# Resnet을 이용한 감정 인식 기반 콘텐츠 반응 분석 시스템

Content Response Analysis System Based on Emotion Recognition Using Resnet

저자
(Authors)

최원재, 정병훈, 김정선
Wonjae Choi, Byunghun Jeong, Jungsun Kim

출처
(Source)

한국정보과학회 학술발표논문집 , 2018.12, 281-283(3 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07613578

APA Style

최원재, 정병훈, 김정선 (2018). Resnet을 이용한 감정 인식 기반 콘텐츠 반응 분석 시스템. 한국정보과학회 학술발표논문집, 281-283

# Resnet을 이용한 감정 인식 기반 콘텐츠 반응 분석 시스템

최원재 ○ 정병훈 김정선
wonjaechoi2014@gmail.com, byunghun.jeong@gmail.com, kimjs@hanyang.ac.kr
한양대학교 컴퓨터공학과

# Content Response Analysis System Based on Emotion Recognition Using Resnet

Wonjae Choi ○ Byunghun Jeong Jungsun Kim
Department of Computer Science and Engineering, Hanyang University

**요 약**

누구나 영상 제작자가 될 수 있는 환경에서 시청자의 영상 콘텐츠의 '좋아요' 클릭은 단순하지만 큰 파급력을 가지고 있고, 영상 제작자에겐 시청자의 영상에 대한 호감 여부를 제공하는 피드백이 된다. 하지만 이러한 방식은 시청자가 직접 버튼을 눌러야 하는 단점과 부정 사용자에 의한 조작이 쉬운 문제점이 있고, 영상 제작자에게는 피드백으로써 제공되는 정보가 객관성이 부족하고, 시청자의 매 순간의 반응을 담고 있지 않은 문제점이 있다. 따라서 본 논문은 영상 시청 중 시청자의 얼굴 표정을 시각 처리에 뛰어난 성능을 보이는 합성곱 신경망 중 하나인 Resnet을 통해 감정을 예측하여, 영상 제작자에게는 매 순간 시청자의 반응을 담은 피드백정보를 제공하고, 영상 시청자에게는 콘텐츠 선호도 측정을 통한 자동 투표 서비스를 제공함으로써 기존 '좋아요' 클릭의 문제점을 해결하고자 한다.

## 1. 서 론

최근 유튜브, 아프리카TV, 인스타그램, 페이스 북 등 누구나 영상 제작자가 될 수 있는 플랫폼이 발달하였다. 이러한 플랫폼에서 영상 콘텐츠에 대한 시청자의 '좋아요' 클릭은 단순하지만 큰 파급력을 가지고 있다[1]. 하지만 이러한 방식은 시청자가 직접 '좋아요' 버튼을 눌러 생기는 저조한 투표율과 부정 사용자에 의한 허위 조작이 쉬운 문제점이 있다. 또한 콘텐츠 제작자는 영상 콘텐츠에 대한 피드백을 '좋아요'와 댓글에 의존하게 되는데, 이러한 '좋아요'와 댓글은 영상 콘텐츠에 대한 객관적인 피드백이 부족하고, 시청자의 매 순간의 반응을 담고있지 않다.

따라서 본 논문에서는 영상 콘텐츠 시청 중 시청자의 얼굴 표정으로 감정을 예측하여 매 순간의 감정 예측 데이터 기반으로 시청자에게는 자동 투표 서비스를 제공하고, 제작자에게는 시청자의 감정 기반 피드백을 제공하여 문제를 해결하는 감정 인식 기반 콘텐츠 반응 분석 시스템을 제안한다.

감정 예측은 심층 신경망 중 시각 처리에 뛰어난 성능을 보이는 합성곱 신경망을 이용하였다. 기존의 심층 신경망의 경우 망의 깊이가 깊어질수록 정확도가 낮아지는 현상때문에 정확도를 높이는 것에 한계가 있었다[2]. 본 논문에서는 이를 Residual Learning으로 해결한 Resnet(Residual neural network)[3]을 사용하여 감정 예측의 정확도를 높였다.

본 논문에서 제안하는 시스템은 Resnet을 통해 영상을 시청하는 시청자의 매 순간의 감정을 예측하여 영상에 대한 시청자 별 감정 그래프를 생성한다. 이 데이터를 통해 영상 제작자에게는 매 순간의 시청자의 반응을 담은 데이터를 제공하고, 영상 시청자에게는 콘텐츠 선호도 측정을 통한 자동 투표 서비스를 제공함으로써 기존 '좋아요' 클릭의 문제점을 해결하고자 한다.

