훈련 및 성능 평가를 위해 IEMOCAP,[11] 종단간 음성인식 사전훈련을 위해 LibriSpeech[12] 데이터 셋을 사용하였다. 감정인식 성능 평가 결과 감정인식 가중정확도 70.6 % 및 비가중정확도 71.6 %를 달성하여, 종단간 음성인식을 이용한 전이학습 및 다중작업학습 방법이 감정인식 모델의 성능을 개선하는 것을 확인하였다.

II. 신경망 구조

2.1 트랜스포머(Transformer)

트랜스포머는 어텐션 메커니즘을 통해 순차적 데이터를 병렬로 처리할 수 있도록 설계된 신경망 모델이다. RNN처럼 각 시점의 출력을 얻기 위해 이전 단계의 출력을 이용하는 대신, 전체 입력 시퀀스로부터 현재 시점의 입력과 관련된 정보를 수집한다. Fig. 1은 트랜스포머의 구조를 나타내는데, 그림의 좌측은 인코더, 우측은 디코더를 나타낸다. 본 연구에서는 디코더는 제외하고 인코더로만 구성된 트랜스포머를 사용하였다. 트랜스포머 인코더는 동일한 구조의 블록을 쌓아서 만들어진다. 트랜스포머 인코더 블록은 멀티 헤드 어텐션 레이어와 순방향 레이어의 순차적인 연결로 구성된다. 각 레이어의 입력과 출력 사이에는 잔류 연결이 적용되는데, 즉 레이어의 출력 시퀀스에 입력 시퀀스가 더해지고 여기에 레이어 정규화를 거친 결과가 다음 레이어로 전달된다. 잔류 연결을 활용하기 위해 각 레이어의 입력과 출력 시퀀스의 벡터는 모두 동일한 차원으로 설정된다.

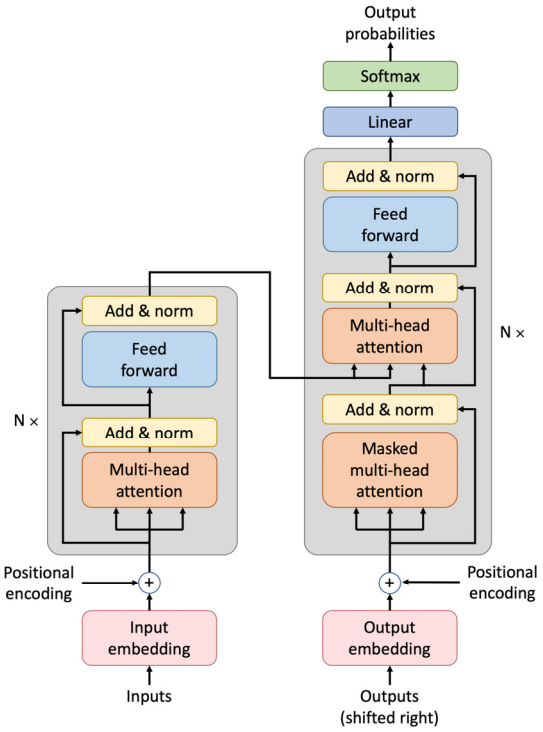

Fig. 1. (Color available online) The transformer model architecture.[8]

멀티 헤드 어텐션 레이어는 전체 입력 시퀀스로부터 현재 시점의 데이터와 연관성이 높은 정보를 모으는 트랜스포머의 핵심 역할을 수행한다. 우선 각 입력 벡터는 투사 레이어를 통해 쿼리, 키, 그리고 밸류라 불리는 벡터들로 변환되는데, 각각의 벡터는 지정된 헤드의 수만큼 만들어진다. 각 헤드에 대해 투사 레이어가 독립적으로 존재하므로 모든 쿼리, 키 밸류는 서로 다른 벡터로 변환된다. 어텐션 메커니즘은 각 헤드마다 쿼리, 키, 밸류를 통해 독립적으로 수행된다. 트랜스포머 인코더의 어텐션 매커니즘은 해당 시점의 쿼리와 시퀀스의 모든 키, 밸류 벡터를 통해 다음과 같은 방법으로 출력을 결정한다.

