특집논문 (Special Paper)
방송공학회논문지 제23권 제3호, 2018년 5월 (JBE Vol. 23, No. 3, May 2018)
https://doi.org/10.5909/JBE.2018.23.3.351
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

준 지도학습과 여러 개의 딥 뉴럴 네트워크를 사용한 멀티 모달 기반 감정 인식 알고리즘

김 대 하[a)], 송 병 철[a)‡]

Multi-modal Emotion Recognition using Semi-supervised Learning and Multiple Neural Networks in the Wild

Dae Ha Kim[a)] and Byung Cheol Song[a)‡]

요 약

인간 감정 인식은 컴퓨터 비전 및 인공 지능 영역에서 지속적인 관심을 받는 연구 주제이다. 본 논문에서는 wild 환경에서 이미지, 얼굴 특징점 및 음성신호로 구성된 multi-modal 신호를 기반으로 여러 신경망을 통해 인간의 감정을 분류하는 방법을 제안한다. 제안 방법은 다음과 같은 특징을 갖는다. 첫째, multi task learning과 비디오의 시공간 특성을 이용한 준 감독 학습을 사용함으로써 영상 기반 네트워크의 학습 성능을 크게 향상시켰다. 둘째, 얼굴의 1 차원 랜드 마크 정보를 2 차원 영상으로 변환하는 모델을 새로 제안하였고, 이를 바탕으로 한 CNN-LSTM 네트워크를 제안하여 감정 인식을 향상시켰다. 셋째, 특정 감정에 오디오 신호가 매우 효과적이라는 관측을 기반으로 특정 감정에 robust한 오디오 심층 학습 메커니즘을 제안한다. 마지막으로 소위 적응적 감정 융합 (emotion adaptive fusion)을 적용하여 여러 네트워크의 시너지 효과를 극대화한다. 제안 네트워크는 기존의 지도 학습과 반 지도학습 네트워크를 적절히 융합하여 감정 분류 성능을 향상시켰다. EmotiW2017 대회에서 주어진 테스트 셋에 대한 5번째 시도에서, 제안 방법은 57.12 %의 분류 정확도를 달성하였다.

Abstract

Human emotion recognition is a research topic that is receiving continuous attention in computer vision and artificial intelligence domains. This paper proposes a method for classifying human emotions through multiple neural networks based on multi-modal signals which consist of image, landmark, and audio in a wild environment. The proposed method has the following features. First, the learning performance of the image-based network is greatly improved by employing both multi-task learning and semi-supervised learning using the spatio-temporal characteristic of videos. Second, a model for converting 1-dimensional (1D) landmark information of face into two-dimensional (2D) images, is newly proposed, and a CNN-LSTM network based on the model is proposed for better emotion recognition. Third, based on an observation that audio signals are often very effective for specific emotions, we propose an audio deep learning mechanism robust to the specific emotions. Finally, so-called emotion adaptive fusion is applied to enable synergy of multiple networks. The proposed network improves emotion classification performance by appropriately integrating existing supervised learning and semi-supervised learning networks. In the fifth attempt on the given test set in the EmotiW2017 challenge, the proposed method achieved a classification accuracy of 57.12%.

Keyword : Emotion recognition, Multi-task learning, Semi-supervised learning, Multi modal signal, EmotiW 2017 challenge

Ⅰ. 서 론

인간과 컴퓨터 상호 작용에 대한 인간의 감정을 인식하는 문제는 오랫동안 컴퓨터 비전 및 기계 학습의 연구 분야에서 연구되었다. 이는 감정 인식이 인간과 로봇 상호 작용 또는 인간과 기계 상호 작용에서 가장 중요한 기술 중 하나이기 때문이다. 예를 들어 현재의 AI (인공 지능) 로봇은 인간의 감정 상태를 감지하여 기초적인 수준이지만 인간의 반응에 적절하게 대응할 수 있다. 최근에는 인간의 감정을 파악하기 위해 영상 및 음성을 기본 정보로 사용하고 보조 정보로 ElectroEncephaloGram (EEG) 신호와 같은 생체 정보까지 사용하고 있다[1]. AV 정보는 일반적으로 다른 정보보다 획득하기 쉽기 때문에 오디오 및 비디오 (AV) 정보를 기반으로 한 인간의 감정 인식이 널리 연구되고 있다[2,3].

지금까지 AV 정보의 감정을 인식하기 위해 handcraft feature를 이용한 기계 학습 방법이 많이 연구되어 왔다[4-7]. 예를 들어, Emotion Recognition in Wild (EmotiW) 2013 대회에서 대부분의 기계 학습 기법은 handcraft feature를 기반으로 학습한다. 이후 2012 ImageNet challenge에서 등장한 AlexNet[9]은 딥 러닝 기술에 중요한 촉매 역할을 하였다. 따라서 오늘날에는 다양한 분류 문제를 해결하기 위해서 딥 러닝 네트워크를 기반으로 하는 연구가 진행되고 있다[10-13].

