Regular paper | Journal of the KIECS. pp. 685-692, vol. 19, no. 4, Aug. 31. 2024, t. 126, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.4.685

스펙트로그램을 이용한 CNN 음성인식 모델

정원석* · 이행우**

Speech Recognition Model Based on CNN using Spectrogram

Won-Seog Jeong* · Haeng-Woo Lee**

요 약

본 논문에서는 명령어 음성신호의 인식 성능을 개선하기 위한 새로운 합성곱 신경망(CNN: Convolutional Neural Network) 모델을 제안한다. 이 방법은 입력신호의 단구간 푸리에 변환(STFT: Short-Time Fourier Transform) 후 스펙트로그램 이미지를 구하고 CNN 모델을 이용한 지도학습을 통하여 명령어 인식 성능을 개선하였다. 입력신호를 단시간 구간별로 푸리에 변환한 다음 스펙트로그램 이미지를 구하고 CNN 딥러닝 모델을 이용하여 다중 분류 학습을 수행한다. 이는 시간영역 음성신호를 특성이 잘 표현되도록 주파수영역으로 변환하고 변환 파라미터에 대한 스펙트로그램 이미지를 이용하여 딥러닝 훈련을 수행함으로써 명령어를 효과적으로 분류한다. 본 연구에서 제안한 음성인식시스템의 성능을 검증하기 위하여 Tensorflow와 Keras 라이브러리를 사용한 시뮬레이션 프로그램을 작성하고 모의실험을 수행하였다. 실험 결과, 제안한 심층학습 알고리즘을 이용하면 92.5%의 정확도를 얻을 수 있는 것으로 확인되었다.

ABSTRACT

In this paper, we propose a new CNN model to improve the recognition performance of command voice signals. This method obtains a spectrogram image after performing a short-time Fourier transform (STFT) of the input signal and improves command recognition performance through supervised learning using a CNN model. After Fourier transforming the input signal for each short-time section, a spectrogram image is obtained and multi-classification learning is performed using a CNN deep learning model. This effectively classifies commands by converting the time domain voice signal to the frequency domain to express the characteristics well and performing deep learning training using the spectrogram image for the conversion parameters. To verify the performance of the speech recognition system proposed in this study, a simulation program using Tensorflow and Keras libraries was created and a simulation experiment was performed. As a result of the experiment, it was confirmed that an accuracy of 92.5% could be obtained using the proposed deep learning algorithm.

키워드

Speech Recognition, Deep Learning, Spectrogram, Short Time Fourier Transform, Convolutional Neural Network
음성 인식, 심층 학습, 스펙트로그램, 단구간 푸리에 변환, 합성곱 신경망

* ㈜AICube
** 교신저자 : 남서울대학교 지능정보통신공학과
• 접 수 일 : 2024. 06. 03
• 수정완료일 : 2024. 07. 08
• 게재확정일 : 2024. 08. 12
• Received : Jun. 03, 2024, Revised : Jul. 08, 2024, Accepted : Aug. 12, 2024
• Corresponding Author : Haeng-Woo Lee
Dept. Intelligent Information and Communication Engineering, Namseoul University
Email : hwlee@nsu.ac.kr

Ⅰ. 서 론

음성인식시스템은 특정 요구사항을 해결하기 위해 인간과 기계의 상호작용을 이용하는 방식으로 음성 패턴을 식별하고 변환하기 위해 컴퓨팅 도구에서 사용하는 기술 및 알고리즘을 나타낸다. 음성인식시스템은 특히 의료, 로봇공학, 홈자동화 기술 등 다양한 용도와 응용분야에서 다양성과 기능성으로 인해 특별한 관련성이 많으며 이러한 유형의 인터페이스를 처리하는 장치를 점점 더 정확하고 다루기 쉽게 되었다. 음성인식 모델은 일반적으로 음성신호 획득, 신호 전처리, 음성신호 패턴인식 및 분류의 단계를 거친다. 신뢰할 수 있는 오차 범위를 달성하려면 음성 패턴의 특징을 추출하는 기술을 사용해야 한다. 가장 많이 사용되는 기술은 Mel 계수에서 얻은 스케일로그램, 주로 화자를 식별하는데 사용되는 MFCC(: Mel Frequency Cepstral Coefficients), 오디오 샘플의 볼륨 변화에 민감한 웨이블릿 변환, 음성 패턴에 대한 주파수 영역의 정보를 제공하는 푸리에 변환 등이다[1-2]. 음성인식 시스템의 장점은 많은 양의 어휘를 어려움 없이 처리할 수 있고, 처리시간이 짧아 사용자가 성능을 만족스럽게 평가할 수 있다는 것이다.