$$s_{ij} = \frac{q_i^T k_j}{\sqrt{d_k}}, \tag{1}$$

$$w_{ij} = \frac{\exp(s_{ij})}{\sum_j \exp(s_{ij})}, \tag{2}$$

$$a_i = \sum_j w_{ij} v_j. \tag{3}$$

이때 q_i, k_i, v_i는 각각 i 시점의 쿼리, 키, 밸류 벡터이며, d_k는 이 벡터들의 차원을 나타낸다. 이를 통해 각 헤드에서 독립적으로 입력 시퀀스의 모든 시점 i에 대한 a_i를 구한다. 멀티 헤드 어텐션 레이어의 최종 출력은 각 시점에서 모든 헤드의 출력을 연결하는 것으로 완성된다. 이때 얻어지는 출력과 입력의 차원이 동일하도록 쿼리, 키, 밸류의 차원이 정해진다.

멀티 헤드 어텐션 레이어는 입력 시퀀스의 각 시점 간 상호 작용을 통해 출력을 구하는 반면, 순방향 레이어는 각 시점의 벡터에 대해 독립적으로 연산을 수행한다. 입력 시퀀스의 모든 벡터는 동일한 선형 레이어에 의해 보다 높은 차원의 벡터로 변환된 후 Rectified Linear Unit(ReLU) 활성함수가 적용되고, 또 다른 선형 레이어에 의해 입력과 동일한 차원의 벡터로 변형되어 다음 레이어로 전달된다.

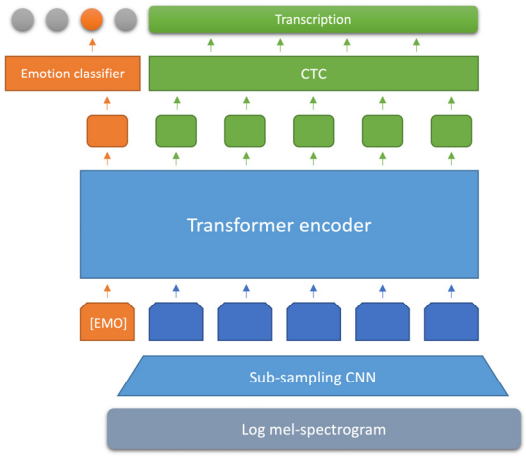

Fig. 2. (Color available online) The proposed model architecture.

2.2 모델 구조

Fig. 2는 제안된 모델의 구조를 나타내며, 입력 특징으로 로그 멜-스펙트로그램을 사용한다. 입력된 특징은 우선 CNN을 이용한 서브샘플링 과정을 거치게 된다. 서브샘플링 CNN을 구성하는 합성곱 레이어들은 모든 영역에서 연산을 수행하는 대신, 합성곱 연산 사이 일정한 간격을 두어 출력의 크기를 축소하도록 만든다. 이를 통해 입력 특징의 인접한 정보를 효율적으로 표현하고, 시퀀스의 길이를 축소하여 트랜스포머에 요구되는 메모리 및 연산량을 줄일 수 있다.

서브샘플링 CNN의 출력은 투사 레이어를 통해 정해진 입력 차원의 벡터 시퀀스로 변형된 후, 위치 임베딩과 더해져 트랜스포머 인코더로 전달된다. 트랜스포머의 멀티 헤드 어텐션 레이어에는 어떠한 마스크도 적용 되지 않아 모든 시점에서 입력 시퀀스전체를 참조할 수 있도록 설계되었다. 마지막으로 트랜스포머의 출력 시퀀스를 선형 레이어와 소프트맥스 레이어의 순차적 연결로 만들어진 분류 레이어로 전달하여 음성인식 및 감정인식을 수행한다.

