UAV 및 모바일 기기를 위한 얼굴 표정 인식 네트워크

최은지 박병준 윤경로†
건국대학교
yoonk@konkuk.ac.kr†

Face Expression Recognition Network for UAV and Mobile Device

Eunji Choi Byeongjun Park Kyoungro Yoon†
Konkuk University

요 약

최근 자동화의 필요성이 증가함에 따라 얼굴 표정 인식 분야(face expression recognition)가 인공지능과 이미지 처리 분야에서 활발히 연구되고 있다. 본 논문에서는 기존 인공신경망에서 요구되었던 고성능 GPU 환경과 높은 연산량을 극복하고자 모델 경량화(Light weighted Model) 기법을 적용하여 드론 및 모바일 기기에서 적용될 수 있는 얼굴 표정 인식 신경망을 제안한다. 제안하는 방법은 미세한 얼굴의 표정 인식을 위한 방법으로, 입력 이미지의 receptive field 를 늘려 특징 맵의 표현력을 높이는 방법을 제안한다. 또한 효과적인 신경망의 경량화를 위하여, 파라미터의 연산량을 줄일 때 발생하는 문제점을 극복하기 위한 방법을 제시한다. 따라서 제안하는 네트워크를 적용하면 많은 연산량과 느린 연산속도로 인해 제한되었던 네트워크 환경을 극복할 수 있을 뿐만 아니라, UAV(Unmanned Aerial Vehicle, 무인항공기) 및 모바일 기기에서 신경망을 이용한 실시간 얼굴 표정 인식을 할 수 있다.

1. 서론

자동 얼굴 표정인식(FER) 분야는 인간과 컴퓨터의 상호작용(HCI), 스마트 홈 IoT, UAV 및 로봇틱스(Robotics), 고급 운전자 지원 시스템 영역에 다양한 응용 프로그램을 제공한다. 얼굴 표정인식은 자동화를 위한 AI 와 이미지 처리 분야에서 활발한 연구 분야 중의 하나이다. 최근, 인공지능 분야에서는 UAV 와 모바일 IoT 기기 등 온 디바이스(On device) 네트워크에 대한 연구가 진행되고 있다. 그러나, 기존 네트워크에서 요구하는 높은 연산량과 느린 연산 속도로 인해, 그 적용 성과는 미미한 실정이며, 이에 따라 효과적인 모델 경량화 방법이 요구된다.

모델 경량화 기술은 적은 연산과 효율적인 구조로 설계된 알고리즘 연구를 말하며, 모델이 가지는 파라미터의 수를 줄이는 것을 목적으로 한다. 적은 연산과 함께 효율적인 구조를 설계하기 위해서, 불필요한 파라미터를 줄이고 특징 추출 효율을 높여 정확도를 최대한 보존하는 것이 중요하다. 또한, 성능 대비 추론 속도가 중요한 응용 프로그램을 위해 지연시간, 에너지 소모량 감소시키는 것이 필요하다.

본 논문에서는 기존 인공신경망에서 고성능 GPU 환경과 높은 연산량을 요구하는 문제를 해결하여 UAV 와 모바일 기기에서도 실시간으로 얼굴 표정인식이 가능한 네트워크를 제안한다. 이 논문에서는 receptive field 를 늘려 특징 맵의 표현력을 유지하는 법을 제시하고, Residual block 과 SE block 의 교차된 구조의 Shallow Network 를 구성하여, 파라미터를 줄이면서 성능을 유지하는 효율적인 네트워크를 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 데이터 셋과 모델 경량화에 대한 관련 연구를 살펴본 후, 3 절에서는 본 네트워크에서 사용된 모델 경량화 기법과 네트워크 구조에 대해 기술한다. 4 절에서는 제안한 기법의 성능을 실험을 통해서 확인한다. 마지막으로 5 절에서는 본 논문에 대한 결론을 맺는다.