음성인식시스템의 상업적 활용 중에는 애플의 Siri, 마이크로소프트의 Cortana, 구글의 Assistant 및 Now, 아마존의 Alexa, 그리고 삼성의 Bixby와 같은 가상 비서 등이 있다. 또한 기본 응용 프로그램 중에는 고립된 단어 및 연결된 숫자와 같은 음성인식 도구 개발을 위한 Villamil, 짧은 문장 인식, 의학에서의 다양한 적용 등 음성인식은 적용대상이 광범위하며 학제적인 연구분야가 되었다. 현재는 가정 내에서 언어 프로세스를 이해하고 실행하는 과정을 통해 자동화와 편리성을 지원할 수 있기 때문에 음성인식시스템이 필요한 곳이 바로 홈오토메이션분야이다. 본 연구에서는 단구간 푸리에 변환으로부터 스펙트로그램(spectrogram) 이미지를 이용하여 CNN 모델을 기반으로 명령어 분류용 음성인식시스템을 설계하고 구현한다.

논문의 내용은 Ⅱ절에서 스펙트로그램과 CNN 모델에 대해 살펴보고, Ⅲ절에서는 딥러닝 명령어 분류시스템의 구조를 설명한다. 그리고 Ⅳ절에서 이 시스템에 대한 시뮬레이션 및 그 결과에 대하여 기술하고, 끝으로 Ⅴ절에서 결론을 도출한다.

Ⅱ. 스펙트로그램과 CNN 모델

스펙트로그램은 음성 신호의 주파수 성분을 시간에 따라 표현한 그래프이다. 이를 위해 단구간 푸리에 변환을 사용한다. 합성곱 신경망(CNN)은 이미지 처리에 탁월한 성능을 보이는 신경망으로 음성 정보를 인식하는 데에도 활용된다. 이 연구에서는 스펙트로그램 이미지를 이용한 CNN 기반 명령어 인식 기법을 수행한다. 이 방법은 명령어에 대한 주파수상의 특징 벡터를 효과적으로 추출하기 위해 STFT를 사용하며, STFT로 검출된 특징 벡터들은 스펙트로그램 이미지로 변환되어 CNN을 이용해 명령어별로 분류된다. 이 방법은 효과적으로 명령어를 인식할 뿐만 아니라 소리 기반의 다양한 자동화 시스템에도 활용될 수 있다.

푸리에 변환은 디지털화된 시계열 신호가 다양한 진폭과 주파수를 가진 여러 삼각함수들의 중첩으로 이루어져 있다고 간주하고, 해당 삼각함수들의 주파수 및 각 주파수별 진폭의 분포를 분석하는 것에 중점을 두고 있다. 해당 분포는 시간에 따라 계속 변하지만, 아주 짧은 시간 동안 거의 일정하다(quasi-stationary)고 간주하며, 이를 응용하여 해당 시간 간격을 창함수(window function)화하여 주파수 분포를 구하고, 이 분포가 시간에 따라 변하는 양상을 스펙트럼으로 시각화하는 기법이다. 즉 특정 파동을 여러 개의 조각으로 잘게 잘라 조각 하나하나를 주파수별 여러 순수 삼각함수로 분리한 데이터이다.

스펙트로그램은 시간상 진폭 축의 변화를 시각적으로 볼 수 있는 파형과 주파수상 진폭 축의 변화를 시각적으로 볼 수 있는 스펙트럼의 특징이 모두 결합된 구조로, 시간축과 주파수축 상의 진폭의 차이를 색상으로 나타내는 것이다. 보통 RGB 3색이라면 파랑-초록-빨강 순으로 신호가 강하게 나타난다. 창함수의 길이, 창함수의 중첩 길이, 창함수의 종류를 파라미터라 부르며, 각 파라미터를 어떻게 조정하느냐에 따라 스펙트로그램은 다양하게 변화한다. 이는 음향학 및 음성학에서 매우 많이 쓰이며, 특히 음성인식을 위한 말소리 분석에 필수적으로 이용된다.