감정 인식 연구에서는 인간의 감정을 주로 분노, 혐오, 공포, 행복, 슬픔, 놀람, 중립 등 7 가지 범주로 분류하였다[14]. 물론 인간의 얼굴 행동 단위 (AU)[16]를 통해보다 상세한 감정 분류를 할 수 있지만, EmotiW challenge에서는 일상 생활에서 주로 느끼는 감정으로 일곱 감정을 분류하는 문제를 다룬다. EmotiW challenge[25]에서 비디오 기반 감정 인식은 하나의 정지 영상에 대한 감정 인식보다는 일정 길이의 비디오 클립에 대한 감정 인식을 목표로 한다. 또한 연구실 환경에서 제작 된 CK+[17] 데이터 셋과 달리 EmotiW는 영화 또는 리얼리티 TV 프로그램으로 구성된 데이터 세트를 사용한다. 따라서, 비디오 내 인물의 표정 및 주변 환경의 변화는 매우 다양하다. 최근 EmotiW 대회에서 보여준 감정 인식 방법의 대부분은 딥 러닝 네트워크를 사용한다. 예를 들어, 2016 년 challenge에서 1위를 수상한 팀의 네트워크의 경우 convolutional 3D[18]와 Convolutional Neural Network (CNN)-Recurrent Neural Network (RNN)을 동시에 사용하고, 오디오 분류기는 서포트 벡터 머신 기반(SVM)을 사용하여 기존보다 높은 감정 분류 정확도를 달성하였다. 2위의 네트워크[3]에서는 영상 데이터와 LBP (Local Binary Pattern)를 함께 사용하여 딥 러닝 네트워크를 학습하였다. Bargal et al.[19]은 3개의 딥 러닝 네트워크를 병렬로 결합하여 얻은 deep feature를 이용하여 비디오의 감정 분류를 진행하였다. 요약하면, 최신 감정 인식 방법은 다양한 딥 러닝 네트워크를 융합하는 경향을 보인다[2,3,19].

하지만 지금까지의 네트워크는 오직 지도학습 기반의 알고리즘과 이들을 단순 가중치 합산을 통하여 최종 결과를 얻어내었다. 하지만 본 논문에서는 이미지 기반 네트워크로서 준 지도학습 네트워크[20]와 보조 네트워크[21]로 구성된 3 차원 (3D) CNN을 제안한다. 본 논문에서는 얼굴 특징점 정보를 효과적으로 활용하기 위한 새로운 특징 생성 방법과 그에 따른 네트워크 구성 방법을 제안한다. 다음으로 우리는 오디오 신호에 적합한 세 가지 딥 러닝 네트워크를 통합한 오디오 기반 네트워크를 제안한다. 마지막으로 적응적 감정 융합 기법은 서로 다른 네트워크 간의 시너지를 극대화하기 위해 사용한다.

Ⅱ. 제안 알고리즘

그림 1에서 볼 수 있듯이 multi modal 신호 기반 제안 기법은 이미지 기반 네트워크, 얼굴 특징점 기반 네트워크, 그리고 오디오 기반 네트워크로 구성된다. Multi modal 신호에 적합한 다중 네트워크로부터 emotion score가 얻어진다면, emotion score의 중요성에 따라 불균등 한 가중치를 부여하는 적응적 감정 융합이 적용된다. [23]과 [37]에 따르

a) 인하대학교 전자공학과(Department of Electronic Engineering College of Engineering, Inha University)
‡ Corresponding Author : 송병철(Byung Cheol Song)
E-mail: bcsong@inha.ac.kr
Tel: +82-32-860-7413
ORCID:http://orcid.org/0000-0001-8742-3433
※ 이 논문은 2017년도 정부(과학기술정통부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1C1B2003044).
※ This research was supported by National Research Foundation of Korea Grant funded by the Korean Government (2016R1A2B4007353).
· Manuscript received March 22, 2018; Revised April 26, 2018; Accepted April 26, 2018.

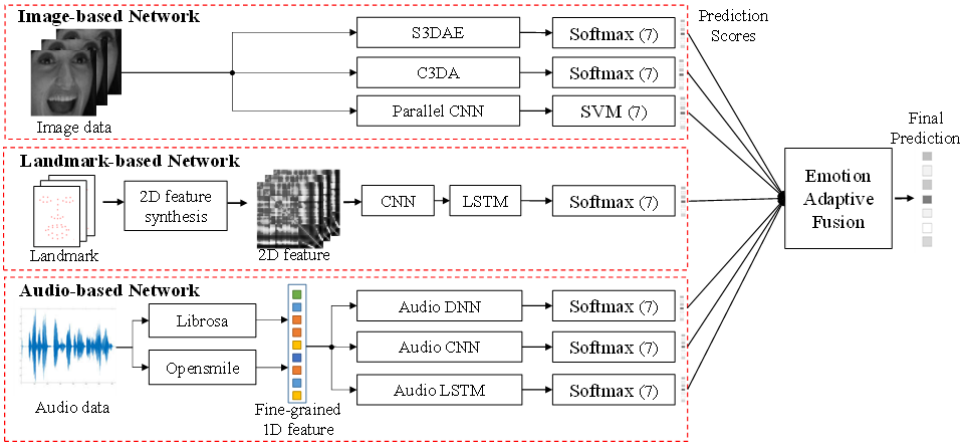


그림 1. 제안 알고리즘의 구성도
Fig. 1. The flowchart of proposed algorithm

면 독립된 네트워크 학습이 아닌 여러 네트워크를 그룹화하여 학습할 때 네트워크의 성능이 향상된다는 것을 확인할 수 있다. 이러한 연구를 기반으로 우리는 이미지 기반 네트워크의 기본 프레임 워크로 Convolutional 3D (C3D)[18]를 사용한다. 또한 auxiliary network[21]를 사용하여 딥 러닝 네트워크가 자주 직면하는 gradient vanishing problem을 완화한다. 둘째, 얼굴 특징점 정보를 2D 영상 특성으로 변환하기 위한 새로운 변환 모델을 제안한다. 2D기반 얼굴 특징점 정보는 CNN-Long Short Term Memory (LSTM)을 사용하여 학습한다. 셋째, 오디오 기반 딥 러닝 네트워크를 설계하여 표정에서 파악하기 힘든 정보를 배경 음악 및 인물에서 나오는 소리를 사용하여 파악한다. 따라서 오디오 기반 네트워크는 이미지 기반 네트워크의 한계를 보완 할 수 있다. 마지막으로, 다중 네트워크로부터의 score의 감정 적응 융합은 네트워크의 최종 분류 정확도를 더욱 향상시키는데 도움을 준다.

