얼굴 표정 인식을 위한 Densely Backward Attention 기반 컨볼루션 네트워크

서현석, Cam-Hao Hua, 이승룡
경희대학교 컴퓨터공학과

e-mail : {shs,hao.hua,sylee}@oslab.khu.ac.kr

Convolutional Network with Densely Backward Attention for Facial Expression Recognition

Hyun-Seok Seo, Cam-Hao Hua, Sung-Young Lee
Dept. of Computer Science and Engineering, Kyung-Hee University

요 약

Convolutional neural network(CNN)의 등장으로 얼굴 표현 인식 연구는 많은 발전을 이루었다. 그러나, 기존의 CNN 접근법은 미리 학습된 훈련모델에서 Multiple-level 의 의미적 맥락을 포함하지 않는 Attention-embedded 문제가 발생한다. 사람의 얼굴 감정은 다양한 근육의 움직임과 결합에 기초하여 관찰되며, CNN 에서 딥 레이어의 산출물로 나온 특징들의 결합은 많은 서브샘플링 단계를 통해서 class 구별와 같은 의미 정보의 손실이 일어나기 때문에 전이 학습을 통한 올바른 훈련 모델 생성이 어렵다는 단점이 있다. 따라서, 본 논문은 Backbone 네트워크의 Multi-level 특성에서 Channel-wise Attention 통합 및 의미 정보를 포함하여 높은 인식 성능을 달성하는 Densely Backwarnd Attention(DBA) CNN 방법을 제안한다. 제안하는 기법은 High-level 기능에서 채널 간 시멘틱 정보를 활용하여 세분화된 시멘틱 정보를 Low-level 버전에서 다시 재조정한다. 그런 다음, 중요한 얼굴 표정의 묘사를 분명하게 포함시키기 위해서 multi-level 데이터를 통합하는 단계를 추가로 실행한다. 실험을 통해, 제안된 접근방법이 정확도 79.37%를 달성 하여 제안 기술이 효율성이 있음을 증명하였다.

1. 서론

최근 딥러닝 기술의 등장으로 시각 데이터처리에 많은 진보를 이루었으며, 컴퓨터 비전 분야에서 널리 사용되고 있다. 특히, 딥러닝 분야에서 잘 알려진 Convolutional Neural Network(CNN)[1]-[4]는 다양한 인식 기반 문제를 해결하고 시각 데이터 처리에 뛰어난 성능을 보이고 있어 얼굴 표정 인식(FER: Facial Expression Recognition)[5]에 많이 활용되고 있다.

FER 은 인간-컴퓨터상호작용 (HCI: Human-Computer Interaction) 분야와 관련된 다양한 응용 프로그램에서 활용되어 왔으며, 현재까지도 활발한 연구가 진행되고 있다[6]. 최근에는 실험실 환경[7]과 실제 환경[8]에서 수집되는 이미지 수가 비약적으로 증가함에 따라 CNN 을 이용해 인지 프로세스를 진행하는 연구들이 증가하고 있다. 특히, FER 분야에서는 CNN 기반 접근법을 통해 많은 문제들을 해결하고 있다.

얼굴표정의 감정은 사람 얼굴의 눈, 눈썹, 코, 입 모양 등 다양한 근육의 형태의 조합으로 나타난다. 이러한 다양한 상황 인지를 위해 최근 연구들은 얼굴 특징뿐만 아니라 높은 감정 인지를 위해 다양한 근육이 조합되는 상황을 인지하기 위해 CNN 은 다양한 단계에서 서로 다른 근육들의 여러 개의 네트워크 통합하여 감정을 인식한다.

CNN Backbone 네트워크 통합 과정에서 의미적 정보를 포함하여 네트워크를 구성을 통해 전이학습 수행하면 FER 에 대한 성능을 높일 수 있다.

따라서, 본 논문은 FER 에서 높은 정확도를 달성할 수 있도록 사전에 훈련된 CNN 모델에서 다양한 단계의 특징맵에 의미를 포함하여 전이학습을 수행하는 Densely Backward Attention(DBA)-Network 를 제안한다. 제안하는 기법은 High-level 계층에서의 특징과 Low-level 계층에서의 특징 등 다양한 의미 정보를 집계하여 DBA 방식을 활용하여 해당 얼굴 표정을 인식한다.

