Manifold Learning 을 통한 표정과 Action Unit 간의 상관성에 관한 연구

김선빈*, 김현철*
*고려대학교 컴퓨터학과
e-mail : {lsmgame, harrykim}@korea.ac.kr

A Study in Relationship between Facial Expression and Action Unit

Sunbin Kim*, Hyeoncheol Kim*
*Dept. of Computer Science and Engineering, Korea University

요 약

표정은 사람들 사이에서 감정을 표현하는 강력한 비언어적 수단이다. 표정 인식은 기계학습에서 아주 중요한 분야 중에 하나이다. 표정 인식에 사용되는 기계학습 모델들은 사람 수준의 성능을 보여준다. 하지만 좋은 성능에도 불구하고, 기계학습 모델들은 표정 인식 결과에 대한 근거나 설명을 제공해주지 못한다. 이 연구는 표정 인식의 근거로서 Facial Action Coding Unit(AUs)을 사용하기 위해서 CK+ Dataset 을 사용하여 표정 인식을 학습한 Convolutional Neural Network(CNN) 모델이 추출한 특징들을 t-distributed stochastic neighbor embedding(t-SNE)을 사용하여 시각화한 뒤, 인식된 표정과 AUs 사이의 분포의 연관성을 확인하는 연구이다.

1. 서론

우리가 감정을 표현하는데 있어서, 표정보다 더 강력한 의사소통 방법을 찾아보기 힘들다. 이러한 이유를 차치하더라도, 표정 인식은 기계 학습과 컴퓨터 비전 분야에서 아주 중요한 기술 중에 하나이다.

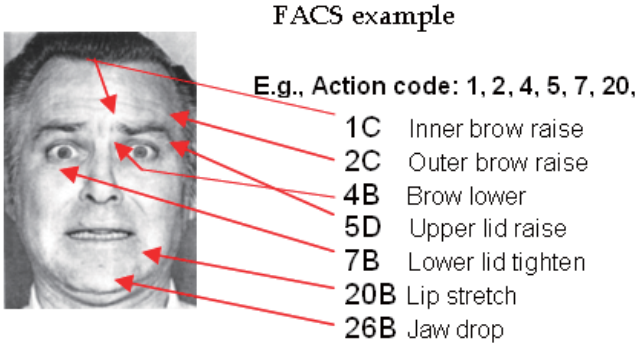
표정 인식에 사용되는 기계학습 모델들은 사람 수준의 높은 성능을 보여준다. 하지만 표정에는 기계학습 모델 뿐만 아니라, 사람들 역시 감정을 읽는데 어려운 표정이나 판단을 내리기 모호한 것들이 존재한다. 사람이 판단을 내리기 어려운 표정들에 대해서, 기계학습 모델이 판단을 내렸을 때, 이 판단에 대해서 사람들이 납득하기 어려운 경우들이 생길 수 있다.

이러한 문제가 생겼을 때, 기계학습 모델들은 높은 정확도를 가진 것과 반하여, 그 판단에 대한 어떠한 근거나 설명을 제공할 수 없다. 이러한 이유로 사람들은 기계학습 모델이 판단을 내렸을 때, 그 결과만을 알 수 있을 뿐, 과정이나 판단의 근거를 갖고 이해하거나 납득할 수는 없다.

이러한 기계학습 모델의 문제점을 해결하기 위해서 많은 시도들이 있다. 해석 가능한 모델(Interpretable Model)과 설명 가능한 모델(Explainable Model)이 그 예시이다. 해석 가능한 모델의 경우, 모델의 판단 과정을 보고 사람들이 해석할 수 있는 형태로 변환해주는 모델을 일컫는다. 설명 가능한 모델의 경우, 기계학습 모델이 판단을 내리는 과정과 결과를 사용하여, 마땅한 근거나 설명을 생성해주는 모델을 말한다.

설명 가능한 모델을 구축하기 위해서는 사람들이 납득할 수 있는 근거의 형태를 갖고 있어야 한다. 이는 [1]에서의 연구처럼 자연어가 될 수 있고, 박스라고 부르는 표현 방식을 근거로 사용할 수도 있다.

표정을 설명할 때, 표정을 여러 가지 얼굴 근육의 움직임의 조합으로 보는 가정을 한다면, Facial Action Coding System(FACS) [2]을 사용하여, 근거로 사용할 수 있다.



(그림 1) Facial Action Coding System(FACS)

이 연구는 FACS 를 표정 인식의 근거로 사용하는데 정당성을 부여하기 위하여, 사람의 8 가지의 공통된 표정(무표정, 화남, 역겨움, 두려움, 행복함, 슬픔, 놀람, 경멸)과 FACS 의 단위인 Action Units(AUs) 간의 관계에 대하여 분포를 시각화하여 연관성을 보여준다.