1. 이미지 기반 네트워크

최근에는 인간 행동 인식을 위한 다양한 spatio-temporal 기반 딥 러닝 네트워크가 연구되고 있다. 예를 들어, convolutional 3D(C3D)는 시공간 학습을 사용하는 대표적인 딥 러닝 학습 방법이다. 그러나 C3D 네트워크는 배치, 프레임, 너비, 높이 및 채널로 구성된 5 차원 tensor를 입력으로 사용하므로 네트워크 학습에 오랜 시간이 소요된다. 또한 마지막 단계의 fully connected (FC) layer가 '오버 피팅'을 유발할 수 있다. 이러한 문제들을 극복하기 위해 C3D 기반의 3D 오토인코더(S3DAE)를 사용한 준 지도 학습 모델을 제안한다 (C3DA). 이는 [18]과 [21]에서 언급한 gradient vanishing problem을 어느정도 완화할 수 있다. 또한 [19]의 네트워크를 응용하여 여러 개의 딥 러닝 네트워크를 병렬로 학습한 뒤 SVM을 사용하여 비디오 내 감정을 분류한다. 한편, 각 프레임에 대해, 얼굴 검출기는 cascade CNN face detector[22]를 사용하여 얼굴 영역을 검출한 뒤 히스토그램 평활화(histogram equalization)를 적용하여 전체 입력 데이터의 전 처리 과정을 수행한다.

1.1 3D Semi-supervised learning with 3D autoencoder (S3DAE)

반 지도 학습은 기본적으로 지도 학습 문제를 라벨이 없는 데이터를 함께 사용하여 지도학습의 학습 성능을 향상

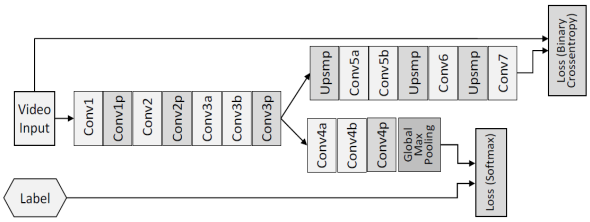


그림 2. 3D 오토인코더 기반 준 지도 학습
Fig. 2. Semi-supervised learning with 3D autoencoder

시키는 머신 러닝 기법이다[38,39].

그림 2와 같이 우리는 오토인코더를 사용하여 라벨이 없는 데이터의 학습을 진행한다. 오토인코더는 기본적으로 feed-forward network로 구성되어 있으며 학습 과정에서의 네트워크의 파라미터 세팅을 위한 transfer learning, weight initialization 기법으로 많이 쓰인다. 대부분의 오토인코더의 경우 mean square error 손실 함수를 사용하여 네트워크를 최적화하지만 본 논문에서는 기존과 비교하여 얼마나 많은 정보가 복원 과정에서 보존되었는지를 분석하기 위해 binary crossentropy 손실 함수를 사용한다. Binary cross-entropy를 사용하기 때문에 네트워크의 마지막 단에 sigmoid 활성화 함수를 사용한다. 오토인코더의 기본 식 및 손실 함수는 아래와 같이 정의된다.

$$z = \sigma(Wx + b) \qquad (Encoder) \tag{1}$$

$$x' = \sigma'(W'z + b') \ (Decoder) \tag{2}$$

$$J(x, x') = -\sum_k^d \left[x_k \log x_k' + (1 - x_k) \log(1 - x_k') \right] \tag{3}$$

여기서 W, b의 경우 네트워크의 학습 파라미터, x는 입력 정보, z는 중간 변수, x'는 복원 정보, 마지막으로 σ는 sigmoid 또는 ReLU와 같은 element-wise 활성화 함수이다.

제안 네트워크는 라벨이 없는 데이터의 학습을 진행하는 오토인코더와 지도 학습 기반 네트워크를 hard parameter sharing[18] 기법을 통하여 하나의 네트워크로 구성한다. 하지만 Hard parameter sharing을 통한 서로 다른 네트워크의 학습이 네트워크의 일반화 효과를 가져다주는 것은 사실이지만 오히려 두 네트워크의 학습이 다른 네트워크의 파라미터 업데이트에 악영향을 줄 가능성도 농후하다. 하지만 제안 네트워크의 학습이 가능한 이유는 본디 오토인코더의 목적이 네트워크의 파라미터를 미리 세팅해 두는 pre-train model로 많이 쓰였기 때문에 오토인코더의 인코더가 3D Convolution 네트워크의 학습에 정규화 효과를 가져다 줄 수 있다.