2. Densely Backward Attention 기반 컨볼루션 네트워크

2.1 제안하는 컨볼루션 네트워크 아키텍처

제안된 아키텍처는 ImageNet[14]으로 사전 훈련된 Backbone CNN 과 DBA 의 관련 경량스트림의 두 부분

※This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016K1A3A7A03951968)

으로 구성된다. 사전에 훈련된 CNN 에 DBA 메커니즘이 부착되어 있는 전반적인 구조를 상세히 기술한다. 그 후, 관심 특징을 추출하는 것(그림 1 의 갈색 모듈)에 대한 설명은 2.2 에서 한다.

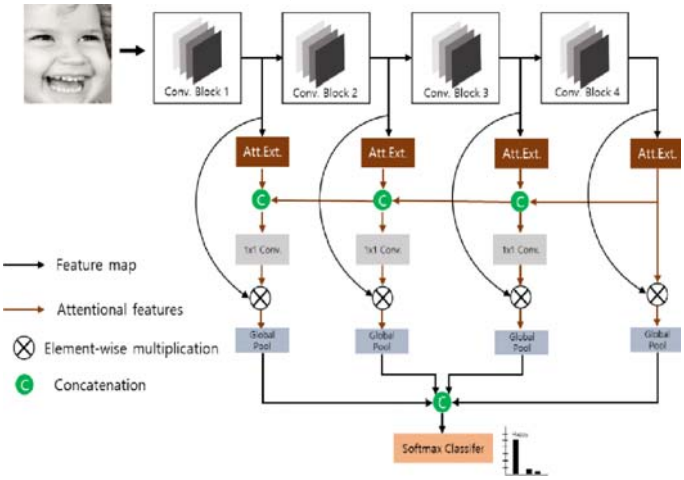

그림 1. 얼굴 표정 인식을 위해 제안된 DBA-Net 의 구조.

그림 1 에서 박스에 있는 컨볼루션 블록은 Backbone CNN 의 기본 구성요소를 나타내며, 나머지는 얼굴 감정 인식을 위한 다차원적 정보를 수집하는 Attention-embedded 흐름을 나타낸다. 'Conv bolock'과 'Att.Ext' 두 개는 여러 개의 컨볼루션 계층과 관심 특징 추출기의 블록을 나타낸다. '1x1 Conv'와 'Global Pool'은 각각 커널 크기가 1x1 컨볼루션 계층과 평균 Global Pooling 계층을 의미한다. VGG[1], ResNet[2], DenseNet[3] 등으로 구성된 각기 다른 CNN 기술을 적용하여 기술의 유연성을 보여줄 수 있다. 전형적으로, 분류 네트워크에서는 각 컨볼루션 블록의 계층들은 시멘틱 레벨에 해당하는 특정 스케일로 획득된 특징을 학습하고 수행한다. 예를 들어, ResNet 과 DesnNet 은 모두 4 개의 기본 컨볼루션 블록들(그림 1 참고)로 구성되어 있다. 이 중 마지막 출력은 입력의 공간 사이즈와 비교하여 각각 4, 8, 16, 32 개의 필터를 적용하는 간격을 가지며, 각 블록에서 비선형 활성 계층과 컨볼루션의 전체 수는 다양하다. 따라서, 계산량이 증가하는 것을 설명 하기 위해, 앞에서 설명한 학습 가능한 블록의 최종 산출물인 4 개의 특징 맵만 주의 추출 단계를 고려한다. 한편, VGG 아키텍처에는 컨볼루션 블록에 대한 명확한 정의가 없으므로, 이전과 동일한 스텝을 갖는 마지막 4 개의 최대-풀링 계층 앞에 있는 Rectified Linear Unit(ReLU) 활성화 계층의 출력이 추가 프로세스에 선택된다.

