CNN 기반 전이학습을 이용한 음성 감정 인식

CNN-based Speech Emotion Recognition using Transfer Learning

저자
(Authors)

최희원, 박승민, 심귀보
Hee Won Choee, Seung Min Park, Kwee-Bo Sim

출처
(Source)

한국지능시스템학회 논문지 29(5), 2019.10, 339-344(6 pages)
Journal of Korean Institute of Intelligent Systems 29(5), 2019.10, 339-344(6 pages)

발행처
(Publisher)

한국지능시스템학회
Korean Institute of Intelligent Systems

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09225683

APA Style

최희원, 박승민, 심귀보 (2019). CNN 기반 전이학습을 이용한 음성 감정 인식. 한국지능시스템학회 논문지, 29(5), 339-344

Journal of Korean Institute of Intelligent Systems
Vol. 29, No. 5, October 2019, pp. 339–344
http://dx.doi.org/10.5391/JKIIS.2019.29.5.339

ISSN(Print) 1976–9172
ISSN(Online) 2288–2324

JKIIS

CNN 기반 전이학습을 이용한 음성 감정 인식

CNN-based Speech Emotion Recognition using Transfer Learning

최희원 · 박승민 · 심귀보†
Hee Won Choee, Seung Min Park†, and Kwee-Bo Sim†

중앙대학교 전자전기공학부
Department of Electrical and Electronics Engineering, Chung-Ang University

요 약

로봇은 사람의 편의를 위해 존재하므로 사람과 로봇의 상호작용은 중요하다. 로봇이 사람의 감정을 파악하는 것은 여러 상호작용 중 하나이다. 최근 사람의 음성으로 감정을 인식하는 음성 감정 인식(speech emotion recognition; SER)분야는 딥러닝 (deep learning)의 접목으로 그 성능이 향상되고 있다. 하지만, 데이터의 부족으로 깊은 신경망을 사용하거나 추가적인 학습 기법을 적용하지 않고서는 높은 정확도를 기대하기 힘들다. 본 논문에서는 데이터가 부족할 때 사용하는 학습 기법 중의 하나인 전이학습 (transfer learning)을 SER에 적용한 효과를 확인한다. 딥러닝을 적용하기 위해 합성곱 신경망 (convolutional neural networks; CNN) 구조를 사용한다. 전이학습에 음성 감정 데이터가 아닌 일반 소리 데이터를 사용하여 데이터 개수에 대한 한계를 없앴다. 전이학습 중 특징 추출기 (feature extractor)로써 사용한 경우와 미세조정 (fine tuning)을 한 경우로 나누어 결과를 확인한다. 그 결과, 미세조정한 경우 수렴 시간이 약 20% 줄었고, 특징 추출기로써 사용한 경우 약 20%에서 70% 줄었다. 정확도는 특징 추출기로써 사용한 경우 오히려 정확도가 감소하는 경우가 발생하였고 증가한 경우 약 3% 증가했다. 미세조정을 한 경우 정확도가 평균적으로 약 7% 향상되었다.

키워드 : 음성 감정 인식, 전이학습, 딥러닝, 합성곱 신경망

Abstract

Interaction between human and robot is important because robots exist for the convenience of people. Robot grasping human emotions is one of many interactions. The field of SER (speech emotion recognition) has been improved by combining deep learning. The lack of data makes it difficult to expect high accuracy without using deep neural networks or applying additional learning techniques. In this paper, we confirm the effect of applying the transfer learning, which is one of the learning methods used when there is insufficient data, to SER. For deep learning, CNN (convolutional neural networks) architecture is used. By using general sound data instead of speech emotion data for the transfer learning, the limit on the number of data is eliminated. The results are verified by dividing transfer learning into two case, using as a feature extractor and fine-tuning. As a result, convergence time was reduced by about 20% when fine-tuning, and about 20% to 70% when used as a feature extractor. Accuracy of the feature extractor is rather reduced when it is used as a feature extractor and increased by about 3% when it is increased. On the average, the accuracy was improved by about 7% when fine-tuning.