Mel 스펙트럼은 소리의 대표적인 특징으로 딥러닝을 이용한 음성인식 및 오디오 분류 문제에서 많이 사용된다. 오디오 신호에 STFT를 수행하면 주파수를 x축으로 하는 스펙트럼이 생성되고, 이때 y축의 크기

를 제곱하면 전력 스펙트럼이 된다. 여기에 log 스케일을 적용하여 데시벨 단위로 변환한 것을 Log 스펙트럼이라고 하며 이것을 세로로 세워서 프레임마다 쌓으면 스펙트로그램이 된다. Mel 스펙트로그램은 주파수는 mel 주파수로, 전력은 log 크기로 나타낸다. 다음 그림은 음성에서 추출한 Mel 스펙트로그램이다.

인공신경망(ANN: Artificial Neural Network)은 딥러닝의 기초가 되는 신경망[3-6]으로 인간의 신경망 구조에 기반한다. 딥러닝의 한 종류인 합성곱 신경망 즉, CNN은 패턴이나 물체를 인식하는 시각처리과정을 모방한 것이다. CNN에는 컨볼루션 레이어가 있어 합성곱 연산을 이용하여 특징(feature)을 추출한다. 이 과정을 통해 입력 이미지의 공간적인 정보를 잃지 않는다. 즉, 이미지에서 인접한 픽셀들의 정보를 잃지 않아 입력 데이터의 특징을 파악할 수 있다. CNN은 사용자의 의도에 따라 컨볼루션 레이어와 풀링 레이어를 중복 배치하여 심층 신경망을 구성하고 종단에 배치된 완전연결층을 통하여 분류작업을 한다. 입력 데이터는 컨볼루션 계층에서 특징을 추출하고 풀링 계층에서 파라미터 개수를 줄여 데이터의 잡음이나 왜곡을 해소하는 효과를 얻는다. 다중 분류에서는 종단의 완전연결층에서 항목별로 확률값을 계산하는 softmax 함수를 활성화 함수로 사용한다. CNN은 주로 이미지 학습에 사용되는 신경망[7-11]이다. 그림 1에서와 같이 본 연구에서는 이미지가 아닌 명령어 즉, 오디오에서 추출한 특징인 Mel 스펙트럼을 학습하는 오디오 분류에 대한 실험을 수행한다.

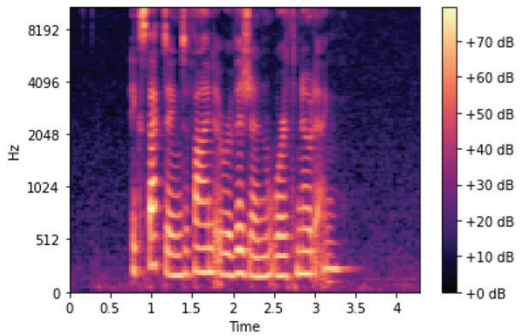


그림 1. 스펙트로그램 이미지
Fig. 1 Spectrogram image

딥러닝 모델은 그림 2와 같은 CNN 구조를 이용한다. 입력은 일정 구간 음성에 대한 스펙트로그램 이미지로서 각각 16, 32, 64, 128 개의 커널(kernel)로 이루어진 4개의 컨볼루션 레이어를 거치게 된다. 각 합성곱 계층 다음에는 풀링 계층(pooling layer)을 배치하여 과적합을 방지하도록 한다. 그다음 1차원 평탄화 과정을 거친 후 2개의 완전연결층을 통과한다. 완전연결층 사이에도 과적합을 방지하기 위해 50% dropout을 실시하며 최종단에는 softmax 활성화함수를 사용하여 다중분류를 수행한다.