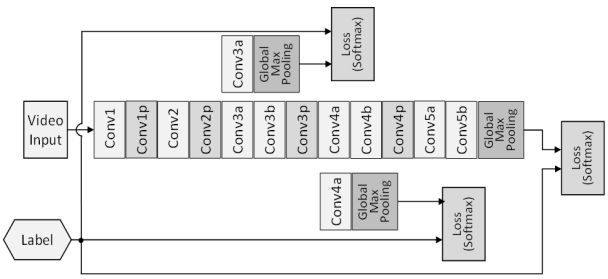


그림 3. 3D 보조 네트워크를 사용한 Convolutional 3D 네트워크
Fig. 3. Convolutional 3D with auxiliary network

1.2 Convolutional 3D with auxiliary network (C3DA)

C3D에는 프레임 정보를 포함한 5D 정보가 입력 데이터로 사용되므로 학습과정에서의 loss값의 수렴이 어려운 경향을 보인다. 특히 마지막 FC layer에서의 vanishing gradient problem을 완화하기 위해 우리는 유명한 GoogleNet [24]과 같이 보조 네트워크[21]를 포함한 multi-task 네트워크를 구성한다. C3DA는 그림 3과 같이 3번째와 4번째 convolution layer에서 보조 네트워크를 삽입한다. 따라서 기존 네트워크 대비 네트워크의 역전파 과정에서 손실되는 gradient를 완화할 수 있기 때문에 parameter 학습을 보다 원활하게 진행할 수 있다.

1.3 Parallel CNN network

마지막 이미지 기반 네트워크로는 [19]에서 아이디어를 얻어 총 3층의 병렬 네트워크를 구성하여 얼굴 영상에서의 deep feature를 얻어낸다. 기존의 VGG[11]를 대체하기 위해 R-VGG13 및 R-VGG16을 제안하여 기존 VGG 대비 보다 효율적인 학습을 진행한다. 또한 우리는 기존의 Residual Network 대신에 Wide Residual Network (WRN)[26] 사용하여 향상된 feature learning 성능을 얻는다. 이렇게 총 3 개의 네트워크를 병렬로 연결하여 얼굴의 deep feature를 얻는다. 동시에 Xception[27]을 사용하여 추가적인 deep feature를 얻는다. 마지막에는 Support Vector Machine(SVM) [7]을 사용하여 얻어낸 4352 차원의 deep feature를 분류하여 7가지 감정을 구분한다.

2. 얼굴 특징점 기반 네트워크

얼굴 특징점은 얼굴에 존재하는 key point의 위치 정보를 나타내며 감정 인식을 위한 강력한 도구로 자주 사용되었

다[28]. 기존의 얼굴 특징점 기반 감정 인식 방법은 기본적인 DNN 또는 SVM을 사용하여 감정을 분류하였다[29, 30]. 그러나 종래의 방법들은 wild 환경에서의 얼굴 특징점 위치 정보를 표준화 하기 힘들다. 따라서 우리는 얼굴 특징점 위치-이미지 변환 모델을 제안한다. 얼굴 특징점 위치 정보를 2D 이미지로 변환하여 CNN-LSTM 네트워크를 사용하여 비디오 내 감정을 분류한다. 따라서 얼굴 특징점 기반 네트워크는 섹션 2의 이미지 기반 네트워크를 보완하는 또 다른 방식을 제안한다.

2.1 Landmark based network

얼굴 특징점 위치 변화는 표정의 변화를 의미한다. 따라서 우리는 연속적인 프레임에서 각 특징점들의 상대적 거리 변화를 나타내는 2D feature를 제안한다. 먼저, i 번째 표식과 j 번째 표식 사이의 L2 거리, 즉 $P(i, k)$와 $P(j, k)$를 계산 한 후, $(k - 1)\ th$ 프레임의 결과와 비교한다. 그 변화는 식 (4)와 같이 2D 특징점 feature로 정의된다.

$$L(i,j,k) = \| P(i,k) - P(j,k) \|_2 - \| P(i,k-1) - P(j,k-1) \|_2 \quad (4)$$

2.2 CNN LSTM network

CNN-LSTM 기법을 사용하여 sequential 2D landmark feature을 통해 감정을 분류한다. 우리는 VGG16을 CNN 모델로 채택하고 LSTM 모델 기반의 stacked LSTM을 채택한다. LSTM의 vector dimension은 순차적으로 128, 128, 64, 그리고 32로 구성되고 20% 비율의 dropout을 사용한다. 만약 2D 데이터가 CNN에 $M \times M \times (N - 1)$을 입력하면 $256 \times (N - 1)$의 특징 벡터가 생성된다. 다음으로 VGG16의 $256 \times (N - 1)$ 피처가 LSTM에 입력된다. 마지막으로 softmax 분류기를 사용하여 총 7 가지 비디오 내 감정을 분석한다 (그림 1).

3. 음성 기반 네트워크

배경 소리가 지배적인 비디오의 경우 감정 정보는 얼굴의 표정 변화가 아닌 주로 오디오 정보[2,30]에서 얻어진다. 예를 들어 슬픔이나 두려움 감정의 경우 얼굴 표정 변화를 알기 힘들지만 배경 소리 같은 오디오 정보는 감정을 파악하는데 중요한 단서를 제공한다. 그러나 오디오 기반 감정 인식에 관한 이전 연구에서는 단순한 학습 구조를 채택하여 감정을 분류하였다. 하지만 우리는 다양한 딥 러닝 네트워크를 사용하여 음성 정보를 효과적으로 분석한다. 음성 신호 기반 네트워크는 각각 audio DNN, audio CNN, 그리고 audio LSTM 네트워크를 사용한다. 비디오에서 audio feature를 추출하기 위한 도구로는 Opensmile[31]과 Librosa[32] 파이썬 기반 라이브러리를 사용한다.

Audio DNN의 구성 그림 4와 같다. FC layer와 batch normalization을 사용하여 네트워크를 구성한다. Fully connected layer는 64차원의 노드로 구성한다. 데이터 셋에 비해 네트워크의 capacity가 작기 때문에 네트워크의 오버 피팅 문제가 발생할 수 있다. 따라서 오버피팅을 완화하기 위해 dropout 비율을 0.5로 설정한 뒤 각 레이어에 적용한다. 마지막 층은 softmax 분류함수를 사용하여 emotion score를 출력한다. Audio CNN에서는 주로 1D convolution과 max pooling을 사용하여 네트워크를 구성한다. Convolution filter 수는 차례로 32, 64, 그리고 128를 사용한다. 마지막 FC layer의 경우 256 차원의 노드를 사용한다. 자세한 구성은 그림 5 와 같다. 마지막으로 audio LSTM은 3층의 stacked LSTM으로 구성한다. LSTM에서의 vector dimension은 전부 577을 사용한다.

