JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 1: 27~32, Jan. 2019

소프트맥스를 이용한 딥러닝 음악장르 자동구분 투표 시스템

배준[1] · 김장영[2*]

Deep Learning Music genre automatic classification voting system using Softmax

June Bae[1] · Jangyoung Kim[2*]

[1]Ph.D. Student, Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea
[2*]Assistant Professor, Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

요 약

인간이 가진 뛰어난 능력 중의 하나인 곡 분류 과정을 딥러닝 알고리즘을 통해 구현하는 연구는 단일데이터를 이용한 유니모달 모델, 멀티모달 모델, 뮤직비디오를 이용한 멀티모달 방식 등이 있다. 이 연구에서는 곡의 스펙트로그램을 짧은 샘플들로 분할하여 각각을 CNN으로 분석한 뒤 그 결과를 투표하는 시스템을 제안하여 더 좋은 결과를 얻었다. 딥러닝 알고리즘 중 CNN이 RNN에 비해 음악 장르 구분에 있어 우수한 성능을 보였으며 CNN과 RNN을 같이 적용했을 때 성능이 좋아짐을 알 수 있었다. 음악샘플을 나누어 각각의 CNN 결과를 투표하는 시스템이 이전 모델에 비해 좋은 결과를 나타내었고 이 모델에 Softmax 레이어를 추가한 모델이 가장 좋은 성능을 보였다. 디지털 미디어의 폭발적인 성장과 수많은 스트리밍 서비스 속에서 음악장르의 자동분류에 대한 필요는 점점 증가하고 있는 추세이다. 향후 연구에서는 미분류 곡의 비율을 낮추고 최종적으로 미분류된 곡들의 장르구분에 대한 알고리즘을 개발할 필요가 있을 것이다.

ABSTRACT

Research that implements the classification process through Deep Learning algorithm, one of the outstanding human abilities, includes a unimodal model, a multi-modal model, and a multi-modal method using music videos. In this study, the results were better by suggesting a system to analyze each song's spectrum into short samples and vote for the results. Among Deep Learning algorithms, CNN showed superior performance in the category of music genre compared to RNN, and improved performance when CNN and RNN were applied together. The system of voting for each CNN result by Deep Learning a short sample of music showed better results than the previous model and the model with Softmax layer added to the model performed best. The need for the explosive growth of digital media and the automatic classification of music genres in numerous streaming services is increasing. Future research will need to reduce the proportion of undifferentiated songs and develop algorithms for the last category classification of undivided songs.

키워드 : 딥러닝, 소프트맥스, 음악장르구분, CNN, RNN

Key word : Deep learning, Music Classification, Softmax, CNN, RNN

Received 30 October 2018, **Revised** 12 November 2018, **Accepted** 20 November 2018

*** Corresponding Author** Jangyoung Kim (E-mail: jykim77@suwon.ac.kr, Tel: +82-31-229-8345)
Assistant Professor, Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.1.27 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

인간이 가진 능력 중에 뛰어난 것 중에 하나가 음악을 분류하는 것이라고 할 수 있을 것이다. 몇 초 안에 우리는 지금 듣고 있는 음악이 클래식인지 랩, Rock, 혹은 EDM(Electronic Dance Music) 인지 말할 수 있다. 스마트폰 등에 수많은 곡들을 저장해 놓을 수 있는 디지털 미디어 장비에 일반인들은 평균 7,000 여개 정도의 노래들 보관하고 있다고 한다. 노래 하나의 장르를 구분하는데 약 3초가 걸린다고 한다면 모든 노래들의 장르를 구분하는데 약 6시간이 걸릴 것이다. 또한 수작업으로 일일이 노래의 태그를 수정한다고 한다면 최소 10시간 이상의 작업이 필요할 것이다.

딥러닝을 이용한 음악장르 구분법에 대한 기존연구는 단일데이터를 입력으로 하는 유니모달 모델, 멀티모달 딥러닝 방식, 뮤직 비디오를 이용한 멀티모달 방식 등이 있다 [1].

이 연구에서는 음악의 스펙트로그램을 여러개의 샘플로 나누어 각각의 CNN(Convolutional Neural Network) 결과를 투표 시스템을 이용해 곡의 장르를 구분하는 방법을 제안한다. 그리고 Softmax 레이어 방식을 추가해 딥러닝 곡 분류의 성능을 높이는 방식에 대해 알아보기로 한다.