표정 인식을 학습한 Convolutional Neural

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2017R1A2B4003558)

Network(CNN)가 추출한 특징을 Manifold learning 기술 중에 하나인 t-distributed stochastic neighbor embedding(t-SNE)을 사용하여, 2 차원에 Embedding 하여 표정과 AUs 간의 관계를 시각화한 뒤, 그 분포의 유사성을 통해서 연관성에 대해서 밝혀보고자 한다.

2. 관련 연구

[3]의 연구에 의하면 사람들은 무표정을 제외하 7 가지 감정을 표현할 때, 문화권과 인종에 관련 없이 모두 같은 표정을 사용한다고 한다. 그러므로 7 가지 감정을 표현할 때 사용하는 표정 근육은 모두 공통적이라는 것이다.

이러한 사회 실험을 바탕으로 [3]의 저자는 FACS [2]라는 표정 근육의 움직임을 각각 코드를 부여한 시스템을 구축한다. FACS 는 얼굴 근육의 움직임을 아주 작은 단위로 쪼개어 각각의 코드를 부여한 시스템이다. 이 각각의 코드 혹은 단위를 Action Units(AUs)이라고 부른다.

이 시스템에서 부여한 이 표정 근육 움직임의 단위 즉, Action Units(AUs)의 조합을 통해서 7 가지의 공통된 표정을 표현할 수 있다고 주장한다.

표정을 설명할 때, 표정을 여러 가지 얼굴 근육의 움직임의 조합으로 보는 가정을 한다면, Facial Action Coding System(FACS) [2]을 사용하여, 근거로 사용할 수 있다.

하지만, 이는 인간의 직관으로 표정 근육에 대한 코드를 부여한 것이므로, 실제로 이 근육 움직임들이 표정과 직접적으로 연관성이 있다고 확실하게 말할 수 없다.

이러한 이유로 Explainable Model 에서 표정 인식 판단의 근거나 설명으로서, AUs 를 사용하는 것에 대한 정당성으로 AUs 과 인식된 표정 간의 연관성을 파악하려고 한다.

연관성을 파악하기 위해서, 표정 인식을 학습한 Convolutional Neural Network(CNN) 모델을 사용하여, 특징을 추출한 뒤, 추출된 특징을 Manifold Learning 중에 하나인 t-distributed stochastic neighbor embedding(t-SNE)를 사용하여 2 차원 공간에 Embedding 한 뒤, 이 분포를 사용하여, 실험을 진행하였다.

3. 실험

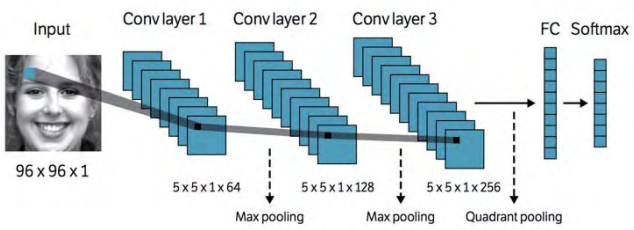
표정과 Action Unit 사이의 분포를 그리기 위해서 우선 이미지 데이터로부터 특징을 추출하였다. 특징을 추출하는 데는 CNN 모델을 사용하였다. CNN 모델은 CK+ 데이터셋[5]을 사용하여 학습한 뒤, 이 데이터셋을 사용하여 특징을 추출하였다. CK+ 데이터셋은 7 가지의 표정과 Action Unit 이 Labeling 된 593 개의 순서를 가진 이미지 묶음으로 이뤄진 데이터셋이다. 실험을 위해서 CK+ 데이터셋 중에 7 가지 표정과 Action Unit 이 잘 부여된 이미지 묶음을 추출한 뒤, 이미지 묶음 중 가장 첫 번째 이미지를 무표정으로 가정하고, 마지막 3 장의 이미지를 표정이 부여된 이미지라 가정하여 1307 장의 이미지를 추출하여 사용하였다.

