멀티모달 딥러닝을 이용한 비디오 감정 회귀 분석

Video Emotion Regression Analysis using Multi-modal Deep Learning

저자 (Authors) 김하연, 이인권
Hayeon Kim, In-Kwon Lee

출처 (Source) 한국컴퓨터그래픽스학회 학술대회 , 2018.7, 43-44(2 pages)

발행처 (Publisher) 한국컴퓨터그래픽스학회
Korea Computer Graphics Society

URL http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07515118

APA Style 김하연, 이인권 (2018). 멀티모달 딥러닝을 이용한 비디오 감정 회귀 분석. 한국컴퓨터그래픽스학회 학술대회, 43-44

멀티모달 딥러닝을 이용한 비디오 감정 회귀 분석*

김하연[0], 이인권
연세대학교 컴퓨터과학과
qoocrab@gmail.com, iklee@yonsei.ac.kr

Video Emotion Regression Analysis using Multi-modal Deep Learning

Hayeon Kim[0], In-Kwon Lee
Dept. of Computer Science, Yonsei University

요약

비디오 콘텐츠 분류 및 검색에서 감정은 중요한 구성 요소 중 하나이기 때문에 비디오 감정 인식은 꾸준히 연구되고 있는 분야이다. 본 논문은 최근 다양한 인식 분야에서 좋은 성과를 거둔 딥러닝(deep learning)을 이용한 멀티 모달 학습 구조의 새로운 비디오 감정 회귀 분석 모델을 제안한다. 그리고 본 논문에서 제안한 모델이 이전 다른 형태의 멀티 모달 구조의 회귀 분석 모델들보다 성능의 향상이 있었음을 보인다.

1. 서론

비디오는 대표적인 복합 데이터로, 영상과 소리로 이루어져 있다. 특히 감정과 같은 복합적인 정보 기반의 요소를 기계학습 방식으로 분석하고자 할 때는 두 개 이상의 데이터 모달리티(modality)를 학습하는 멀티모달 학습(multimodal learning) 방식으로 주로 접근한다.

현재 비디오 감정 분석을 위해 멀티모달 러닝 구조로 접근하는 모델들은 주로 피쳐 (feature) 수준에서의 융합 방법과 (feature-level fusion) 결정 수준에서의 융합 방법 (decision-level fusion)으로 나누어진다. 전자는 모든 피쳐들을 하나의 벡터로 연결하여 분류기(Classifier) 혹은 Regressor에 넣어 인식하는 방법이고, 후자는 각 요소의 분류 혹은 회귀 결과를 합치는 방법이다.

본 논문에서는 최근 이미지 분류 및 인식에서 좋은 성능을 보였던 DenseNet [1]과 사운드 데이터를 학습시켜 레이블이 지정되지 않은 비디오를 분류하는 모델 SoundNet [2]을 fine tuning하고 피쳐 수준에서의 융합 방법을 이용한 회귀 분석 모델을 제안한다. 그리고 감정 비디오 데이터셋인 LIRIS-ACCEDE 데이터베이스를 이용한 실험 결과로 제안된 방법의 우수성을 입증한다.

2. 제안 방법

본 연구에서 제안하는 모델은 그림 1과 같이 총 3개의 스트림으로 나누어진다. 첫 번째 옵티컬 플로우 스트림에서는 정밀한 움직임 특성을 추출할 수 있는 dense optical flow [9]를 전체 프레임 영역에서 추출한 후 ImageNet 데이터셋에 사전 훈련된 5 layer의 ConvNet 모델에 통과시킨다. 두 번째 플로우 스트림에는 10프레임 간격으로 샘플링된 연속 RGB 프레임을 ImageNet 데이터셋으로 사전 훈련된 DenseNet 모델에서 fine tuning한다. 두 스트림에서 얻은 피쳐를 하나의 특징 벡터로 연결한 후 1 layer ConvNet과 글로벌 풀링에 통과시킨다.

