공학 석사 학위논문

패턴학습블록 기반
음성감정인식 경량모델

전남대학교 대학원

인공지능융합학과

임 현 택

2024 년 2 월

공학 석사 학위논문

패턴학습블록 기반
음성감정인식 경량모델

전남대학교 대학원

인공지능융합학과

임 현 택

2024 년 2 월

패턴학습블록 기반
음성감정인식 경량모델

이 논문을 공학 석사학위 논문으로 제출함

전남대학교 대학원
인공지능융합학과
임 현 택

지도교수 양 형 정

임지영의 공학 석사 학위논문을 인준함

심사위원장 정 희 용 (인)

심 사 위 원 유 석 봉 (인)

심 사 위 원 양 형 정 (인)

2024 년 2 월

목 차

그림 목차

표 목차

약어

PLB	Pattern Learning Block
EMO-DB	The Berlin Emotional Database
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
SER	Speech Emotion Recognition
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
HCI	Human Computer Interaction
DCNN	Dilated Convolution
DWSCNN	Depth-wise Separable Convolution
FCN	Fully Connected Network
DCC	Dilated Causal Convolution
CBAM	Convolutional Block Attention Module
MLP	Multi-Layer Perceptron
GAP	Global Average Pooling
MFCC	Mel-Frequency Cepstral Coefficient

패턴 학습 블록 기반 음성감정인식 경량 모델

임 현 택

전남대학교 대학원 인공지능융합학과

(지도교수: 양형정)

(초록)

음성감정인식은 음성 신호에서 인간의 감정과 정서적 상태를 추론하여 인간과 기계 간의 상호작용을 개선하는 역할을 한다. 현재 음성감정인식이 직면한 주요 과제 중 하나는 실시간 및 가벼운 모델을 필요로 하는 영역에서 응용이 무시되고, 실시간 응용프로그램에서 처리 능력이 부족하다는 것이다. 또한, 고성능 모델 대부분은 학습 가능한 많은 매개변수에 의존한다.

본 연구에서는 서로 다른 시점의 감정 패턴을 분석하는 Pattern Learning Block(PLB)를 제안한다. PLB 를 통해 여러 시점의 감정 정보와 상황 특성을 고려하여 정방향 및 역방향 시간 흐름의 감정 패턴을 학습한다. 또한, 음성의 채널 및 공간 신호에서 발생하는 풍부한 정보를 지속 활용하여 표현을 풍부하게 하였다.

약 95,000 개의 매개변수로 구성된 제안 모델은 세가지 음성감정 데이터셋(EMODB, RAVDESS, IEMOCAP)의 전반적인 실험에서 매개변수 감소와

정확도 향상을 보여주며 최신 성능을 달성하였다. 추가로 제안한 모델의 구성 요소가 성능에 미치는 영향을 평가하기 위한 실험, 메모리가 제한된 환경에서의 추론 속도를 관측하였다. 경량화 된 모델은 적은 계산 비용과 빠른 추론 속도로 실시간 처리 능력 또한 확보하여 음성 기반 챗봇, 기분 모니터링 등 실시간 시스템에 활용 가능하다. 추후 제안 모델이 실시간 및 하드웨어 제약이 있는 영역에서 감정인식에 대한 이식성을 높일 수 있을 것으로 기대한다.

1. 서론

가. 연구 배경

음성감정인식(SER: Speech Emotion Recognition)은 인간의 감정을 이해하고 분석하는 수단으로 주목을 받고 있다. 인간의 감정 특성을 분류하긴 어렵지만, 인간의 감정을 기계가 이해할 수 있도록 학습시키는 연구는 빠르게 발전되고 있다. 대화에서 감정을 정확하게 식별하는 능력은 콜센터, 우울증 및 고통 분석 시스템, 커넥티드 카 등 다양한 응용 분야가 있다[1].

음성에 포함된 감정을 인식하기 위해 수작업으로 수집된 다양한 특징들이 연구에 적용되었다. 이러한 특징의 정확도는 상대적으로 높지만, 전문적인 지식과 비용이 많이 소모된다[2]. 이를 완화하기 위해 최근 낮은 수준의 특징에서 높은 수준의 특징을 추출하는 딥러닝이 도입되고 있다[3].

최근 많은 딥러닝 기반 SER 모델이 제안되었다. 주파수 영역의 특징 추출을 위해 합성곱신경망(CNN: Convolution Neural Network), RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory), 트랜스포머(Transformer)를 조합한 모델로 합리적인 결과를 달성하였다[4-6]. 그러나 고성능 모델 대부분은 학습 가능한 많은 매개변수에 의존하므로, 실시간 및 가벼운 모델을 필요로 하는 HCI(Human Computer Interaction) 영역에서 사용이 제한된다[6].

많은 매개변수는 일반적인 작업에 딥러닝 모델을 활용하는 데 어려움을

겪는다. 딥러닝 모델의 깊이가 커짐에 따라 매개변수 수는 증가하고, 과적합을 발생시킨다[7]. 즉, 모델의 크기는 복잡성을 결정하는 주요 요인이다. 또한, RNN 의 순차적인 특성은 메모리 제약이 있는 환경에 적용하기 어렵고[8], 합성곱신경망은 풀링(Pooling)으로 인해 음성의 시간적 구조가 점차 손실된다[9]. 또한, 경량화 된 단일 구조는 더 나은 일반화를 위해 음성 신호에서 사용할 수 있는 채널 및 공간 신호에 대한 풍부한 정보가 무시된다[10].

나. 연구 목적 및 내용

본 논문에서는 서로 다른 시점의 감정 패턴을 학습하는 Pattern Learning Block(PLB)를 제안한다. 먼저, 시간의 정방향 및 역방향 흐름의 감정 정보를 추출하고, 채널 및 공간 신호 정보를 부각하기 위해 Convolutional Block Attention Module(CBAM)을 적용하였다. 이후, 다양한 시간적 위치에서의 특징 정보를 집계하기 위해 시간의 정방향 및 역방향 흐름을 통합하였다. 제안 모델에서는 음성 신호의 채널 및 공간 신호 보존을 위해 CBAM 을 반복 적용하였다.

제안 모델은 화자의 발음 속도, 정지 구간 등 개인의 특성을 반영하기 위해 확장된 합성곱(DCNN: Dilated Convolution)을 구성하여 시간 패턴에 서로 다른 크기의 필터를 적용하였다. 주의 기법은 특징의 차원이 큰 경우, 많은 계산

비용이 발생한다[11]. 깊이별 분리 합성곱(DWSCNN: Depth-wise Seperable Convolution)을 적용해 공간적, 채널적 특징을 모두 고려하여 매개변수 및 연산량을 감소하였다. 이후, 감정의 중요한 특성 표현을 캡슐에 담아 각 캡슐의 중요도를 동적으로 조정하였다.

제안 모델의 성능 평가를 위해 CNN 기반의 감정인식 모델과 성능 및 자원의 효율성을 비교하였다. 또한, 성능과 효율성의 Trade-off 를 평가하였다. 마지막으로 구성된 각 모듈이 성능에 미치는 영향과 자원이 제한된 환경에서의 추론 속도를 평가하기 위한 추가 실험을 수행한다.

다. 논문 구성

본 논문의 구성은 다음과 같다.

- 2 장에서는 CNN 기반의 음성감정인식과 경량화를 주제로 수행된 연구와 한계를 기술하고, 특징 추출에서의 합성곱, 주의 기법의 구조 및 역할을 설명한다.
- 3 장에서는 서로 다른 시점의 감정 패턴을 학습하는 제안 모델에 대해 서술한다.
- 4 장에서는 실험을 위한 환경, 데이터 및 전처리에 대한 설명과 수행한 실험결과에 대해 기술한다.

- 마지막으로 5 장에서는 결론 및 향후 연구에 대하여 서술한다.