Key Words : Speech Emotion Recognition, Transfer Learning, Deep Learning, Convolutional Neural Networks

Received: May. 16, 2019
Revised: Sep. 20, 2019
Accepted: Oct. 4, 2019
†Corresponding authors
kbsim@cau.ac.kr

1. 서 론

오늘날 사람과 의사소통하는 인공지능 (artificial intelligence)의 개발로 의사소통의 중요한 요소인 사람의 감정을 인식하기 위한 연구가 활발히 진행 중이다. 감정인식을 위한 데이터 중에서도 음성 데이터를 이용하는 음성 감정 인식(Speech Emotion Recognition; SER)은 주목받는 연구주제이다. 딥러닝 (deep learning)의 발전으로 SER 또한 발전하고 있다. 하지만, 데이터의 부족으로 높은 정확도를 얻어내기에 어려움이 있다. 데이터 부족을 해결하기 위해서 컴퓨터 비전 (computer vision)영역에서는 전이학습 (transfer learning)이나 데이터 증대 (data augmentation)기법 등을 사용하는 것이 일반적이다. 이 두 기법을 SER에도 적용할 수 있는데 CNN (convolutional neural networks)과 LSTM (long short term memory)을 이용하여 데이터 증대 기법을 적용한 사례가 있다[1]. 하지만 데이터 증대 기법을 사용한 결과 약 1% 내외의 정확도 향상만 있었다. DBN (deep belief networks)을 이용하여 다른 언어 데이터간 전이학습을 적용한 사례도 있다[2]. 이 연구에서는 두 개의 독일어 데이터, 두 개의 영어 데이터, 그리고 한 개의 이탈리아어 데이터를 사용했다. 결과는 흥미롭게도 정확도가 오히려

본 논문은 본 학회 2019년도 춘계학술대회에서 선정한 우수 논문입니다.

감소하는 경우도 있었고, 학습한 데이터가 더 많은 경우가 더 적은 경우보다 정확도가 낮은 경우도 있었다.

본 연구에서는 SER에서의 데이터 부족을 해결하기 위해 CNN을 기반으로 음성 감정 데이터와 특징은 같지만, 레이블(label)이 다른 일반 소리 데이터를 사용한 전이학습을 적용하여 그 성능을 비교한다. 이 방법은 이미지 분류 문제에서 먼저 사용되었다. 이미지 분류 문제에서 데이터 부족을 해결하기 위해 이미지넷(ImageNet)이라는 대량의 일반 데이터를 전이학습 시키는 것은 자주 사용되는 방법이다[3][4]. 본 연구에서 이와 비슷하게 SER에서도 일반 소리 데이터를 전이학습에 사용하는 경우 성능 향상 효과가 있는지 확인해보았다.

CNN에서의 전이학습은 두 가지 학습 방법으로 나눌 수 있다. 첫 번째, 특징 추출기(feature extractor)로 사용하는 것과 두 번째, 전이학습 후 미세조정(fine tuning)을 하는 것이다[5]. 특징 추출기로 사용하는 경우, 마지막 분류 층만 본 데이터로 학습하기 때문에 시간은 단축되지만, 정확도를 높이기에는 무리가 있을 수 있다. 미세조정은 모든 층을 다시 본 데이터로 학습시키기 때문에 시간 단축 효과는 적지만 정확도를 높일 수 있다.

음성 감정 데이터를 CNN에서 학습시키기 위해서는 데이터의 특징을 이미지화하는 과정이 필요하다. 오디오 데이터의 특징 중 스펙트럼 특징을 사용하는 것이 일반적이다. 오디오의 스펙트럼 특징을 이용하여 학습시키는 것은 몇몇 연구에서도 효과가 증명되었다[6][7]. 본 연구에서는 CNN을 사용하므로 스펙트럼을 2D로 나타내기 위해 스펙트럼에 시간 축을 더한 mel spectrogram을 사용한다. SER에서 CNN에 mel spectrogram을 사용하여 효과를 본 연구가 있다[1][8].

