Journal of the Korea Convergence Society
Vol. 10. No. 9, pp. 23-29, 2019

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2019.10.9.023

심층 CNN을 활용한 영상 분위기 분류 및 이를 활용한 동영상 자동 생성

조동희[1], 남용욱[2], 이현창[3], 김용혁[4*]
[1]광운대학교 소프트웨어학부 학생, [2]광운대학교 컴퓨터과학과 박사,
[3]광운대학교 컴퓨터과학과 석사과정, [4]광운대학교 소프트웨어학부 교수

Image Mood Classification Using Deep CNN and Its Application to Automatic Video Generation

Dong-Hee Cho[1], Yong-Wook Nam[2], Hyun-Chang Lee[3], Yong-Hyuk Kim[4*]
[1]Student, Dept. Computer Science, Kwangwoon University
[2]Master's Course, Dept. Computer Science, Kwangwoon University
[3]Ph. D, Dept. Computer Science, Kwangwoon University
[4]Professor, Dept. Computer Science, Kwangwoon University

요 약 본 연구에서는 영상의 분위기를 심층 합성곱 신경망을 통해 8 가지로 분류하고, 이에 맞는 배경 음악을 적용하여 동영상을 자동적으로 생성하였다. 수집된 이미지 데이터를 바탕으로 다층퍼셉트론을 사용하여 분류 모델을 학습한다. 이를 활용하여 다중 클래스 분류를 통해 동영상 생성에 사용할 이미지의 분위기를 예측하며, 미리 분류된 음악을 매칭시켜 동영상을 생성한다. 10겹 교차 검증의 결과, 72.4%의 정확도를 얻을 수 있었고, 실제 영상에 대한 실험에서 64%의 오차 행렬 정확도를 얻을 수 있었다. 오답의 경우, 주변의 비슷한 분위기로 분류하여 동영상에서 나오는 음악과 크게 위화감이 없음을 확인하였다.

주제어 : 융합, 기계학습, 다중 클래스 분류, 감정 분류, 합성곱 신경망, 다층 퍼셉트론

Abstract In this paper, the mood of images was classified into eight categories through a deep convolutional neural network and video was automatically generated using proper background music. Based on the collected image data, the classification model is learned using a multilayer perceptron (MLP). Using the MLP, a video is generated by using multi-class classification to predict image mood to be used for video generation, and by matching pre-classified music. As a result of 10-fold cross-validation and result of experiments on actual images, each 72.4% of accuracy and 64% of confusion matrix accuracy was achieved. In the case of misclassification, by classifying video into a similar mood, it was confirmed that the music from the video had no great mismatch with images.

Key Words : Convergence, Machine Learning, Multi-class Classification, Mood Classification, Convolutional Neural Network, Multilayer Perceptron

*This research was supported by the MIST (Ministry of Science and ICT), under the National Program for Excellence in SW (2017-0-00096), supervised by the IITP (Institute for Information & communications Technology Promotion)
*Corresponding Author : Yong-Hyuk Kim (yhdfly@kw.ac.kr)
Received August 16, 2019 Revised September 5, 2019
Accepted September 20, 2019 Published September 28, 2019

1. 서론

미디어(media)란 인간 사회에서 감정이나 자신의 의사 또는 객관적 정보를 주고받을 수 있도록 마련된 수단을 지칭하는 말이다. 그중 음성과 영상 등으로 이루어진 다양한 정보를 다루는 멀티미디어(multimedia)는 디지털화된 미디어의 복합체라고 할 수 있는데, 정보의 양이 많기 때문에 이를 처리하기가 매우 복잡하고 까다롭다.

따라서 이러한 멀티미디어들을 기계학습 알고리즘으로 분석하고 모델링하여 여러 분야에 적용하는 연구가 활발히 진행되고 있다[1]. 그 중, 사람의 감정을 추출하는 기술로, 어떤 감정을 가졌는지를 판단하여 분석하는 감성 분류(sentimental analysis)를 사용하면, 개인을 위한 추천 알고리즘이나 광고와 같은 다양한 분야에 적용할 수 있다.

본 연구는 심층 합성곱 신경망(convolutional neural network; CNN)과 다중 클래스 분류(multi-class classification)를 사용하여 영상의 분위기를 8 가지로 분류하고, 이에 맞는 배경 음악을 매칭하여 동영상을 생성하는 모델을 제안한다. 본 논문의 구성은 다음과 같다.