컨볼루션 블록들 사이의 피드-포워드 흐름에 따라, 추출된 특징 맵들의 공간 해상도는 절반으로 감소되며, 깊이 사이즈는 빠르게 증가한다. 또한 이후 계층의 결과는 앞에서 얻은 것과 비교하여 채널 차원에 더 의미 있는 맥락을 포함한다. 그것들은 하위 계층에서 추출한 특징에 대해 역순으로 재보정(필요없는 정보를 제거)하는데 활용할 수 있다. 이러한 작동에 의해 검토된 low-level 특징 맵의 공간 디테일은 애매모호함을 제거하는데 도움이 되고, 완전한 시멘틱 정보가 된다. 결과적으로는, FER 의 성능을 향상시키기 위해 많은 시멘틱 정보를 넣고, 얼굴 형태가 잘표현된 미세한 패턴의 특징 맵을 포함하는게 좋다. 다음 내용들은 각각 관심 특징을 추출하기 위한 알고리즘과 해당 결과가 관심 특징 맵에 어떻게 포함되는지를 보여준다.

2.2 Attentional Feature Extractor

관심 추출 메커니즘의 (Attentional Feature Extractor) 주요 목적은 더 나은 학습 과정을 위해서 중요한 특징들을 보다 집중적으로 훈련할 수 있도록 하는 것이다. 기존의 관심 추출 메커니즘은[4] 특징 가중치를 각 채널에서 전체적인 정보에 맞춰서 스스로 보정하는 방식을 활용한다. 그러나, 제안된 네트워크는 핵심 CNN 에서 선택한 특징 맵의 교차채널과 함께 스스로 능력을 향상시키고, 다양하고 많은 디테일을 low-level 에 포함시키기 위해 활용된다.

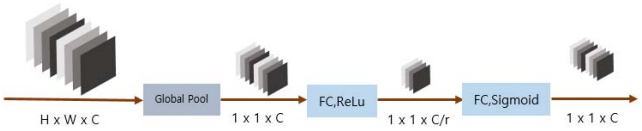

그림 2. 관심 특징 추출기

그림 2 는 제안하는 관심 특징 추출기의 예시로, 처음에는 $H \times W \times C$ 크기의 관련 특징 맵의 각 채널이 공간적으로 평균화하여 C 길이 벡터를 생성한다. f_i는 앞서 말한 컨볼루션 블록들로부터 출력된 특징 맵이며, 여기서 $i = 1, 2, 3, 4$는 Backbone 네트워크에서의 피드포워드 방향에 해당한다. 즉, f_1는 관심 있는 최하위 레벨, f_2는 최상위 레벨의 특징 맵을 나타낸다. 그리고, $g_i \in R^C$로 표현된 벡터는 채널간 상호의존성에 대한 정보를 수행한다.

$$g_i = [G_1(f_1), \cdots, G_c(f_i), \cdots, G_C(f_i)]^T \qquad (1)$$

수식 1 에서 G(.)는 고려된 특징 맵 f_i의 c^{th} 채널을 다음과 같은 방정식으로 처리하는 Global Pool(GP) 연산자를 나타낸다.

$$G_c(f_i) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} [f_i(h, w)]_c \qquad (2)$$

수식 2 에서 h=1,…,H 와 w=1,…,W 는 특징 맵의 f_i의 픽셀 좌표이며, $[.]_c$는 관련 특징 맵 또는 벡터의 c^{th} 채널을 의미한다. 결과적으로, 깊이별 시멘틱 정보는 C 길이 벡터들 g_i에 종합적으로 인코딩된다.

