웨어러블 센서를 이용한 비디오 시청자 감정 상태의 지속적인 학습 및 예측 연구

박경화[1*] 김병희[2*], 장병탁[1,2]
[1]서울대학교 뇌과학 협동과정
[2]서울대학교 컴퓨터공학부
{kwpark, bhkim, btzhang}@bi.snu.ac.kr

Lifelong Learning to Estimate Emotions of a Video Watcher with Wearable Sensors

Kyung-Wha Park[1*] Byoung-Hee Kim[2*] Byoung-Tak Zhang[1,2]
[1]Brain Science Program, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University

요 약

비디오 시청자가 느끼는 감정을 예측하는 문제는 최근 감성기반 인지컴퓨팅(affective computing) 분야의 핫 이슈중 하나이다. 본 논문에서는 멀티모달 특성을 가지는 비디오 정보를 주요 요인으로 감정을 예측하는 기존의 멀티모달 퓨전 방식의 접근법을 확장하여, 시청자 상태에 대한 지속적인 학습을 위한 Lifelong 학습 기반 기계학습 기법을 결합한 멀티모달 Lifelong 감정 예측 방법을 제안한다. 개인 시청자의 지속적인 감정 예측을 안정적으로 수행하기 위한 실용적인 프로토콜로서 웨어러블 센서를 이용한 감정 반응 수집과 사람들의 공통 반응에 대한 모델을 조합하는 방법을 제시한다. 최근 공개된 대규모 감정 반응 데이터베이스를 이용하여 온라인 멀티모달 감정 예측법의 기능을 확인하였다.

1. 서 론

감성 컴퓨팅(Affective computing)은 감정에 대한 계산학적 기술이며, 가트너 그룹이 매년 발표하는 신흥 기술 목록에 2013년도에 포함되기 시작한 이래 학술 분야의 응용 분야에서 모두 주목하는 기술이 되었다. 특히, 인공지능 및 기계학습 기법의 새로운 응용 분야로서 관련 연구자들의 주목을 받고 있다. 최근 주목받는 딥러닝 전문가 중 한 명인 Yann Lecun도 최근 SNS 상에서 데이터 기반 감정 예측의 중요성에 대해 언급한 바 있다.

기존의 감성 컴퓨팅 연구 중 감정 예측 문제는 크게 이미지, 음악, 동영상 등의 컨텐츠 자체에 표현된 감정을 예측하는 문제와, 컨텐츠를 감상하는 사람의 감정 반응을 예측하는 문제로 구분된다. 두 사례 모두에서 기계학습 기반의 예측기법은 표준적인 방법론으로 자리잡았으며, 멀티모달 자극에 대한 반응 모델링의 특성을 반영하여 멀티모달 퓨전 기법이 제안되었다[1]. 그러나, 감성 컴퓨팅이라는 용어가 소개된 이후[2] 지난 십 수 년의 연구에도 불구하고, 감정 반응 예측 연구는 감정이 개별 편차가 크고 동일한 개인에서도 상황에 따른 반응이 큰 특성으로 인한 한계를 벗어나지 못하고 있다. 또한, 기존의 연구는 일회적인 자료 수집에 기초한 정적인 모델링 위주여서, 역동적인 감정 반응에 대한 체계적인 접근으로서는 제한적이었다.

한편, 스마트 기기와 웨어러블 기기에 장착된 다중 센서를 활용하여 사용자의 감정을 비롯한 다양한 상태를 예측하는 문제에 대한 관심이 뜨겁다. 다양한 센서를 이용한 상태 예측 문제에 대한 접근법은 필연적으로 동적인 모델링에 기반하여야 한다.

최근 기계학습 분야에서는 장기적이고 지속적인 학습 기법으로서 Lifelong 학습 방법이 소개되었다. 이는 기존의 정적인 설정에 기반한 1회성 모델 학습 위주의 접근법뿐만 아니라, 온라인/증분적 학습 기법의 한계를 넘고자 하는 의미 있는 시도로서 주목받고 있다.