III. 신경망 훈련 방법

3.1 사전훈련

본 연구에서 제안하는 사전훈련 과정은 종단간 음성인식 모델을 훈련하는 과정으로 볼 수 있다. 인코더-디코더 모델,[13] RNN Transducer(RNN-T),[14] Connectionist Temporal Classification(CTC)[15] 등 여러 종단간 음성인식 모델 구조 중, 여기서는 CTC를 사용하여 사전훈련을 진행하였다. CTC는 공백 레이블을 도입하여 레이블 시퀀스의 길이를 모델 출력 시퀀스와 동일하게 만든다. 이때 모든 가능한 공백 레이블 배치를 고려하여, 모델의 출력과 레이블 시퀀스 사이 가능한 모든 정렬에 대해 손실 함수를 최소화 하도록 훈련한다. CTC를 통해 별도의 디코더 네트워크 없이 인코더 네트워크만으로 ASR 모델 훈련이 가능한데, 이는 사전훈련과 미세조정 사이 공유하는 파라미터 수를 최대화하여 효율적인 전이학습이 가능하도록 한다.

CTC는 출력 시퀀스가 각 시점 사이 독립이라는 가정 하에 동작한다. 즉 특정 시점의 출력을 전후 시점의 출력과는 무관하게 입력 시퀀스만을 참조하여 결정한다. 이러한 특징 때문에 종단간 음성인식 모델 중 CTC는 외부 언어 모델이나 복잡한 디코딩 알고리즘이 없이는 다른 방법들에 비해 낮은 성능을 보인다. 그러나 여러 연구를 통해 인코더-디코더, RNN-T 등 다른 종단간 음성인식 모델에서 CTC를 인코더의 보조 손실 함수로 사용하는 경우 성능 향상에 도움이 되는 것이 확인되었다.[16] 이러한 연구 결과들을 바탕으로 CTC를 통한 훈련이 감정인식 성능 개선에도 도움이 될 것이라는 가정을 통해 사전훈련 방식으로 채택하였다.

한편 종단간 음성인식 모델 훈련을 위한 텍스트 레이블은 Byte Pair Encoding(BPE)를 통해 서브워드 단위로 분절하여 훈련에 사용하였다. 종단간 음성인식 모델은 일반적으로 문자 또는 서브워드를 레이블의 단위로 사용한다. 서브워드를 사용할 경우 분류의 범주가 증가함에 따라 요구되는 메모리 및 연산량이 증가하나, 빠르게 발음되며 빈번하게 사용되는 연속적인 문자열을 효율적으로 처리할 수 있는 장점이 있다. 또한 모델이 빈번하게 사용되는 일부 단어들을 직접 예측함을 통해, 디코딩 없이도 어휘에 대한 정보가 감정인식에 보다 직접적으로 활용되는 것을 기대할 수 있다.

3.2 미세조정

미세조정 단계에서는 감정 정보 수집을 위한 특수 임베딩 벡터가 추가된다. 임베딩 벡터의 값은 신경망을 통해 훈련 가능한 파라미터로 구성된다. 입력 음성 특징이 여러 단계를 거쳐 트랜스포머의 입력으로 전달되기 전, 입력 시퀀스의 가장 앞에 임베딩 벡터가 연결된다. 사전훈련과 미세조정 단계에서 트랜스포머의 입출력을 다음과 같이 정리 할 수 있다.

$$\mathbf{x}=[x_1,x_2,\cdots,x_N], \tag{4}$$

$$\tilde{\mathbf{x}}=[x_e,x_1,x_2,\cdots,x_N], \tag{5}$$

$$\mathbf{z}=[z_1,z_2,\cdots,z_N], \tag{6}$$

$$\tilde{\mathbf{z}}=[z_e,z_1,z_2,\cdots,z_N]. \tag{7}$$

$\mathbf{x}$, $\mathbf{z}$는 사전훈련 단계, $\tilde{\mathbf{x}}$, $\tilde{\mathbf{z}}$는 미세조정 단계에서 각각 트랜스포머의 입력 및 출력 시퀀스를 나타낸다. x_i는 i 시점의 입력 특징, z_i는 i 시점의 출력 특징이며, x_e는 감정인식을 위한 임베딩 벡터, z_e는 그에 대응하는 출력 벡터를 의미한다.