2. 관련 연구 및 배경 지식

가. 합성곱신경망 기반의 음성감정인식

Nantasri[12]은 자원이 제한된 환경에서 메모리 및 계산 능력 한계에 주목했다. MFCC 의 델타(Delta) 및 델타-델타 계수와 연결된 MFCC 의 평균값을 3 개의 FCN(Fully Connected Network)으로 구성된 모델의 입력으로 사용하였다. 약 164,300 개의 매개변수를 가지는 모델은 EMODB, RAVDESS 데이터셋에서 각각 정확도 87.80%, 82.30%를 달성하였다. Atsavasirilert[13]은 컴퓨팅 및 메모리 자원이 제한된 환경에서 감정을 인식하는 연구를 진행하였다. 멜스펙트로그램(Mel-spectrogram)을 입력으로 AlexNet[14]의 일부 구성 요소만 사용하는 네트워크를 제안하였다. 220,000 개의 매개변수를 지닌 모델은 EMODB 데이터셋에서 85.54%의 정확도를 달성하였다. Tang[15]는 종단간(end-to-end) 음성기반 감정인식을 위해 DCC(Dilated Causal Convolution)를 사용해 매개변수를 크게 줄이면서 큰 수용 필드를 유지했다. DCC 와 최대풀링(max-pooling)의 조합으로 구성된 모델은 약 430,000 개의 매개변수로 IMEMOCAP 데이터셋에서 65.80%의 정확도를 달성하였다. 그러나 음성은 시간 영역에 걸쳐 다중 스케일 표현을 가지는 신호이므로 FCN 과 합성곱신경망 기반 단일 아키텍처의 조합[12, 13, 15]은 음성감정인식 모델에 적합하지 않다[16]. 따라서, 더 다양한 데이터셋에서 일반화 능력을 평가하는 연구가 추가로 필요하다.

나. RNN 기반의 음성감정인식

Bautista[17]는 병렬로 실행되는 네트워크의 조합을 사용하여 비병렬 하이브리드 모델에 비해 더 적은 매개변수를 유지하면서 성능은 높였다. 멜스펙트로그램 이미지를 이용해 시간적 특징 표현을 위한 트랜스포머 및 공간적 특징을 모델링 하기 위한 합성곱을 병렬 구성하였다. J.Pengxu[18]는 CNN 을 통해 학습된 공간 정보를 LSTM 에 입력하여 장기적인 정보를 학습하는 직렬 구조이다. 시간적 특징을 부여하기 위해 적용된 RNN 계열의 네트워크는 모델의 매개변수 대부분을 차지한다. 이러한 특성은 더 높은 계산 비용을 요구하고 실시간 처리 능력이 감소하게 된다.

다. 주의(Attention) 기법

합성곱신경망은 일반적으로 풀링(Pooling)과 함께 사용되며, 음성의 시간적 구조가 점차 손실된다[9]. 시계열 특성을 고려하여 입력 순서를 기억하고 맥락을 파악하는 RNN 계열의 네트워크가 적용되었다[5]. 그러나, 네트워크의 구조상 입력된 특징이 클수록 서로 멀리 떨어진 특징에 대한 정보가 줄어들면서 제대로 된 예측을 할 수 없게 되는 의존성 문제가 발생하게 된다. 또한, 순차적으로 연산을 진행하여 병렬화가 불가능하므로 연산 속도가 감소한다.

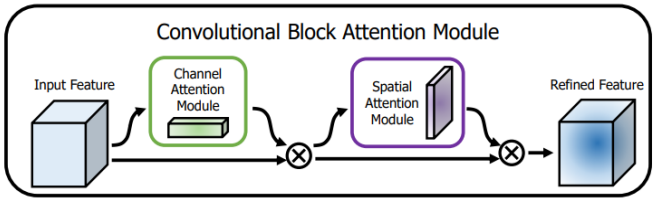
주의(Attention)[8] 기법은 다른 특징과의 관련성과 특징 벡터 간 중요하고 유사한 정보를 위한 문맥(Context) 벡터를 확보한다. 문맥 벡터는 은닉상태벡터

Q(Query)와 K(Key), V(Value)를 이용해 [식 1]과 같이 계산된다.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad [식\ 1]$$

QK^T 연산 직후 행렬의 크기와 벡터가 커지게 된다. 많은 수의 특징을 가진 벡터는 소프트맥스 함수를 적용하면 특징의 확률값 대부분이 0 에 가까워진다. 각 특징의 확률값이 0 에 가까운 것을 막기 위해 $\sqrt{d_k}$ 로 스케일링(Scaling) 한다. 시계열 데이터의 정확한 예측을 위해 입력 특징을 순차적으로 제공하고 현재 순서가 아닌 데이터의 정보를 숨기기 위해 마스킹을 사용한다. 정보의 모든 부분을 확인하기 때문에 RNN 보다 훨씬 먼 거리에 있는 특징을 추출할 수 있다. 또한 각 네트워크 구성 요소의 계산량이 줄어들고, 병렬 처리가 가능한 계산 영역이 많아진다[8].

CBAM(Convolutional Block Attention Module)[19]은 채널, 공간 정보에 대한 주의 맵(Attention map)과 입력 특징을 곱해 모델이 어디의 무엇에 집중해야 하는지에 대한 정보를 부각하며 적응적으로 특징을 개선한다. CBAM 은 채널 주의(Channel Attention)와 공간 주의(Spatial Attention)로 구성된다. [그림 1]은 CBAM 의 구성을 나타낸다.



[그림 1. Convolutional Block Attention Module(CBAM)]

채널 주의는 입력 특징 F의 채널 관계를 활용해 중요한 특징 정보를 담은 채널 주의 맵을 생성해 무엇이 중요한지 집중한다. 효율적인 계산을 위해 입력 특징의 공간 차원을 1×1로 압축한다(즉, $Channel\times1\times1$). 이후, 공간 정보 통합을 위해 평균풀링과 최대풀링을 적용하여 특징의 중요한 단서를 확보한다. 채널 어텐션 M_c는 [식 2]와 같이 계산된다.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \qquad [식\ 2]$$

풀링으로 인코딩 된 2 개의 특징을 각각 MLP(Multi Layer Perceptron)하여 비선형성을 추가하고 은닉층을 통해 수축시킨다. 평균풀링과 최대풀링을 거친 벡터는 같은 채널로부터 나온 벡터들이기 때문에 동일한 MLP 를 거쳐 더해지고 확률화되어 시그모이드 값이 된다. M_c 는 서로 다른 C(채널)개의 특징을 고려하여 어떤 특징이 중요한지 표현된 확률값이다. 이 값을 입력 특징 F에 곱해 $F`$을 생성한다.

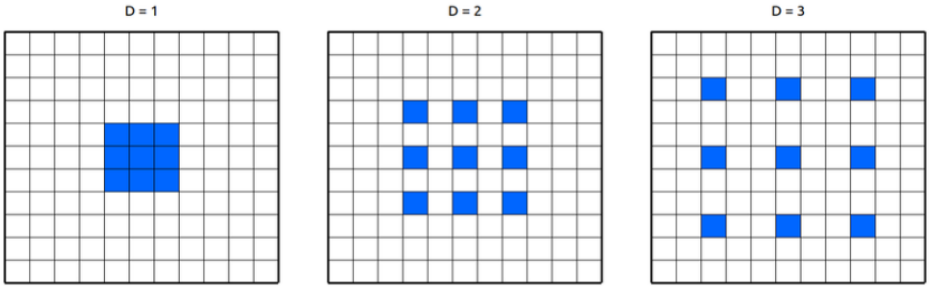
공간 주의는 $F`$으로부터 어느 부분에 집중할지 집계하는 단계이다. 하나의 특징에 대해 각 채널이 갖고 있는 정보 중에서도 판별적인 특징과 평균적인 값 모두 고려하기 위해 평균풀링과 최대풀링을 적용한 값을 연결(Concatenate)하여 채널 차원을 압축한다. 합성곱신경망에서 압축된 특징의 가중치를 학습하기 때문에 특징과 가중치가 혼합될수록 더 판별적인 특징에 집중할 수 있게 된다. 공간 어텐션 M_s는 [식 3]과 같이 계산된다.