2 절에서는 영상 분위기 분류에 대한 선행 연구들에 대하여 정리한다. 3 절에서는 분위기 분류 기준과, 데이터 수집을 위한 웹 크롤링(Web crawling) 및 데이터 전처리에 관해 설명한다. 4 절에서는 본 논문에서 제안하고자 하는 모델에 사용된 기계학습 기법과 파라미터 설계, 그리고 분류 실험에 관해 설명한다. 5 절에서는 실험 결과로 교차 검증과 오차 행렬에 대해 설명한다. 마지막으로 6 절에서는 결론 및 향후 연구에 대해 소개한다.

2. 관련 연구

2.1 감성 분석

딥러닝을 활용하여 영상에서의 얼굴 인식뿐만 아니라, 감성을 분석하는 연구가 다양한 방식으로 진행되어 왔다. Lee 등[2]의 연구에서 다중 스레드(thread) 기술을 이용한 HOG(histogram of gradients) 알고리즘을 이용하여 얼굴 인식과 신원 식별 그리고 인물의 감정을 분석했다. Kim[3]의 연구에서는 기존의 감정 차원 모델 중 가장 대표적인 Russell의 모델[4]에 기반하여 영상의 감성을 분석하고자 했다. Russell의 2 차원 정서 모형은, '각성(arousal) 및 정서가(valence)에 의해 표현될 수 있는데, (arousal + valence +) (arousal + valence −) (arousal − valence +) (arousal − valence −)에 따라 4 가지의 감성 분류가 가능하다. 이를 바탕으로 영상을 일정한 길이의 영상 프레임으로 나누어 CNN과 이진 교차엔트로피를 이용한 감성 분석을 진행했다.

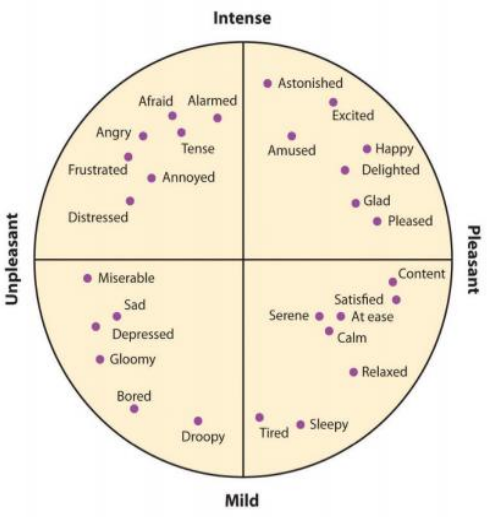

Fig. 1. Russell's valence-arousal model of emotion [4]

2.2 심층 합성곱 신경망

심층 합성곱 신경망은 여러 합성곱과 풀링(Pooling)이 반복적으로 구성된 많은 신경망 층으로 이루어진다. 심층 합성곱 기반의 모델들은 특징적인 표현들을 학습하여 좋은 성능을 보여 주었기 때문에, 이미지 분류, 물체 탐지, 안면인식 등 여러 영역에 적용되어왔다. Ko 등[5]과 Lee[6]의 연구에서는 심층 합성곱 신경망을 사용하여 각각 안면 검증 연구와 화물차의 차종분류 연구를 진행했다. 그리고 Ramadhani 등[7]의 연구에서도 심층 합성곱 신경망을 사용하여 고용 소득을 예측했다.

2.3 동영상의 배경 음악 동기화

기계학습 기반의 연구가 활발히 진행되면서 학습에 필요한 데이터를 수집하는 작업이 중요해지고 있다. 그중에서 음악 데이터에 대해, Lee 등[8]의 연구에서는 수많은 웹 문서를 자동으로 돌아다니며 각종 정보를 수집하는 프로그램인 웹 크롤링을 통해 분위기를 기반으로 음악을 검색하거나 분류하는 방법을 제안한다. Lee[9]의 연구에서는 음악의 파형 특징점과 동영상의 움직임 특징점을 동적계획법 매칭 또는 그래프 순회를 사용하여 동기화하거나 새로운 음악을 생성하는 연구를 진행했다.