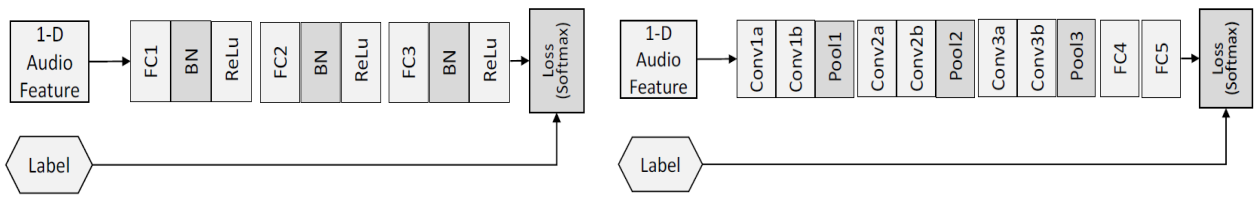


그림 4 & 5. 음성 DNN 네트워크 & 음성 CNN 네트워크
Fig. 4 & 5. Structure of the audio DNN network and CNN network

4. 적응적 감정 융합

제안 기법은 앞에서 언급한 총 7개의 network를 통하여 7개의 감정에 대한 prediction score를 얻을 수 있다. 일반적으로 여러 개의 네트워크 결과의 융합은 네트워크의 앙상블 효과를 기대할 수 있다. 우리는 앙상블 효과를 최대화하기 위해 적응적 감정 융합이라는 방법을 제안한다. 각 네트워크는 class에 따라 추정 정확도가 다를 수 있다. 따라서 validation data를 통해 emotion class에 추정 정확도를 확인하였고 이를 이용하여 7개 network 각각의 score weight W_k를 결정한다. score weight W_k는 vector 형태로 network 각각의 class별 정확도에 따라 각 성분이 adapative하게 결정된다. 7 개의 네트워크를 통한 final prediction score는 식 (5)과 같이 결정된다.

$$S_{total} = \Sigma_{k=1}^{7} W_k S_k \quad (5)$$

Ⅲ. 실험 및 결과

1. 데이터 셋

챌린지 데이터 셋. EmotiW 2017 대회에서 제공한 AFEW 6.0[15]은 학습 데이터 셋 (773 비디오 클립), 유효성 검사 데이터 셋 (383 비디오 클립) 및 테스트 데이터 셋 (653 비디오 클립)의 세 부분으로 구성된다. 올해 60 개의 비디오 클립이 작년 테스트 데이터 세트 (593 개의 비디오 클립)에 새로 추가되었다. 새로운 테스트 비디오의 대부분은 시트콤과 같은 TV 프로그램에서 가져온 것으로, 기존의 학습 데이터의 구성과 특성이 매우 다르다. 따라서 작년 대비 향상된 난이도의 대회가 진행된다.

자체 데이터 셋. 새로 추가 된 테스트 비디오에 효과적으로 대응하기 위해 YouTube와 같은 공개 도메인에서 834개의 짧은 비디오 클립을 수집하여 네트워크 학습에 사용하였다. 수집 데이터 셋은 대부분 행복, 중립, 공포와 같은 감정을 표현하는 비디오 클립으로 구성하였다 (표 1 참조).

표 1. 수집 데이터 셋에 대한 감정 카테고리 분포
Table 1. Emotion category distribution of additional dataset

Total	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
834	30	12	32	416	316	9	19

2. 구현 세부사항

우리는 앞서 구성했던 각각의 딥 러닝 네트워크들의 학습을 통하여 파라미터를 최적화 시켰다. 우리는 Keras[33] 텐서플로 기반의 High-level deep learning library을 사용하여 딥 러닝 네트워크 구축 및 파라미터 최적화 과정을 진행하였다. Keras의 손쉬운 딥 러닝 네트워크의 Tensor 관리와 간편한 모듈화를 장점으로 손쉽게 네트워크를 구성할 수 있었다.

이미지 기반 네트워크. 앞의 1.1 절에서 언급했듯이 이미지 기반 네트워크에서 C3D 기반의 network의 경우 각 비디오 클립의 프레임 중 맨 앞에서 40프레임만을 딥 러닝 네트워크의 입력으로 사용하였고, 40 프레임보다 적은 비디오 클립의 경우 맨 끝 이미지를 padding하였다. 입력 이미지 크기로는 Semi-Supervised network에서만 112x112 픽셀 사이즈를 사용하였고, 나머지 네트워크에서는 224x224 픽셀 사이즈를 사용하였다. Semi-Supervised network에서만 SGD with momentum을 사용하였고, 나머지 네트워크에서는 adam optimizer를 사용하였다. 모든 네트워크 학습의 진행 단위는 epoch 단위로 진행하였으며, weight decay 값으로는 0.0005, weight initialization의 경우 he initialization[34], Batch normalization의 경우 [35]를 사용하였다. 활성화 함수의 경우 Semi-supervised learning의 맨 마지막 단을 제외하고는 전부 ReLU 활성화 함수를 사용하였다.

