CNN 기반의 얼굴 표정 인식

*최인규, *안하은, **송 혁, **고민수, *유지상
*광운대학교 전자공학과, **한국전자부품연구원
cig2982@kw.ac.kr, mysco227@kw.ac.kr, hsong@keti.re.kr, kmsqwet@keti.re.kr, jsyoo@kw.ac.kr

CNN-based facial expression recognition

*In-Kyu Choi, *Ha-Eun Ahn, **Hyok Song, **Min-Soo Ko, *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. 다섯 가지 주요 표정의 얼굴 영상을 CNN 구조에 스스로 학습시켜 각각의 표정 패턴에 적합한 특징 지도(feature map)를 형성하고 이 특징 지도를 통해 들어오는 입력 영상을 적합한 표정으로 분류한다. 기존의 CNN 구조를 본 논문에서 이용한 데이터 셋에 알맞게 convolutional layer 및 node의 수를 변경하여 특징 지도를 형성하고 학습 및 인식에 필요한 파라미터 수를 대폭 감소시켰다. 실험 결과 제안하는 기법이 높은 얼굴 표정 분류 성능을 보여준다는 것을 보였다.

1. 서론

컴퓨터는 인간의 일상 생활에 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 다섯 가지 얼굴 표정 영상에 대한 데이터 셋을 구축하고 기존의 CNN 모델을 확보한 데이터 셋에 적합한 구조로 변형하고 학습시켜 입력 영상을 올바른 표정으로 분류하는 것을 보여준다.

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. 학습 데이터에 적합하고 다양한 변이에도 잘 적응하는 고수준의 특징을 추출하기 때문에 영상 인식 분야에 적용되어 높은 성능을 보인다.

본 논문에서는 2012년에 krizhevsky가 제안한 AlexNet을 참고로 한다.[1] AlexNet의 구조는 아래 그림 1과 같다. AlexNet은 5개의 convolutional layer와 3개의 fully-connected layer로 구성되어 있다. AlexNet은 100만장 이상의 영상을 학습하고 1000 가지의 부류로 분류하는 구조이기 때문에 보다 적은 학습 데이터를 이용하고 다섯 가지의 표정으로 분류하는 목적에 맞게 구조를 변경하는 것이 효율적이다.

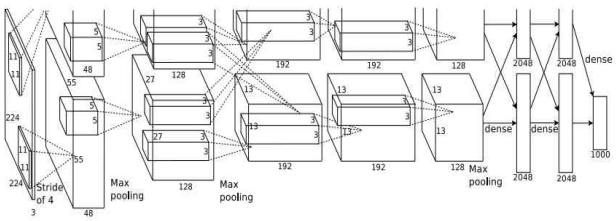

그림 1. AlexNet의 구조

AlexNet 구조에서 입력 영상의 크기는 변경하지 않고 convolutional layer에서는 특징 지도의 채널 수 그리고 fully-connected layer에서는 노드의 수를 감소시키면서 데이터를 학습하여 가장 높은 성능을 보이는 최적의 구조를 찾고자 한다. 또한 우리가 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 학습할 입력 영상을 1 채널의 흑백 영상으로 변형하도록 하였다.

얼굴 표정 데이터 셋은 연구 목적으로 공개된 얼굴 데이터베이스를 통해 수집하였다[2-4]. 그리고 연구원들의 도움을 받아 얼굴 영상을 '무표정, 행복함, 슬픔, 화남, 놀람'에 해당하는 각각의 다섯 가지 표정으로 분류하고 4,433장의 학습 영상과 498장의 시험 영상으로 분리한다. 또한 얼굴 표정을 인식 시 필요 없는 배경정보를 제거하기 위하여 얼굴 중심으로 영상을 잘라내는 작업을 수행한다. 그림 2는 수집한 데이터 셋에서 얼굴을 중심으로 잘라낸 몇 가지 영상을 보여준다.

그림 2. 얼굴을 중심으로 잘라낸 영상의 예

실험 환경은 i5-4670 CPU 3.40GHz, RAM 8GB, Geforce GTX 680와 같고 학습은 batch 크기가 128인 stochastic gradient descent을 이용하여 진행하였다. 학습율(learning rate)은 초기 값을 0.01로 시작하여 총 epoch를 200에서 epoch가 [50, 100, 150]이 될 때 1/10만큼 감소하도록 설정하였다.

표 1은 AlexNet에서 마지막 단의 분류를 위한 노드 수만 5개로 변경한 구조에 대하여 입력 영상 채널 수에 따른 인식률을 보여준다. 1채널의 흑백영상을 이용하면 특징 지도를 형성하기 위해 필요한 파라미터의 수가 줄어들 뿐만 아니라 인식률도 향상되는 것을 알 수 있다.

표 1. AlexNet 구조에서 입력 영상 채널 수에 따른 성능 비교

입력 영상 채널 수	인식률 (%)
1	79.8
3	77.6

표 2는 입력 영상이 1채널일 때 기존의 AlexNet 구조와 특징 지도의 채널 수와 full-connected layer의 노드 수를 적절하게 변경한 구조에 대해서 시험한 결과를 보여준다. 각 숫자는 5개의 convolutional layer와 3개의 fully-connected layer의 특징 지도 채널의 수와 노드 수를 의미한다. 특징 지도의 채널의 수와 노드 수를 줄여 개선한 구조가 표정 분류 성능이나 필요한 파라미터의 수를 볼 때 더 효과적이라는 보여준다. 이를 통해 얼굴 표정을 분류할 때 너무 많은 특징이 필요하지 않다는 것과 이런 특징을 이용하여 최종적으로 어떤 부류로 분류할 때 너무 많은 노드의 수가 과적합(over-fitting) 문제를 발생시켜 성능을 떨어뜨릴 수 있음을 알 수 있다.

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

특징 지도 채널 수 / 노드 수	인식률 (%)	파라미터 용량 (MB)
96-256-384-384-256 / 4096-4096-5	79.8	217
48-128-96-384-256 / 2048-2048-5	85.7	91

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. 또한 객체 인식과는 달리 1채널의 흑백 영상을 입력으로 사용함으로써 얼굴 표정 인식에는 색상 정보가 중요하지 않다는 것을 확인하였다. 데이터 셋을 여러 공개된 데이터베이스를 통해 수집하였기 때문에 영상이 취득된 환경이 각기 다르고 구분하기 어려운 애매한 표정의 영상이 존재한다. 따라서 정제된 데이터 셋과 환경에 강인한 영상 인식 기술을 이용하면 더 나은 성능을 보일 것이라 판단된다.

ACKNOWLEDGMENT

본 논문은 미래창조과학부 SW컴퓨팅산업원천기술개발사업 (과제번호 R0190-15-1115)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