특징 연결과 깊이별 분리 컨볼루션을 이용한 효율적인 얼굴 감정인식 CNN

Efficient CNNs with Feature Concatenation and Depthwise Separable Convolution for Facial Expression Recognition

저자 (Authors)
이명오, 윤의녕, 고승현, 조근식
Myeong Oh Lee, Ui Nyoung Yoon, Seunghyun Ko, Geun-Sik Jo

출처 (Source)
한국정보과학회 학술발표논문집 , 2019.6, 754-756(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763318

APA Style
이명오, 윤의녕, 고승현, 조근식 (2019). 특징 연결과 깊이별 분리 컨볼루션을 이용한 효율적인 얼굴 감정인식 CNN. 한국정보과학회 학술발표논문집, 754-756

특징 연결과 깊이별 분리 컨볼루션을 이용한 효율적인 얼굴 감정인식 CNN*

이명오O 윤의녕 고승현 조근식
인하대학교 컴퓨터공학과
eremo2002@naver.com, entymos@hotmail.com, kosehy@gmail.com, gsjo@inha.ac.kr

Efficient CNNs with Feature Concatenation and Depthwise Separable Convolution for Facial Expression Recognition

Myeong Oh LeeO Ui Nyoung Yoon Seunghyun Ko Geun-Sik Jo
Department of Computer Engineering, Inha University

요 약

최근 딥러닝 기반의 컨볼루션 신경망은 컴퓨터비전 분야의 여러 문제에서 좋은 성능을 보여주었으며 다양한 형태의 컨볼루션 신경망 아키텍처가 제안되어왔다. 특히 얼굴 표정에서 감정을 인식하기 위한 문제 역시 딥 컨볼루션 신경망을 사용한 연구가 활발히 진행되고 있다. 본 논문에서는 사람의 얼굴 표정에서 나타나는 감정을 인식하기 위한 효율적인 컨볼루션 신경망을 새롭게 제안한다. 본 논문에서는 경량화된 컨볼루션 신경망을 설계하기 위해 깊이별 분리 컨볼루션을 사용하여 파라미터 수와 연산량을 감소시키고 특징 연결과 채널 정보의 압축 및 복원을 통해 특징의 재사용성과 채널 정보를 강화하였다. 제안하는 모델의 학습 파라미터 개수는 0.48 M(Million)으로 기존 모델에 비해 적은 파라미터 개수로 FER2013, RAF-single 데이터셋에서 각각 70.69%, 85.4%의 정확도를 달성하였다.

1. 서 론

최근 컨볼루션 신경망은 컴퓨터비전 분야의 다양한 이미지 인식 문제에서 높은 성능을 보여주고 있으며 정확도를 향상시키기 위해 더 많은 레이어를 쌓는 다양한 구조의 네트워크들이 등장하였다[1, 2]. 얼굴 표정에서 감정을 인식하기 위한 FER(Facial Expression Recognition)문제 역시 다양한 구조의 컨볼루션 신경망이 제안되었으며 레이어를 더 깊게 쌓아 성능을 높이려는 연구가 진행되어왔다. 그러나 CNN에서 레이어가 깊어질수록 파라미터 개수와 연산량이 기하급수적으로 증가하는 문제가 발생하게 된다. 이러한 문제는 모바일 디바이스, 임베디드 장치와 같은 제한된 리소스를 가지는 환경에서 더 심각한 문제가 될 수 있다.

따라서 본 논문에서는 사람의 얼굴 표정에서 나타나는 감정 상태를 인식하기 위한 효율적인 컨볼루션 신경망을 새롭게 제안한다. 제안하는 감정인식 CNN 모델은 네트워크의 복잡도를 줄이기 위해 깊이별 분리 컨볼루션과 특징 연결을 위한 Skip Connection 그리고 채널 정보의 압축 및 복원을 사용한 Dense Block을 사용하였으며 이를 기반으로 적은 파라미터 개수로 효율적인 감정 인식 모델을 설계하였다.