설계 의도에 맞게 모델을 훈련하기 위해, 트랜스포머 인코더의 출력 시퀀스 $\tilde{\mathbf{z}}$를 z_e와 $\mathbf{z}$로 분리한다. 이중 z_e는 분류 레이어로 전달되어 감정 분류를 수행하게 된다. 감정 분류는 일반적인 교차 엔트로피 손실함수를 통해 훈련된다.

다중작업학습을 위한 손실함수 L은 감정 분류를 위한 교차 엔트로피 L_{ce}와 종단간 ASR 훈련을 위한 CTC 손실함수 L_{ctc}의 가중합으로 나타낼 수 있다.

$$L=L_{ce}(\hat{y_e},y_e)+\alpha L_{ctc}(\hat{\mathbf{y}},\mathbf{y}). \tag{8}$$

이때 $\hat{y_e}$은 감정인식을 위한 분류 레이어를 통과한 z_e, $\hat{\mathbf{y}}$는 ASR을 위한 분류 레이어를 통과한 $\mathbf{z}$를 나타내며, y_e와 $\mathbf{y}$는 각각 감정 레이블과 서브워드 단위로 분절된 ASR 레이블을 의미한다. 두 손실 함수는 하이퍼 파라미터 α를 통해 균형이 조절된다.

IV. 실험 환경 및 결과

4.1 실험 데이터

사전훈련 단계에서 ASR 훈련을 위해 낭독체 영어 음성으로 구성된 LibriSpeech 데이터셋의 전체 훈련 데이터 약 970 h 분량을 모두 사용하였다. 음성 데이터로부터 80차 로그 멜-스펙트로그램을 25 ms의 윈도우 크기, 10 ms의 윈도우 간격으로 추출하고, 발화 단위 평균 정규화를 적용하였다. SentencePiece[17]와 훈련 데이터의 전사 텍스트를 이용하여 5000개의 서브워드로 구성된 BPE 모델을 만들고, 이를 통해 레이블을 생성하였다.

감정인식 모델 훈련 및 평가를 위해 IEMOCAP 데이터셋을 사용하였다. 전체 데이터 중 기존 연구의 방법에 따라 기쁨, 흥분, 슬픔, 분노, 중립의 다섯 범주로 분류된 5531 발화를 훈련 및 평가에 사용하였고, 이 중 흥분에 해당하는 데이터를 기쁨 범주에 포함시켜 데이터를 네 개의 감정 범주로 정리하였다. 훈련 데이터의 부족을 보완하기 위해 모든 실험은 5겹 교차검증을 통해 진행하였다. 음성 특징은 사전훈련과 동일한 방법으로 추출하였고, 서브워드 단위의 텍스트 레이블은 사전훈련 단계에서 생성된 BPE 모델을 통해 생성하였다.

4.2 실험 설정

제안한 모델은 크게 서브샘플링 CNN, 트랜스포머 인코더, 둘 사이를 연결하는 투영 레이어, 그리고 출력 생성하기 위한 분류 레이어로 구성된다. 우선 서브샘플링 CNN은 출력 채널의 수가 512인 두 개의 합성곱 신경망으로 구성되는데, 각각의 필터 크기가 3, 5, 합성곱 사이 간격은 2, 3으로, 시간 및 주파수 축에 대해 1/6로 크기가 축소된다. 투영 레이어는 서브샘플링 CNN의 결과를 트랜스포머 입력 차원인 512로 변환한다. 트랜스포머 인코더는 18개 블록의 연결로 만들어진다. 각 블록은 8개 헤드를 가지며, 키, 쿼리, 밸류 벡터의 차원이 64인 멀티 헤드 어텐션 레이어와 은닉층의 차원이 2048인 순방향 레이어로 구성된다. IEMOCAP 데이터에 대해 4개 범주의 감정이 정의되었으므로, 감정인식을 위한 분류 레이어는 512

차원의 벡터를 입력으로, 4차원의 벡터를 출력으로 한다. BPE 모델은 5000개의 서브워드로 정의되었으므로, ASR을 위한 분류 레이어는 512차원의 벡터를 입력으로, 5000차원의 벡터를 출력으로 한다.