그 후, 채널 전체의 기본 상관관계를 적절하게 모델링하기 위해, 벡터 g_i를 ReLU 활성 함수에 의해 중심이 되는 2 개의 Fully Connected(FC) 층으로 공급한다. 이러한 훈련 가능한 계층의 크기는 각각 C/r 및 C와 동일하며, 여기서 r 은 계산 복잡도를 줄이기 위해 본 연구에서는 압축 비율을 16 으로 고정하여 진행한다. 이러한 학습 절차는 아래와 같이 제시될 수 있다.

$$G_{int_i} = W_{fc_2}\left(ReLU\left(W_{fc_1 g_i} + b_{fc_2}\right)\right) + b_{fc_2} \quad (3)$$

수식 3 에서 $\left\{W_{fc_1} \in R^{C\times\frac{C}{r}},\ b_{fc_1} \in R^{C}\right\}$ 와 $\left\{W_{fc_2} \in R^{\frac{C}{r}\times C}, b_{fc_2} \in R^{\frac{C}{r}}\right\}$는 각각 첫 번째 및 두 번째 FC 레이어의 학습 가능한 매개 변수이고 g_{int_i}는 C 길이의 중간의 깊이별 특징 벡터다. 그 다음, Sigmoid 활성 함수를 이용하여 아래 공식에 따라 벡터 g_{int_i}의 반응을 (0, 1) 범위에서 다시 가중치를 부여한다.

$$g_{att_i} = \frac{1}{1 + e^{-\left[g_{int_i}\right]_c}} \quad (4)$$

`수식 4 에서 $g_{att_i} \in R^{C}$는 Backbone CNN 에서 추출한 관심 특징 맵에서 정보 컨텍스트를 풍부하게 하기 위해 사용되는 채널별 관심 특징 벡터를 나타낸다.

2.3 Densely Backward Attention(DBA) schemer

앞서 설명한 관심 특징 추출기로 각각 선택된 4 개의 특징맵 f_1, f_2, f_3, f_4에 해당하는 4 개의 관심 특징 벡터 g_{att1}, g_{att2}, g_{att3}, g_{att4} (그림 1 에서 Att. Ext 와 갈색화살표로 표시)로 추론한다. 다양한 Backbone 모델에 관련하여, 벡터 g_{att_i}는 표 1 에 보이는 거와 같이 다른 길이의 C 를 갖는다. 큰 차원을 가진 g_{att_i}는 더 유익한 특수성으로 구성되며, 이것은 이전의 컨볼루션 블록으로부터 학습되어있는 미세하게 분해된 특징을 재교정하는데 사용 할 수 있다. 따라서, 감정 예측에 미세한 패턴의 특징을 포함시키기 위해서는, Backbone CNN 의 상위 레벨의 특징으로부터 채널별 시멘틱 디테일을 역으로 생각하여 구현한다. 또한, 각 관심 특징 벡터 g_{att_i}는 특정 레벨의 서술적 통계를 포함하고 있기 때문에, 그러한 다양한 레벨 시멘틱 특징을 통합하기 위해 밀집된 결합 방법을 선택하는데, 이 방법은 low-level 수준의 특성의 디테일한 공간보다 유연하게 개선시키기 위함이다.

그림 1 에서 나타난 것과 같이 관심 특징을 추출한 후 다음과 같은 작업이 수행된다.

$$\begin{aligned} f_{att4} &= f_4 \otimes g_{att4} \\ f_{att3} &= f_3 \otimes W_{113}(C[g_{att3},\ g_{att4}]) \\ f_{att2} &= f_2 \otimes W_{112}(C[g_{att2},\ g_{att3},\ g_{att4}]) \\ f_{att1} &= f_1 \otimes W_{111}(C[g_{att1},\ g_{att2},\ g_{att3}, g_{att4}]) \end{aligned} \quad (5)$$

수식 5 에서 $\otimes$는 요소별 곱셈 연산자를 나타낸다. C[.]는 벡터 연속된 결합을 나타내며, $\{W113,\ W112,\ W111 \in R^{D\times1\times1\times C}\}$는 크기가 $1\times1\times$D 인 C 컨볼루션 필터의 학습 가능한 매개변수를 나타낸다. C 는 고려된 특징 맵 f_i의 채널 크기인 반면, D 의 값은 관심 벡터 $g_{att_i}, g_{att_{i+1}}, \cdots, g_{att_4}$에 의해 연결된 출력의 차원에 따라 달라진다.