3. 데이터

본 연구에서는 8 가지 감정 데이터로 분류하기 위해 Fig. 1의 Russell의 감정 모델을 참고했다[4]. Fig. 1에서 각 사분면의 분위기 중, 대표적인 2 가지를 선별해서 Fig. 2의 분위기 모델을 생성했다. 분위기의 선택 기준은 인간의 감정을 포괄하고, 주로 사용되는 분위기까지 고려하였다. 그리고 Fig. 2의 이웃한 분위기들이 어느 정도 공통된 뉘앙스를 가진다는 점에서 색의 스펙트럼 같은 자연스러운 분위기의 분포를 표현하고자 했다.

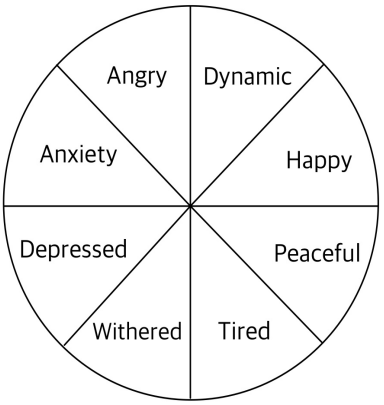

Fig. 2. Mood classification based on Russell's model

분위기 데이터 수집은 공공데이터와 웹 크롤링[10]을 통해 이루어졌다. 일반적으로 이미지 분류에 활용되는 데이터는 본 논문의 목적에 부합하지 않아 훈련 데이터와 테스트 데이터를 직접 크롤링하는 것이 가장 적합하였다. 먼저, JPG, PNG 확장자를 비롯한 이미지 파일을 각 분위기 당 1,200 개, 총 9,600 개의 데이터를 수집했다. 데이터의 전처리 과정에서는 임의의 데이터에 대해 저자 4 명의 판단 기준을 적용했다. 하나의 이미지는 여러 가지 분위기를 포함하고 있기 때문에, 각 분위기와 가장 근접한 이미지만 학습시키고자 했다. 따라서 데이터를 보고 분위기에 대한 의견을 종합했을 때, 4 명 중 3 명 이상의 의견이 일치한 경우에만 학습 데이터로 분류했다. 학습 데이터로 분류된 데이터를 학습 기준에 맞게 128×128 크기로 재조정하는 것을 마지막으로, Fig. 3과 같이 감정마다 400 개씩 총 3,200 개의 학습 데이터를 준비했다.

분류된 분위기에 맞는 음악 데이터는 유튜브(YouTube)에서 직접 선별하였다. 본 연구를 실제로 서비스하기 위

Fig. 3. Example of training images

해서는 각 분위기에 맞는 음악이 준비되어 있어야 한다. 음악을 분위기에 맞게 자동 생성할 수 있으면 가장 좋지만, Russell의 감정분석표의 분위기에 맞는 음악을 자동 생성해 주는 시스템이 아직 없기 때문에 Fig. 2의 각 분위기에 맞는 음악을 직접 선별하였다. 물론 실제 서비스에 선정한 음악을 사용하려면 저작권 소유자에게 저작권료를 지불해야 한다. 사용한 음악은 이 링크[1]에 정리하였으며, 영상 분위기의 선별 기준과 동일하게 4 명 중 3 명 이상이 동영상의 배경음으로 사용하는 것에 동의할 때, 해당 음악을 동영상에 사용하는 음악으로 선별하였다.

Table 1. Data for each type of emotion

	Before preprocessing	After preprocessing			
Type of emotion	The number of				
	Data		Train	Test	Music
Angry	1,200	400	320	80	4
Anxiety	1,200	400	320	80	4
Depressed	1,200	400	320	80	4
Dynamic	1,200	400	320	80	4
Happy	1,200	400	320	80	4
Peaceful	1,200	400	320	80	4
Tired	1,200	400	320	80	4
Withered	1,200	400	320	80	4
Total	9,600	3,200	2,560	640	32