이 논문에서는 웨어러블 센서를 기초로 한 사용자의 감정 상태에 대한 지속적인 학습 기법으로서 [그림 1]의 구조를 가지는 멀티모달 Lifelong 학습 프레임웍을 제안한다. 그리고 개인 시청자의 지속적인 감정 예측을 안정적으로 수행하기 위한 실용적인 프로토콜로서 웨어러블 센서를 이용한 감정 반응 수집과 사람들의 공통 반응에 대한 모델을 조합하는 방법을 제시한다. 이러한 프로토콜 중 비디오 시청자의 공통적인 감정 반응에 대한 멀티모달 Lifelong 모델링 예시로서, Arousal-Valence 반응을 수집한 LIRIS-ACCEDE[3] DB를 활용한 시뮬레이션 결과를 보인다.

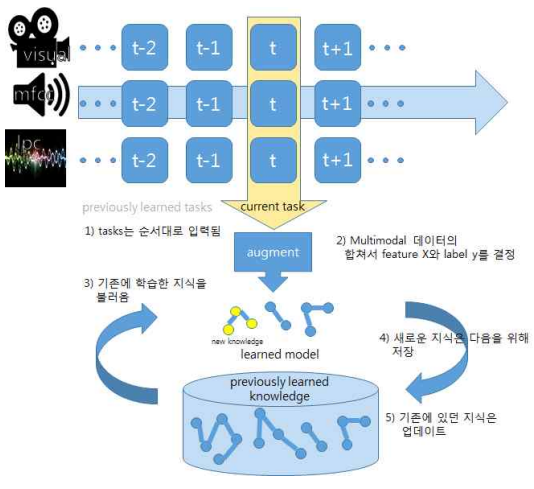

그림 1. 멀티모달 Lifelong 학습 프레임워크

2. 관련 연구

2.1 비디오 시청자의 감정 반응에 대한 감성기반 컴퓨팅

최근 감성 기반 컴퓨팅에서 비디오 시청자의 감정 반응에 대한 모델링은 흥분도(arousal)와 긍정/부정도(valence)의 2차원 감정 공간을 기초로 하는 추세이다. 연구자들은 비디오 시청자의 감정 반응을 설문 등의 self-report를 통해 명시적으로 수집하거나, 비명시적인 다양한 생리학적 신호를 수집하여 대규모의 데이터베이스를 구축한 후 기계학습 기법을 이용하여 컨텐츠 대비 감정 반응을 예측하는 모델링을 시도하고 있다. 명시적 감정 반응 예측 성능은 arousal과 valence 모두 이진 분류 문제 설정의 경우에도 70%를 넘지 못하고 있으며[3], 연구자들은 유리천장을 언급하기도 한다[4]. 비디오에 내재된 영상, 소리 등의 멀티모달 정보를 특징값(feature)로 두는 경우, 각 모달별 특징값과 예측 모델을 구분하는 멀티모달 퓨전 접근법이 적용되며, 특징값 추출 단계에서 퓨전하는 것을 사전 퓨전, 모달별 예측 모델의 출력값을 입력으로 한 단계의 예측을 추가로 수행하는 것을 사후 퓨전이라고 한다[1].

2.2 지속적인 기계학습 기법(Lifelong learning)

최근 ELLA (Efficient Lifelong learning algorithm) [5]와 GO-MTL [6] 등 다중작업 학습(multitask learning)을 Lifelong learning으로 확장한 여러 연구가 소개되었다. 온라인 멀티 태스크 학습을 기초로, 연속적으로 입력되는 데이터에서 지식의 구조를 sparse coding을 통해 인코딩한 후 기존에 학습한 지식이라면 해당 지식을 개선하고, 새로운 것이라면 지식에 추가하는 식으로 지속적으로 학습해나간다.