마지막 스트림에는 각 비디오 파일에서 얻은 sound waveform을 추출한 후 라벨이 없는 200만개의 동영상으로 미리 훈련된 SoundNet 모델에서 피쳐를 얻는다. 그리고 이전 글로벌 풀링에 통과시킨 특징 벡터와 결합한다. 이 최종 특징 벡터를 회귀 계층(FC Layer)의 입력으로 사용하였다. 그리고 loss 함수는 MSE (mean squared error)를 사용하였고 optimizer는 Adam (adaptive moment)을 사용하였다.

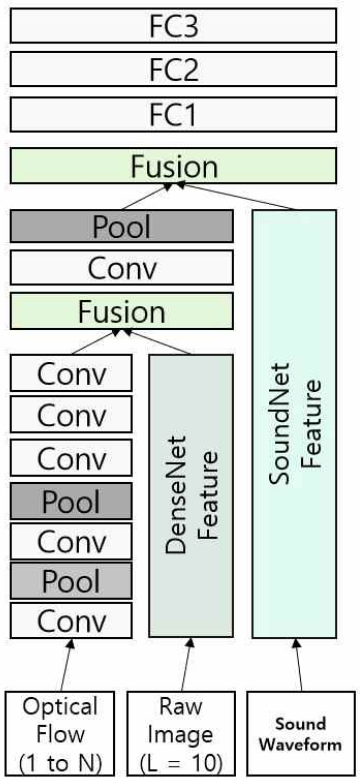

그림 1: 제안 모델

* 구두 발표논문
* 본 논문은 연구초기결과의 요약논문 (Extended Abstract)임
* 이 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임(과제번호 SRFC-IT1601-04).

3. 실험 및 결과

3.1. 감정 디스크립터

본 연구에서는 비디오에 사용자의 주관적인 감정 평가를 매기기 위해 차원적(dimensional) 접근의 디스크립터를 사용한다. 특히 Valence, Arousal로 구성된 2차원 레이블 집합을 이용해 감정을 측정한다. Valence란 즐거움의 정도 혹은 긍정/부정의 감정을 나타내는 척도이고, Arousal는 감정의 활성화 즉 흥분의 척도를 나타낸다.

3.2. 데이터 세트

MediaEval 2016[3]에서 제공하는 LIRIS-ACCEDE Discreate 데이터셋을 실험에 사용하였다. 160개의 장편 영화 및 단편 영화에서 추출한 9800개의 클립으로 이루어진 트레이닝 데이터셋과 테스트를 위한 별도의 1200개의 클립이 주어진다. 트레이닝 셋의 경우 원본은 0위부터 9799위까지 Arousal 값과 Valence 값 순위가 매겨져 있다. 해당 데이터셋을 사용하였던 MediaEval 2016에서는 랭킹 값을 이용해서 Gaussian process regression 모델로 회귀 분석을 위한 감정 스코어 값을 추출했다. 본 실험에서는 이 MediaEval 2016 ground truth 값을 기준으로 회귀 분석을 진행하였다.

3.3. 실험 조건

9800개의 영상 중 2450개의 영상을 Validation Set으로, 7350개를 Train Set으로 사용하였다. 또한, 성능 증가를 위해서 Data Augmentation을 적용하였다. 224 × 224 사이즈로 Random cropping을 하였고, 무작위 좌우 반전(left-right flipping)을 적용하였다.

3.4. 실험 결과 및 분석

우리의 실험 결과를 같은 데이터셋을 사용한 MediaEval 2016에서 제안한 비디오 회귀 분석 작업과 비교하였다. 또한 회귀 분석 결과에 대한 척도로 평균 제곱 오차 (MSE, Mean Squared Error)를 채택하였다. 그 비교 결과는 표1에서 확인할 수 있다.