예를 들어, VGG-16 모델을 사용할 경우 W111, W112, W113 의 값 D 는 각각 1024, 1280 ,1408 이다(표 1 의 C 값에 기초). 요약하자면, 학습 가능한 1×1 컨볼루션 계층의 기능은 고려되었던 특징 맵의 재보정을 위해 연결된 특징 벡터의 치수를 효과적으로 줄이는 것이다. 게다가, 이러한 densely backward 방법에 의해 얕은 층에서 발견된 특징은 더 효율적으로 개선되어 더 높은 수준의 표현을 얻을 수 있다. 그런 다음, 가중치가 변경된 특징 맵 f_{att_i}는 아래와 같이 다양한 레벨에서 전체적으로 필요한 각각의 컨텍스트를 수집 및 통합하기 위해 GP 모듈을 통과한다.

$$f_{fer} = C[G(f_{att1}), G(f_{att2}), G(f_{att3}), G(f_{att4})] \quad (6)$$

수식 6 에서 f_{fer}는 DBA-Net 의 핵심적인 특성이며 G 는 (1)과 (2)에 정의된 GP 연산자를 나타낸다. 마지막으로, Softmax 분류기를 사용하여 사전 정의된 감정 표현에 대한 라벨로 해당 얼굴 표정을 인식한다.

<표 1>
추출된 관심 특징 벡터 $\boldsymbol{g_{att_i}}$의 길이 C
$(\boldsymbol{i = 1, 2, 3, 4})$, 각각의 Backbone CNN

BACKBONE CNN	$\boldsymbol{g_{att1}}$	$\boldsymbol{g_{att2}}$	$\boldsymbol{g_{att3}}$	$\boldsymbol{g_{att4}}$
VGG-16[1]	128	256	512	512
RESNET-101[2]	256	512	1024	2048
DENSENET-161[3]	384	768	2112	2208

참고: C 값은 Backbone 네트워크에서 추출된 해당 특징 맵 $\boldsymbol{f_1, f_2, f_3, f_4}$의 크기와도 동일하다.

3. 실험 방법

3.1 데이터셋 준비

RAF-DB[8]는 Real-world Affective Faces Database 로 약 3 만개의 다양한 얼굴 이미지를 가진 대규모 데이터베이스다. 이 데이터베이스는 나이, 성별, 다양한 인종, 머리 모양, 조명 상황 등 매우 다양하게 있다. 본 논문에서는 싱글 라벨로만 실험한다. 각 이미지는 분노, 혐오, 두려움, 행복, 무표정, 슬픔, 놀라움 등 7 가지 기본 감정 중 하나만 분류된다. 얼굴 영역 주위로 100×100의 해상도로 자른 뒤, 12,271 개의 데이터셋과 3,068 개의 테스트셋으로 실험한다.

3.2 구현

제안된 모델과 모델 평가는 Pytorch[15], Scikit-learn[16] 프레임워크를 사용한다. 기존 작업과 마찬가지로, 64 배치사이즈로 설정한 후, 색조와 채도는 임의로 변경되게 하고, 트레이닝 배치 회전은 (20°, 20°) 범위로 했다. 제안된 모델은 오버피팅 방지를 하기 위해 가중치 감소를 0.0005 로 설정했다. 트레이닝에 대해서는 Learning Rate 는 0.005 로 초기화하고 Softmax 손실 함수를 사용하여 주어진 Ground-Truth 라벨로 DBA-Net 의 매개변수의 성능을 평가한다. 그런 다음, 트레이닝 가능한 매개 변수와 관련하여 계산되었던 손실을 최소화하기 위해, 0.9 의 momentum 을 갖는 경사 하강법과 Learning Rate 감소를 위해 'poly'-스타일을 함께 사용되는 [17]의 최적화 방법을 사용한다. 그리고, 학습 과정은 NIVIA 1080 TI GPU 1 개에서 50 개의 epoch 으로 진행한다.