2. 관련 연구

2-1. CK+48 데이터셋

그림 1. CK+48 데이터셋 [1]

그림 1 은 CK+ 48 데이터 셋의 샘플 이미지를 보여준다. CK+48 데이터셋은 Anger, Contempt, Sadness, Surprise, Disgust, Fear, Happy 의 7 가지 얼굴 표현 클래스를 포함한다.[1] 전체 데이터 셋의 개수는 981 개의 라벨링 된 이미지로, 589 개의 test 이미지, 196 개의 validation 이미지, 196 개의 test 이미지로 구성되어 있다.

2-2. MobileNetV1

MobileNetV1 에서는 Depthwise Separable Convolution 을 통한 효율적인 연산 방법을 제안한다[2]. 그림 2 는 일반 convolution 과 Depthwise Separable Convolution 의 구조를 보여준다. Depthwise Separable Convolution 은 하나의 채널에 대해, 하나의 filter 를 적용하는 Depthwise Convolution 과 Depthwise convolution 을 통과한 아웃풋을 N 개의 1x1 convolution 을 통해 새로운 하나의 채널로 합쳐 주는 Pointwise Convolution 로 구성되어 있다. Depthwise Separable Convolution 은 이 구조로 인해, 일반 Convolution 에 비해 약 1/9 배 적은 연산량을 갖는다.

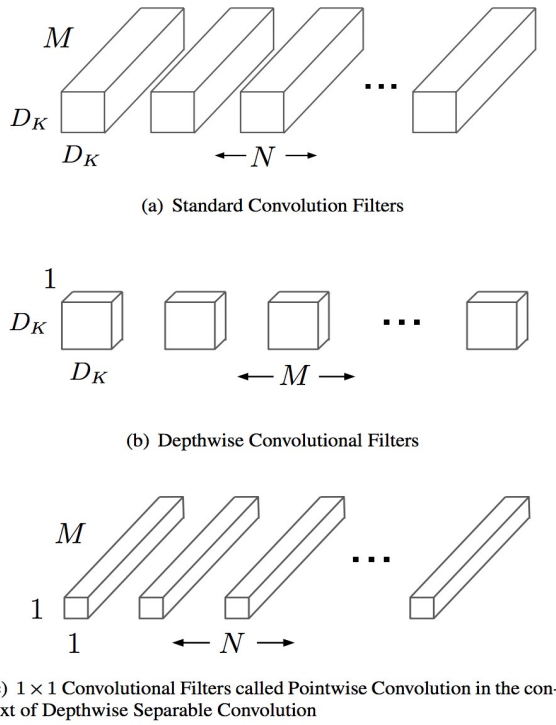

그림 2. 일반 Convolution 과 Depthwise Separable convolution 의 차이점 [2]

2-3. SE Net

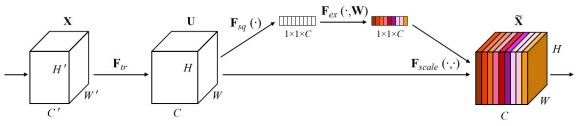

그림 3. Squeeze and Excitation Block [3]

그림 3 은 SE Net 의 Squeeze and Excitation Block 의 구조를 보여준다. SE Net 는 Squeeze and Excitation Block 을 통해 채널 정보를 압축하고, 채널 간의 의존성을 계산하여 채널의 중요도를 스케일 하는 방법을 제시한다[3].

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j)$$

그림 4. Squeeze Module[3]

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})),$$

그림 5. Excitation Module[3]

그림 4 와 그림 5 는 Squeeze Module 과 Excitation Module 의 수식을 보여준다.

Squeeze Module 은 C 개 채널의 2 차원 특징 맵을 Global Average Pooling 을 통해 평균을 계산하고, 각 채널의 중요한 정보를 가지는 1x1 사이즈의 특징 맵으로 변환한다.