2. 이론적 배경

2.1 Transfer Learning

전이학습(transfer learning)은 같은 문제를 해결하는 기존에 학습된 것과 비슷한 데이터를 학습시킬 때, 새로운 학습 과정을 빠르게 하기 위해 고안되었다[9]. 데이터의 레이블은 다르지만, 도메인(domain)이 연관 있는 경우, 처음부터 다시 학습시킬 필요 없이 기존의 가중치를 재사용하는 것이다. 현재까지의 전이학습을 조사한 논문의 내용[10]을 인용하면 도메인은 데이터의 특징이라고 볼 수 있고 과제는 레이블 분류라고 볼 수 있다. 기존 학습에 사용된 데이터는 소스(source)이고 새로 학습시켜야 할 데이터는 타겟(target)이다. 기존 데이터의 특징은 소스 도메인이고 새로운 데이터의 특징은 타겟 도메인이다. 소스 도메인과 타겟 도메인이 같은지와 소스 과제와 타겟 과제가 같은지에 따라 분류할 수 있다.

본 연구에서의 전이학습은 여러 접근 방식 중 네트워크 기반의 접근 방식이다. 소스 도메인으로 학습된 네트워크를 타겟 도메인의 네트워크로 가져오는 것이다. 이런 방식의 전이학습은 두 가지 방식으로 나눌 수 있다.

첫 번째는 특징 추출기(feature extractor)이다. CNN 구조의 마지막 층인 분류층 이전 층들을 소스 데이터로 학습시킨다. 이 층들은 타겟 데이터로 학습하지 않는다. 학습된 층들은 특징 추출기가 되어 타겟 데이터의 특징을 추출한다. 분류층은 타겟 데이터로 학습하여 타겟 레이블에 대한 분류 작업, 즉 타겟 과제를 할 수 있도록 한다.

두 번째는 미세조정(fine tuning)이다. 분류층 이전 층들을 소스 데이터로 학습시킨 후, 다시 한 번 타겟 데이터로 분류층을 포함한 모든 층을 학습시킨다. 이미 학습된 분류층 이전 층들을 타겟 데이터의 특징 추출에 더 적합하게 미세하게 조정하는 것이다.

2.2 Mel Spectrogram

본 연구에서는 1D인 소리 데이터를 2D로 변환할 필요가 있어 소리 데이터를 mel spectrogram으로 변형하여 사용한다. 본 연구에서 사용된 mel spectrogram은 정확히는 log scaled mel spectrogram이다. log scaled mel spectrogram을 얻기 위해 먼저 소리가 없는 부분을 자른다. 즉, 음성 신호가 없는 부분을 제거하는 것이다. 그 후 STFT(short time fourier transform)을 계산한다. STFT는 windowing과 FFT(fast fourier transform)를 해주는 것이다. 식 (1)은 STFT를 정의한 식이다[11].

$$X_m(\omega) = \sum_{n=-INF}^{INF} x(n)\omega(n-mR)e^{-j\omega n} \quad (1)$$

여기서 $x(n)$은 시간 n에 대한 입력 신호이고, $\omega(n)$은 window function이다. window function은 hann function을 사용하였다. Hann function은 식 (2)와 같다. R은 hop size로 연속된 프레임 사이의 샘플 개수이고, 512로 설정하였다. FFT window 크기는 2048로 설정하였다. Power spectrogram을 구하기 위해 식 (3)과 같이 STFT 결과의 크기에 제곱을 해준다.

$$\omega_{hann}(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \quad (2)$$

$$spectrogram\{x[n]\}(m,\omega) \equiv |X_m(\omega)|^2 \quad (3)$$

그 후, 사람의 청각 특성을 반영하여 주파수 영역을 매핑(mapping)해주는 mel filter를 적용한다[12]. mel filter bank는 128개로 설정하였다. 마지막으로 power spectrogram을 데시벨(decibels)로 변환하기 위해 로그(log)변환을 해준다.

그림 1은 가공되지 않은 신호를 나타낸 것이고, 그림 2는 위의 과정을 거쳐 log scaled mel spectrogram이 된 그림이다.