성능 비교를 위해 CNN-BLSTM-SA 기반 베이스라인 모델을 훈련 및 평가하였다. 모델 구조는 크게 CNN, Bidirectional Long Short-term Memory(BLSTM), Self Attention(SA), 그리고 분류 레이어로 구성된다. CNN은 128채널의 합성곱 신경망과 크기 3의 맥스 풀링 레이어가 2번 반복되는 구조이다. 그 결과는 128차원의 은닉 차원으로 구성된 BLSTM 레이어 2개로 전달된다. BLSTM 레이어의 출력은 8개의 헤드로 구성된 셀프 어텐션 모듈을 통해 단일 벡터로 변환되고, 분류 레이어로 전달되어 감정 분류를 수행한다. 모델의 입력은 80차 로그 멜-스펙트로그램으로 트랜스포머 모델과 동일한 방법으로 추출되었고, 훈련을 위해 교차 엔트로피 손실함수를 사용하였다.

모든 신경망 모델은 PyTorch[18]를 통해 구현되었으며, 훈련을 위해 Adam 옵티마이저를 사용하였다. 트랜스포머 모델의 사전훈련은 학습률 3e-4, 가중치 감쇠율 1e-6, 배치 크기 64로 전체 훈련 데이터를 25회 반복 훈련하였다. 트랜스포머 모델의 미세조정은 학습률 1e-4, 가중치 감쇠율 1e-6, 배치 크기 32로 각 훈련 데이터를 20회 반복 훈련하였다. CNN-BLSTM-SA 모델은 학습률 1e-4, 가중치 감쇠율 1e-6, 배치 크기 40으로 각 훈련 데이터를 20회 반복 훈련하였다. 다중작업학습 적용 시 두 손실 함수의 균형을 위한 파라미터 alpha의 값은 실험을 통해 0.6으로 결정하였다.

모델의 감정인식 성능은 가중정확도, 비가중정확도의 두 가지 지표로 평가하였다. 가중정확도는 모든 데이터에 대해 정확히 분류한 비율을 계산하는 반면, 비가중정확도의 경우 각 범주의 데이터에 대한 정확도를 각각 계산한 뒤 이를 평균하여 구한다. IEMOCAP 데이터는 각 감정 범주의 데이터 양에 불균형이 존재하므로, 특정 범주에 치우쳐서 분류하지 않는지 확인하고자 두 지표에 대한 성능을 모두 살펴보았다.

Table 1. Experimental results on IEMOCAP dataset. Weigted Accuracy (WA), UA (Unweighted Accuracy (UA), Transfer Learning (TL), Multi-Task Learning (MTL).

Model	WA (%)	UA (%)
CNN-BLSTM-SA	58.1	59.8
Transformer	55.2	55.5
Transformer (TL)	67.1	67.9
Transformer (TL+MTL)	70.6	71.6

4.3 실험 결과

Table 1에 IEMOCAP 데이터에 대한 감정인식 실험 결과를 정리하였다. 우선 트랜스포머의 경우 전이학습, 즉 사전학습 없이 감정인식 모델로 훈련하는 경우 베이스라인인 CNN-BLSTM-SA 모델보다 나쁜 결과를 보여주었다. 이는 약 10 h 분량의 IEMOCAP 데이터로는 트랜스포머 모델의 파라미터들을 모두 훈련하기에 부족하기 때문인 것으로 판단된다.

CTC 손실함수를 통해 ASR 사전훈련된 모델에 IEMOCAP 데이터로 미세조정한 경우, 트랜스포머 모델은 CNN-BLSTM-SA 모델의 성능 대비 가중정확도 및 비가중정확도 측면에서 각각 15.5 % 및 13.5 % 개선된 성능을 보여주었다. 이를 통해 ASR을 이용한 전이학습이 음성감정인식 성능 개선에 크게 도움이 되는 것을 확인할 수 있었다.