2. 관련 연구

2.1 깊이별 분리 컨볼루션(Depthwise Separable Convolution)

효율적인 컨볼루션 신경망을 설계하기 위해 깊이별 분리 컨볼루션을 사용한 많은 연구 진행되고 있다[3, 4, 5, 6]. 깊이별 분리 컨볼루션은 채널 방향의 학습과 공간 방향의 학습을 분리한 구조로 깊이별 컨볼루션과 1x1 컨볼루션 두 단계로 구성되어 있다. 깊이별 분리 컨볼루션에서는 먼저 입력 특징의 깊이별로 컨볼루션 연산을 하게 된다. 이때 생성되는 출력 특징의 수는 입력 특징의 깊이에 따라 결정되며 깊이별 컨볼루션에서 생성된 출력 특징들을 1x1 컨볼루션을 통해 선형적으로 조합된 새로운 특징으로 만들어 다음 레이어의 입력 특징으로 사용하게 된다. 기존의 컨볼루션 연산을 채널 단위의 연산과 이를 다시 결합하는 1x1 컨볼루션 두 단계로 나누어 연산하기 때문에 깊이별 분리 컨볼루션에서는 파라미터 수를 대폭 줄일 수 있다.

2.2 특징 연결을 위한 Skip Connection

CNN에서 레이어가 더욱 깊어질수록 그래디언트 소실(Gradient Vanishing)로 인해 학습이 제대로 되지 않는 문제와 이전 레이어의 정보가 다음 레이어로 잘 전달되지 못하는 특징 정보의 손실 문제가 발생하게 된다. 이러한 문제를 보완하기 위해 Densely connected networks[7]은 현재 레이어를 이전 모든 레이어와 연결한 Skip Connection을 통해 이전 레이어의 모든 특징을 현재 레이어의 특징과 연결하였다. DenseNet의 Skip Connection 구조는 이전 레이어의 특징이 Connection을 통해 다음 레이어로 전달되기 때문에 연속된 컨볼루션 레이어로 인한 특징 정보의 손실을 방지할 수 있고 특징의 재사용성을 증가시킬 수 있다. 또한, 특징 정보의 Propagation을 강화하여 특징 정보의 흐름을 극대화하면서 컨볼루션 레이어의 수를 줄일 수 있으며 그래디언트가 소실되는 문제 또한 방지할 수 있다.

2.3 채널 정보의 압축 및 복원

최근 컨볼루션 신경망을 설계하는데 있어서 파라미터를 줄이거나 효과적으로 특징을 추출하기 위한 목적으로 많은 네트워크들[3, 4, 5]이 특정 구조의 Block 및 모듈을 사용하여 성능을

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2017-0-01642)

개선시키고 있다. 특히 SENet[8]은 Global Average Pooling을 사용하여 특징의 채널 단위로 정보를 압축한 뒤 이를 다시 복원하여 각 채널 간의 Dependency를 계산하는 Squeeze Excitation Block을 사용하여 각 채널에서 중요한 정보를 강화하였으며 Squeeze Excitation Block을 통해 적은 파라미터 증가량으로도 기존 컨볼루션 신경망의 성능을 크게 개선시켰다.

3. 제안하는 모델

3.1 깊이별 분리 컨볼루션을 이용한 Dense Block

본 논문에서는 사람의 얼굴 표정에서 나타나는 감정 상태를 분류하기 위해 깊이별 분리 컨볼루션과 특징 연결을 위한 Skip Connection 및 Squeeze Excitation Block을 결합한 효율적인 Dense Block을 새롭게 제안한다. 제안하는 Dense Block에선 기존의 컨볼루션 레이어 대신 깊이별 분리 컨볼루션을 사용하여 네트워크가 깊어질수록 급증하는 파라미터 수와 연산 비용의 문제를 방지하였다. 또한, 특징 연결을 위한 Skip Connection 구조를 사용하여 이전 레이어의 특징을 현재 레이어의 특징과 연결하여 특징 정보의 손실 문제를 방지하고 특징의 재사용성을 증가시켜 네트워크 전체에서 특징 정보의 흐름을 극대화하였다. 나아가 Squeeze Excitation Block을 추가하여 깊이별 분리 컨볼루션에서 계산된 특징의 채널 정보를 압축 및 복원하여 채널 간의 Dependency를 계산하고 각 채널에서 중요한 정보를 강화한 특징을 추가적으로 연결하였다. 본 논문에서 제안하는 Dense Block의 구조는 아래 그림 1과 같다.