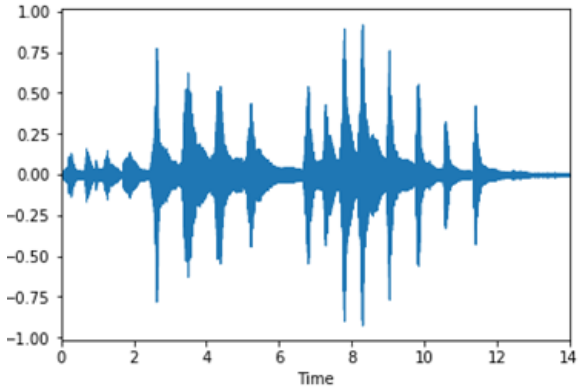

그림 1. 가공되지 않은 신호

Fig. 1. Raw signal

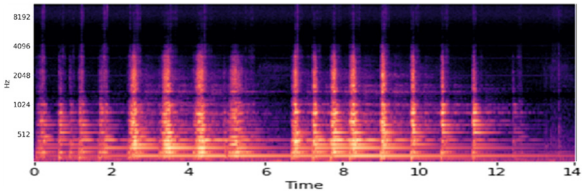

그림 2. 로그 스케일 멜 스펙트로그램

Fig. 2. Log scaled mel spectrogram

3. 데이터 설명

3.1 일반 소리 데이터

일반 소리 데이터는 전이학습 과정 중 사전학습 (pre-training)에 쓰이는 데이터이다. 본 연구에서는 일반 소리 데이터로 DCASE2018 데이터를 사용했다[13]. 이 데이터는 감정 레이블과는 무관한 버스 소리, 트럼펫, 첼로 등 총 41개의 레이블로 구성되어 있다. bit-depth는 16이고, sampling rate는 44.1kHz이다. 테스트용 데이터는 9,400개이고, 학습용 데이터는 총 9,473개이다. 본 연구에서는 학습용 데이터만 사용하였다.

3.2 음성 감정 데이터

본 연구에서는 총 세 개의 음성 감정 데이터를 사용했다. 표 1에 나타낸 것처럼, 첫 번째는 IEMOCAP (interactive emotional dyadic motion capture database)이다[14]. IEMOCAP 데이터는 네 개의 감정 레이블(중립, 행복, 분노, 그리고 슬픔)을 갖는다. 데이터 개수는 총 5,531개이다. 두 번째는 EMO-DB (Berlin database of emotional speech)이다[15]. EMO-DB 데이터는 여섯 개의 감정 레이블(중립, 행복, 분노, 슬픔, 역겨움, 그리고 지루함)을 갖는다. 데이터 개수는 총 466개이다. 그리고 마지막으로, SAVEE (surrey audio-visual expressed emotion)이다[16]. SAVEE 데이터는 네 개의 감정 레이블(중립, 행복, 분노, 그리고 슬픔)을 갖는다. 데이터 개수는 총 480개이다.

표 1. 음성 감정 데이터

Table 1. Speech emotion data

Database	Language	Number of data	Labels
IEMOCAP	English	5,531	Neutral, anger, happiness, sadness
EMO-DB	German	466	Neutral, anger, happiness, sadness, disgust, boredom
SAVEE	English	480	Neutral, angry, happy, sad

4. CNN 기반 전이학습 적용한 성능 비교 방법

4.1 합성곱 신경망 구조

본 연구에 사용된 CNN 구조는 LeNet-5 구조[17]를 참고하여 그림 3과 구현하였다. 특징을 추출하는 convolution layer와 추출한 특징의 크기를 줄이는 pooling layer를 하나의 블록으로 두 블록을 붙인다. 그다음, 앞의 두 블록에 의한 결과를 1D로 정렬하는 fully connected layer 두 개를 쌓은 구조이다. pooling layer에서는 max pooling을 사용하였다. convolution layer와 첫 번째 fully connected layer의 활성화 함수로는 ReLu (rectified linear unit)함수를 사용하였다. 마지막 fully connected layer는 다중 클래스 분류 문제에 많이 쓰이는 softmax 함수를 사용한 분류층이다. 입력 데이터의 크기는 (높이, 너비, RGB channel) 순으로 (100, 300, 3)이다.