4. 기계학습을 통한 실험

4.1 모델링

1) https://whehd16.github.io/

Table 2. Summary of models

Type of layer	Shape of output	Parameters	Dropout(0.25)	(16, 16, 64)	0
Convolution2D	(128, 128, 32)	896	Convolution2D	(16, 16, 64)	36,928
Activation(Relu)	(128, 128, 32)	0	Activation(Relu)	(16, 16, 64)	0
Convolution2D	(128, 128, 32)	9,248	Convolution2D	(16, 16, 64)	36,928
Activation(Relu)	(128, 128, 32)	0	Activation(Relu)	(16, 16, 64)	0
MaxPooling2D	(64, 64, 32)	0	MaxPooling2D	(8, 8, 64)	0
Dropout(0.25)	(64, 64, 32)	0	Dropout(0.25)	(8, 8, 64)	0
Convolution2D	(64, 64 ,64)	18,496	Flatten	(4,096)	0
Activation(Relu)	(64, 64, 64)	0	Dense(512)	(512)	2,097,664
Convolution2D	(64, 64, 64)	36,928	Activation(Relu)	(512)	0
Activation(Relu)	(64, 64, 64)	0	Dropout(0.5)	(512)	0
MaxPooling2D	(32, 32, 64)	0	Dense(64)	(64)	32,832
Dropout(0.25)	(32, 32, 64)	0	Activation(Relu)	(64)	0
Convolution2D	(32, 32, 64)	36,928	Dropout(0.5)	(64)	0
Activation(Relu)	(32, 32, 64)	0	Dense(8)	(64)	520
Convolution2D	(32, 32, 64)	36,928	Activation(softmax)	(8)	0
Activation(Relu)	(32, 32, 64)	0	Total parameters		2,344,296
MaxPooling2D	(16, 16, 64)	0			

8 가지의 분위기를 다중 클래스 분류[11]하기 위해 Table 1의 데이터를 사용했다. 실험에 사용된 모델은 여러 덴스(dense) 층과 활성화 함수를 이용한 다층 퍼셉트론(multilayer perceptron)[12]으로 구성되는 심층 합성곱 신경망 모델이다. 활성화 함수로는 정류된 선형 유닛(rectified linear unit; ReLU)[13]을 사용하여 기울기 소실(vanishing gradient)[14] 문제를 줄이고, 모델의 마지막 계층에서는 소프트맥스 – 교차엔트로피 손실함수(softmax – cross entropy loss function)[15]를 사용하여 손실을 줄인다. 학습 과정에서 과적합을 피하고자 신경망에 필요한 매개변수의 수를 줄이는 드롭아웃(dropout)[16]을 활용하였다. Table 2를 통해 모델의 생성 구조를 확인할 수 있다. 학습에 사용된 에포크(Epoch) 수는 1,000 그리고 배치 크기 (Batch Size)는 64이다.

4.2 분류

실험에 사용된 데이터는 전처리 과정에서 학습 데이터로 사용되진 않았지만, 남은 일반 이미지들을 저자 4 명 중 2 명 이상의 동의로 20 개씩 총 160 개로 생성한 것이다.

먼저 실험에 사용할 데이터에서 프레임별로 이미지를 추출하여 모델에 적용할 수 있도록 배열화한다. 추출된 프레임의 RGB 색상을 모두 고려하며, 실험 데이터와 규격에 맞는 128×128 크기로 재조정한다. 조정된 데이터들은 각각 학습된 모델을 거쳐 8 개의 분위기 중 하나로 예측되고, 동영상으로 생성된다. Fig. 4를 통해 영상의 프레임별 분위기를 예측한 결과, ‘Peaceful’ 분위기로 분류됨을 확인할 수 있고, 분류된 분위기에 맞는 음악의 자동 매칭 예시를 확인할 수 있다.

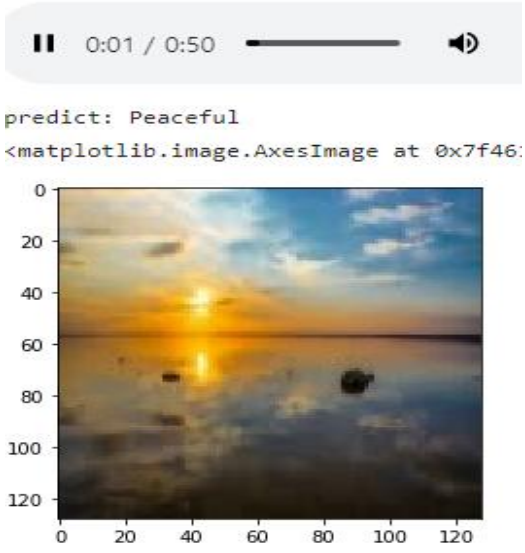

Fig. 4. Extracting a frame image with matched music (captured from our developed program)

5. 실험 결과

본 논문에서 제안하는 모델의 10겹 교차 검증[17]의 결과는 Table. 3과 같다. 각 교차 검증의 결과인 'cross_val_score'의 평균값은 72.4%이다. 일반 합성곱 신경망 모델의 검증 결과는 평균 70.0%로, 심층 합성곱 신경망 모델의 실험 결과가 높음을 확인할 수 있다.

