얼굴 표정 인식을 위한 Convolutional Neural Networks

*최인규, **송 혁, *유지상
광운대학교 전자공학과, 한국전자부품연구원
cig2982@kw.ac.kr, hsong@keti.re.kr, jsyoo@kw.ac.kr

Convolutional Neural Networks for Facial Expression Recognition

*In-Kyu Choi, **Hyok Song, *Jisang Yoo
*Department of Electronic Engineering, Kwangwoon University
**Korea Electronics Technology Institute

요약

본 논문에서는 딥러닝 기술 중의 하나인 CNN(Convolutional Neural Network) 기반의 얼굴 표정 인식 기법을 제안한다. 제안한 기법에서는 획득한 여섯 가지 주요 표정의 얼굴영상들을 학습 데이터로 이용할 때 분류 성능을 저해시키는 과적합(over-fitting) 문제를 해결하기 위해서 데이터 증대 기법(data augmentation)을 적용한다. 또한 기존의 CNN 구조에서 convolutional layer 및 node의 수를 변경하여 학습 파라미터 수를 대폭 감소시킬 수 있다. 실험 결과 제안하는 데이터 증대 기법 및 개선한 구조가 높은 얼굴 표정 분류 성능을 보여준다는 것을 확인하였다.

1. 서론

컴퓨터는 인간의 일상생활에 중요한 일부분이 되었을 뿐 아니라, 다양한 형태로 편리성을 제공하고 있다. 앞으로도 컴퓨터와 인간과의 밀접성 및 상호작용은 계속해서 증가할 것으로 보인다. 이에 따라 인간과 컴퓨터와의 상호 작용(Human-Computer Interaction, HCI)에 대한 연구가 인간 공학, 산업 공학, 심리학, 컴퓨터 과학 등 여러 학문 분야에서 진행되고 있다. 인간과 컴퓨터 간의 자연스러운 상호 작용을 위해서 컴퓨터는 사용자의 의도를 종합적으로 판단하고 그에 맞는 반응을 해야 한다. 감정은 인간의 마음 상태를 표출하는 가장 중요한 요소로 사용자의 만족을 극대화하기 위해서는 사용자의 감정 인식이 중요하다. 감정의 형태를 나타내는 중요한 수단의 하나가 얼굴 표정이므로 얼굴 표정을 분류하는 기술이 필요하다.

최근에 하드웨어의 발전과 빅데이터 안에서 데이터를 기반으로 스스로 학습하고 패턴을 찾아 사물을 구별하는 딥러닝(deep learning) 기술이 주목받고 있다. 복잡한 문제에 대해서 성능이 급격하게 저하되는 기존의 기계학습 모델과는 달리 딥러닝은 깊은 신경망(deep neural networks) 모델을 이용하여 주어진 데이터에 알맞은 고수준의 특징을 추출함으로써 기존의 기계학습의 기술적 한계를 극복할 수 있는 방법론이다. 그 중에서도 인간의 시각 처리 과정을 모방하기 위해 개발된 CNN(convolutional neural networks)은 영상 인식 분야에 다양하게 적용되어 높은 성능을 보이고 있다.

따라서 본 논문에서는 인간의 여섯 가지 얼굴 표정 영상에 대한 데이터 셋을 구축하고 기존의 CNN 모델을 확보한 데이터 셋에 적합한 구조로 변형하고 학습시켜 입력 영상을 올바른 표정으로 분류하는 기법을 제안한다.

2. 본론

기본적인 CNN의 구조는 convolutional layer와 fully-connected layer로 이루어진다. 복수의 convolutional layer를 차례대로 지나면서 특징을 추출하고 추상화하면서 점차 고수준의 특징을 추출한다. 그리고 full-connected layer에서 추출한 고수준의 특징으로부터 최종 분류 결과를 결정한다.

본 논문에서는 2012년에 krizhevsky가 제안한 AlexNet을 참고로 한다[1]. 수집한 데이터를 효율적으로 학습하고 높은 정확도의 표정분류를 위해 convolutional layer에서는 특징 지도의 채널 수 그리고 fully-connected layer에서는 노드의 수를 변경하여 최적의 구조를 찾고자 한다.

얼굴 표정 데이터 셋은 연구 목적으로 공개된 얼굴 데이터베이스를 통해 수집한다[2-8]. 데이터 셋은 '무표정, 행복함, 슬픔, 화남, 놀람, 역겨움' 등의 여섯 가지 표정으로 구성하며 표정 별로 각각 대략 9:1의 비율로 학습 영상과 시험 영상으로 분리한다. 학습 영상과 시험 영상의 표정 별 영상의 수는 아래의 표 1과 같다. 각기 피부색이 다른 사람의 얼굴을 보고 어떤 표정인지 판단할 때 피부색은 고려하지 않고 눈, 눈썹, 코, 입 등의 모양이나 위치 등으로 판단을 하는 점에 착안하여 수집한 영상들을 1 채널의 흑백 영상으로 변환한다. 또한 얼굴 표정인식 시에 필요 없는 배경정보를 제거하기 위하여 얼굴 중심으로 영상을 잘라내는 전 처리 작업을 수행한다.