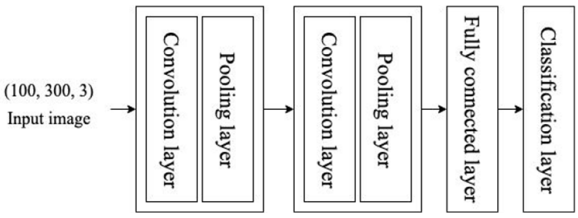

그림 3. 합성곱 신경망 구조

Fig. 3. CNN architecture

CNN 구조의 두 번째 블록과 분류층 이전에 드롭아웃 (dropout)[18]을 적용하였다. 이 기법은 적은 수의 데이터를 학습시킬 때, 오버피팅 (overfitting)을 방지하기 위하여 네트워크 일부 즉, 특징 감지기 (feature detector)를 생략하는 것이다. 본 연구에서 사용하는 음성 감정 데이터의 개수가 적기 때문에 드롭아웃 기법을 적용하였다. 이 기법을 고안한 논문에서는 드롭아웃 비율을 50%로 하였지만 본 연구에서는 데이터의 절반을 생략하기에는 CNN 구조가 층이 깊지 않고 드롭아웃을 두 번을 적용하기 때문에 비율을 20%로 정하였다.

4.2 실험 디자인

본 논문에서는 전이학습을 적용한 경우와 적용하지 않은 경우의 성능을 비교한다. 전이학습을 적용한 경우는 특징 추출기로

사용하는 경우와 미세조정을 하는 경우로 나뉜다. 세 경우 모두 IEMOCAP, EMO-DB, 그리고 SAVEE 데이터를 사용했다. 사전학습에는 DCASE2018 데이터를 사용했다. DCASE2018 데이터가 소스가 되고 음성 감정 데이터가 타겟이 된다. 두 데이터 모두 소리 데이터이고 멜 스펙트토그램이라는 같은 특징을 추출하였으므로 소스 도메인과 타겟 도메인이 일치하고 두 데이터의 레이블이 다르다는 점에서 소스 과제와 타겟 과제가 다르다고 할 수 있다. 본 연구의 전이학습은 네트워크 기반의 전이학습이므로 네트워크 층간의 가중치 전이를 하였다. 특징 추출기로 사용하는 경우 분류층 이전까지 DCASE2018 데이터로 학습시킨 후 분류층만 음성 감정 데이터로 학습시켰다. 미세조정을 하는 경우 분류층 이전까지 DCASE2018 데이터로 학습시키고 음성 감정 데이터로 분류층 이전 층을 다시 학습시키고 분류층 또한 학습시켰다. 성능 확인을 위한 테스트 데이터는 학습 데이터와 별개로 무작위로 선정하였다. 각각의 테스트 데이터 개수는 IEMOCAP 2000개, EMO-DB 100개, 그리고 SAVEE 100개이다.

본 연구를 통한 성능 향상을 확인하기 위하여 정확도와 수렴 시간을 계산하였다. 정확도와 수렴 시간을 구하기 위해서 수렴 기준을 정해야 한다. 학습에는 k-fold 교차검증 (cross validation)을 적용하였다. k 값은 4로 정하였다. 가중치 학습은 교차검증에서 발생하는 검증 손실 (validation loss)를 최소화하는 것을 목적으로 하였다. 검증 손실이 5번의 배치 학습 동안 감소하지 않으면 이미 수렴하였다고 간주하고 학습을 중단하였다. 학습을 중단한 지점을 기준으로 가장 작은 검증 손실 값을 갖는 시점을 수렴 시점으로 정하여 그때까지의 시간을 수렴 시간으로 하였다. 그 수렴 시점의 가중치를 그 학습의 수렴 가중치로 하여 테스트 데이터로 평가하여 정확도를 계산하였다.

모든 실험은 아나콘다 (anaconda)와 주피터 노트북 (jupyter notebook) 환경에서 진행되었다. 실험에 사용한 딥러닝 프레임워크 (framework)는 텐서플로우 (tensorflow)이고, 실험에 사용된 GPU는 NVIDIA Geforce 940MX 8GB이다.

