얼굴사진 기반 감정인식 모델의 특성 분석[1)]

김민경 양지윤 최유주*
서울미디어대학원대학교 인공지능응용소프트웨어학과
muzzcats@naver.ac.kr, njs10919@gmail.com, yjchoi@smit.ac.kr

Feature Comparison of Emotion Recognition Models using Face Images

MinGeyung Kim, Jiyoon Yang, Yoo-Joo Choi*
Department of AI Software Engineering, Seoul Media Institute of Technology
*Corresponding Author

요 약

본 논문에서는 얼굴사진 기반 감정인식 심층망, 음성사운드를 기반한 감정인식 심층망을 결합한 앙상블 네트워크 구축을 위한 사전연구로서 얼굴사진 기반 감정을 인식하는 기존 딥뉴럴 네트워크 모델들을 입력 데이터 처리 방법에 따라 분류하고, 각 방법의 특성을 분석한다. 또한, 얼굴사진 외관 특성을 기반한 감정인식 네트워크를 여러 구조로 구성하고, 구성된 방법의 성능을 비교하여, 우수 성능을 보이는 네트워크를 선정하여 추후 앙상블 네트워크의 구성 네트워크로 사용하고자 한다.

1. 서론

메타버스 시대로의 도입과 산업 분야 곳곳에 일어나는 디지털 트랜스포메이션 산업 고도화와 더불어 발전된 인간과 컴퓨터 상호기술이 요구되고 있고, 이에 따라 사용자의 감정을 예측하고, 예측된 감정에 따른 서비스를 제공하기 위한 감성 컴퓨터(Affective computing)에 관한 관심이 모아지고 있다.

감성, 감정을 예측하는데 있어서 얼굴사진과 음성 데이터들이 활용되고 있고, 특히, 얼굴사진을 이용한 감정인식의 정확도를 높이기 위한 다양한 얼굴 데이터베이스[1]와 네트워크 모델[1-3]들이 제시되어 왔다.

본 연구는 얼굴사진과 사운드 정보를 기반한 감정인식을 위한 앙상블 네트워크 구성을 최종목표로 하고, 앙상블 네트워크 구축을 위한 사전연구로서 얼굴사진을 기반으로 감정을 인식하는 기존 딥뉴럴 네트워크 모델을 분류하고 특성을 분석한다. 또한, 얼굴사진을 기반한 감정인식 네트워크를 여러 구조로 구성하고, 구성한 네트워크의 성능을 비교 분석함으로써, 추후 앙상블 네트워크 구성 네트워크로 활용하고자 한다.

1) 본 연구는 문화체육관광부 “관광서비스 혁신성장 연구개발사업”(R2022020105)의 지원에 의하여 수행되었음

2. 얼굴사진 기반 감정인식 심층망 분류

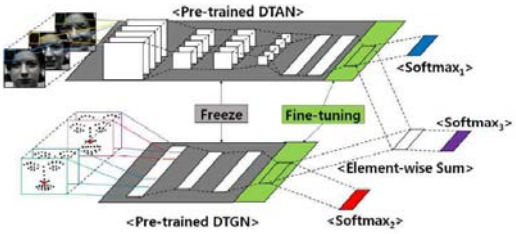
얼굴사진을 기반으로 감정을 추정하는 딥러닝 네트워크는 입력데이터의 구성형태에 따라 다음과 같이 구분할 수 있다.

첫째, 얼굴영상을 미리 학습된 네트워크를 적용하여 잠재벡터(latent vector)로 추출하여 이를 입력으로 감정레이블과 매칭시키는 네트워트 형태이다[2].

둘째, 연속 얼굴 이미지들과 각 이미지로부터 추출한 얼굴 특징점들(facial landmarks)을 추출하여 이를 입력으로 감정인식을 학습하는 형태이다[3].