표 1. 수집한 데이터 셋의 표정 별 영상의 개수

	무표정	행 복	슬 픔	화 남	놀 람	역겨움	총 합
학 습	912	920	419	498	512	451	3,712
시 험	88	88	46	55	57	50	384

분류 성능을 저해시키는 과적합 문제(over-fitting)를 해결하기 위해서 학습 영상의 수를 증가시키는 데이터 증대(data augmentation) 기법을 이용한다. 기준 영상에 대하여 시계 방향, 반시계 방향으로 각각 5°, 10°, 15° 만큼 회전 연산한 영상을 획득한다. 그리고 회전된 영상들과 기준 영상을 각각 수평 반전하여 하나의 기준 영상을 열네 장의 영상으로 증가시킨다. 그림 1은 데이터 증대 기법을 적용한 결과를 보여준다.

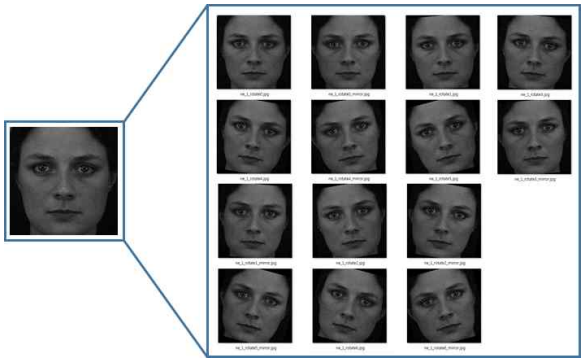

그림 1. 데이터 증대 기법을 적용한 결과

표 2는 기존의 AlexNet 구조와 개선한 구조에 대해서 시험 영상을 입력하여 실험한 결과를 보여준다. 각 숫자는 5개의 convolutional layer와 3개의 fully-connected layer의 특징 지도 채널의 수와 노드 수를 의미한다. 개선한 구조가 표정 분류 성능이나 필요한 파라미터의 수를 볼 때 더 효과적이라는 보여준다.

표 2. 특징 지도 채널 수와 노드 수에 따른 성능과 파라미터 용량 비교

	특징 지도 채널 수 / 노드 수	데이터 증대 적용 여부	인식률 (%)	파라미터 용량 (MB)
(a)	96-256-384-384-256 / 4096-4096-6	O	95.05	217
(b)	36-96-144-96-128 / 1024-1024-6	O	96.88	6.2
(c)	36-96-144-96-128 / 1024-1024-6	X	88.8	6.2

표 3. 표 2의 (b)결과에 대한 confusion matrix

	0	1	2	3	4	5	총 합	인식률 (%)
0	82	3	1	2	0	0	88	93.18
1	0	88	0	0	0	0	88	100
2	2	0	44	0	0	0	46	95.65
3	0	1	1	51	0	2	55	92.73
4	0	0	0	0	57	0	57	100
5	0	0	0	0	0	50	50	100
							384	96.88

표 3은 데이터 증대 기법을 적용한 학습 영상을 개선한 구조에 학습한 결과에 대한 confusion matrix이다. Confusion matrix를 보면 여섯 가지 표정 각각의 인식률을 확인할 수 있을 뿐만 아니라 어느 표정으로 오 분류되었는지 확인할 수 있다. 표정 마다 차이는 있지만 여섯 가지 표정 모두 92% 이상의 높은 인식률을 보여주는 것을 확인할 수 있다.

3. 결론

본 논문에서는 기존의 CNN 모델인 AlexNet의 구조를 적절히 변경하여 연구 목적의 공개 데이터베이스를 통해 구축한 데이터 셋에 대한 성능 및 수용성을 향상시켰음을 결과를 통해 확인하였다. 과 적합 문제를 해결하기 위해서 회전 및 반전 연산을 이용하여 학습 영상의 수를 증가시켰고 시험 영상을 입력하여 나온 분류 정확도를 통해 그 효과를 확인하였다.

ACKNOWLEDGMENT

본 논문은 미래창조과학부 SW컴퓨팅산업원천기술개발사업 (과제번호 R0190-16-1115)을 지원받아 수행한 결과입니다.

참 고 문 헌

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, 2012

[2] W. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs" Journal of Experimental Psychology: General, 142(4):1323 - 1334, 2013

[3] S. Setty and et al, "Indian Movie Face Database: A Benchmark for FaceRecognition Under Wide Variation". In NCVPRIPG, 2013

[4] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression" in Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010

[5] Ma, Correll, and Wittenbrink (2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. Behavior Research Methods, 47, 1122-1135.

[6] http://pics.stir.ac.uk/ESRC/

[7] J. Van der Schalk, S. T. Hawk, A. H. Fischer, and B. J. Doosje, (2011). Moving faces, looking places: The Amsterdam Dynamic Facial Expressions Set (ADFES), Emotion, 11, 907-920. DOI: 10.1037/a0023853

[8] D. Lundqvist, A. Flykt, and A.Öhman (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.