5. 시뮬레이션 및 결과

본 연구에서는 SER에서의 전이학습의 효과를 확인하기 위하여 전이학습을 적용하는 경우와 적용하지 않는 경우로 나누어 실험을 진행하였다. 데이터를 CNN에 학습시키기 위하여 멜 스펙트로그램을 구하고 이를 이미지로 변환하였다. 그 결과, 그림 4, 5와 같은 결과가 나왔다. 미세조정을 한 경우는 모든 경우 정확도가 향상하였다. IEMOCAP은 0.95%, EMO-DB는 14%, SAVEE는 7% 상승하였다. 특징 추출기로 사용한 경우 EMO-DB 데이터로 실험한 결과 3% 증가한 것 이외에는 정확도가 평균 약 5% 감소하였다. 사전학습에 쓰인 DASE2018 데이터 개수에 비해 상대적으로 적지 않은 IEMOCAP 데이터를 사용한 경우 미세조정한 경우 0.95% 증가하고 특징 추출기로 사용한 경우 2.7% 감소하여 증감 폭이 다른 데이터에 비해 작았다. 수렴 시간은 특징 추출기로 사용한 경우와 미세조정을 한 경우 모두 감소하였다. 특징 추출기로 사용한 경우가 미세조정을 한 경우보다 수렴 시간이 더 감소하였다. EMO-DB에서 특징 추출기 실험의 경우 수렴 시간이 약 4분의 1로 감소하였다.

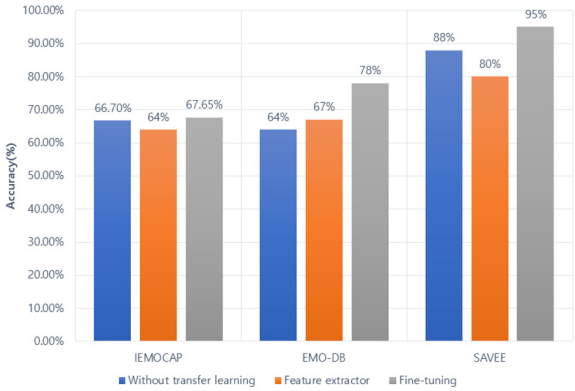

그림 4. 정확도 결과

Fig. 4. Accuracy results

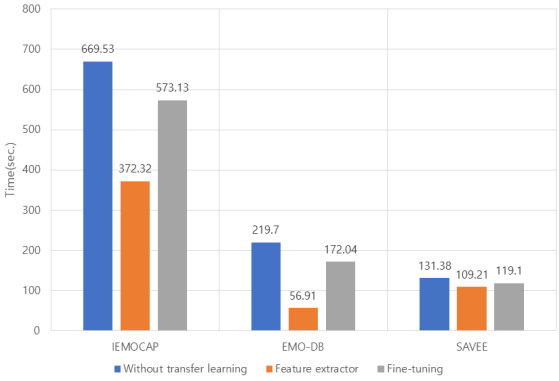

그림 5. 수렴 시간 결과

Fig. 5. Results of time to converge

6. 결론 및 향후 연구

본 논문에서는 음성 감정 인식(SER)의 성능 향상을 위하여 전이학습의 사전학습에 일반 소리 데이터인 DCASE2018을 사용하여 SER의 성능을 향상시키는 방법을 제안하였다. 제안한 방법의 성능 평가를 위해서 본 논문에서는 3개의 다른 음성 감정 데이터 (IEMOCAP, EMO-DB, SAVEE)를 사용하여 실험하였으며, 정확도와 수렴 시간을 서로 비교하여 그 성능을 평가하였다. 그 결과, 특징 추출기만을 사용한 경우는 수렴 시간은 감소하였지만, 정확도가

약간 감소하였고, 미세조정을 한 경우에는 정확도가 향상되고 수렴 시간도 줄어들어 전체적으로 성능이 향상되었다. 실험결과를 통하여 알 수 있는 바와 같이 일반 소리 데이터로 사전학습을 하여 전이학습을 적용하는 것이 SER의 성능을 향상시킨다는 것을 알 수 있었다. 하지만, 특징 추출기만을 사용하는 경우 정확도가 감소하는 경우가 발생하여 전이학습을 사용하려면 미세조정을 하는 것이 더 좋은 결과를 얻을 수 있었다. 이미지넷을 SER의 사전학습에 사용한 연구에서도 전이학습 중 미세조정을 하여 성능이 개선된 보고도 있다[4].