랜드 마크 기반 네트워크. Landmark-based 네트워크는 AFEW 6.0에서 함께 제공하는 49개의 point로 이루어진 landmark vector를 이용하였으며 추가적인 학습 데이터 셋에 대해서는 [36]을 이용하여 landmark를 추출하였다. CNN의 입력으로 49x49의 2D feature를 147x147 size로 3배 up-sampling한다. 또한 비디오 클립에 따라 landmark가 정확히 추출이 되지 않는 예외 경우가 존재한다. 이럴 경우에는 해당 비디오 클립은 학습 데이터로 이용하지 않으며 결과 예측 단계에서도 landmark가 추출되지 않는 clip은 fusion과정에서 해당 network의 score weight를 제거한다.

오디오 기반 네트워크. 앞의 4.1 절에서 언급했듯이 비디오 기반 네트워크에서는 OpensMile 및 Librosa python library를 사용하여 총 577차원의 오디오 feature를 얻을 수 있었다. 이렇게 fine-grained feature를 feature 일반화 과정을 거친 후 각각의 네트워크의 입력으로 사용하였다. 상세 학습 스펙은 3절과 동일하다.

3. 실험 결과

평가 데이터 셋 결과. 1, 2번 째 시도에서는 네트워크의 학습을 AFEW 6.0의 학습 데이터로만 학습하여 평가 데이터 셋으로 개략적인 평가 기준을 설정하였다. 그리고 테스트 데이터 셋에 대한 결과 평가에는 학습(train) 데이터 및 평가(validation) 데이터를 전부 사용하여 학습을 진행하였다. 첫 번째 제출에는 C3DA 네트워크, Parallel CNN 네트워크, 그리고 Audio 1-D CNN 네트워크 의 총 3개의 융합 네트워크를 구성하였다. 평가 데이터 셋 기준 40.21%의 정확도를 달성하였다. 이후 두 번째 제출 시기에는 기존 네트워크의 ensemble 과 파라미터 세팅에 변화를 주어 42.04%를 달성하였다. 3번 째 제출부터 자체 제작한 데이터 셋을 추가하여 네트워크를 학습하였다. S3DAE, Audio 1-D LSTM, 그리고 추가로 만든 데이터 셋의 일부만을 AFEW 6.0의 학습 데이터와 함께 학습하였는데 그 이유는 우리가 추가로 만든 데이터 셋이 실제로 Test Set에 유효한지 판별해 보기 위함이다.

그리고 4번 째 제출 시기에는 Parallel CNN 네트워크의 deep feature와 [27]의 deep feature를 동시에 사용하는 방식으로 Parallel 네트워크의 감정 분류과정을 진행하여 평가 데이터 셋 기준 47.78%의 결과를 달성하였다. 마지막으로 전체 추가 데이터 셋을 사용하여 네트워크를 학습시키고 최종적으로 2D Landmark 네트워크와 Audio DNN의 추가를 통해 최종 융합 네트워크를 구성하여 평가 데이터 셋 기준 50.39%의 정확도를 달성하였고, baseline 대비 약 10 퍼센트의 성능 향상을 달성하였다.

테스트 셋 결과. 5번째 EmotiW 대회에서는 총 7번의 제출 기회가 주어진다. 앞선 절에서 설명한 대로 우리는 평가 데이터 셋을 사용하여 네트워크의 개략적인 평가를 진행하였고, 이를 토대로 테스트 데이터 셋에 대한 네트워크 학습을 진행하였다. 테스트 시기에는 AFEW 6.0 의 학습 데이터 뿐만 아니라 평가 데이터도 함께 사용하여 학습하였다. 그리고 3번 째 제출 시기부터는 추가로 구성한 데이터 셋까지 사용하여 학습을 진행하여 기존의 학습 데이터에 비해 2배 이상의 데이터를 사용하여 학습을 진행하였다. 1, 2번 째 시기에서 테스트 데이터에 대한 정확도는 각각 46.55%, 47.32%로 평가 데이터 시기에서 확인한 상승폭과 유사한 경향을 확인하였다. 그리고 3번 째 시기 이후에는 각각 50.84%, 52.06%, 57.12%의 정확도를 달성하였고, 이는 평가 데이터 셋 의 정확도에 비해 큰 상승폭을 이루었다.

이에 대한 결과에 대한 이유는 크게 두 가지를 들 수 있다. 첫째, 우리의 추가 데이터 셋 구성이 테스트 데이터 셋

표 2. 각 네트워크에 대한 클래스 별 감정 인식 정확도
Table 2. Per-class emotion recognition accuracy for each network

Network	Ang.	Dis.	Fear	Hap.	Neu.	Sad	Sur.
S3DAE	51.56	2.50	2.17	17.46	58.73	32.79	4.35
C3DA	46.88	0.00	4.35	49.20	65.08	9.83	34.78
Parallel CNN	45.31	5.00	4.35	57.14	73.02	19.67	19.56
Audio DNN	73.44	0.00	10.87	52.38	50.79	18.03	2.17
Audio CNN	60.94	7.50	28.26	34.92	42.86	14.75	10.87
Audio LSTM	54.69	0.00	26.09	66.67	50.79	14.75	2.17
Landmark	57.37	0.00	0.00	35.71	17.42	28.26	3.33
Total network	**81.25**	**0.00**	**21.74**	**76.19**	**85.71**	**27.87**	**23.91**

	Ang.	Dis.	Fea.	Hap.	Neu.	Sad	Sur.
Ang.	82.81	0.00	0.00	4.69	7.81	0.00	4.69
Dis.	12.50	0.00	2.50	27.50	35.00	10.00	12.50
Fea.	21.74	0.00	21.74	13.04	26.09	6.52	10.87
Hap.	4.76	0.00	0.00	76.19	19.05	0.00	0.00
Neu.	3.17	0.00	0.00	7.95	85.71	3.17	0.00
Sad	27.87	1.64	0.00	0.00	39.34	27.87	3.28
Sur.	15.22	0.00	0.00	13.04	43.48	4.35	23.91