Ⅱ. 관련연구 및 기존연구 문제점

음악 장르 구분에 대한 연구는 MFCC, Spectral Centroid 등 신호 처리 이론을 바탕으로 한 음악의 특성을 추출해내는 방식이 주를 이루었다. 이런 추출 특성을 머신러닝을 이용해 장르를 구분하는 방법이 연구되었다. SVM(Support Vector Machine) 방식이 주를 이루었고 KNN(K-Nearest Neighbors), GMM(Gaussian Mixture Model) 등 여러 방식이 이용되었다.

최근에는 머신러닝의 발전으로 입출력이 하나의 모델에서 이루어지는 end to end model 형태의 딥러닝 연구가 이루어지고 있다. 여러 계층을 가진 Deep Neural Network 구조의 딥러닝은 모델 자체에서 특성을 추출, 분류를 하는 방식이다.

1990년대 인공신경망 모델은 역전파(Back propatation) 알고리즘을 이용한 것으로 큰 기대를 모았으나 은닉계층이 많아지면 성능이 떨어지는 단점이 있었다. 2000년대 이후 모델이 개선되고 GPU의 성능이 비약적으로 발전되면서 다시 주목받고 있다. 큰 문제점이었던 과적합 문제를 드롭아웃(dropout) 등의 방법으로 정규화 성능을 개선했다[2].

CNN은 1990년대 후반 필기인식 연구에서 시작되었다. 인간의 시각처리를 구현하여 이미지 처리에 알맞은 모델이다. 최근 연구로 모델의 레이어를 늘려 정확도를 늘렸고 특히 영상인식분야에서 뛰어난 성능을 발휘한다.

RNN(recurrent neural network)은 순차 데이터 처리에 이용되는 모델로 Deep neural network의 중간 레이어 값을 재귀해 사용해 데이터 특성을 순차적으로 추출하는데 뛰어난 성능을 보여준다. 현재 LSTM(Long short term memory)와 GRU(Gated Recurrent Unit) 등 모델이 발전되어 음성인식 및 자연어 처리에 이용되고 있다 [3].

기존 딥러닝을 이용한 음악 장르 분류 학습 모델은 음악의 스펙트로그램을 이용한 CNN 방식과 음악의 순차적 시그널 데이터를 입력으로 하는 RNN 방식, 그리고 이 두 방식을 결합한 멀티모달 방식이 있다. 이 방식들은 음악 장르를 구분하는데 있어서 변화가 적은 짧은 음악을 분류하는 데는 효과적이나 하나의 곡 안에서 변화가 많은 곡을 분류하는 데는 적합하지 않을 수 있다. 왜냐하면 한 곡 안에서 클래식처럼 조용한 부분과 록처럼 강한 부분이 같이 있다면 그 곡의 시작부분 30초 정도를 분석하는 것만으로는 장르 분류에 오류가 있을 수 있다.

이러한 오류를 줄이기 위하여 전체 곡을 작은 샘플로 나누고 각각의 샘플을 CNN 분석하여 그 결과들의 총합으로 장르 구분을 하는 투표 시스템을 제안한다.

Ⅲ. 딥러닝 음악 장르 분류 투표 시스템 모델

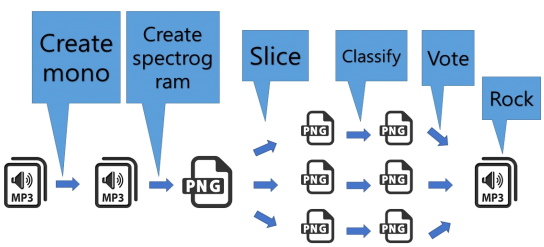

Fig. 1 Deep Learning Music genre automatic classification voting system flow chart (proposed model overview)

딥러닝 음악 장르 분류 투표 시스템은 우선 음원의 데이터량을 줄이기 위해 스테레오파일을 모노파일로 만든 후 이를 스펙트로그램으로 변환하여 PNG 파일로 만든다. 이를 여러개의 일정한 크기의 스펙트로그램으로 분할하여 각각을 CNN 분석을 하고 그 결과별 투표를 하여 그 곡의 장르를 결정하는 방식이다. 이 시스템의 세부사항은 다음 항에서 설명하고 있다 (그림 1).