<표 2> RAF-DB[8]의 정확성 비교 테스트셋 및 최신 기술과의 비교

네트워크명	정확성(%)
DLP-CNN [8]	74.20
3DMFA [12]	75.73
ResiDen [11]	76.54
MRE-CNN [10]	76.73
Capsule-based Net [9]	77.48
Double Cd-LBP [18]	78.60
SPDNet [13]	79.43
DBA-Net (VGG-16)	78.81
DBA-Net (ResNet-101)	79.33
DBA-Net (DenseNet-161)	79.37

제안된 DBA-Net 은 표 2 에서 표시된 정량적 결과를 기반으로 최신 기술에 비해 경쟁력 있는 결과를 보여준다. 제안된 아키텍쳐에서 VGG-16 을 핵심 네트워크로 적용해봄으로써 비교된 값(SPDNet[13]을 제외)을 보면 0.21-4.61%의 높은 결과를 볼 수 있다. 또한 ResNet-101 과 DenseNet161 보다 깊은 신경망 Backbone 네트워크를 사용하면 제안된 접근 방식의 성능이 지속적으로 향상되는 것을 볼 수 있고, 이는 SPDNet[13]의 성능과 비교할 때 정확도가 0.06-0.1% 낮은 모습을 볼 수 있다. 데이터셋에서 이러한 인상적인 성능은 밀집된 백워드 구조에서 채널별 관심 메커니즘을 활용해 low-level 과 high-level 의 기능을 통합하는 이점을 보여준다.

4. 결론

본 연구는 전이학습을 위한 의미정보 손실을 방지하는 Backbone 컨볼루션 네트워크 DBA-Net 을 제안한다. 제안된 접근 방법은 얼굴 감정이 여러 수준에서 추출된 다른 근육의 융합에 의해 표현된다는 가설에 따라 효율적인 방법으로 low-level 및 high-level 에서의 특징을 통합한다. 이러한 접근법으로, FER 의 성능을 향상시키기 위해 사전 훈련된 분류 기반 CNN 내에서 관심 특징이 역으로 활용된다. 실험을 통하여 제안하는 컨볼루션 네트워크 알고리즘이 기존의 알고리즘보다 높은 성능을 나타냄을 입증하였다.

5. 참고문헌

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 2261–2269.

[4] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[5] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, “Weightadapted convolution neural network for facial expression recognition in human-robot interaction,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, pp. 1–12, 2019.

[6] S. Li and W. Deng, “Deep facial expression recognition: A survey,” CoRR, vol. abs/1804.08348, 2018.

[7] P. Lucey, J. F. 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 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, June 2010, pp. 94–101.

[8] S. Li and W. Deng, “Reliable crowdsourcing and deep localitypreserving learning for unconstrained facial expression recognition,” IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 356–370, 2019.

[9] S. Ghosh, A. Dhall, and N. Sebe, “Automatic group affect analysis in images via visual attribute and feature networks,” in 2018 25th IEEE International Conference on Image Processing (ICIP), Oct 2018, pp. 1967–1971.

[10] Y. Fan, J. C. Lam, and V. O. Li, “Multi-region ensemble convolutional neural network for facial expression recognition,” in International Conference on Artificial Neural Networks. Springer, 2018, pp. 84–94.

[11] S. Jyoti, G. Sharma, and A. Dhall, “Expression empowered residen network for facial action unit detection,” in 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), May 2019, pp. 1–8.

[12] F. Lin, R. Hong, W. Zhou, and H. Li, “Facial expression recognition with data augmentation and compact feature learning,” in 2018 25th IEEE International Conference on Image Processing (ICIP), Oct 2018, pp. 1957–1961.

[13] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, “Covariance pooling for facial expression recognition,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2018, pp. 480–4807.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, “Imagenet large scale visual recognition challenge,” CoRR, vol. abs/1409.0575, 2014.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in NIPS Autodiff Workshop, 2017.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoderdecoder with atrous separable convolution for semantic image segmentation,” in Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018, pp. 833–851.

[18] F. Shen, J. Liu, and P. Wu, “Double complete d-lbp with extreme learning machine auto-encoder and cascade forest for facial expression analysis,” in 2018 25th IEEE International Conference on Image Processing (ICIP), Oct 2018, pp. 1947–1951.