Excitation Module 에서는, Squeeze Module 에서 얻은 z 값을 W1 가중치와 곱해주고, δ(ReLU)로 활성화 해준다. 또한, δ로 얻은 값은 W2 가중치와 곱해준 다음 시그모이드 함수로 활성화하여 채널의 중요도를 스케일 하게 된다.

3. 네트워크 구조

Input	Structure	*
(46, 46, 3)	Conv 3x3	S=1, D=2
(44, 44, 8)	Conv 3x3	S=1
(44, 44, 8)	Residual Block	S=2
(22, 22, 16)	SE Block	S=1
(22, 22, 16)	Residual Block	S=2
(11, 11, 32)	Residual Block	S=2
(6, 6, 64)	SE Block	S=1
(6, 6, 64)	Residual Block	S=2
(3, 3, 128)	Conv 3x3	S=2
(1, 1, 7)	Global Average Pool	
*D = dilated rate, D(첫 번째 레이어, 두 번째 레이어) *S = Stride		

표 1. 본 논문의 네트워크 구조

표 1 은 본 논문의 네트워크 구조를 보여준다. 본 논문에서 제안하는 네트워크는 크게 세 단계로 나누어 네트워크를 설명할 수 있다.

첫 단계에서, Dilated convolution 은 특징 맵의 추출하고 표현력을 높이기 위해 적용된다[4]. Dilated convolution 은 필터 내부에 zero padding 을 추가하는 방식으로 연산량을 늘리지 않으며, receptive field 의 크기를 크게 만들어 spatial dimension 의 손실을 줄인다.

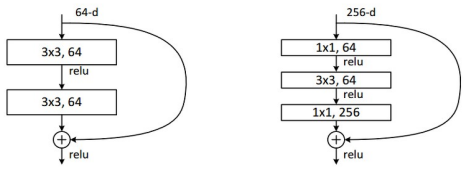

Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152.

그림 4. ResNet 의 Residual Block [5]

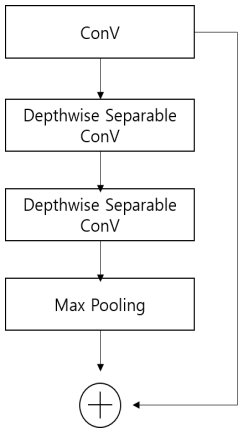

그림 5. 본 네트워크의 Residual Block

그림 4 와 그림 5 는 skip connection 구조를 사용하고 있는 ResNet 의 Residual Block 구조와, 본 네트워크의 Residual Block 구조를 보여준다.

두 번째 단계는, Residual block 과 SE block 의 교차된 구조로 구성된다. 본 네트워크에서 사용한 Residual Block 은 Depthwise Separable layer 이전의 가중치와 풀링 이후의 가중치를 skip connection 구조로 이어 vanishing gradient 를 방지하였으며[5][6], Depthwise Separable Convolution 을 통해 연산 효율을 높였다. 또한, SE block 을 통해, 각 채널들의 중요한 정보만 추출하고, 각 채널에서 중요도에 따라 스케일된 값을 다음 layer 로 넘겨준다.

마지막 단계에서, 특징 맵의 사이즈를 줄이기 위해 3x3 Convolution 필터와 padding 1, stride 2 를 사용하였다. 또한, 많은 연산량을 요구하는 FC 층 대신 Global Average Pooling 을 사용하여 연산효율을 높였다.[7]

4. 실험 결과

4-1. 기존 논문과 비교

4 장에서는 정확도와 파라미터 양의 비교하기 위하여, 동일 데이터 셋을 학습한 다른 네트워크[8]와의 비교를 진행한다.

본 실험은 Lr 0.001, SGD optimizer, Epoch 200 환경에서 진행하였으며, Loss 는 Cross Entropy Loss 를 사용하였다. 실험 결과는 Batch size 를 8 로 설정하고, ELU 를 사용할 때, 97.83(%)로 본 네트워크 실험에서 결과가 가장 좋았다. Vanilla Convolution Network 와 비교할 때, 정확도는 1.47% 부족한 결과를 보이나, 1/230 배 가량 적은 파라미터를 가진다.