Table 3. Results of 10-fold cross validation

Step	Epoch	Batch size	cross_val_score(%)
1	60	64	75.0
2	60	64	74.4
3	60	64	70.0
4	60	64	73.8
5	60	64	68.8
6	60	64	73.4
7	60	64	72.8
8	60	64	73.4
9	60	64	71.9
10	60	64	70.0
Average cross_val_score(%)			72.4

Table 3가 실험 데이터에 대한 교차 검증이라면, Fig. 6은 4.2절의 실제 영상 데이터 대한 실험 오차 행렬[18]이며, 평균 정확도는 64%이었다.

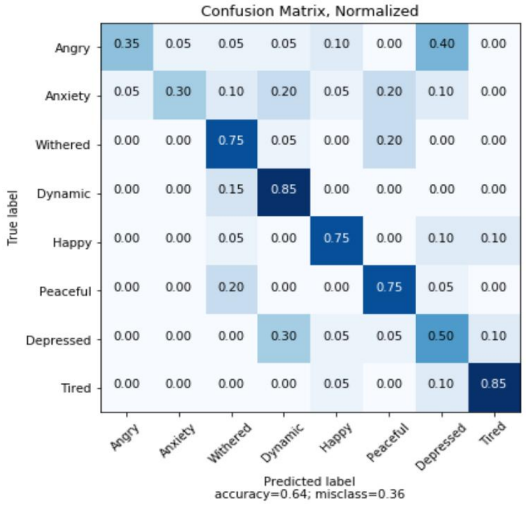

Fig. 5. Confusion matrix of our model

예를 들어 'Tired'로 분류된 데이터에 대해 누군가는 분류된 분위기 그대로 피곤함을 느낄 수 있는 반면에, 'Depressed'를 느낄 수도 있다. 이는 'Tired'로 분류된 데이터도 'Depressed' 분위기의 음악 데이터를 연결해도 이질감 없이 어울린다는 것을 의미한다. 즉, 다중 클래스 분류를 통해 예측된 결과는 이웃한 분위기마다 어느 정도 상호보완이 되므로 Fig. 5 상에 나타난 오차를 어느 정도 줄일 수 있다고 판단할 수 있다. 본 연구에서 생성한 동영상은 이 링크[2]에서 확인할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 심층 CNN의 다중 클래스 분류를 통해 영상의 8 가지 분위기 예측과 자동 음악 매칭 모델을 제안했다. 학습데이터에 대한 교차 검증 결과, 평균 72.4%의 정확도를 얻었고, 실제 데이터에 대한 오차 행렬에서는 평균 64%의 정확도를 얻었다. 분위기와 감정 분류에 있어서 인간의 100% 의견 일치는 힘들며, 예측 정확도가 60%~70%이면 상당히 유망한 수준의 결과라고 알려져 있다[19, 20]. 실험을 통해 예측된 결과는 분류된 다른 분위기와 상호 보완되는 부분이 있으므로, 오차는 어느 정도 보완됨을 알 수 있다.

향후 연구에서 더 세분화되고 다양한 감정 분류와 다중 클래스 분류 모델을 개선 시켜 다양한 분위기의 음악을 적용하는 연구가 가능할 것이다.

REFERENCES

[1] Y. Yan, M. Chen, M. L. Shyu & S. C. Chen (2015, December). Deep learning for imbalanced multimedia data classification. *2015 IEEE International Symposium on Multimedia*. (pp. 483-488). DOI : 10.1109/ISM.2015.126

[2] M. K. Lee, D. H. Kim, D. Y. Choi, and B. C. Song. (2017). Emotion recognition system based deep learning. Journal of the *Korean Society Of Broad Engineers, 16-18*.

[3] S. H. Kim. (2016). *Sentiment classification for videos using deep learning algorithms*. Master dissertation. Seoul University, Seoul.

[4] J. A. Russell. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39(6)*, 1161-1178 DOI : 10.1037/h0077714

[5] D. H. Ko, H. K. Moon, J. W. Jun, J. M. Yu & M. G. Jeon.