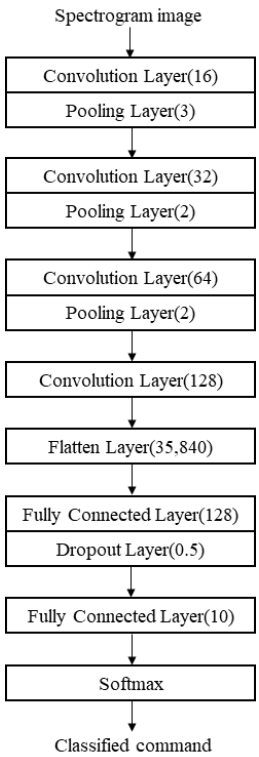


그림 2. CNN 구조
Fig. 2 CNN structure

III. 딥러닝 명령어 분류시스템

딥러닝 명령어 분류과정에서는 음성을 명확하고 간결하게 기술하는 특징 벡터를 사용하는 것이 효과적이다. 따라서 스펙트로그램 이미지를 기반으로 한 음성인식과정은 그림 3과 같이 여러 기능블록으로 이루

어진다. 첫 번째 블록은 음성신호의 샘플링 주파수, 비트폭 등과 같은 기본 특성을 갖는 음성신호를 생성하는 단계이다. 두 번째 블록에서는 단구간 음성신호에 대한 해밍창 적용, 정규화, 평균화 등 전처리가 이루어진다. 세 번째 블록에서는 음성신호의 전처리에서 얻은 단구간 데이터들에 대해 이산푸리에변환을 수행하여 스펙트로그램 이미지를 생성하고 학습 데이터 세트를 구축한다. 네 번째 블록에서는 CNN을 사용하여 스펙트로그램 이미지로부터 특징을 추출하는 과정이 이루어진다. CNN을 사용하면 이미지 학습 중에 중요한 특징을 자동으로 검색할 수 있으므로 특징 추출은 CNN을 기반으로 수행한다. 각 CNN 계층마다 풀링 레이어가 추가되어 특징 데이터를 단축화한다. 그리고 다섯 번째 블록에서는 1차원 평탄화를 거친 후 완전연결층인 FCN(: Fully Connected Layer) 출력에 대해 softmax 함수를 이용하여 테스트 세트의 스펙트로그램 이미지를 분류한다.

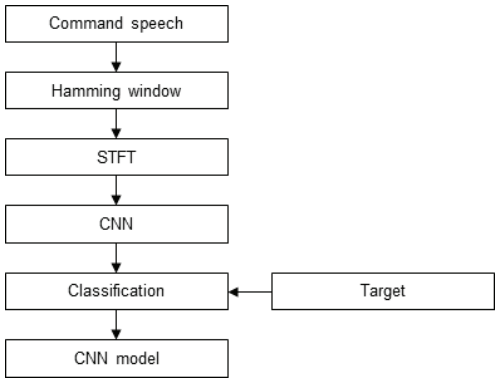


그림 3. 명령어 분ㅊ류 학습과정
Fig. 3 Filter bank of dyadic wavelet transform

명령어 분류시스템은 그림 4와 같이 구성된다. 입력 데이터는 명령어에 대한 스펙트로그램 이미지와 목표값으로 해당 명령어 분류번호가 주어진다. 각 입력신호의 단구간 푸리에변환을 계산하기 위해 신호를 음성의 통계적 특성이 변하지 않는 16ms 구간으로 나누고 256 샘플에 대해 해밍(Hamming) 윈도우 함수와 곱한다. 그다음 이 블록들에 대해 푸리에변환을 하고 전구간 통합하여 2차원 스펙트로그램 이미지를 얻는다. CNN 학습은 정확도가 최고점에 도달할 때까지 수십 회의 에포크(epoch) 동안 반복 수행되며 학습이 종료되면 모델 파라미터가 산출된다.

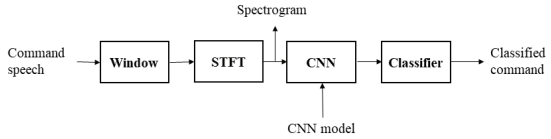


그림 4. CNN 명령어 분류시스템
Fig. 4 CNN command classification system

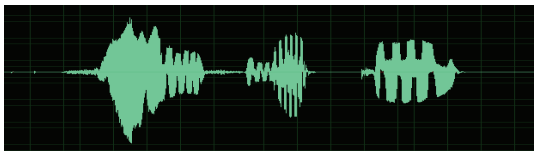
본 명령어 분류시스템을 구현하는데 가장 적합한 사양이 표 1에 나열되어 있다. 이 값들은 여러 시뮬레이션을 통해 가장 성능이 우수한 것으로 선정되었다.