2.1 미리 학습된 모델을 통한 잠재벡터 추출 기반 방법

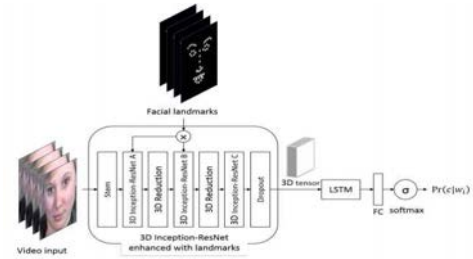
Jung과 Lee등[2]의 방법에서는 미리 학습된 DTAN(Deep Temporal Appearance Network) 모델과 DTGN(Deep Temporal Geometry Network) 모델을 통하여 단계별 잠재벡터를 추출하고, 이를 소프트맥스 함수의 입력으로 사용하여 확률값을 추출하고, 3개의 오차함수를 정의하여 학습단계에서 사용하고, 추론과정에서는 세 번째 오차함수만을 적용하는 방법을 사용한다. [그림 1]은 [2]에서 제시한 네트워크 구조를 보여주고 있다.



[그림 1] Joint fine-tuning method [2]

2.2 연속 이미지와 얼굴 특징점을 이용한 방법

Hasani와 Mahoor 등[3]은 연속된 얼굴이미지와 얼굴 특징점들을 3D Inception-ResNet의 입력으로 사용하고, 이의 출력을 LSTM 의 입력으로 사용하는 네트워크를 구성하고 얼굴 감정인식을 수행하였다. [그림 2]는 [3]에서 제안한 네트워크 구조를 설명하고 있다.



[그림 2] 3D Inception-ResNet을 이용한 얼굴 감정인식[3]

3. 이미지 분류 네트워크 기반 감정인식 실험

본 연구팀은 실시간 처리가 가능한, 얼굴의 외관(appearance) 특징 기반 감정인식 네트워크와 얼굴 특징점 기반 감정인식 네트워크를 결합한 정확도가 높은 앙상블 네트워크 구축을 목표로 하고 있다. 이를 위한 사전 연구로서 실시간 처리가 가능한 효율적인 이미지 분류 네트워크를 기반으로 감정인식 실험을 수행하였다. 본 논문에서는 효율적인 이미지 분류 네트워크로서 MobileNetV2 와 EffectiveNet을 사용하고, 각 네트워크를 2계층의 완전연결계층(Dense Layer)과 연결하여 감정인식 성능실험을 수행하였다. 학습 및 테스트 데이터로 FER2013 데이터셋을 사용하였다.

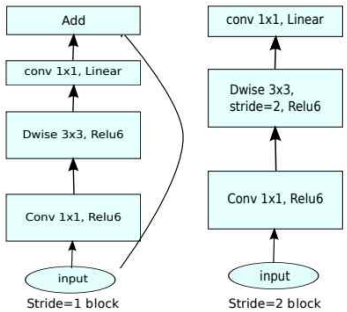
3.1 데이터셋

FER2013 데이터셋은 대략 3만장의 다양한 표정의 48x48 고정사이즈의 RGB이미지이다. 기본 감정 6개와 중립(Neutral)의 감정 1개 총 7개의 감정으로 레이블을 가진다. 즉, Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral로 구성되어 있다. [그림 3]는 FER2013 데이터셋의 샘플영상을 보여주고 있다.

[그림 3] FER2013 데이터셋의 샘플 [1]

3.2 적용 네트워크

(1) MobileNetV2

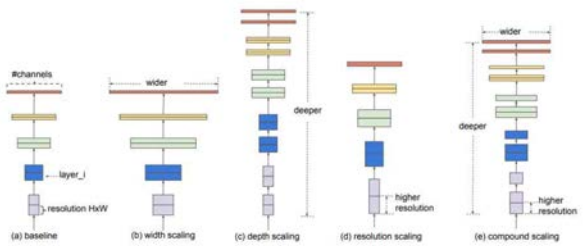


[그림 4] MobileNet V2의 convolution block [4]

MobileNetV2는 ReLU 활성함수를 통과하게 되면 정보가 손실되는 점에 착안하여 손실의 최소화를 위해 제안되었고, 표현력을 유지하기 위해 채널이 적은 레이어에서 비선형성을 제거하는 것이 중요하다는 성질을 이용하여 성능을 향상시켰다. [그림 4]는 MobileNetV2의 block구조를 나타낸 그림이다. Inverted Residucals와 Linear Bottlenecks를 사용하는 구조를 기반으로 중간 확장 계층에서 기능을 필터링하기 위해 깊이별 경량시킨 Depthwise separable convolution을 사용했다.