3.1. 데이터셋 정리

너무나 다양한 음악 장르가 있기 때문에 각 음악의 서브 장르를 통합하여 몇 개의 대표 장르로 구분하는 작업이 필요하다. 클래식을 예로 들면 교향곡, 실내악 등을 구분하지 않고 크게 클래시컬 음악으로 통합하여 구분하였다. 음악을 6개의 장르- 하드코어, 록, 일렉트로, 클래식, 재즈 그리고 랩- 으로 나누어 각 장르별 음악을 수집한 뒤 그 음원으로부터 중요한 정보를 추출하기 시작했다. 노래는 아주 긴 연속된 값의 집합이다. 일반적 샘플링 레이트는 44,100Hz로 이는 1초당 44,100개의 값이 담겨있다는 것이다. 스테레오 사운드의 경우는 이것의 2배의 값을 가지게 된다 [4].

그렇다면 3분짜리 스테레오 노래는 7,938,000 개의 샘플값을 가지게 되는데 이것은 매우 큰 정보량으로서 우리는 이것을 다룰 수 있을 정도로 줄여야만 한다, 우선 스테레오 채널 중 하나만 택함으로써 많은 양의 정보를 줄일 수 있다.

다음에 Fourier's Transform을 이용해 오디오 데이타를 프리퀀시로 변환하여 스펙트로그램(spectrogram)으로 바꿔준다(1). 이는 시간의 흐름에 따른 모든 주파수의 변화를 PNG 파일로 나타내준다 [5].

$$x(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{N2\pi k}{N} n, K=0,\ldots,N-1} \quad (1)$$

$$\hat{X} = \overline{F} X$$

$$X = \frac{1}{m} F \hat{X}$$

다음에 Nyquist-Shannon sampling theorem에 따라 44100Hz의 샘플링 레이트를 22050Hz로 다시 만들어준다. 이는 더 적은 해상도로 스펙트로그램을 만들 수 있게 해준다[6].

여기서 우리는 초당 50 픽셀의 스펙트로그램을 사용했는데 이는 장르구분에 충분한 데이터 량이다 [7].

다음은 프로세싱을 거친 후의 스펙트로그램의 예이다(그림 2).

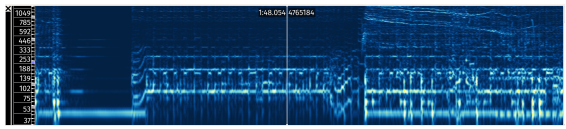

Fig. 2 Spectrogram of a song (X:Time,Y:Frequency)

X축은 시간, Y 축은 프리퀀시를 나타낸다. 위로 갈수록 높은 주파수, 아래로 갈수록 낮은 주파수를 표시한다, 이 실험에서는 128개의 주파수 단계의 스펙트로그램을 사용했는데 이는 장르구분에 필요한 정보- 음정의 차이와 주파수 차이-를 충분히 갖고 있다.

3.2. 데이터 프로세싱

다음으로 다뤄야 할 문제는 노래의 길이이다. 이 문제에 대하여 2가지 접근방식이 있다. 하나는 데이터를 순차적으로 입력해나가는 RNN으로 음악은 시간에 따른 순차 데이터라고 볼 수 있다. RNN은 과거 정보를 회귀하는 신경망으로 시간 레이어가 많아질수록 지난 정보가 잘 보관되지 않는 vanishing gradient로 인해 학습에 문제가 생기는 단점이 있다 [8].

대신에 여기서는 인간처럼 짧은 구절만으로도 노래를 분류할 수 있는 방법을 연구해보기로 한다. 인간이 3초안에 음악을 분류할 수 있다면 왜 컴퓨터는 할 수 없는가?

다른 하나의 방법은 스펙트로그램을 일정한 길이로 잘라서 각 조각을 장르를 표시하는 개개의 샘플로 생각한다. 편의를 위해 128X128 픽셀의 정사각형으로 조각을 자르면 개개의 조각은 2.56초의 데이터를 가지게 된다. 이 데이터의 패턴은 일정하게 유지되어야하는데, 이미지를 회전시키거나 뒤집거나 하는 가공은 불가하다 [9]. 소리의 스펙트로그램은 대칭적 이미지가 아니기 때문이다 (그림 3).