표 1. 명령어 분류시스템의 사양
Table 1. Spec. of commands classification system

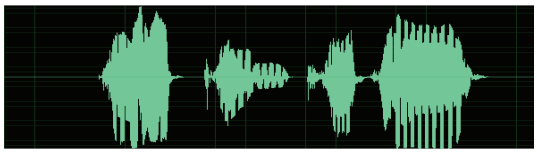
Specifications	Values
Sampling frequency	16 kHz
Data resolution	16 bit
Transform window	Hamming
Window length	192
Window overlap	96
Number of CNN layer	4 layers
Optimization algorithm	Adam
Number of epoch	100
Command length	1 sec
Number of commands	> 10

Ⅳ. 모의실험 및 분석

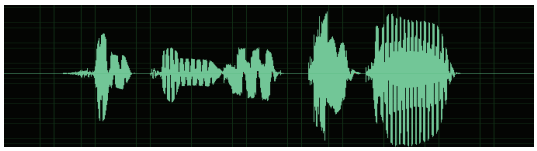
본 논문에서 제안한 명령어 분류시스템의 성능을 검증하기 위해 Tensorflow와 Keras 라이브러리를 이용하여 시뮬레이션 프로그램을 작성하였다. 입력신호는 '하이 스마트' '에어콘 켜줘' '선풍기 켜줘' '선풍기 꺼줘' 등 4개 음성신호가 사용되며 각 음성마다 16-bit, 16kHz로 샘플링된 123개 데이터가 제공된다. 이 시스템은 다중 분류모델의 지도학습에 해당하며 입력 음성신호의 일부 파형을 그림 5에서 보여주고 있다.



(a) 하이 스마트



(b) 에어콘 켜줘



(c) 선풍기 켜줘

(d) 선풍기 꺼줘

그림 5. 명령어의 다양한 입력 파형
Fig. 5 Waveforms of the command

구현방법의 성능을 평가하기 위하여 분류 정확도(accuracy)를 사용하였다. 정확도는 목표값인 명령어로의 분류 성공 여부를 나타낸다. 그림 6에서 훈련(train) 데이터와 확인(validation) 데이터에 대한 분류 정확도 및 손실 특성을 보여주고 있다.

100 에포크 동안 시뮬레이션 결과, 초기 20 에포크 이내에 정확도가 최대값까지 급격히 증가하고 이후에도 지속적으로 최대값 부근을 유지하며, 이때 평균 정확도는 92.5%를 달성하는 것으로 나타났다. 반면에 손실은 초기 20 에포크 이내에 최소값까지 빠르게 감소하며, 여기서 훈련 데이터와 확인 데이터의 곡선이 큰 차이 없이 비슷한 특성을 나타낸다.

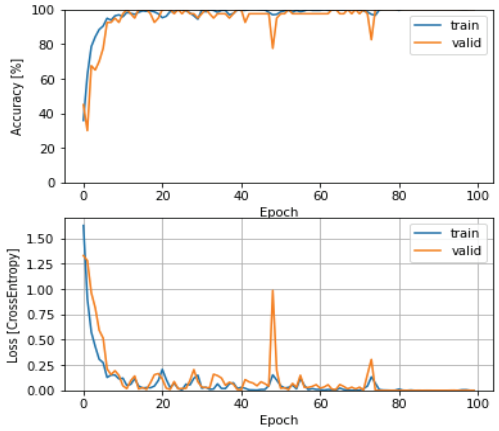


그림 6. 명령어의 정확도 및 손실 학습특성
Fig. 6 Curve of accuracy and loss for commands

그림 7은 명령어들 간 혼동(confusion) 행렬의 평가를 보여주고 있다. 예상되는 바와 같이 '선풍기 켜줘' 명령어와 '선풍기 꺼줘' 명령어 간 구별에서 약간의 혼동이 발생하는 것을 볼 수 있다.