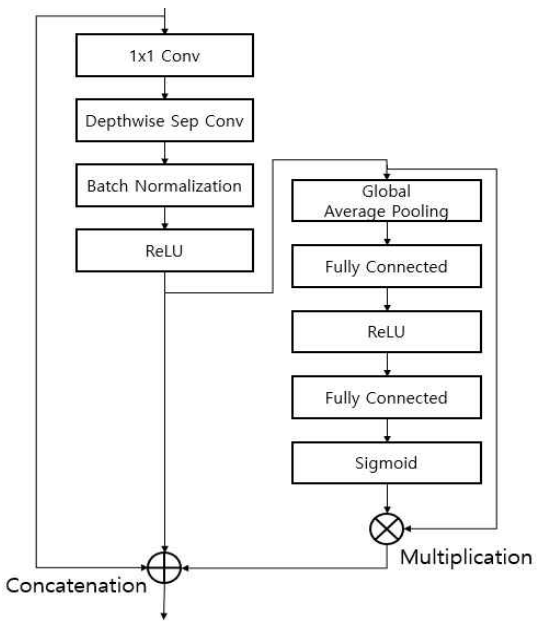

그림 1 제안하는 Dense Block의 구조

3.2 네트워크 구조

학습에 사용되는 입력 이미지는 Grayscale 형태의 이미지를 입력받으며 1x1 컨볼루션을 제외한 모든 컨볼루션 레이어에서 커널의 크기는 3x3을 사용하였다. 또한 깊이별 분리 컨볼루션 레이어를 사용한 뒤 Batch Normalization[9]과 ReLU 활성화 함수를 적용하였으며 각 스테이지 내의 모든 레이어는 특징 연결을 위한 Skip Connection으로 연결되어 있다. 이후 Transition Layer에서 1x1 컨볼루션 및 2x2 Average Pooling을 통해 채널 정보와 공간적인 사이즈를 절반으로 다운 샘플링하였다. 또한, 연속된 완전연결층(Fully-Connected Layer)에서 발생하는 파라미터의 개수와 연산 비용이 네트워크의 전체적인 복잡도에 매우 큰 비중을 차지한다는 점과 특징의 위치정보가 손실될 수 있는 문제를 방지하기 위해 연속된 완전연결층 대신 Global Average Pooling을 사용하였다. Dense Block과 Transition Layer를 사용한 네트워크 구조는 아래의 표1과 같다.

표 1 제안하는 컨볼루션 신경망 구조

Stage	Layer
Input	
Stage 1	[Dense Block] x3
	Transition Layer
Stage 2	[Dense Block] x3
	Transition Layer
Stage 3	[Dense Block] x3
	Transition Layer
Stage 4	[Dense Block] x3
Classification Layer	Global Average Pooling
	7-D Fully Connected, softmax

4. 실험

4.1 학습 데이터

본 논문에서는 감정 인식을 위한 학습 데이터셋으로 7가지 감정 클래스를 가지고 있는 FER2013[10], RAF-single[11] 데이터셋을 사용하였다. FER2013 데이터셋은 총 35,887개의 48x48 Grayscale 이미지로 구성되어 있으며 train 28,709개 validation 3,589개 test 3,589개로 나누어져 있다. RAF-single 데이터셋은 총 15,339개의 100x100x3 RGB 컬러 이미지 및 Grayscale 이미지로 구성되어 있으며 train 12,771개 test 3,068개로 나누어져 있다. FER2013 데이터셋의 경우 입력 이미지를 1x1 컨볼루션을 통해 채널을 증가시킨 뒤 Dense Block의 입력 특징으로 사용하였으며 RAF-single 데이터셋의 경우 3x3 컨볼루션을 통해 특징의 사이즈를 다운 샘플링 한 뒤에 Dense Block에 입력하였다.

그림 2 FER2013, RAF-single 데이터셋 샘플

4.2 학습데이터 전처리

네트워크를 학습하기 위해 사용되는 데이터셋에서 이미지의 각 화소값을 0과 1사이의 값을 갖도록 정규화하였으며 RAF-single 데이터셋의 경우 컬러 이미지 역시 Grayscale로 변환하였다. 또한 학습 데이터에 대해서 Rotation, Shift, Horizontal Flip의 Data Augmentation을 사용하였다.