## 2. 감정 인식 기반 콘텐츠 반응 분석 시스템의 구성 및 동작 원리

본 논문에서 제안하는 감정 인식 기반 콘텐츠 반응 분석 시스템의 구성도는 그림 1과 같다. 영상 콘텐츠 시청 중 시청자의 카메라로부터 이미지를 전달받은 뒤 이미지로부터 시청자의 얼굴을 인식하여 얼굴 이미지를 감정 예측 모델에 입력으로 사용한다. 감정 예측 모델은 감정 별 예측 확률을 출력하게 되고, 이 출력 값을 그래프로 시각화하여 영상 콘텐츠 제작자에게 피드백으로 제공한다. 또한 출력 값을 이용하여 영상에 대한 콘텐츠 선호도를 측정하여 영상 콘텐츠를 자동 투표하게 된다.


그림. 1 감정 인식 기반 콘텐츠 반응 분석 시스템 구조

### 2.1 감정 예측 모델

감정 예측 모델은 심층 신경망 중 시각 처리에 뛰어난 성능을 보이는 합성곱 신경망을 사용하였다. 심층 신경망의 경우 망의 깊이가 깊을수록 학습 데이터 속에 존재하는 대표적인 특징을 잘 추출할 수 있어 학습 결과가 좋아진다. 하지만 기존의 심층 신경망은 망의 깊이가 깊어질 경우 정확도가 낮아지는 현상인 vanishing/exploding gradients 문제가 발생하여 망의 깊이를 깊게 하여 정확도를 높이는 것에 한계가 있었다. 따라서 본 논문에서는 Residual Learning 이라는 방법을 제안해 해결하여 더 깊은 레이어를 쌓아 정확도를 높인 Resnet을 사용하였다. Resnet은 152개의 깊은 레이어를 쌓아 매우 높은 정확도를 갖는 CNN 아키텍쳐이다. Residual Learning은 평범한 CNN망에서 입력 값 x로부터 최적의 출력 값 H(x)를 얻기

위해 레이어를 설정하는 것과 달리 출력 값 H(x)와 입력 값 x의 차를 얻는 것으로 목표를 바꾸었다. 따라서 출력 값과 입력 값의 차이(Residual)가 0이 되는 방향으로 학습하게 되며, 이를 통해 입력과 출력이 연결되기 때문에 forward와 backward path가 단순해져 깊은 망도 쉽게 최적화가 가능하게 되고, 늘어난 깊이로 정확도를 크게 향상시킬 수 있다.

데이터 셋은 Kaggle의 Face Expression Recognition Challenge[4]의 7가지 감정 레이블을 가진 표정 이미지를 사용하였다. 감정 레이블은 행복, 슬픔, 놀람, 화남, 공포, 역겨움, 무표정이다. 이 중 화남, 공포, 역겨움의 감정은 비슷한 표정을 가지고 있어 구분하여 예측하기 어렵고, 모두 선호하지 않음으로 분류할 수 있기 때문에 이 세가지 감정을 모두 공포로 분류하였다. 따라서 수정된 감정 레이블은 행복, 슬픔, 놀람, 공포, 무표정 5가지이다. 데이터 셋은 48 x 48의 흑백 얼굴 표정 이미지로 구성되어 있다. 그중 학습용 데이터의 개수는 28,709개이며, 이중 10%인 2,870개의 데이터를 검증용 데이터로 사용하였고, 테스트용 데이터의 개수는 7,178개이다. 얼굴 표정 데이터 셋의 세부 사항은 표 1과 같다.

표. 1 감정 레이블 별 데이터 셋 세부사항

| Emotion | Training Images | Test Images |
|---|---|---|
| Happy | 7,215 | 1,774 |
| Sad | 4,830 | 1,247 |
| Surprise | 3,171 | 831 |
| Fear | 8,528 | 2,093 |
| Neutral | 4,965 | 1,233 |

### 2.2 감정 예측

감정 예측을 하기 위해선 먼저 입력 이미지로부터 얼굴을 인식하여야한다. 본 논문에서는 얼굴인식을 위해 OpenCV의 Haar Cascade 필터[5]를 사용한다. 필터를 사용해 추출된 얼굴 이미지를 학습된 감정 예측 모델의 입력에 넣기 위해 48 x 48 크기의 흑백 이미지로 변환한다. 변환된 이미지를 모델에 입력하게 되면 5가지 감정에 대한 예측 확률을 출력하게 된다. 그림 2는 카메라를 통해 입력 이미지로부터 얼굴을 인식한 뒤 감정 예측 모델이 5가지 감정에 대한 예측 확률을 출력하고 있는 것을 나타낸다. 실제 시스템에서는 영상 시청 시 영상과 지속적으로 출력되는 해당 화면을 동시에 나타내는 것이 시청자의 영상 시청을 방해하기 때문에 보이지 않게 설정하였다.

그림2. 감정 예측 결과 출력

그림 3은 입력 이미지로부터 감정 예측 모델을 통해 출력 값이 나오는 동작 과정이다.