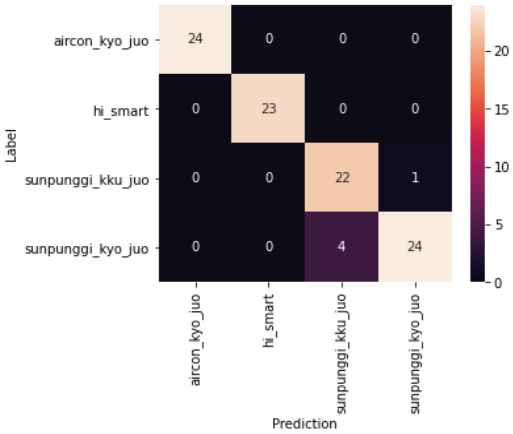


그림 7. 명령어들 간 혼동행렬 평가
Fig. 7 Evaluation of confusion matrix for commands

Ⅴ. 결 론

명령어 음성의 인식 성능을 개선하기 위하여 우수한 명령어 인식시스템의 개발이 요구되고 있다. 본 논문에서는 스펙트로그램 이미지와 딥러닝 기술을 적용한 새로운 명령어 분류시스템을 제안하였다. 음성 데

이터를 푸리에 변환하여 얻은 스펙트로그램 이미지에 대하여 CNN 신경망을 이용한 심층학습으로 음성인식 성능을 향상시킬 수 있다.

본 음성인식시스템은 입력신호의 단구간 푸리에 변환 후 스펙트로그램 이미지를 구하고 CNN 다중 분류 지도학습을 통하여 상당한 성능 개선을 달성하였다. 연구 결과, 제안한 시스템은 92.5%의 정확도 성능을 나타내는 것으로 확인되었다.

References

[1] D. Kim, A. Lee, G. Lee, S. Kim, and B. Lee, "Kiosk for the visually impaired using voice recognition," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 5, 2022, pp. 873-882. https://doi.org/10.13067/JKIECS.2022.17.5.873

[2] D. Jang, and J. Kim, "Two-way interactive algorithms based on speech and motion recognition with Generative AI technology," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 19, no. 2, 2024, pp. 397-402. https://doi.org/10.13067/JKIECS.2024.19.2.397

[3] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, *Handwritten digit recognition with a back-propagation network.* Burlington,: Morgan Kaufmann Publishers Inc., 1990.

[4] S. Lawrence, C. Giles, A. Tsoi, and A. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Network*, vol. 8, 1997, pp. 98-113. https://doi.org/10.1109/72.554195

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, 2015, pp. 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *In Proceedings of the Advances in Neural Information Processing Systems 2012, Lake Tahoe, NV, USA*, Dec. 2012, pp. 3-8. https://doi.org/10.1145/3065386

[7] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 22, 2014, pp. 1533-1545. https://doi.org/10.1109/TASLP.2014.2339736

[8] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," In *Proceedings of the 25th ACM International Conference on Multimedia,* Mountain View, CA, USA, Oct. 2017, pp. 23-27. https://doi.org/10.1145/3123266.3123371

[9] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE,* Calgary, AB, Canada, Apr. 2018, pp. 15-20. https://doi.org/10.1109/ICASSP.2018.8461870

[10] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," *In Proceedings of the 2018 International Conference on Digital Health,* Lyon, France, Apr. 2018, pp. 23-26. https://doi.org/10.1145/3194658.3194671

[11] X. Liu, J. van de Weijer, and Bagdanov, "A.D. Exploiting unlabeled data in cnns by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, 2019, pp. 1862-1878. https://doi.org/10.1109/TPAMI.2019.2899857

저자 소개

정원석(Won-Seok Jeong)

1991 세명대학교 전자공학과(공학사)
2023년 수원대학교 글로벌창업대학원 창업경영학과(석사)

2006년~2017년 ㈜테크엘 IoT사업본부 수석연구원
2017년~2019년 ㈜비엔컴 신사업 총괄 이사
2021년~현재 ㈜AICube 대표이사
※관심분야 : IT 모듈제작, 키오스크, 배경잡음 제거

이행우(Haeng-woo Lee)

1985 광운대학교 전자공학과(공학사)
1987년 서강대학교 대학원 전자공학과(공학석사)

2001년 전북대학교 대학원 전자공학과(공학박사)
1987년~1998년 한국전자통신연구원 선임연구원
2001년~현재 남서울대학교 지능정보통신공학과 교수
※관심분야 : VLSI 설계, 딥러닝, 음향잡음 소거