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F`); MaxPool(F`)])) \quad [식\ 3]$$

라. 합성곱신경망

합성곱신경망은 행렬로 표현된 필터의 각 요소가 데이터 처리에 적합하도록 학습하는 네트워크이다. 일반적인 합성곱신경망 기반 네트워크에서는 필터의 크기를 3x3, 5x5, 7x7 과 같이 점차 키워가며 학습한다[14]. 필터의 크기가 커진다는 것은 정보의 양이 많다는 것이고, 성능이 높아질 확률 또한 높아진다. 그러나 학습해야 할 양이 많아져 연산량이 증가하게 된다.

확장된 합성곱[20]은 RF 를 크게 만들어 많은 영역을 커버하면서 연산량의 증가는 가져오지 않는 효과적인 방법이다. 확장 비율(Dilation rate)이 커질수록 필터가 출력하는 특징의 간격이 점차 멀어지게 된다. 이로 인해, 필터는 커졌지만 연산량은 증가하지 않는다. [그림 2]는 확장 비율을 각각 1, 2, 3 으로 설정했을 때의 모습이며, 필터 크기는 3x3 으로 설정하였다.



[그림 2. 확장 비율별 수용 필드]

깊이별 분리 합성곱은 MobileNet[21]에서 처음 제안된 병목 블록으로,

합성곱 연산을 좀 더 효율적으로 수행할 수 있도록 개선된 방법이다. 일반적인 합성곱신경망은 입력에 대해 합성곱 연산을 수행하는데, 이 때 필터가 입력 전체에 대해 적용된다. 이에 반해, 깊이별 분리 합성곱은 입력 데이터의 각 채널마다 필터를 적용하는 깊이별 합성곱과 1x1 크기의 필터를 사용해 채널 간의 상호작용을 고려하는 포인트별 합성곱으로 나누어 처리된다. 표준 합성곱, 깊이별 합성곱, 포인트별 합성곱의 곱셈 연산량은 각각 [식 4, 5, 6]으로 계산된다. 여기서 K, C, M은 각각 필터 크기, 입력 채널 수, 출력 채널 수이다.

$$Standard\ CNN = C(K^2 \times M) \quad [식\ 4]$$

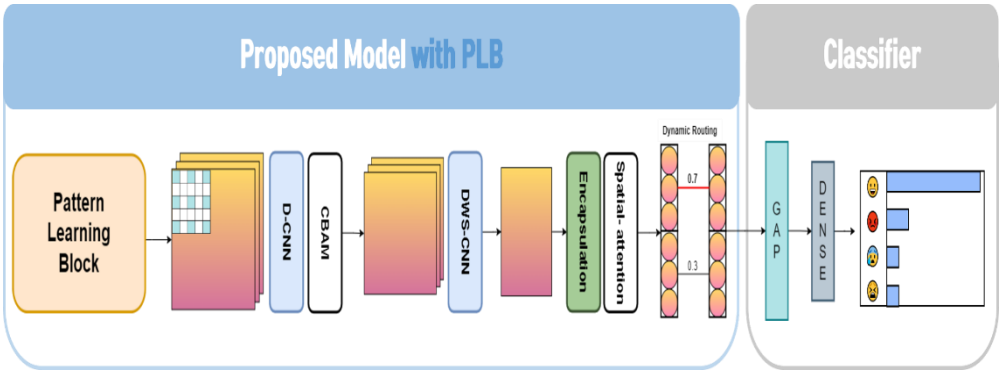
$$Depth - wise\ Convolution = K \times K \times C \quad [식\ 5]$$

$$Point - wise\ Convolution = 1 \times 1 \times C \times M \quad [식\ 6]$$

결과적으로 깊이별 분리 합성곱은 $C(K^2 + M)$이 되어 곱셈에서 덧셈으로 바뀐다. 보통 K보다 출력 채널 수가 더 크기 때문에 약 K^2만큼 전체 매개변수 수 및 연산량이 줄어들게 된다.

3. 제안 방법

본 논문에서는 실시간 환경에서 음성 기반 감정인식을 위한 경량 모델 구조를 제안한다. [그림 3]는 본 논문에서 제안하는 모델 구조이다.

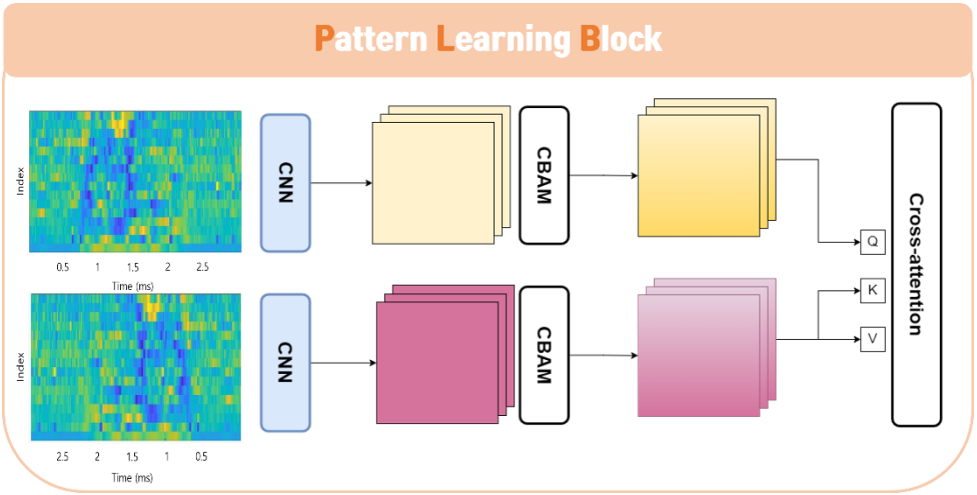


[그림 3. 제안 모델 구조]

[그림 3]과 같이 패턴학습블록(Pattern Learning Block, PLB)를 통해 감정의 패턴을 학습한다. 확장된 합성곱(DCNN: Dilated Convolution)에서 개인의 특성을 반영하고 깊이별 분리 합성곱(DWSCNN: Depth-wise Seperable Convolution)을 통과하며 매개변수 및 연산량을 감소시켰다. 이후, 감정의 중요한 특징 정보를 캡슐에 담고 캡슐의 중요하고 유사한 정보 확보를 위해 자기 주의(Self-attention) 및 CBAM 의 공간 주의(Spatial-attention)를 계산한다. 동적 라우팅(Dynamic Routing)을 통해 감정에 대한 중요도를 조정하였다. 마지막으로 분류를 위해 전역평균풀링(GAP: Global Average Pooling) 후 Dense 계층을 거쳐 각 클래스에 대한 확률값을 계산한다.

가. 패턴학습블록(Pattern Learning Block)

합성곱신경망(CNN: Convolution Neural Network)은 풀링으로 인해 음성의 시간적 구조가 점차 손실된다[9]. 본 연구에서는 서로 다른 시점의 감정 패턴을 추출하기 위한 패턴학습블록을 제안하였다. 제안 모델에서는 [그림 4]와 같이 서로 다른 시점의 감정을 패턴학습블록의 입력으로 사용하여 각각의 시간 패턴을 추출한다. 추가로 CBAM 을 적용하여 음성 신호에서의 채널 및 공간 정보를 부각하였다.



[그림 4. 패턴 학습 블록(Pattern Learning Block, PLB)]

[그림 4]의 합성곱신경망은 4 개의 추가 계층이 구성된다. 배치 정규화(BN: Batch Normalization)[22]를 통해 각 배치(Batch)마다 변형된 분포가 나오지 않도록 평균과 분산을 조정한다. Elu(Exponential linear unit) 활성화 함수는 입력 값이 음수일 때 미분값이 항상 0 이 되는 Dying ReLU 현상을 해결했으며, 값을 (0, 0) 중심으로 모이게 하는 zero-centered 특징을 가진 함수이다[23].