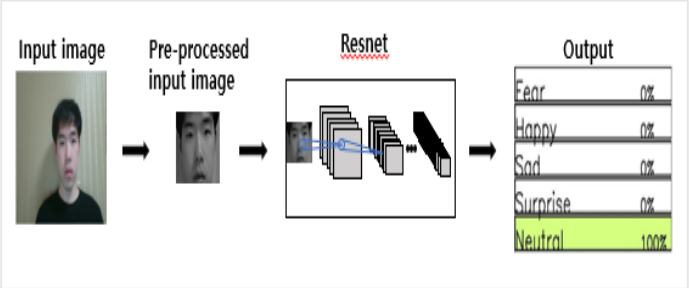


그림. 3 감정 예측 과정

Resnet모델을 통해 학습된 모델의 감정 예측 정확도는 표 2와 같다.

표. 2 학습된 모델의 감정 별 예측 정확도

| Emotion | Test accuracy |
|---|---|
| Happy | 79.8% |
| Sad | 41.2% |
| Surprise | 77.1% |
| Fear | 63.2% |
| Neutral | 56.9% |

## 3. 핵심 기능

본 논문에서 제안하는 감정 인식 기반 콘텐츠 반응 분석 시스템에는 두가지 핵심 기능이 있다. 먼저 영상 제작자에게 영상에 대한 시청자 별 감정 예측 그래프를 통해 매 순간 시청자의 반응을 담은 분석 피드백을 제공하는 기능과 시청자에게 콘텐츠 선호도 측정을 통한 자동 투표 서비스를 제공하는 기능이 있다.

### 3.1 분석 피드백 제공 기능

영상 시청 시 감정 예측 모델을 통해 매 순간의 감정 별 예측 값이 텍스트 파일에 저장된다. 하지만 시청자의 카메라 프레임이 모두 동일하지 않기 때문에 텍스트 파일의 길이가 일정하지 않다. 또한 중요하지 않은 찰나의 순간의 표정 변화가 출력 값에 쓰여진다. 이러한 문제를 해결하기 위해 시청자의 카메라의 1초당 프레임을 계산한 후, 계산된 프레임 단위로 감정 확률을 평균을 내어 감정 예측 텍스트 파일을 만든다. 카메라의 1초당 프레임은 식 (1)을 통해 구할 수 있다. P는 콘텐츠의 재생시간, L은 저장된 텍스트의 라인 수를 의미한다.

$$\text{카메라 1초당 프레임} = \frac{L}{P} \qquad (1)$$

이러한 방식으로 시청자가 콘텐츠를 시청하였을 시 1초 단위의 감정 확률 예측 데이터를 생성한다. 또한 시청자 평균 데이터를 제공하기위해 영상 시청 후 데이터가 생성될 때 마다 예측 데이터를 누적 시켜 조회 수로 나눈다.

감정 예측 데이터로부터 영상 제작자에게 제공할 분석 피드백을 만들기 위해 영상 제작자가 영상의 의도한 감정에 대해 예측 확률의 최대값 구간을 계산해야한다. 최대값 구간의 조건은 다음과 같다. 1. 2초 이상 지속된 감정, 2. 최대값으로 예측된 확률이 0.5 이상.

1초 순간의 최대값은 시청자가 느낀 감정이 아닌 영상 시청 시 나타나는 순간의 표정이기에 배제한다. 예측 감정 확률의

최대 값이 0.5 이하일 경우 불분명하게 예측된 확률로 판단하여 배제한다. 이 기준을 통해 만들어진 각 감정 별 최대값 구간들은 시작점과 끝점으로 이루어진 쌍의 리스트로써 저장 된다. 이 분석 데이터를 영상 제작자에게 가시성 있게 제공하기 위해 본 논문에서는 데이터 시각화 라이브러리인 python의 matplotlib을 사용하여 시간에 따른 감정 확률 값을 보여주는 그래프의 형태로 변환하여 제공한다. 그림 4와 5는 matplotlib 라이브러리를 사용하여 감정 확률 예측 데이터와 시청자 평균 감정 별 확률 예측 데이터의 그래프이다. 그래프에서 영상 제작자가 영상의 의도한 감정에 대한 최대값 구간은 하이라이트가 되어 강조된다.