본 연구에서 사전학습에 사용한 데이터 개수는 5,531개로 이미지넷의 데이터 개수에 비하면 훨씬 적은 개수다. 만약 사전학습에 사용될 일반 소리 데이터가 더 확보된다면 지금보다 성능이 더욱더 향상될 것으로 기대된다. 학습에 CNN을 사용했는데 멜 스펙트로그램은 시간 축을 가지고 있기 때문에 시간에 따른 변화를 사용하는 LSTM을 추가로 적용한다면 성능이 더 향상될 것으로 기대된다.

References

[1] C. Etienne, G. Fidanza, A. Pettrovskii, "Speech Emotion Recognition with Data Augmentation and Layer-wise Learning Rate Adjustment", *arXiv preprint arXiv:1802.05630*, 2018.

[2] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, "Transfer Learning for improving speech emotion classification accuracy", *arXiv preprint arXiv:1801.06353*, 2018.

[3] M. Oquab, L. Bottou, I. Laptev, J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks," *IEEE conference on computer vision and pattern recognition*, pp. 1717-1724, 2014

[4] H. W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 443-449, 2015

[5] CS231n Convolutional Neural Networks for Visual Recognition, "Transfer Learning and Fine-tuning Convolutional Neural Networks", Available: http://cs231n.github.io/transfer-leraning/, 2017, [Accessed: Jan. 2019]

[6] Q. Mao, M. Dong, Z. Huang, Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," *IEEE Transactions on Multimedia*, vol. 16, no 8, pp. 2203-2213, Dec. 2014

[7] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011

[8] A. Satt, S. Rozenberg, R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," *Proc. Interspeech 2017*, pp. 1089-1093, 2017

[9] L. Y. Pratt, "Discriminability-Based Transfer between Neural Networks," NIPS Conference: Advances in Neural Information Processing Systems 5, pp. 204-211, 1993

[10] S. J. Pan, Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2010

[11] J. B. Allen, L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558-1564, 1977

[12] S. S. Stevens, J. Volkmann, E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, vol. 8, Issue 3, pp. 185-190, 1937

[13] E. Fonseca, M. Plackal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, X. Serra, "General-purpose Tagging of Freesound Audio with Audioset Labels: Task Description, Dataset, and Baseline," *DCASE2018 Workshop*, 2018

[14] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Recsources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech," *Proc. Interspeech*, vol. 5, pp. 1517-1520, Sep. 2005

[16] S. Haq, P. J. B. Jackson, "Multimodal Emotion Recognition," *In W. Wand(ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press*, ISBN 978-1615209194, chapter 17, pp. 398-423, 2010 2014

[17] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gadient-Based Learning Apllied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998

[18] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, "Improving neural networks by preventing co-adaption of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012

저 자 소 개

최희원(Hee Won Choee)

2016년~현재 : 중앙대학교 전자전기공학부 학사과정

관심분야 : Machine Learning, Deep Learning, Intelligent System

E-mail : choee97@naver.com

박승민(Seung Min Park)

2010년 : 중앙대학교 전자전기공학부 공학사

2019년 : 중앙대학교 대학원 전자전기공학과 석박사통합과정 공학박사

관심분야 : Machine Learning, Brain-Computer Interface, Deep Learning, Robotics, Intelligent System

E-mail : sminpark@cau.ac.kr

심귀보(Kwee-Bo Sim)

1990년 : The University of Tokyo 전자공학과 공학박사

1991년~현재 : 중앙대학교 전자전기공학부 교수

2006~2007년 : 한국지능시스템학회 학회장

관심분야 : BCI(뇌-컴퓨터 인터페이스), 감정인식, 의도인식, 유비쿼터스 지능형로봇, 지능시스템, 컴퓨테이셔널 인텔리전스, 유비쿼터스 컴퓨팅 및 센서 네트워크, 소프트 컴퓨팅(신경망, 퍼지, 진화연산), 인공면역시스템, 지능형 감시시스템, 패턴인식, 기계학습, 사물인터넷(IoT), 빅데이터, 딥러닝 등.

E-mail : kbsim@cau.ac.kr