최대풀링은 정해진 필터 크기에서 가장 큰 값만 추출하는 것으로 매개변수의 수, 모델 크기, 추론 시간을 감소시킨다. 특징 추출 네트워크에 공간 드롭아웃[24]을 채택해 전체 2D 특징을 삭제하여 특징 간 독립성을 향상하였다.

나. CBAM(Convolutional Block Attention Module)[19]

합성곱신경망을 경량화를 위한 합성곱신경망으로 대체하면 정확도 손실이 있을 수 있다[25]. 또한 음성은 단일 스케일 신호가 아니며 인간의 말에서 계속해서 감정 변화가 발생하기 때문에[16] 단일 구조는 음성감정 모델링에 적합하지 않다. 이에 제안 모델에서는 음성 신호에서 발생하는 풍부한 정보를 확보하기 위해 CBAM(Convolution Block Attention Module)[19]을 적용하였다.

CBAM 은 입력 특징에 어떤 것이 의미 있는지 학습하는 채널(Channel) 주의와 중요한 부분이 어디에 있는지 학습하는 공간(Spatial) 주의를 계산한다. 채널 주의와 공간 주의는 각각 [식 2]와 [식 3]을 통해 계산된다. CBAM 은 효율적인 계산을 위해 입력 특징의 공간 차원을 압축하고, 공간 정보를 통합해 특징의 중요한 단서를 확보한다. 또한, 판별적인 특징과 평균적인 값을 모두 고려하여 어떤 특징이 중요한지 확률값으로 표현한다.

제안 모델의 전반적인 과정에서 CBAM 이 지속된다. 이를 통해 음성의 채널 및 공간 신호에서 발생하는 정보를 지속적으로 활용하였다. 이에 그치지 않고 CBAM 의 주의(Attention) 가중치를 공유하는 방법[26]을 적용하여 추가적인

매개변수가 사용되지 않게 하였다.

다. 교차 주의(Cross-attention) 기법[8]

음성과 텍스트를 교차 주의(Cross-attention 하는 방법[6]을 응용하여 시간의 정방향 및 역방향 패턴을 병합하였다. 학습 단계에서 시간의 흐름에 따른 정보를 가정(Inductive bias)으로 사용하여 여러 시점의 감정 정보를 고려하였다. 제안 모델은 감정의 시간적 측면을 고려하여 장거리 감정 의존성을 학습한다.

라. 확장된 합성곱(Dilated Convolution)[20]

화자마다 발음 속도, 숨쉬는 구간 등 다양한 차이가 존재한다. 감정인식을 위해서 이러한 화자의 특성을 반영하는 시간이 필요하다[27]. 제안 모델에서는 시간의 척도를 고려하기 위해 교차 주의 이후에 확장된 합성곱(Dilated Convolution, DCNN)을 채택하였다. 확장된 합성곱은 필터가 수용하는 영역인 수용 필드(Receptive Field)를 크게 만들어 통합된 정방향과 역방향의 시간 패턴에서 전체적인 특징, 문맥적(context)인 특징을 추출할 수 있게 된다.

마. 깊이별 분리 합성곱(Depth-wise Seperable Convolution)[21]

이후 특징 추출 과정에서도 자기 주의 기법과 음성의 공간 신호 정보를 활용하기 위해 CBAM 의 공간 주의를 계산한다. 이전 계층까지 여러 번의

합성곱신경망을 통과하며 특징의 크기가 거대해진 상태이다. 주의 기법은 특징의 크기가 클수록 많은 계산 비용이 발생한다[11]. 이를 위해 깊이별 분리 합성곱(Depth-wise Seperable Convolution, DWSCNN)을 채택하였다. 깊이별 분리 합성곱은 음성의 채널 및 공간 신호를 고려하여 매개변수 및 연산량을 감소시킨다.

바. 동적 라우팅(Dynamic Routing)[28]

동적 라우팅(Dynamic Routing)을 위해 스칼라(수치) 값을 벡터(수치+방향) 값으로 조정하였다. 이전 출력층의 특징 차원을 (H, W, C)에서 (L, 64)로 변경한(Reshape) 형태이다. H, W, C, L 은 각각 높이, 너비, 채널, 벡터의 길이를 의미한다. 64 개의 특징을 가진 벡터(캡슐)는 감정에서 중요한 속성(음의 높낮이, 떨림 등)을 표현한다. 변경된 차원에는 채널에 대한 정보가 없어 이후 과정부터 CBAM 의 공간 주의(Spatial-attention)만을 적용하였다. 각 캡슐은 감정에서 중요한 속성을 표현한다. 캡슐 내 특징의 상관관계를 활용할 수 있도록 자기 주의(Self-attention) 기법을 적용하였다. 동적 라우팅(Dynamic Routing)은 학습 과정에서 캡슐에 대한 점수를 동적으로 계산하여 캡슐의 가중치를 조정한다. 자기 주의가 적용된 캡슐의 특징은 동적 라우팅의 반복으로 더 높은 수준의 음성 특징을 생성한다. 동적 라우팅은 CapsNet[28]에서 처음 제안된 기법으로 다수의 층이 입력 특징의 위치, 크기를 학습한다. 알고리즘 1 은 동적 라우팅의 동작 과정이며 u, w, R, b 는 각각 이전 네트워크의 출력, 가중치, 반복 횟수,

편향이다. $Softmax$는 입력 값의 모든 요소를 0 이상 1 이하로 만들고 총합이 1 이 되도록 변환하는 함수이다. 가중치는 임의의 값으로 초기화 되었다.

알고리즘 1. 동적라우팅

given: u, w

R←3

b←0

$\hat{u}$ ←u*w

for R iterations do

c←softmax(b)

v←c + $\hat{u}$

b←b + $\hat{u}$ ° v

return v

알고리즘 1 의 밑줄 친 부분에서 모델의 예측값($\hat{u}$)과 실제값(v)을 곱해 코사인(Cosine) 유사도를 계산한다. 큰 코사인 유사도는 위치 정보가 유사하다는 의미이다. 코사인 유사도를 모델의 예측값에 더해 어떤 샘플의 예측값을 더 크게 전파할지 결정한다.

본 논문에서 제안하는 TPMNet 의 구성은 [표 1]과 같다. Module, Layer, Hyper params, Param #은 각각 네트워크 이름, 네트워크 구성 요소, 구현에 필요한 초매개변수 및 값, 매개변수의 수를 의미한다. 마지막 Softmax 는 각 클래스에 대한 확률 값을 반환하는 활성화 함수로 데이터셋 별로 클래스의 수가 상이하므로, 약간의 차이가 있다. 제안 모델은 약 95,000 개의 매개변수를

가진다.