4.3 실험결과 및 비교

본 논문에서 제안하는 모델의 파라미터 수는 약 0.48 M(Million)개로 FER2013 데이터셋에서 70.69%를 달성하였으며 RAF-single 데이터셋에서 85.4%의 정확도를 달성하였다. 표 1에서 제안하는 모델과 Tang의 모델[12] 사이의 파라미터 수를 비교했을 때 약 25배 가량 적은 수의 파라미터를 사용하였으며 단일 모델에서 72.7%로 가장 높은 정확도를 보인 Paramerdorfer의 VGG 모델[15]보다 약 3.8배 가량 파라미터 수

를 감소시켰다. 표 2의 RAF-single 데이터셋을 사용한 실험 결과에서 제안하는 모델은 Archarya의 모델[16]보다 14배 가량 적은 수의 파라미터를 사용하였다. 또한 Li의 모델[11]과 비교했을 때 약 156배, Vielzeuf의 모델[17]과 비교했을 때 4배 가량 적은 수의 파라미터를 사용했음에도 불구하고 5.4%, 0.7% 더 높은 정확도를 달성하였다.

표2 FER2013 데이터셋 테스트 결과

Method	Accuracy	#Params
Tang [12]	71.2%	12.0 M
Guo et al. [13]	71.33%	2.6 M
Kim et al. [14]	71.86%	1.7 M
Pramerdorfer et al. [15]	72.7%	1.8 M
제안하는 모델	70.69%	**0.48 M**

표3 RAF-single 데이터셋 테스트 결과

Method	Accuracy	#Params
Acharya et al. [16]	87.0%	6.7 M
Vielzeuf et al. [17]	80.0%	2 M
Li et al. [11]	84.7%	74.9 M
제안하는 모델	85.4%	**0.48 M**

5. 결론 및 향후 연구

본 논문에서는 얼굴 표정에서 나타나는 감정을 인식하기 위한 효율적인 컨볼루션 신경망을 설계하기 위해 깊이별 분리 컨볼루션과 특징 연결 및 채널 정보의 압축 및 복원을 위한 Squeeze Excitation Block을 조합하여 새로운 모듈을 제안하였다. 또한 완전연결층(Fully-Connected Layer)을 Global Average Pooling으로 대체하여 완전연결층에서 발생하는 파라미터를 감소시킬 수 있었다. 이를 통해 기존 모델들에 비해 매우 적은 수인 약 0.48 M(Million)개의 파라미터를 가지는 경량화된 컨볼루션 신경망을 설계하였으며 FER2013, RAF-single 데이터셋에서 각각 70.69%, 85.4%의 정확도를 달성하였다.

향후 연구에서 얼굴 표정에서 감정을 인식할 때 주요한 단서가 되는 눈과 입 랜드마크에 집중하여 효과적으로 감정을 인식할 수 있는 어텐션 모듈을 적용한 효율적인 컨볼루션 신경망을 통해 더 높은 성능을 얻을 수 있을 것으로 예상한다.

참 고 문 헌

[1] S. Alizadeh and A. Fazel. "Convolutional Neural Networks for Facial Expression Recognition," arXiv:1704.06756, 2017.

[2] E. Barsoum, C. Zhang, C. C. Ferrer and Z. Zhang. "Training deep networks for facial expression recognition with crowd sourced label distribution," In ICMI, pages 279-283, 2016.

[3] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," arXiv:1602.07360, 2016.

[4] F. Chollet. "Xception: Deep learning with Depthwise Separable Convolutions," arXiv:1610.02357, 2016.

[5] X. Zhang, X. Zhou, M. Lin, and J. Sun. "Shufflenet: An extremely efficient convolutional neural network for mobile devices," arXiv:1707.01083, 2017.

[6] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[7] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger. "Densely connected convolutional networks," In CVPR, 2017.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. "Squeeze-and-Excitation Networks," arXiv:1709.01507, 2017.

[9] S. Ioffe, and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." arXiv:1502.03167, 2015.

[10] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, et al. "Challenges in representation learning: A report on three machine learning contests," In International Conference on Neural Information Processing, Springer, pages 117–124, 2013.

[11] S. Li, W. Deng, and J. Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[12] Y. Tang, "Deep learning using linear support vector machines," arXiv:1306.0239, 2013.

[13] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp 1–6, 2016.

[14] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, pp 48–57, 2016.

[15] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," arXiv:1612.02903, 2016.

[16] D. Acharya, Z. Huang, D. Paudel and L. V. Gool, "Covariance Pooling For Facial Expression Recognition." arXiv:1805.04855, 2018.

[17] V. Vielzeuf, C. Kervadec, A. Lechervy, S. Pateux, and F. Jurie, "An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets," In Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018.