3. 멀티모달 Lifelong 학습 프레임워크

3.1 Lifelong 학습

Lifelong 학습을 수행하는 에이전트는 일련의 감독 학습 작업 $Z^{(1)}, Z^{(2)}, ..., Z^{(T_{\max})}$을 순차적으로 받는다. 각 작업 $Z^{(t)} = (\hat{f}^{(t)}, \mathbf{X}^{(t)}, \mathbf{y}^{(t)})$는 인스턴스 공간 $X^{(t)} \subseteq \mathbb{R}^d$에서 레이블의 집합 $Y^{(t)}$(이진 분류 문제에서는 $Y^{(t)} = \{-1, 1\}$와 회귀 문제에서는 $Y^{(t)} \subset \mathbb{R}$)으로의 매핑 $\hat{f}^{(t)} : X^{(t)} \mapsto Y^{(t)}$으로 정의된다. 각 작업별로 n_t개의 학습 데이터 $\mathbf{X}^{(t)} \in \mathbb{R}^{d \times n_t}$와 레이블 $\mathbf{y}^{(t)} \in Y^{(t)^{n_t}}$가 주어진다. 전체 작업의 수 $T_{\max}$, 작업의 확률분포 및 작업의 순서에 대한 사전 정보는 없는 것으로 가정한다.

이와 같은 Lifelong 학습 문제 설정 하에서의 멀티모달 퓨전은 두 가지 방향으로 전개된다. 사전 퓨전의 경우, 각 인스턴스는 $X^{(t)} = \{X_{m_1}^{(t)}, X_{m_2}^{(t)} \ldots, X_{m_p}^{(t)}\}$와 같이 구성이 된다. 사후 퓨전의 경우, 각 작업은 $Z^{(t)} = \{Z_{m_1}^{(t)}, Z_{m_2}^{(t)}, ..., Z_{m_p}^{(t)}\}$와 같이 p개의 동기화된 부분 작업과 각 태스크별 매핑을 결합하는 최종 매핑 함수 $F^{(t)} = \{f_{m_1}^{(t)}, f_{m_2}^{(t)}, ..., f_{m_p}^{(t)}\}$를 구성한다.

각 작업의 구조적 모델은 GO-MTL[5] 및 ELLA[4]와 같이 은닉 변수 기반의 컴포넌트 공유 모델을 적용할 수 있다. 작업별 예측 함수는 $f^{(t)}(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}^{(t)})$와 같이 파라미터 벡터 $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^d$로 동작을 조절한다. 파라미터 벡터는 작업 간에 공유되는 k 개의 은닉 모델 컴포넌트($\mathbf{L} \in \mathbb{R}^{d \times k}$의 각 열)의 선형 결합으로 표현한다. 선형 결합시 적용되는 가중치 벡터를 $\mathbf{s}^{(t)} \in \mathbb{R}^k$로 표현할 때 $\boldsymbol{\theta}^{(t)} \in \mathbf{L}\mathbf{s}^{(t)}$이다.

이러한 설정에서 적용 가능한 학습의 목표 중 하나는 전체 T개의 작업에 대한 예측 오류율을 최소화하는 방향이며, 오류 함수 $\mathcal{L}$이 정의되어 있을 때, 식 (1)과 같은 목적함수로 표현된다:

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \left\{ \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}\left(f\left(\mathbf{x}_i^{(t)}; \mathbf{L}\mathbf{s}^{(t)}\right), y_i^{(t)}\right) + \mu\|\mathbf{s}^{(t)}\|_1 \right\} + \lambda\|\mathbf{L}\|_F^2 , \quad (1)$$

최적화 과정 및 매 작업 수행 시점에서의 모델 갱신 과정은 ELLA의 방식을 따를 수 있다.

멀티모달 퓨전 과정에서, 사전 퓨전을 적용하는 경우 파라미터 벡터의 크기가 각 모달 별 feature 벡터의 길이의 총합으로 표현되는 것 외에는 동일한 프레임웍을 적용할 수 있다. 사후 퓨전을 적용하는 경우에는 각 모달에 해당하는 부분 작업별로 별도의 은닉 모델 컴포넌트와 가중치 벡터를 둘 수 있으며, 이 경우 모달 별로 별개의 지식 저장소를 유지하는 것으로 해석할 수 있다. 퓨전 단계에서 적용할 상위 개념의 공통 지식 저장소를 각 모달 별 지식 저장소의 정보를 취합하여 구성하는 선택이 가능하다.