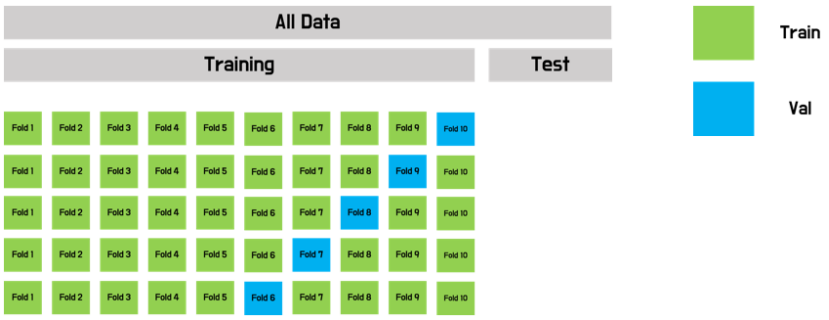
[표 1. 제안 모델 구성]

Module	Layer	Hyper params	Param #
CNN	Conv2D	filter=64, kernel_size=3	1,416
	BatchNorm2D	axis=1	
	Activation	'elu'	
	AveragePooling2D	pool_size(2,2)	
	SpatialDropout2D	rate=0.2	
CNN	Conv2D	filter=64, kernel_size=3	1,416
	BatchNorm2D	axis=1	
	Activation	'elu'	
	AveragePooling2D	pool_size(2,2)	
	SpatialDropout2D	rate=0.2	
DCNN	Conv2D	filter=64, kernel_size=3, rate=2	37,300
	BatchNorm2D	axis=1	
	Activation	'elu'	
	AveragePooling2D	pool_size(2,2)	
	SpatialDropout2D	rate=0.2	
CBAM	Channel-attention	filter=64	4,294
	Spatial-attention		
Spatial-attention2	Spatial-attention		102
DWSCNN			25,600
Attention	Dot-product attention		1
	LayerNorm		128
Dynamic Routing			24,576
GAP			0
Softmax	Dense	num_class= 7 [4, 8] *클래스 개수에 따라 매개변수 수 상이	455
			95,288

4. 실험 결과 및 분석

가. 실험 환경

제안 모델의 실험에서 음성감정인식을 위한 데이터셋을 사용하여 학습한다. 데이터셋은 모델의 견고성 및 일반화 능력 평가를 위해 10-Fold 교차 검증을 수행하였다. [그림 5]는 실험에서 사용된 10-Fold 교차 검증에 대한 도식이다.



[그림 5. 10-Fold 교차검증]

딥러닝 모델 학습에서는 Adam optimizer($\beta_1 = 0.97$, $\beta_2 = 0.93$, $\epsilon = 1e-8$)[29]와 50 Epoch 이후 20 Epoch 마다 학습률을 0.15 씩 감소시키는 학습률 조정 스케줄러를 사용하였다. 초매개변수는 Epoch 300, 배치 크기 32, 초기 학습률은 0.001 로 설정하였다. 모델 검증은 각 학습이 완료되었을 때 수행되었으며, 실제값과 모델이 예측한 값의 손실이 적은 방향으로 학습되었다. 장비는 Intel i7-12700K, 32GB RAM, NVIDIA GeForce RTX 3080Ti 로 구성된 데스크탑 PC 를 사용하였고, Windows10 환경에서 Tensorflow 프레임워크를

기반으로 구현하였다.

나. 데이터셋 및 전처리

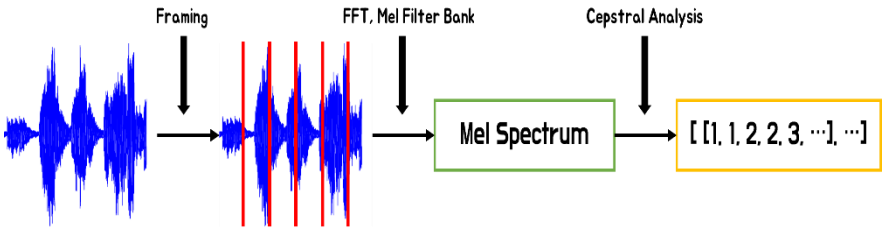
제안 모델의 매개변수 및 성능 입증을 위해 3 가지 데이터셋을 사용한다. EMO-DB[30]는 독일인 남녀(남:5, 여:5)가 7 가지 감정을 담은 10 문장을 발화한 음성파일이다. RAVDESS[31]는 영국인 남녀(남:12, 여:12)가 8 가지 감정을 발화한 말뭉치이다. IEMOCAP[32]은 총 302 개의 비디오에 대해 세션당 2 명의 화자가 9 가지 감정을 연기한 말뭉치이다. 이 중에서 IEMOCAP 은 많은 연구에서 4 개의 감정만을 사용한다[27]. 본 연구에서는 제안된 모델의 비교 및 평가를 위해 기존 연구와 동일하게 4 개의 감정 분류를 수행하였다. 데이터셋별 감정 구성과 개수는 표 2 에 명시하였다. 표 2 의 Dataset, Class, Num.classes 는 각각 데이터셋, 감정, 감정의 개수를 의미한다.

[표 2. 감정 구성과 개수]

Dataset	Class	Num. classes
EMO-DB	angry	81
	boredom	69
	disgust	46
	fear	71
	happy	62
	neutral	127
	sad	79
RAVDESS	angry	96

	calm	192
	disgust	192
	fear	192
	happy	192
	neutral	192
	sad	192
	surprise	192
IEMOCAP	angry	1103
	neutral	1708
	sad	1084
	happy+ excited	1636

실험에 사용한 데이터셋은 특징 추출을 위해 librosa 라이브러리[33]로부터 추출한 MFCC(Mel-Frequency Cepstral Coefficient) 특징을 사용하였다. 특징 추출 과정에서 데이터 증강은 고려되지 않았다. 특징을 생성하는 동안 39 개의 멜필터뱅크는 MFCC 를 계산하는데 사용되었다. [그림 6]은 MFCC 출 과정이다.



[그림 6. MFCC 추출 과정]

또한, 이전 연구[34]와 동일한 전처리 방식을 적용하였다. 음성 데이터의 신호 길이가 100,000 보다 짧으면 패딩(padding)을 사용하여 0 으로 채우고,

길면 100,000 까지의 신호만 사용한다. 패딩을 적용한 샘플과 그렇지 않은 샘플의 수는 표 3 에 명시하였다.

[표 3. 패딩된 샘플과 그렇지 않은 샘플]

Dataset	samples	padding	non-padding
EMO-DB	535	507	28
RAVDESS	1,440	1,406	34
IEMOCAP	5,531	3,509	2,022

다. 성능 평가 지표

제안 모델은 모델의 매개변수의 수, Weighted Average Recall(WAR)을 통해 평가된다. 또한, 모델의 적절한 성능 평가를 위해 혼동행렬(Confusion Matrix)과 Scikit-learn 에서 제공하는 Classification Report 를 사용한다. Classfication Report 는 감정의 평가 지표를 보기 쉽게 시각화한다. 표 4 는 혼동행렬을 나타내며 혼동 행렬의 True Positive(TP), False Positive(FP), False Negative(FN), True Negative(TN)을 사용하여 제안 모델의 각 감정에 대한 성능을 계산한다.

[표 4. 혼동행렬(Confusion Matrix)]

		실제 정답	
		Positive	Negative
모델 예측	Positive	TP	FP
	Negetive	FN	TN

분류 모델에서 사용되는 WAR 은 [식 7]을 통해 계산된다. K, M_k, N 은 클래수 수, 감정 K의 음성 수, 음성 데이터샘플의 수이다. $TP_{ki}, TN_{ki}, FN_{ki}$는 각각 음성에 대한 감정 의 TP, TN, FN 을 나타낸다.

$$WAR = \sum_{k=1}^{K} \frac{M_k}{N} \times \frac{\sum_{i=1}^{M_k} TP_{ki}}{\sum_{i=1}^{M_k} (TP_{ki} + FN_{ki})} \quad [식\ 7]$$

정밀도(Precison)는 모델이 양성으로 예측한 데이터 중 실제 데이터의 값이 양성인 데이터의 비율을 나타낸다. 정밀도는 [식 8]를 통해 계산된다. 재현율(Recall)은 실제 양성으로 나타난 데이터 중 모델이 양성 데이터로 예측한 데이터의 비율을 나타낸다. 재현율은 [식 9]을 통해 계산된다. F1 스코어(F1 score)는 정밀도와 재현율의 조화 평균을 이용하여 계산하는 지표이다. 제안 모델이 양성 예측을 잘하면서 동시에 실제 양성인 데이터의 비율을 잘 유지하고 있는지 평가한다. F1 스코어는 [식 10]을 통해 계산된다.