(a)

	Ang.	Dis.	Fea.	Hap.	Neu.	Sad	Sur.
Ang.	68.37	1.02	2.04	3.06	22.45	2.04	1.02
Dis.	15.00	7.50	0.00	7.50	55.00	15.00	0.00
Fea.	12.86	0.00	42.86	7.14	21.43	2.85	12.86
Hap.	5.56	0.00	0.00	67.36	24.30	2.08	0.70
Neu.	6.22	0.00	4.14	9.84	75.13	4.15	0.52
Sad	11.25	1.25	11.25	7.50	33.75	35.00	0.00
Sur.	17.86	0.00	17.86	10.71	39.29	3.57	10.71

(b)

	Ang.	Dis.	Fea.	Hap.	Neu.	Sad	Sur.
Ang.	74.50	0.00	3.10	3.10	18.40	1.00	0.00
Dis.	20.00	0.00	5.00	15.00	40.00	20.00	0.00
Fea.	27.10	0.00	34.30	1.4	21.40	15.70	0.00
Hap.	6.90	0.00	0.70	82.60	6.30	3.50	0.00
Neu.	8.30	0.00	2.10	5.7	69.90	14.00	0.00
Sad	12.50	0.00	7.50	13.80	25.00	41.30	0.00
Sur.	21.40	0.00	21.40	7.10	35.70	14.30	0.00

(c)

그림 6. AFEW 6.0에 대한 confusion matrices (a) 평가 (b) 테스트 데이터 셋 (c) 비슷한 구조의 네트워크를 사용한 다른 팀의 테스트 데이터 셋 결과
Fig. 6. Confusion matrices from the results at the 5th submission (a) validation (b) test dataset (c) Test dataset results from another team using a similar network

의 비디오 클립을 타겟으로 제작되어서 평가 데이터 셋에는 강인하게 대응하지 못했지만 테스트 데이터 셋에서는 효과를 발휘했다는 점이다. 두번째, 3번 째 제출 시기부터 여러 가지 네트워크의 융합을 통하여 네트워크의 시너지 효과를 달성하였다는 점이다. 이와 같이 우리의 네트워크 및 데이터 셋에 대한 이해를 기준으로 우리는 5번 째 제출까지 점진적인 성능 향상을 달성하였다. 각각의 제출 시기에 따른 최종 정확도 및 네트워크에 대한 설명은 표 2에 정리하였다.

그림 6의 (a), (b)를 바탕으로 우리는 검증 데이터 집합에서 잘 인식되지 않았던 슬픔과 두려움 비디오 클립을 잘 분류하였다. 이를 통해 우리의 Audio 기반 네트워크가 비디오 내 인물의 표정 변화 이외의 정보를 효율적으로 분석하였다는 것을 확인 할 수 있다. 또한 우리와 비슷한 네트워크 구조를 가진 논문[40]의 결과를 그림 6의 (c)에서 확인할 수 있다. 우리의 결과와 비교하면 행복, 화남과 같이 좀 더 역동적인 감정 표현에 강인함을 확인할 수 있다.

Ⅳ. 결 론

제안 기법은 video의 이미지 정보 및 음성 정보를 다양한 방식의 딥 러닝 학습을 적용하여 Emotion recognition in the Wild (EmotiW) 2017 Challenge의 비디오 기반 감정 인식의 문제를 해결하였다. 제안 네트워크는 기존의 대회 논문에서 다루지 않았던 Semi-supervised learning을 통하여 네트워크를 학습하였고, robust feature인 landmark 정보를 2D feature로 변환하여 CNN의 입력으로 이용하였다. 또한

표 3. 총 5번의 평가 및 테스트 데이터 셋 실험 결과
Table 3. Validation and Test dataset recognition accuracy of 5th submission

Submission #	Validation (%)	Test (%)	Method
1	40.21	46.55	C3DA + Audio CNN + Parallel CNN Network
2	42.04	47.32	(Submission 1) + network ensemble & parameter setting
3	46.21	50.84	(Submission 2) + S3DAE + Audio LSTM + (a part of) our own training dataset
4	47.78	52.06	(Submission 3) + Parallel CNN Network with xception
5	50.39	57.12	(Submission 4) + Landmark-based + Audio DNN + our own training dataset

기존의 단순한 음성 신호 분석에 그치지 않고 다양한 네트워크를 통해 효율적으로 음성 신호를 분석하였다. 각 network의 class adaptive fusion을 통해 인식 성능을 향상시켰으며 그 결과 Validation Set에서 50.39%, Test Set에서 57.12%의 정확도를 달성하였다.

참 고 문 헌 (References)

[1] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, "EEG-based emotion recognition in music listening," *Proceeding of IEEE Transactions on Biomedical Engineering*, 57(7), pp.1798-1806, 2010.

[2] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks. *Proceeding of the 18th ACM International Conference on Multimodal Interaction*, pp.445-450, 2016, doi:10.1145/2993148.2997632.

[3] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: towards robust emotion recognition in the wild," *Proceeding of the 18th ACM International Conference on Multimodal Interaction*, pp.472-478, 2016, doi:10.1145/2993148.2997639.

[4] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," *Proceeding of IEEE Computer Society Conference on* (Vol. 1), pp.886-893, 2005, doi:10.1109/CVPR.2005.177.