<표 1> CK+ 데이터셋 내의 AU 과 감정 분포표

	AU 1	AU 2	AU 4	AU 5	AU 6
Angry	0/0.0%	0/0.0%	120/19.9%	18/3.0%	24/4.0%
Disgust	3/5.0%	3/5.0%	3/5.0%	0/0.0%	0/0.0%
Fear	3/0.5%	0/0.0%	108/17.3%	0/0.0%	54/8.7%
Happy	66/16.8%	30/7.6%	63/16.0%	48/12.2%	9/2.3%
Sadness	0/0.0%	0/0.0%	0/0.0%	0/0.0%	198/31.7%
Surprise	75/23.6%	18/5.7%	66/20.8%	0/0.0%	0/0.0%
Contempt	243/20.6%	243/20.6%	3/0.3%	210/17.8%	0/0.0%
	AU 7	AU 9	AU 12	AU 15	AU 17
Angry	96/15.9%	9/1.5%	3/0.5%	9/1.5%	117/19.4%
Disgust	0/0.0%	0/0.0%	15/25.0%	12/20.0%	15/25.0%
Fear	99/15.9%	174/27.9%	6/1.0%	6/1.0%	120/19.2%
Happy	18/4.6%	0/0.0%	6/1.5%	0/0.0%	9/2.3%
Sadness	21/3.4%	0/0.0%	201/32.2%	0/0.0%	0/0.0%
Surprise	3/0.9%	0/0.0%	0/0.0%	66/20.8%	78/24.5%
Contempt	0/0.0%	0/0.0%	9/0.8%	3/0.3%	0/0.0%
	AU 20	AU 23	AU 24	AU 25	AU 27
Angry	0/0.0%	108/17.9%	99/16.4%	0/0.0%	0/0.0%
Disgust	0/0.0%	3/5.0%	6/10.0%	0/0.0%	0/0.0%
Fear	0/0.0%	6/1.0%	21/3.4%	27/4.3%	0/0.0%
Happy	75/19.1%	0/0.0%	0/0.0%	69/17.6%	0/0.0%
Sadness	3/0.5%	0/0.0%	0/0.0%	201/32.2%	0/0.0%
Surprise	0/0.0%	9/2.8%	3/0.9%	0/0.0%	0/0.0%
Contempt	3/0.3%	3/0.3%	0/0.0%	246/20.9%	216/18.3%

표 1 은 CK+ 데이터셋 내의 AU 과 표정 간의 관계에 대한 표이다. 표 1 에서도 연관성을 보임을 분포를 통해서 알 수 있다.

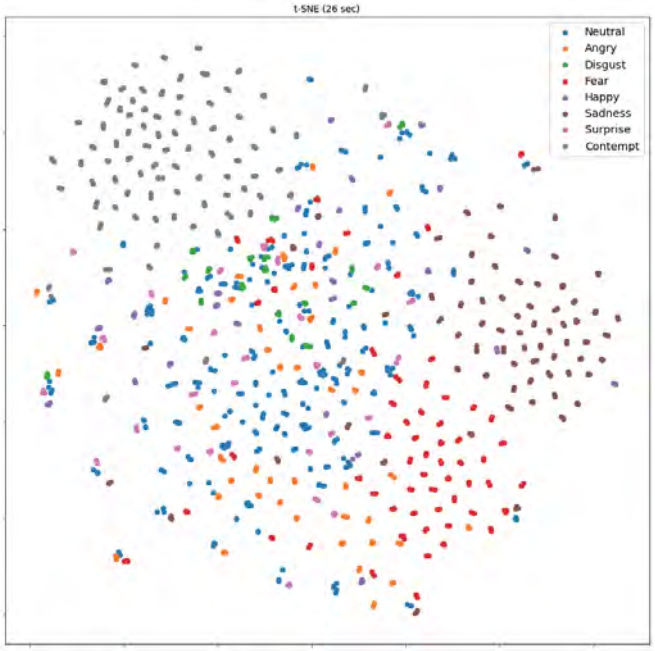
특징을 추출하기 위해 사용한 CNN 모델은 [4]의 구조를 차용하였다. 구조는 그림 2 와 같다.



(그림 2) 특징 추출을 위해 사용된 CNN 구조 [4]

CNN 은 이미지를 분류하기 위해서 적절한 Filter 를 학습한 뒤, Layer 를 지나면서 점점 추상적인 표현방식을 추출하여, 마지막으로 Softmax layer 를 지나면서 최종 판단을 내린다. 이러한 성질을 이용하여, Quadrant Pooling 의 출력 값을 CNN 이 뽑은 특징으로 가정하고 실험에 사용하였다.