$$Precision = \frac{TP}{TP + FP} \quad [식\ 8]$$

$$Recall = \frac{TP}{TP + FN} \quad [식\ 9]$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [식\ 10]$$

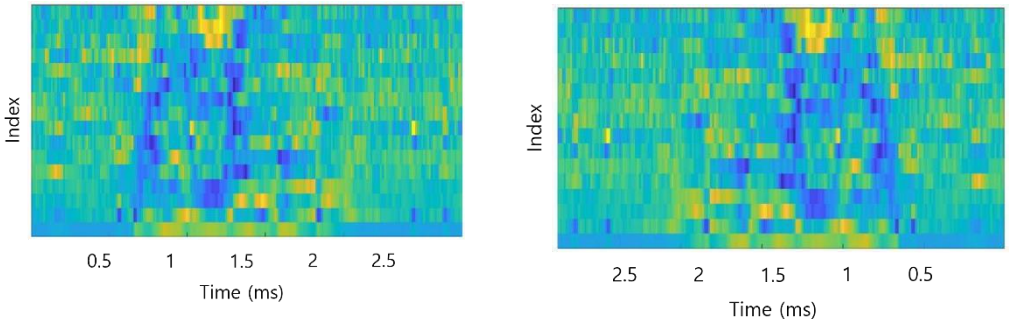
실시간 및 경량 모델을 필요로 하는 영역에서는 매개변수의 수가 적고 성능이 뛰어난 모델이 요구된다. 따라서 매개변수 감소 대비 WAR 을 평가하는 성능 지표가 필요하다. 이를 평가하기 위해 비용편익비율 지표를 활용한다. [식 11]은 비용편익비율을 계산하는 방법이며, 비용은 매개변수의 수, 편익은 WAR 을 대입하였다.

$$Benefit-cost\ ratio = \frac{benefits}{costs} \qquad [식\ 11]$$

라. 실험 결과

1) 시간의 정방향($T_{1\sim N}$) 및 역방향 흐름($T_{N\sim 1}$)

제안 모델에서는 입력으로 서로 다른 시점의 특징을 활용한다. [그림 7]과 같이 앞서 추출된 MFCC 의 원본 특징과 시간의 축을 반전시킨 특징을 의미한다.



(a) 정방향 흐름($T_{1\sim N}$) (b) 역방향 흐름($T_{N\sim 1}$)

[그림 7. 입력 특징]

[그림 4]와 같이 제안 모델의 패턴학습블록(Pattern Learning Block, PLB)에서 서로 다른 시점의 특징을 교차 주의(Cross-attention) 기법을 통해 통합한다. 이에 따라 통합된 데이터와 그렇지 않은 데이터의 성능을 비교 실험하였다. 표 5, 6, 7 은 통합, 정방향 흐름($T_{1\sim N}$), 역방향 흐름($T_{N\sim 1}$)을 사용했을 때의 성능이며 min, max, avg 는 10 개의 폴드 중 최소, 최대, 평균 WAR 을 의미하고, std 는 데이터 구간 간 성능의 표준편차를 의미한다. 표 3 의 파란색, 붉은색 표기는 각각 성능 향상, 성능 감소를 의미하며 해당 구간에서 성능이 가장 높은 모델은 볼드체로 나타낸다. 결과적으로 EMO-DB, RAVDESS 데이터셋에서 10 개 폴드의 최소 성능이 향상되었다. 즉, 어떤 폴드의 데이터가 입력되어도 EMO-DB, RAVDESS 각각 최소 85.11%, 83.75%의 성능을 보장한다는 의미이다. 또한, 모든 데이터셋에서 폴드 간 성능 표준편차(std)가 감소하였다.

[표 5. 제안된 패턴학습블록(Pattern Learning Block, PLB)]

	Proposed			
	min(%)	max(%)	avg(%)	std
EMO-DB	**85.11(↑2.45)**	90.76	88.18(↓0.91)	**1.75(↓1.24)**
RAVDESS	**83.75(↑3.50)**	87.50	85.56(↓0.71)	**1.46(↓1.30)**
IEMOCAP	63.17(↓1.22)	66.20	65.20(↓1.27)	**0.91(↓0.71)**

[표 6. 정방향 흐름($T_{1\sim N}$)]

	$T_{1\sim N}$

	min(%)	max(%)	avg(%)	std
EMO-DB	83.09	**96.05**	**89.09**	4.06
RAVDESS	80.00	90.00	86.43	2.76
IEMOCAP	63.52	**69.27**	66.07	1.82

[표 7. 역방향 흐름($T_{N\sim1}$)]

	$T_{N\sim1}$			
	min(%)	max(%)	avg(%)	std
EMO-DB	82.66	92.93	88.25	2.99
RAVDESS	80.25	**90.62**	86.27	2.77
IEMOCAP	**64.39**	69.09	66.47	1.62

2) 제안 모델 분류 성능 평가

제안 모델의 분류 성능 평가를 위해 혼동행렬(Confusion Matrix)과 Scikit-learn 에서 제공하는 Classification Report 의 결과를 분석하였다. [그림 8]의 (a), (b), (c) 는 각각 EMO-DB, RAVDESS, IEMOCAP 의 Classification Report 이며, [그림 10]의 (a), (b), (c) 는 각각 EMO-DB, RAVDESS, IEMOCAP 의 혼동행렬이다. [그림 8]의 Classification Report 에서 재현율(Recall)이 가장 낮은 감정은 EMO-DB, RAVDESS, IEMOCAP 에서 각각 fear, fear, happy+ excited 이다. [그림 9]의 혼동행렬(Confusion Matrix)을 살펴보면 EMO-DB, RAVDESS 의 fear 감정에 대한 잘못된 예측 대부분이 fear 를 neutral 로

분류했음을 확인할 수 있다.

	precision	recall	f1-score	support
angry	0.90	0.81	0.86	81
boredom	0.88	0.93	0.90	69
disgust	0.98	0.87	0.92	46
fear	0.78	0.76	0.77	71
happy	0.89	0.92	0.90	62
neutral	0.91	0.91	0.91	127
sad	0.82	0.92	0.87	79
accuracy			0.88	535
macro avg	0.88	0.87	0.88	535
weighted avg	0.88	0.88	0.88	535

(a) EMO-DB

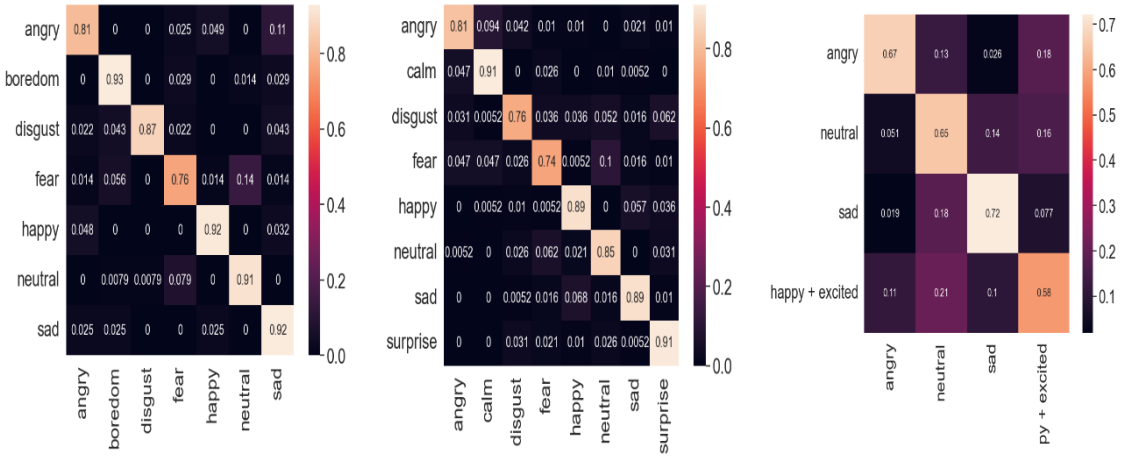
	precision	recall	f1-score	support
angry	0.76	0.81	0.78	96
calm	0.90	0.91	0.90	192
disgust	0.86	0.76	0.81	192
fear	0.81	0.74	0.78	192
happy	0.86	0.89	0.87	192
neutral	0.80	0.85	0.83	192
sad	0.89	0.89	0.89	192
surprise	0.85	0.91	0.88	192
accuracy			0.85	1440
macro avg	0.84	0.85	0.84	1440
weighted avg	0.85	0.85	0.85	1440