Model	Parameter	Acc(%)
Vanilla Convolution Network	15,781,215	99.30
Ours (Batch 8, ELU,* D(2,0))	66,823	97.83
Ours (Batch 8, ELU,* D(2,2))	66,823	97.46
Ours (Batch 8, ELU, *D(2,4))	66,823	97.42
Ours (Batch 8, ReLU)	66,823	96.59
Ours (Batch 16, ELU)	66,823	95.43
Ours (Batch 32, ELU)	66,823	93.77

표 2. 파라미터와 정확도 비교

표 2 는 기존 논문과의 성능 비교와, 실험 셋팅에 따른 정확도를 보여준다.

4-2. Confusion Matrix

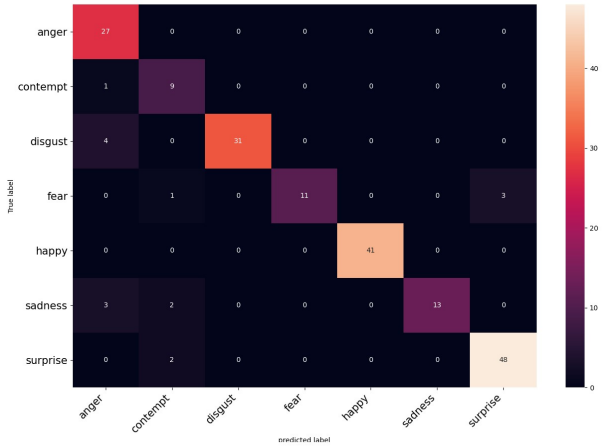

그림 6. Confusion Matrix

그림 6 은 클래스별 Confusion Matrix 결과를 보여준다. Confusion Matrix 결과에 따르면, True label 과 Predicted label 일치하는 결과가 Surprise, Happy, Disgust 순으로 높다, 추론에 사용한 CK+48 데이터 셋은 Anger 이미지 27 장, Contempt 이미지 10 장, Disgust 이미지 35 장, Fear 이미지 15 장, Happy 이미지 41 장, Sadness 이미지 18 장, Surprise 이미지 50 장이며, 이 결과는 클래스별 이미지 개수의 편향으로, 이미지 개수가 많은 클래스의 정확도가 높게 측정된 경향이 있다는 것을 알 수 있다.

5. 결론

얼굴 표정 인식 분야는 자동화에 대한 수요에 따라, 인공지능과 이미지 처리 분야에서 활발히 연구하고 있다. 그러나, 많은 자원이 요구되는 인공 신경망의 환경으로 인해, 적용 성과는 미미하다. 본 모델은 CK+48 데이터 셋을 사용한 실험에서 66,823 의 파라미터를 가지면서, 97.83%의 성능을 보여주었다. 이것은 같은 데이터 셋으로 학습한 Vanila Convolution Network 보다 파라미터를 1/230 배 줄인 결과이다. 우리는 본 논문을 통해 자원이 제한적인 UAV 및 모바일 기기에서 사용 가능한 얼굴 표정인식 네트워크를 제안한다.

감사의 글

본 연구는 2020 년도 산업통상자원부(산업용 무인비행장치 전문인력 양성사업, 과제번호 : N0002431)의 지원으로 수행되었음.

Reference

[1] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010.]

[2] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[3] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Arriaga, Octavio, Matias Valdenegro-Toro, and Paul Plöger. "Real-time convolutional neural networks for emotion and gender classification." arXiv preprint arXiv:1710.07557 (2017).

[7] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[8] Videla, Lakshmi Sarvani, and PM Ashok Kumar. "Facial Expression Classification Using Vanilla Convolution Neural Network." 2020 7th International Conference on Smart Structures and Systems (ICSSS). IEEE, 2020.