[5] T. Ojala, M. Pietikainen, and D. Marwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," *Proceeding of the 12th IAPR International Conference on*, pp.582-585, 1994, doi:10.1109/ICPR.1994.576366.

[6] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern recognition, 29(1), pp.51-59, 1996, doi:10.1016/0031-3203(95)00067-4.

[7] C. Cortes, and V. Vapnik, "Support-vector networks," Machine learning, pp.273-297, 1995.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, " Imagenet: A large-scale hierarchical image database," *Proceeding of the IEEE conference on Computer Vision and Pattern Recognition*, pp.248-255, 2009.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *In Advances in neural information processing systems*, pp.1097-1105, 2012.

[10] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *In Advances in neural information processing systems*, pp.396-404, 1990.

[11] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition". arXiv preprint arXiv:1409. 1556, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceeding of the IEEE conference on computer vision and pattern recognition*, pp.770-778, 2016.

[13] G. Huang, Z. Liu, K. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *Proceeding of the IEEE conference on computer vision and pattern recognition*, 2017.

[14] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, pp.169-200, 1992.

[15] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," 2012, doi:10.1.1.407.4632.

[16] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *Proceeding of the IEEE Transactions on pattern analysis and machine intelligence*, pp.97-115, 2001.

[17] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, Z.,I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *Proceeding of the IEEE conference on Computer Vision and Pattern Recognition Workshops*, pp.94-101, 2010.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *Proceeding of the IEEE international conference on computer vision*, pp.4489-4497, 2015.

[19] S. Bargal, E. Barsoum, C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," *Proceeding of the 18th ACM International Conference on Multimodal Interaction*, pp.433-436, 2016, doi:10.1145/2993148.2997627.

[20] X. Zhu, "Semi-supervised learning literature survey". Computer Science, University of Wisconsin-Madison, 2(3), 4, 2006.

[21] L. Wang, C. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," arXiv preprint arXiv:1505.02496, 2015.

[22] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5325-5334, 2015.

[23] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," arXiv preprint arXiv:1706.05098, 2017.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, ... and A. Rabinovich, "Going deeper with convolutions,". *Proceeding of the IEEE conference on computer vision and pattern recognition*, pp.1-9, 2015.

[25] A. Dhall, R. Goecke, S. Ghosh, J. Hoshi, J. Hoey, T. Gedeon, "From Individual to Group-level Emotion Recognition: EmotiW 5.0", *Proceeding of the 18th ACM International Conference on Multimodal Interaction (in press)*, 2017.

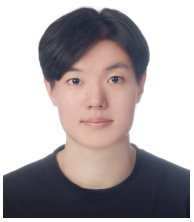
[26] S. Zagoruyko, and N. Komodakis, "Wide residual networks," arXiv preprint arXiv:1605.07146, 2016.

[27] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceeding of the IEEE conference on computer vision and pattern recognition*, pp.1251-1258, 2017.

[28] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," *Proceeding of the IEEE Conference on*

Computer Vision and Pattern Recognition, pp.1859-1866, 2014.
[29] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," *Proceeding of the IEEE International Conference on Computer Vision*, pp.2983-2991, 2015.
[30] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," *Proceeding of the 18th ACM International Conference on Multimodal Interaction*, pp.458-463, 2016.
[31] F. Eyben, M. Wöllmer, B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceeding of the 18th ACM international conference on Multimedia*, pp.1459-1462, 2010.
[32] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," *Proceeding of the 14th python in science conference*, pp.18-25, 2015.
[33] F. Chollet, Keras, http://keras. io, 2015.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceeding of the IEEE international conference on computer vision*, pp.1026-1034, 2015.
[35] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *In International Conference on Machine Learning*, pp.448-456, 2015.
[36] Zhang, Kaipeng et al. "Joint face detection and alignment using multitask cascaded convolutional networks," *Proceeding of IEEE Signal Processing letters*, pp.1499-1503, 2016.
[37] Li, Xi, et al. "DeepSaliency: Multi-task deep neural network model for salient object detection," *Proceeding of IEEE Transactions on Image Processing*, pp.3919-3930, 2016.
[38] Rasmus, Antti, et al."Semi-supervised learning with ladder networks," *Advances in Neural Information Processing Systems*, 2015.
[39] S. Laine, and T. Aila "Temporal Ensembling for Semi-Supervised Learning," arXiv preprint arXiv: 1610.02242, 2016.
[40] V. Vielzeuf, S. Pateux, and F. Jurie. "Temporal multimodal fusion for video emotion classification in the wild." *Proceeding of the 19th ACM International Conference on Multimodal Interaction*, 2017.

저 자 소 개



김 대 하

- 2017년 2월 : 인하대학교 전자공학과 학사 졸업
- 2017년 3월 ~ 현재 : 인하대학교 전자공학과 석사과정
- ORCID : http://orcid.org/0000-0003-3838-126X
- 주관심분야 : 컴퓨터 비전, 인간 감정 인식, 딥러닝

송 병 철

- 1994년 2월 : 한국과학기술원 전기 및 전자공학과 졸업 (공학사)
- 1996년 2월 : 한국과학기술원 전기 및 전자공학과 졸업 (공학석사)
- 2001년 2월 : 한국과학기술원 전기 및 전자공학과 졸업 (공학박사)
- 2001년 3월 ~ 2008년 2월 : 삼성전자 삼성리서치 (구 디지털미디어연구소) 책임연구원
- 2008년 3월 ~ 현재 : 인하대학교 전자공학과 교수
- ORCID : http://orcid.org/0000-0001-8742-3433
- 주관심분야 : 컴퓨터 비전, 딥러닝, 영상 신호처리, 영상 시스템/SoC