(b) RAVDESS

	precision	recall	f1-score	support
angry	0.71	0.67	0.69	1103
neutral	0.62	0.65	0.64	1708
sad	0.65	0.72	0.68	1084
happy + excited	0.63	0.58	0.60	1636
accuracy			0.65	5531
macro avg	0.65	0.66	0.65	5531
weighted avg	0.65	0.65	0.65	5531

(c) IEMOCAP

[그림 8. Confusion Matrix]



(a) EMO-DB (b) RAVDESS (c) IEMOCAP

[그림 9. 혼동행렬(Confusion Matrix)]

3) 제안 모델의 구성 요소가 성능에 미치는 영향

모델의 일부 특징을 제거해 성능에 어떤 영향을 주는지 평가하였다. 표 8, 9, 10 은 EMO-DB, RAVDESS, IEMOCAP 에서 모듈을 제거했을 때의 성능이다. 표 8, 9, 10 의 Remove 열의 구성요소 중 Proposed 는 제안 모델이며, 그 외는

위쪽부터 깊이별 분리 합성곱(DWSCNN: Depth-wise Seperable Convolution), Convolutional Block Attention Module(CBAM), 공간 주의(Spatial-attention), 자기 주의(Self-attention), 동적 라우팅(Dynamic Routing)을 의미한다. #Params, Max WA, Min WA, Avg WA 열은 각각 모델의 매개변수의 수, 10 개의 폴드 중 최소, 최대, 평균 WAR 을 의미한다. 추가로 파란색, 붉은색 표기는 각각 성능 향상, 성능 감소를 의미하며 해당 구간에서 성능이 가장 높은 모델은 볼드체로 나타낸다. 결과적으로 제안 모델의 구성이 EMO-DB, RAVDESS 데이터셋에서 10 개 폴드의 최소, 평균 WAR 이 가장 높음을 알 수 있다.

[표 8. EMO-DB / 모듈이 성능에 미치는 영향]

EMO-DB				
Remove	#Params	Max WA(%)	Min WA(%)	Avg WA(%)
Proposed	95,288	-3.73 90.76	+2.73 85.11	+0.94 88.18
DWSCNN	69,688	92.11	74.03	82.17
CBAM	90,994	94.49	78.79	85.88
Spatial-attention	93,770	92.11	82.38	87.24
Self-attention	95,160	92.93	79.53	84.51
Dynamic Routing	70,712	87.42	70.83	78.66

[표 9. RAVDESS / 모듈이 성능에 미치는 영향]

RAVDESS				
Remove	#Params	Max WA(%)	Min WA(%)	Avg WA(%)
Proposed	95,353	-0.62 87.50	+2.50 83.75	+0.69 85.56
DWSCNN	69,753	78.12	65.62	72.68
CBAM	91,059	88.12	81.25	84.87

Spatial-attention	93,835	85.00	77.50	80.68
Self-attention	95,225	82.50	74.37	78.50
Dynamic Routing	70,777	70.62	65.00	67.62

[표 10. IEMOCAP / 모듈이 성능에 미치는 영향]

IEMOCAP				
Remove	#Params	Max WA(%)	Min WA(%)	Avg WA(%)
Proposed	95,093	-3.76 66.20	-2.52 63.17	-2.20 65.20
DWSCNN	69,493	65.77	59.52	62.72
CBAM	90,799	69.00	63.47	65.07
Spatial-attention	93,575	69.96	65.69	67.40
Self-attention	94,965	67.18	60.91	64.56
Dynamic Routing	70,517	66.66	62.21	64.30

4) 제안 모델과 최신 모델과의 성능 비교

제안 모델은 약 95,000 개의 매개변수를 지닌 모델이다. 문헌 조사 단계에서 1,000,000 개 이상의 매개변수를 지니면서 제안 모델보다 WAR 이 낮은 모델은 성능 비교에서 제외시켰다. 표 11, 12, 13 은 서로 다른 데이터셋에서의 다양한 모델과 성능 비교를 진행하였다. Ref`, Method, WA, #Params, Trade-off 는 각각 참조문헌, 연구 방법, WAR, 매개변수의 수, 비용편익비율을 나타내며, 비용편익비율을 제외한 모든 수치는 참조문헌에 명시된 수치를 사용하였다. 해당 구간에서 성능이 가장 높은 모델은 볼드체로 나타낸다. 결과적으로 제안 모델이

EMO-DB, RAVDESS 데이터셋에서 WAR, 매개변수의 수, 비용편익비율 모두 가장 높은 성능을 달성하였다. 시간의 패턴에 주목한 제안 모델이 음성 기반의 감정인식 모델과 비교하여 충분히 경쟁력 있음을 보여준다.

[표 11. EMO-DB / WAR, 매개변수의 수, 비용편익 비율]

EMO-DB				
Ref`	Method	WAR(%)	#Params	Trade-off
[35]	AG-TFNN	81.86	115,168	71.079
[35]	TFNN	83.54	95,980	87.039
[13]	Sequence of Log Mel-spectrograms	85.54	220,000	38.882
[36]	MLP+ ARN	86.71	575,116	15.077
[37]	DCNN-DTPM	87.31	600,000	14.552
[12]	Mean of MFCCs, Deltas and Delta-Deltas	87.80	164,359	53.420
[34]	RoutingConvNet	87.86	156,384	56.182
	Proposed	**88.18**	**95,288**	**92.541**

[표 12. RAVDESS / WAR, 매개변수의 수, 비용편익 비율]

IEMOCAP				
Ref`	Method	WAR(%)	#Params	Trade-off
[35]	TFNN	50.21	95,980	52.313
[35]	AG-TFNN	51.42	115,168	44.648
[38]	CNN based on log-mel-spectrograms	59.33	194,664	30.478
	Proposed	65.21	**95,093**	**68.570**

[15]	DiCCOSER-CS	65.80	430,000	15.302
[34]	RoutingConvNet	**66.06**	156,189	42.295

[표 13. IEMOCAP / WAR, 매개변수의 수, 비용편익 비율]

EMO-DB				
Ref`	Method	WAR(%)	#Params	Trade-off
[17]	Parallel CNN+ BLSTM+ Attention	65.94	261,288	25.237
[36]	MLP+ ARN	81.02	575,116	14.088
[17]	Parallel CNN+ Transformer	81.33	395,176	20.581
[12]	Mean of MFCCs, Deltas and Delta-Deltas	82.30	164,359	50.073
[34]	RoutingConvNet	83.44	156,449	53.334
	Proposed	**85.56**	**95,353**	**89.732**

5) 자원이 제한된 환경에서의 실시간 능력 평가

자원이 제한된 환경에서 제안 모델의 음성감정인식 추론 시간을 실험하였다. 실험은 잘 알려진 임베디드 보드인 Raspberry Pi 에서 진행되었다. 실시간 능력 평가를 위해 배치 크기를 1 로 설정하고 모든 데이터의 추론 시간의 합을 데이터의 개수로 나눠주었다. 즉, 모든 데이터의 평균 추론 시간으로 평가하였다. 실험 결과 EMO-DB, RAVDESS, IEMOCAP 데이터셋에서 각각 1.42, 1.35, 1.22 초로 평균 1.33 초에 음성감정인식이 이루어지는 것을 확인할 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 음성 기반 감정인식 모델의 한계점 중 합성곱신경망과 RNN 구조의 어려움을 개선하기 위해 시간의 패턴을 추출하는 TPAB(Temporal Pattern Attention Block) 기반의 TPMNet(Temporal Pattern Multi-scale Network)를 제안하였다. 제안 모델은 EMO-DB, RAVDESS, IEMOCAP 총 세가지 데이터셋에서 실험을 수행하였고, 기존 연구에서 제안된 적은 매개변수를 가지는 다양한 모델과 성능을 비교하였다. 그 결과 본 논문에서 제안한 TPMNet 이 모든 데이터셋에서 더 적은 매개변수와 높은 자원 효율성으로 최신 성능을 달성하였고, EMO-DB, RAVDESS 데이터셋에서 WAR 향상을 이루었다. 본 논문의 결과는 음성 기반 감정인식 기술이 실제 응용 분야에서 활용될 수 있는 가능성을 보여주며, 음성 챗봇, 음성 기반 모니터링 등 다양한 분야에서 적용될 수 있음을 시사한다.