	Valence	Arousal
RUC(run1) [4]	0.218	1.479
THU-HCSI [5]	0.214	1.531
BUL [6]	0.231	1.413
Liu's [7]	0.240	1.185
Gan [8]	0.33	0.77
Ours	0.269	0.628

표 1. MediaEval 2016 Movie Task 참가 모델 및 최근 모델과의 결과 에러 비교 (MSE)

우리의 모델이 상대적으로 좋은 성능을 보이고 있음을 볼 수가 있는데 특히 다른 모델들과 비교했을 때 흥분의 척도인 Arousal 값에서 더 좋은 성능을 보이고 있음을 확인할 수 있었다.

화면의 물체나 사람의 움직임 강도가 증가하면 청중의 Arousal 이 증가한다는 연구가 있다[10]. 그러므로 프레임 장면에서 이미지 객체의 정밀한 움직임을 측정한 Optical Flow의 사용으로 인해 성능 향상이 있었던 것으로 보인다. 또한, SoundNet은 오로지 사운드를 통해서 비레이블 비디오 내의 행동에 대해 좋은 인식률을 보이고 있는 모델이다. 200만개의 비레이블 비디오를 미리 학습한 이 모델을 이용하여 추출한 피쳐를 사용하였기에 흥분의 척도인 Arousal이 더 잘 인식된 것으로 보인다. Valence의 경우 상대적으로 낮은 인식률을 보이고 있음을 볼 수 있는데, 이것은 기존의 멀티모달 모델들이 Valence 인식을 위해 추가하는 색상 피쳐, 얼굴 표정 피쳐의 부재 때문으로 판단된다.

4. 결론

본 연구에서는 비디오 감정 인식을 위한 새로운 형태의 멀티모달 딥러닝 구조를 제안하였다. 피쳐 수준의 융합 모델로, SoundNet과 DenseNet을 적용한 모델이다. 이 모델은 이전 최신 비디오 감정 인식 모델들과 비교하였을 때 비교적 적은 레이어 개수나 피쳐 개수를 사용한 간단한 구조임에도 불구하고 특히 Arousal 값에 대한 인식률이 개선되었고 전반적으로 성능이 뛰어난 것을 확인할 수 있었다. 이는 연구 초기 결과로서 그 의의가 있다고 볼 수 있다.

이후 본 연구를 확장하여 LIRIS-ACCEDE Continuous 데이터셋에도 적용하여 모델 성능을 검증해보고, Valence 인식률을 개선하기 위해 사람 얼굴 표정 피쳐를 추가하는 방식을 검토해 볼 것이다.

참고문헌

[1] G. Huang, Z. Liu. "Densely connected convolutional networks." In CVPR, 2017.

[2] Y. Aytar, C. Vondrick, A. Torralba. "SoundNet: Learning Sound Representations from Unlabeled Video." In NIPS, 2016

[3] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A Video Database for Affective Content Analysis." in IEEE Transactions on Affective Computing, 2015.

[4] S. Chen and Q. Jin. Ruc at mediaeval 2016 emotional impact of movies task: Fusion of multimodal features

[5] Y. Ma, Z. Ye, and M. Xu. Thu-hcsi at mediaeval 2016: Emo- tional impact of movies task

[6] A. Jan, Y. F. A. Gaus, F. Zhang, and H. Meng. Bul in medi- aeval 2016 emotional impact of movies task.

[7] Y. Liu, Z. Gu, Y. Zhang, and Y. Liu. Mining emotional fea- tures of movies

[8] Q. Gan, S. Wang, L. Hao, and Q. Ji. "A Multimodal Deep Regression Bayesian Network for Affective Video Content Analyses" In ICCV, 2017.

[9] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Proc. ECCV, 2004

[10] B. H. Detenber, R. F. Simons, and G. G. Bennett Jr, "Roll em!: The effects of picture motion on emotional responses." J. Broadcasting Electron. Media, vol. 42, no. 1, pp. 113-127, 1998.