2) https://youtu.be/u4cl-GkyiwA

(2017). Face Verification based on Deep Convolutional Nerual Network. *Journal of The Korean Institute of Information Scientists and Engineers*

[6] D. G. Lee. (2018). Classification of Trucks using Convolutional Neural Network. *Journal of Convergence for Information Technology, 8(6)*, 375-380
DOI : 10.22156/CS4SMB.2018.8.6.375

[7] A. M. Ramadhani, N. R. Kim & H. R. Choi. (2018). Predicting Employment Earning using Deep Convolutional Neural Networks. *Journal of Digital Convergence, 16(6)*, 151-161.
DOI : 10.14400/JDC.2018.16.6.151

[8] J. Y. Lee, C. B. Moon and B. M. Kim. (2018). Music crawler for mood-based music classification and retrieval systems. Journal of *Korea Information Science Society*, 699-701

[9] C. W. Lee. (2005). D*evelopment of automatic synchronization tool for scene and background music.* Chungcheongbuk-do : INET.

[10] C. Olston & M. Najork. (2010). Web crawling. *Foundations and Trends® in Information Retrieval, 4(3)*, 175-246.
DOI : 10.1561/1500000017

[11] G. B. Huang, H. Zhou, X. Ding & R. Zhang. (2011). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(2),* 513-529.
DOI : 10.1109/TSMCB.2011.2168604

[12] M. Riedmiller. (1994). Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces, 16(3)*, 265-278.
DOI : 10.1016/0920-5489(94)90017-5

[13] A. Krizhevsky, I. Sutskever, & G. E. Hinton. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems,* 1097-1105
DOI : 10.1145/3065386

[14] S. Hochreiter. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02)*, 107-116.
DOI : 10.1142/S0218488598000094

[15] R. A. Dunne & N. A. Campbell. (1997, June). On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. *Proc. 8th Aust. Conf. on the Neural Networks,* Melbourne.
DOI : 10.1.1.49.6403

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15(1)*, 1929-1958.
DOI : 10.1214/12-AOS1000

[17] A. Krogh & J. Vedelsby. (1995). Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems.* (pp. 231-238). Cambridge,MA:MITPress.

[18] M. Sokolova & G. Lapalme. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45(4)*, 427-437.
DOI : 10.1016/j.ipm.2009.03.002

[19] T. Kincl, M. Novák & J. Pribil. (2013, November). Getting inside the minds of the customers: automated sentiment analysis. *ECMLG2013-Proceedings For the 9th European Conference on Management Leadership and Governance: ECMLG 2013.* (pp. 122-128). Klagenfurt : Academic Conferences Limited

[20] V. Gajarla & A. Gupta. (2015). *Emotion detection and sentiment analysis of images.* Atlanta : Georgia Institute of Technology.

조 동 희(Dong-Hee Cho) [학생회원]

· 2015년 3월 ~ 현재 : 광운대학교 소프트웨어학부 학사과정
· 관심분야 : 인공지능, 유전알고리즘, 기계학습
· E-Mail : whehd16@gmail.com

남 용 욱(Yong-Wook Nam) [정회원]

· 2014년 2월 : 광운대학교 컴퓨터소프트웨어학과(공학사)
· 2019년 8월 : 광운대학교 컴퓨터과학과 공학박사
· 관심분야 : 자동 작곡, 최적화 알고리즘, 인공지능
· E-Mail : mitssimvz@gmail.com

이 현 창(Hyun-Chang Lee) [학생회원]

· 2018년 8월 : 광운대학교 컴퓨터소프트웨어학과(공학사)
· 2018년 9월 ~ 현재 : 광운대학교 컴퓨터과학과 석사과정
· 관심분야 : 유전 알고리즘, 최적화 알고리즘
· E-Mail : qzecwxad@naver.com

김 용 혁(Yong-Hyuk Kim) [정회원]

· 1999년 2월 : 서울대학교 전산과학(이학사)
· 2001년 2월 : 서울대학교 전기컴퓨터공학부(공학석사)
· 2005년 2월 : 서울대학교 전기컴퓨터공학부(공학박사)
· 2005년 3월 ~ 2007년 2월 : 서울대학교 반도체공동연구소 연구원
· 2007년 3월 ~ 2017년 2월 : 광운대학교 컴퓨터소프트웨어학과 조교수/부교수
· 2017년 3월 ~ 현재 : 광운대학교 소프트웨어학부 교수
· 관심분야 : 최적화, 진화연산, 지식공학
· E-Mail : yhdfly@kw.ac.kr