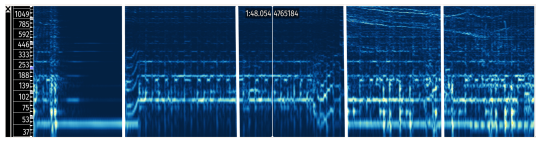

Fig. 3 Divided Spectrogram

3.3. 1 차 음악장르 분류 모델 (그림 4)

모든 곡들을 정사각형의 스펙트로그램 이미지로 만든 후 각 음악장르별로 수십만 개의 데이터셋을 가지게 된다. 이제 Deep Convolutional Neural Network 를 이용해 이 샘플들을 분류하도록 교육시키는데 Tensorflow's wrapper TFLearn를 사용한다. 이미지 데이터 학습에 큰 발전을 가져온 CNN은 데이터를 이미지로 만들고 여기에서 특성을 추출한다. 전형적인 CNN은 convolution 레이어, subsampling 레이어, 그리고 완전 연결 레이어로 이루어져 있다. convolution 레이어는 CNN의 핵심으로서 배움이 가능한 필터들로 이루어져 특성을 추출해내고 subsampling 레이어는 비선형 다운 샘플링(down sampling) 을 통해 2차원 구조의 입력데이터를 효율적으로 분석할 수 있게 한다. 이러한 두 계층을 반복하여 이미지 데이터가 갖고 있는 특성을 잘 추출해내는 것이 가능하다[10].

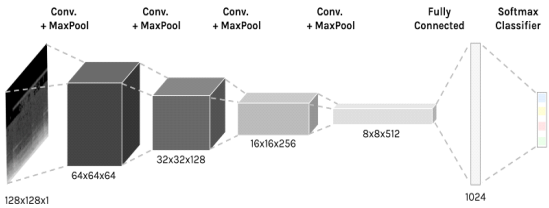

Fig. 4 CNN Structure [11]

3.4. 1차 실험 결과

6개의 장르- 하드코어, 록, 일렉트로, 클래식, 재즈 그리고 랩- 으로 나누어진 2,000개의 노래와 12,000개의 128X128 픽셀의 스펙트로그램 조각들을 이용한 결과 이 모델은 90%의 정확도를 나타내었다. 이것은 노래의 작은 조각들을 사용한 것을 고려했을 때 상당히 좋은 결과라고 할 수 있다. 하지만 이 결과는 조각들의 분류에 대해 이야기하고 있는 것으로 전체 노래를 분류한 것이 아니다.

3.5. 투표 시스템 음악 장르 분류 모델 (그림 5)

이렇게 교육시킨 모델을 가지고 새로운 곡에 적용시켜 보기 위해 다음과 같은 과정을 거친다. 먼저 트레이닝 데이터와 같이 스펙트로그램을 만들고 작은 조각으로 나눈다. 작은 조각들로 나뉘어서 한 번에 전체 곡의 장르를 예측할 수 없기 때문에 '투표' 시스템을 만든다.

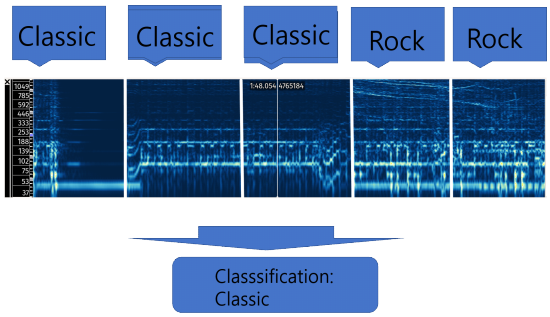

Fig. 5 music classification voting system

그 곡의 각각의 샘플들이 장르에 대해 투표를 하고 가장 많은 표를 받은 장르를 그 곡의 장르로 선택하는 시스템이다. 이로서 장르 예측의 정확도를 높일 수 있었다.

이 시스템으로 새로운 곡의 장르 예측이 가능해졌지만 여기서 투표 시스템을 더욱 발전시키기 위해 다음과 같은 수정을 가했다.