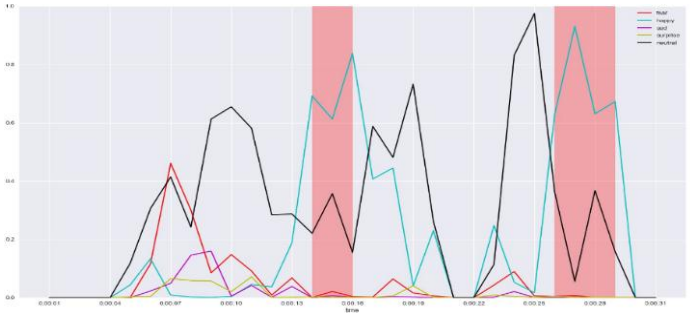

그림. 4 감정 별 확률 예측 그래프

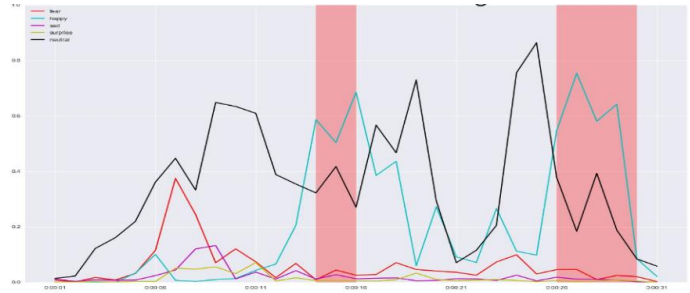

그림. 5 시청자 평균 감정 별 확률 예측 그래프

### 3.2 콘텐츠 선호도 측정을 통한 자동 투표 기능

콘텐츠 자동 투표 서비스는 시청자들이 영상을 본 후 만들어진 감정 별 확률 예측 데이터를 통해 제공된다. 콘텐츠의 선호도는 감정 별 확률 예측 데이터에서 영상 콘텐츠 제작자가 설정한 영상의 의도한 감정을 기준으로 측정된다.

영상 콘텐츠의 선호도 측정에서 영상의 모든 순간이 선호도 측정에 동일한 영향력을 가지고 있지 않다. 또한 감정 카테고리마다 선호도 측정의 큰 영향력을 가지는 중요한 순간들은 다를 것이다. 하지만 영상 제작자에 의도에 따라 같은 감정 카테고리라도 해당 감정을 느끼게 하는 중요한 부분이 다를 수 있기 때문에 영상의 어느 부분이 콘텐츠 선호도에 중요한 영향력을 미치는 부분인지 추측하는 것은 무의미하다. 따라서, 본 논문은 모든 순간을 동일한 영향력을 가지고 있다고 가정한 뒤, 감정 별 확률 예측 데이터에서 영상 콘텐츠 제작자가 설정한 영상의 의도한 감정의 최대 값 구간을 영상 콘텐츠 선호도 측정에 사용하였다. 영상 시청 시 영상 콘텐츠 제작자가 의도한 감정에 일치하는 감정을 많이 보이면 선호하는 것이고, 무표정인 감정을 많이 보이면 선호하지 않는 것이다. 따라서 콘텐츠 선호도는 식 (2)를 통해 구할 수 있다. Fi는 의도된 감정의 최대 값 구간의 횟수, Li은 최대값 구간의 총 길이, Pi는 예측된 감정 확률이다. 마찬가지로 Fn은 무표정으로 예측된 감정의 최대 값 구간의 횟수, Ln은 최대 값 구간의 길이, Pn은 예측된 감정 확률이다.

$$\text{preference} = Fi * Li * \sum Pi - Fn * Ln * \sum Pn \quad (2)$$

이 식을 통해 구한 콘텐츠 선호도가 양수일 경우 해당 영상을 자동 투표하게 된다.

### 4. 결론

플랫폼에서 시청자의 '좋아요'와 댓글로써 제공되는 영상 콘텐츠에 대한 부족한 피드백 문제를 매 순간의 영상 콘텐츠 시청자의 얼굴 표정으로 감정을 예측하여 감정 예측 그래프를 제공함으로써 영상 콘텐츠 제작자에게 매 순간의 시청자의 반응을 담은 피드백 정보를 제공하고, 시청자가 직접 '좋아요' 버튼을 클릭하여 생기되는 부정 투표 문제와 저조한 투표율 문제를 시청 시의 매 순간의 감정을 기반으로 자동 투표함으로써 해결하였다. 이 시스템은 영상 콘텐츠를 시청하는 기기에 카메라가 내장될 시 활용될 수 있으므로, 내장 카메라가 장착되어 있는 모바일 기기를 통해 보편적으로 활용될 것으로 기대된다.

### 참 고 문 헌

[1] Brian Carter, "The Like Economy: How Businesses Make Money With Facebook", pp. 3-8, Que Publishing, 2011.

[2] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio, On the difficulty of training recurrent neural networks, ICML'13 Proceedings of the 30th International Conference on International Conference on Machine Learning, pp. 1310-1318, 2013

[3] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, pp. 770-777, 2015

[4] 케글 얼굴 표정 인식 대회, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

[5] Paul Viola, Michael Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, Accepted Conference on Computer Vision And Pattern Recognition, pp. 511-518, 2001