(2) EfficientNet

기본으로 주어진 기본 네트워크를 Scale-up방법은 3가지가 있는데, 깊이, 넓이, 그리고 해상도를 확장하는 방법이다. 본 네트워크의 목적은 최적의 조합을 AutoML을 통해 찾아 기존보다 훨씬 적은 파라미터수로 좋은 성능을 내도록 하는 것이다. [그림 5]에서 사용하는 model scaling 의 종류를 보여주고 있다.



[그림 5] Model Scaling [5]

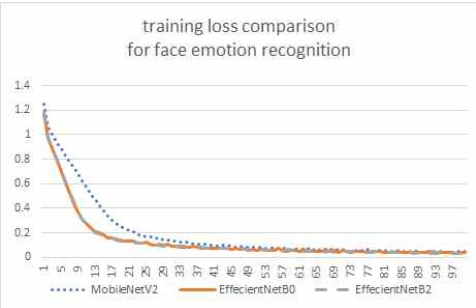
4. 실험방법 및 결과

네트워크는 MobileNetV2, EfficientNetB0, EfficientNetB2을 기본 네트워크로 하고, 각 네트워크 수행 후 2계층의 완전연결계층 (Dense Layer)을 수행하였다. FER2013 데이터셋의 학습데이터를 이용하여 각각 100 에폭 학습 수행 후, 테스트 데이터를 적용하여 성능을 비교 하였다. [표 1]은 각 적용 네트워크에 대한 학습데이터 정확도와 테스트데이터 정확도를 보여주고 있다.

[표1] FER2013 데이셋을 이용한 네트워크 비교실험

Methods	Train Accuracy	Test Accuracy
MobileNetV2 - Dense-Dense	0.96586	0.64698
EfficientNetB0 - Dense-Dense	0.99362	0.65798
EfficientNetB2 - Dense-Dense	0.99010	0.65213

각각의 네트워크를 100에폭 훈련하여 적합한 네트워크를 선별하기 위해서 결과를 분석하였다. 실험결과 EfficientNetB0가 테스트 데이터에 대한 정확도 65.798%로 비교 네트워크 중 가장 높은 정확도를 보여 주었다.



[그림 6] 감정인식을 위한 훈련 손실 그래프

[그림 6]은 각 네트워크의 손실함수를 그래프로 나타낸 것인데, x축은 훈련 에폭(Epochs)을 나타내며 y축은 손실값(Loss)을 나타낸다. 가장 빨리 수렴하는 것은 EfficientNet 계열이며 MobileNetV2도 50 에폭 이후로는 동일하게 수렴하였다.

5. 결론

본 논문에서는 얼굴사진을 기반으로 감정을 인식하는 기존 딥뉴럴 네트워크 모델을 분류하고 특성을 분석하였다. 또한, 실시간 감정인지를 위하여 효율성이 높은 이미지분류 네트워크를 사용하여 FER2013 데이터셋을 이용한 감정인지실험을 수행하였다. 수행결과 65%대의 테스트데이터 판별 정확도를 보였다. 추후 얼굴특징점을 결합한 앙상블 네트워크를 구성하여 판별 정확도를 향상시키고자 한다.

참고문헌

[1] Pramerdorfer, C., & Kampel, M., “Facial expression recognition using convolutional neural networks: state of the art”. arXiv preprint arXiv:1612.02903. ,2016.

[2] Yim, J., Park, S., Kim, J., “Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition”, In Proceedings of the 2015 IEEE International Conference on Computer Vision(ICCV), Santiago, Chile, 7-13, December 2015.

[3] Hasani, B., Mahoor, M.H., “Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks”, In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshop(CVPRW), pp. 2278-2288, Honolulu, HI, USA, 21-26 July 2017.

[4] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[5] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.