3.6. 소프트맥스 레이어를 이용한 투표시스템 개선

소프트맥스(Softmax) 함수는 여러 개의 클래스를 구분할 때 마지막 뉴련의 활성화 함수로 시그모이드를 사용하면 출력값을 공정하게 평가하기 어려울 때 사용한다. 소프트맥스는 뉴런의 출력 값에 지수함수를 적용하되 모든 뉴런에서 나온 값으로 정규화하는 형태를 가진다. 이런 이유로 멀티 클래스 분류인 경우 소프트 맥스 함수를 이용할 때가 많다[12].

$$P_{Rock} = \frac{e^{z}_{Rcok}}{e^{z}_{Jazz} + e^{z}_{Classic} + e^{z}_{Rap}} = \frac{e^{z}_{Rcok}}{\sum_{i=1}^{n} e^{z}}, \quad (2)$$

$$z = w \times x + b$$

장르 구분 시스템의 마지막 레이어에 소프트맥스 레이어를 추가해서 시스템이 장르를 지정하기 보다는 그 가능성을 표시할 수 있도록 하였다. 이를 '분류 확신도 '라고 명명한다 (2). 이를 투표 시스템을 개선하는데 사용하였다. 예를 들면 낮은 확신도의 조각을 투표에서 제외하였다. 또한 투표에서 확실한 승자가 없으면 전체 투표 자체를 무효화하였다 (그림 6).

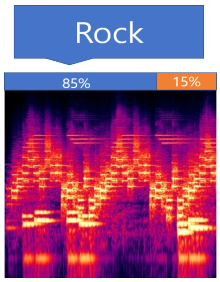

Fig. 6 music classification softmax system

정확도 향상을 위하여 70% 미만의 투표를 받은 곡은 장르 구분에서 제외시켜 장르 미분류곡으로 남겨두었다 (그림 7).

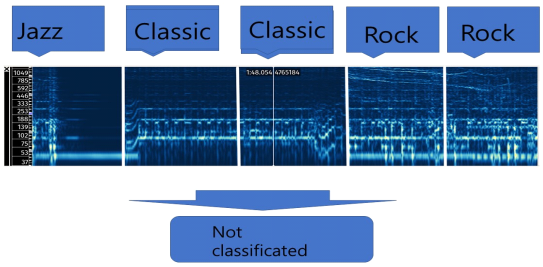

Fig. 7 Truncation of music classification

Ⅳ. 실험결과

분류 확신도에 대한 각 모델들의 실험결과를 표로 나타내면 다음과 같다 (표 1). 이 실험에서는 6개의 각 장르별 - 하드코어, 록, 일렉트로, 클래식, 재즈 그리고 랩- 500개의 총 3,000개의 노래를 스펙트로그램으로 변환하고 128X128 픽셀로 분할하여 34,200개의 스펙트로그램 조각들로 만들었다. 이를 다음 알고리즘들을 이용하여 분류하였고 다음과 같은 결과를 얻었다.

Table. 1 Classification Confidence Rate Comparison

Model	Classification Confidence Rate (%)
CNN	67
RNN	54
CNN+RNN	68
VOTING SYSTEM (suggested model)	73
VOTING SYSTEM+SOFTMAX (suggested model)	76

CNN이 RNN에 비해 음악 장르 구분에 있어 우수한 성능을 보였으며 CNN과 RNN을 같이 적용했을 때 성능이 좋아짐을 알 수 있었다. 음악샘플을 나누어 각각의 CNN 결과를 투표하는 시스템이 이전 모델에 비해 좋은 결과를 나타내었고 이 모델에 소프트맥스 레이어를 추가한 모델이 가장 좋은 성능을 보였다.

Ⅴ. 결론

디지털 미디어의 폭발적인 성장과 수많은 스트리밍 서비스 속에서 음악장르의 자동분류에 대한 필요는 점점 증가하고 있는 추세이다[13]. 이 논문에서는 기존 CNN, RNN 등을 이용한 연구를 발전시켜 소프트맥스 레이어를 이용한 투표 시스템으로 곡 장르분류의 확신도를 높이는 모델을 만들어 실험 결과 기존 방식에 비해 우수한 결과를 도출해내었다.