향후 연구에서는 음성 외 도메인(이미지, 텍스트 등)을 활용한 모델을 구성하고 그 능력을 평가하고자 한다. 또한, 여러 국가의 음성 감정을 동시에 추론하는 교차 언어(cross-lingual)와 다양한 연령층의 감정을 추론하는 교차 연령(cross-age)에 대한 연구도 진행할 예정이다. 제안 논문에서는 세 가지 데이터셋에 포함된 인간의 일부 감정에 대해 실험하였다. 이에 기존 데이터에 속하지 않은 외적 감정에 대한 능력 평가도 진행할 계획이다.

참고 문헌

1. Anvarjon, T., Mustaqeem, and S. Kwon, *Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features.* Sensors, 2020. **20**(18): p. 5212.
2. Er, M.B., *A novel approach for classification of speech emotions based on deep and acoustic features.* IEEE Access, 2020. **8**: p. 221640-221653.
3. Dey, A., et al., *A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition.* IEEE Access, 2020. **8**: p. 200953-200970.
4. 임명진, 박원호, and 신주현, *Word2Vec 과 LSTM 을 활용한 이별 가사 감정 분류.* 스마트미디어저널, 2020. **9**(3): p. 90-97.
5. Parry, J., et al. *Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition*. in *Interspeech*. 2019.
6. Liu, F., et al., *Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition.* Entropy, 2022. **24**(7): p. 1010.
7. Wu, C.W., *ProdSumNet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions.* arXiv preprint arXiv:1809.02209, 2018.
8. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
9. Zhao, Z., et al., *Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition.* Neural Networks, 2021. **141**: p. 52-60.
10. Kwon, S., *MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach.* Expert Systems with Applications, 2021. **167**: p. 114177.
11. Wang, S., et al., *Linformer: Self-attention with linear complexity.* arXiv preprint arXiv:2006.04768, 2020.
12. Nantasri, P., et al. *A light-weight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives*. in *2020 17th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*. 2020. IEEE.
13. Atsavasirilert, K., et al. *A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms*. in *2019 14th International Joint Symposium on Artificial Intelligence and*

Natural Language Processing (iSAI-NLP). 2019. IEEE.
14. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**.
15. Tang, D., et al., *End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network.* EURASIP Journal on Audio, Speech, and Music Processing, 2021. **2021**(1): p. 18.
16. Ye, J.-X., et al., *GM-TCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition.* Speech Communication, 2022. **145**: p. 21-35.
17. Bautista, J.L., Y.K. Lee, and H.S. Shin, *Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation.* Electronics, 2022. **11**(23): p. 3935.
18. Abdelhamid, A.A., et al., *Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm.* IEEE Access, 2022. **10**: p. 49265-49284.
19. Woo, S., et al. *Cbam: Convolutional block attention module*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.
20. Yu, F. and V. Koltun, *Multi-scale context aggregation by dilated convolutions.* arXiv preprint arXiv:1511.07122, 2015.
21. Howard, A.G., et al., *Mobilenets: Efficient convolutional neural networks for mobile vision applications.* arXiv preprint arXiv:1704.04861, 2017.
22. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. pmlr.
23. Clevert, D.-A., T. Unterthiner, and S. Hochreiter, *Fast and accurate deep network learning by exponential linear units (elus). arXiv 2015.* arXiv preprint arXiv:1511.07289, 2016. **2**.
24. Tompson, J., et al. *Efficient object localization using convolutional networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
25. Tan, M. and Q. Le. *Efficientnet: Rethinking model scaling for convolutional neural networks*. in *International conference on machine learning*. 2019. PMLR.
26. Xiao, T., et al., *Sharing attention weights for fast transformer.* arXiv preprint arXiv:1906.11024, 2019.
27. Ye, J., et al. *Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition*. in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP). 2023. IEEE.
28. Sabour, S., N. Frosst, and G.E. Hinton, *Dynamic routing between capsules.* Advances in neural information processing systems, 2017. **30**.
29. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
30. Burkhardt, F., et al. *A database of German emotional speech*. in *Interspeech*. 2005.
31. Livingstone, S.R. and F.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, 2018. **13**(5): p. e0196391.
32. Busso, C., et al., *IEMOCAP: Interactive emotional dyadic motion capture database.* Language resources and evaluation, 2008. **42**: p. 335-359.
33. McFee, B., et al. *librosa: Audio and music signal analysis in python*. in *Proceedings of the 14th python in science conference*. 2015.
34. 임현택, et al., *RoutingConvNet: 양방향 MFCC 기반 경량 음성감정인식 모델.* 스마트미디어저널, 2023. **12**(5): p. 28-35.
35. Pandey, S.K., H.S. Shekhawat, and S. Prasanna, *Attention gated tensor neural network architectures for speech emotion recognition.* Biomedical Signal Processing and Control, 2022. **71**: p. 103173.
36. Kumar, S., et al., *Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance.* Computers, Materials & Continua, 2023. **75**(1).
37. Nagarajan, B. and V. Oruganti, *Deep Learning as Feature Encoding for Emotion Recognition.* arXiv preprint arXiv:1810.12613, 2018.
38. Chauhan, K., K.K. Sharma, and T. Varma. *Speech emotion recognition using convolution neural networks*. in *2021 international conference on artificial intelligence and smart systems (ICAIS)*. 2021. IEEE.

A Light-weight Model for Speech Emotion Recognition based on Pattern Learning Block

Hyun-Taek LIM

Department of Artificial Intelligence Convergence

Graduate School, Chonnam National University

(Supervised by Professor Hyung-Jeong Yang)

(Abstract)

Speech emotion recognition plays a crucial role in inferring human emotions and affective states from speech signals, enhancing interaction between humans and machines. One of the significant challenges faced by current speech emotion recognition is the neglect of applications requiring real-time and light-weight models, as well as inadequate processing capabilities in real-time applications. Furthermore, most high-performance models heavily depend on many trainable parameters.

In this study, we propose the Pattern Learning Block (PLB), which analyzes emotion patterns at different time points. Through the PLB, emotional patterns of both forward and backward time flow are learned, taking into

account emotional information and situational characteristics at multiple points in time. Additionally, we enrich the representation by continuously leveraging abundant information from channel and spatial signals in speech.

The proposed model, with approximately 95,000 parameters, demonstrates parameter reduction and accuracy improvement across comprehensive experiments on three speech emotion datasets (EMODB, RAVDESS, IEMOCAP), achieving state-of-the-art performance. Furthermore, experiments were conducted to evaluate the impact of the proposed model's components on performance and to observe the inference speed in memory-constrained environments. The light-weight model secures real-time processing capabilities with low computational cost and fast inference speed, so it can be used in real-time systems such as speech-based chatbots and mood monitoring system. We anticipate that the proposed model can enhance the transferability of emotion recognition in real-time and hardware-constrained domains in the future.