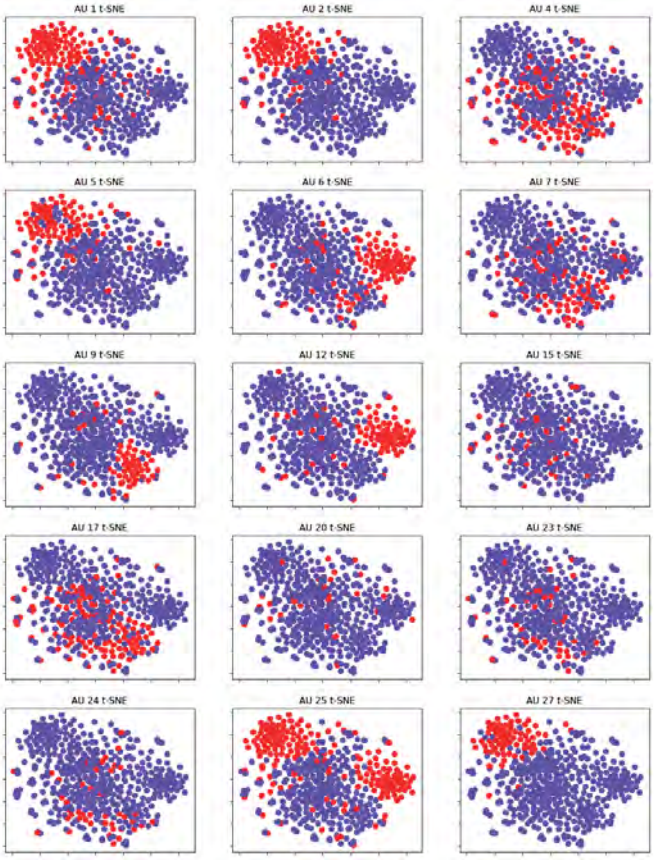
CNN 이 추출한 특징을 Manifold Learning 중에 하나인 t-distributed stochastic neighbor(t-SNE)를 사용하여 2 차원 공간에 Embedding 하여 이를 그래프의 형태로 표현하였다. 표현된 점을 분류된 감정에 따라 색을 달리하여 분포를 확인하였다.



(그림 3) 추출된 특징을 t-SNE 로 시각화한 결과

그림 3 을 통해, 경멸(Contempt) 감정을 가진 표정 특징들은 왼쪽 위에 분포되어 있으며, 슬픔(Sadness)에 대한 감정은 오른쪽 가운데에 군집을 이뤄 분포하고 있음을 알 수 있다.

이를 AU 의 분포와 비교하여 확인해보면 상관성을 알 수 있다. 그림 4 는 t-SNE 를 통해서 2 차원으로 Embedding 한 특징들을 각 AU 을 기준으로 어느 곳에 분포하고 있는지 시각화한 것이다. 그림 4 를 통해서 AU 들이 표정과 어떤 연관성을 갖고 있는지 파악할 수 있다. 그림 4 에 의하면, AU 1 은 경멸(Contempt)에 대해서 Mapping 된 군집과 아주 유사한 경향성을 보인다. AU 2, 5, 27 역시 유사한 군집을 갖는다. 또한, AU 6 의 경우 슬픔(Sadness)에 대해서, AU 1, 2, 5, 27 과 같이 아주 유사한 군집을 갖는다. AU 25 의 경우에는, 경멸(Contempt)와 슬픔(Sadness) 두 표정 모두와 연관성이 있음을 그림 4 를 통해서 알 수 있다. 위의 예시 이외에도 많은 AU 에서 표정과의 연관성을 파악할 수 있다.



(그림 4) t-SNE 의 결과를 AUs 으로 구분한 결과

4. 결론

이 연구는 기계학습 모델이 좋은 성능을 보임에도 불구하고, 그 결과에 대한 근거나 설명을 제공하지 않는 문제를 해결하기 위한 시도이다. 이 연구는 설명 가능한 모델(Explainable Model)을 표정 인식에 적용함에 있어, FACS 의 AUs 을 근거로 사용할 수 있도록 정당성을 부여한다.

FACS 의 AUs 을 표정 인식 판단의 근거로서 사용하기 위해서 표정 인식을 학습한 CNN 모델이 추출한 특징을 Manifold learning 중에 하나인 t-SNE 을 사용하여 2 차원 Manifold 로 시각화하여, 표정과 AUs 분포 간의 관계를 통해서 연관성을 파악할 수 있었다.

그림 3 에서 알 수 있듯이, 현재 추출된 특징을 t-SNE 을 사용하여 2 차원 Manifold 에 Embedding 하였으나, Embedding 된 결과가 아주 명확하게 감정들을 분리되게 군집을 만들지 못하였다. 이러한 이유로 그림 4 에서도 AU 과 표정 사이의 연관성 역시 분명히 파악하기 어려운 부분이 있었다. 이 부분에 대해서 더 보완하여 확실하게 표정들 간의 군집이 확실하게 분류된 값과 매칭되는 경우에는 AU 과 표정에 대해서 좀 더 명확한 연관성을 파악할 수 있을 것으로 예상된다.

참고문헌

[1] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, European Conference on Computer Vision (ECCV), 2016.

[2] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 1, 2

[3] P. Ekman. E motions Revealed, Second Edition: Recognizing Faces and Feelings to Improve Communication and Emotional Life. 2007. 3.20

[4] Pooya Khorrami, Tom Le Paine, and Thomas S. Huang. “Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?.”, In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (ICCVW ’15), IEEE Computer Society, Washington. DC, USA, 2015, pp. 19–27.

[5] 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”, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, 2010, pp. 94–101.