전이학습과 더불어 다중작업학습을 적용한 경우 트랜스포머는 전이학습만 적용된 트랜스포머 모델 대비 가중정확도 및 비가중정확도 관점에서 각각 5.2 % 및 5.4 % 향상된 성능을 보여주었다. 이와 같은 결과를 통해 전사된 텍스트를 직접 입력으로 사용하지 않고, 다중작업학습을 통해 간접적으로 학습된 문맥 정보만으로도 감정인식 성능을 향상시킬 수 있음을 알 수 있었다.

V. 결 론

본 연구를 통해 트랜스포머 모델을 기반으로 전이학습 및 다중작업학습을 적용하였을 때, 두 방법 모두 적용되지 않은 경우 대비 평균 28.5 %의 감정인식 성능 개선 효과를 확인할 수 있었다. 수행한 실험에서는 서브워드 모델 생성 시 음성 인식을 위한 텍스

트만 사용하였으나, 감정 발화 텍스트를 포함시키면 감정에 관련된 어휘들이 서브워드에 포함되어 추가적인 성능 향상 효과를 기대할 수 있다. 제안한 방법은 별도의 디코더가 필요 없는 단순한 모델 구조로 전이학습을 적용할 수 있기 때문에, 동일한 모델을 통해 다양한 사전학습 방법이 감정인식 성능에 미치는 영향을 살펴볼 수 있을 것으로 기대된다. 추후 연구를 통해 음성 신호를 입력으로 하는 여러 사전학습 방법과 비교하여, 제안한 모델이 문맥에 대한 정보를 어떻게 활용하는지 면밀히 살펴보고자 한다.

감사의 글

이 논문은 2019년 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A5A2A03045884)

References

1. H. Hu, M. Xu, and W. Wu, “GMM supervector based SVM with spectral features for speech emotion recognition,” Proc. ICASSP. 413-416 (2007).
2. A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” Proc. ICASSP. 5688-5691 (2011).
3. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” Proc. ICASSP. 5200-5204 (2016).
4. S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” Proc. ICASSP. 2227-2231 (2017).
5. J. Kim, G. Englebienne, K. P. Truong, and V. Eversu, “Towards speech emotion recognition “in the Wild” using aggregated corpora and deep multi-task learning,” Proc. Interspeech, 1113-1117 (2017).
6. S. Yoon, S. Byun, and K. Jung, “Multimodal speech emotion recognition using audio and text,” Proc. SLT. 112-118 (2018).
7. Z. Lu, L. Cao, Y. Zhang, C. Chiu, and J. Fan, “Speech sentiment analysis via pre-trained features from end-to-end ASR models,” Proc. ICASSP. 7149-7153 (2020).
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł.Kaiser, “Attention is all you need,” Proc. NIPS. 6000-6010 (2017).
9. J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. NAACL-HLT. 4171-4186 (2019).
10. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” Proc. NeurIPS. 12449-12460 (2020).
11. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S.l Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources and Evaluation, **42**, 335-359 (2008).
12. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” Proc. ICASSP. 5206-5210 (2015).
13. W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” Proc. ICASSP. 4960-4964 (2016).
14. A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” Proc. ICASSP. 6645-6649 (2013).
15. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” Proc. ICML. 369-376 (2006).
16. S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” IEEE JSTSP. **11**, 1240-1253 (2017).
17. T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” Proc. EMNLP 66-71 (2018).
18. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” Proc. NeurIPS. 8024-8035 (2019).

저자 약력

▸ 박 순 찬 (Sunchan Park)

2016년 2월 : 부산대학교 전자공학과 학사
2018년 2월 : 부산대학교 전기전자컴퓨터 공학과 석사
2018년 1월 ~ 2019년 6월 : LG전자 연구원
2019년 9월 ~ 현재 : 부산대학교 전기전자 공학과 박사과정

▸ 김 형 순 (Hyung Soon Kim)

1983년 2월 : 서울대학교 전자공학과 학사
1989년 2월 : 한국과학기술원 전기및전자 공학과 박사
1987년 1월 ~ 1992년 6월 : ㈜디지콤 선임 연구원
1992년 7월 ~ 현재 : 부산대학교 전자공학 과 교수