기존 방식은 곡들의 일정부분만을 딥러닝하여 음악장르를 구분하였기 때문에 곡의 구성이 복잡한 음악의 장르구분에는 약한 면을 보였으나 소프트맥스 레이어를 이용한 투표 시스템은 곡의 모든 부분을 일정하게 나누어 각각의 딥러닝 결과를 투표하여 곡의 장르를 결정하는 시스템으로 장르구분의 확신도를 높였다.

향후 연구에서는 미분류 곡의 비율을 낮추고 최종적으로 미분류된 곡들의 장르구분에 대한 알고리즘을 개발할 필요가 있을 것이다.

ACKNOWLEDGEMENT

The paper was supported by The research grant of the University of Suwon in 2017.

REFERENCES

[1] S. Kim, D. Kim, and B. Suh, “Music Genre Classification using Multimodal Deep Learning,” *International Journal of Information and Communication Engineering*, vol. 9, no. 4, pp. 358-362, Aug. 2011.

[2] Potla Revathi, “Analytical Hierarchy Process in Fuzzy Comprehensive Evaluation Method,” *Asia-pacific Journal*

of Convergent Research Interchange, vol.1, no.3, pp. 41-52, September 2015.

[3] B. Macfee, "Learning Content Similarity for Music Recommendation," *Journal of latex class files*, vol. 6, no. 1, pp. 1-2, Jan. 2017.

[4] D Cabrera, "A Computer Program for Psycho-acoustical Analysis," *Australian Acoustical Society Conference*, vol. 24, no. 1, pp. 47-54, Mar. 2014

[5] J. C. Na, "Optimization in Cooperative Spectrum Sensing," *Asia-pacific Journal of Convergent Research Interchange*, vol. 3, no. 1, pp. 19-31, March 2017.

[6] D. J. Kim, and P. L. Manjusha, "Building Detection in High Resolution Remotely Sensed Images based on Automatic Histogram-Based Fuzzy C-Means Algorithm," *Asia-pacific Journal of Convergent Research Interchange*, vol. 3, no. 1, pp. 57-62, March 2017.

[7] T. S. Slininger, Y. Xu, and R. D. Lorenz. "Enhancing estimation accuracy by applying cross- correlation image tracking to self-sensing including evaluation on a low saliency ratio machine," *Energy Conversion Congress and Expositionvol*, vol. 22, no. 5, pp. 23-28, May 2016.

[8] L. Maaten, and G. Hinton, "Learning Content Similarity for Music Recommendation Visualizing Data using T-SNE," *Journal of Machine Learning Research*, vol. 9, no. 1, pp. 2579-2605, Nov. 2008.

[9] J. Bae, and J. Kim, "Engine Sound Design for Electric Vehicle by using Software Synthesizer," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 21, no. 8, pp 1547-1552, Aug. 2017.

[10] V. K. Rao, R. Caytiles, "Subgraph with Set Similarity in a Database," *Asia-pacific Journal of Convergent Research Interchange*, vol. 3, no. 2, pp. 29-37, Jun. 2017.

[11] Aphex34, Own work, CC BY-SA 4.0 [Internet]. Available: https://commons.wikimedia.org/w/index.php?curid=45679374.

[12] B. Han, S. Rho, S. Jun, and E. Hwang, "Music emotion classification and context-based music ecommendation," *Multimedia Tools Application*, vol. 47, no. 3, pp. 433-460, May 2010.

[13] J. Bae, J. Kim, and Y. Yang, "Physical modeling synthesizing of 25 strings Gayageum using white noise as exciter," *Journal of the Korea Institute of formation and Communication Engineering*, vol. 22, no. 5, pp. 740-746, May 2018.

배준(June Bae)

연세대학교 정치외교학과 졸업
상명대학교 컴퓨터음악대학원 졸업
수원대학교 컴퓨터학부 박사과정
※관심분야 : 전기차 사운드 디자인, AI 알고리즘 작곡, 머신러닝 플레이리스트 작성, 음성인식, DSP 설계

김장영(Jangyoung Kim)

2005년 2월: 연세대학교 컴퓨터과학 공학사
2010년 5월: Pennsylvania State Univ. 공학석사
2013년 7월: State University of New York 공학박사
2013년 8월: University of South Carolina 교수
2014년 3월: 수원대학교 컴퓨터학부 교수
※관심분야 : Big data, Cloud computing, Networks