각 작업에서 분류를 수행하는 기반 학습 알고리즘으로는 로지스틱 회귀분석을 적용할 수 있다. 사후 퓨전을 위한 최종 매핑 함수는 [1]에서 제시한 다양한 기법을 적용할 수 있다.

3.2 웨어러블 센서 기반의 실용적인 프로토콜

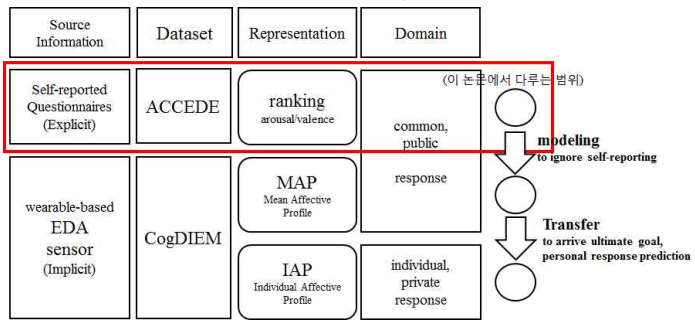

그림 2. 개인화된 감정 반응 예측을 위한 웨어러블 센서 기반의 데이터 구성 및 모델링 프로토콜

개인의 비디오 시청중 감정 반응 예측을 위한 데이터 수집 및 기계학습 모델 구성을 위해서는 웨어러블 센서가 필수적이다. 그러나, 충분한 데이터를 모으는데 필요한 시간과 사용자의 불편함이 장애 요인이다. 또한 개인 데이터만을 이용한 Lifelong 학습은 개인 간의 비교나 연관을 파악하여 소셜 모델링을 하기에는 너무 개인 특이적인 패턴을 보일 우려가 있다. V-A 감정 모델 적용시, 흥분도(A)는 EDA로 측정 가능하지만, 긍정/부정 정도(V)는 센서로 직접적인 측정이 불가하다.

이러한 난점을 해소하는 방법으로 사람들의 공통 반응 모델과의 조합을 제안한다[그림 2]. Self-report에 기반한 공개 V-A 데이터베이스에서 Lifelong 예측 모델을 구성하여 개인 Lifelong 예측 모델에서의 기저 트렌드 정보로 활용하며, 필요시 웨어러블 센서에 기반한 공통 반응 정보를 추가로 활용한다. 현재 ACCEDE, FilmStim[7]과 같

이 self-report에 기반한 대규모 감정 반응 DB와, 다중 센서를 이용한 감정 반응 DB 구축이 활발하다 (CogDIEM[8], Technicolor[9], DEAP[10]). 특히, CogDIEM과 Technicolor는 공통 반응에 대한 지표로서 MAP, 개인적 반응에 대한 지표로서 IAP를 모두 포함하고 있어 표준 모델 구성에 활용 가능하다.

4. 멀티모달 Lifelong 기반 비디오 감정반응 예측 실험

4.1 실험 데이터

감성기반 인지컴퓨팅을 위한 비디오 콘텐츠 데이터로, LIRIS-ACCEDE[3]를 선정했다. ACCEDE는 160편의 영화에서 추출한 9800개의 클립(8~12초 길이)에 대해 크라우드소싱 결과 정리한 시청자 감정 반응의 세기(arousal)와 valence 순위를 제공하는 데이터베이스이다. 우리는 9800개 중에 9600개를 사용했다. 그 절반인 4800개씩의 클립을 각각 학습 및 테스트 데이터로 지정하여 제공한다.

Valence와 arousal 중앙값(median)을 기준으로 이진화하여 Y={-1,1}인 감정 예측 문제를 설정한다. 학습데이터 X의 경우, Y. Baveye 가 저널 논문[3]에서 밝힌 10 종류의 특징값을 비디오 클립별로 추출하여 학습데이터 X를 만들었다. 학습 데이터는 4800개의 클립에서 150개씩 일정한 개수로 나눠서 32개의 task로 구성하였다. 레이블 Y의 불균형이 발생하지 않도록 무작위로 비디오 클립을 섞어 task를 생성했다.

4.2 실험 방법

ELLA는 4가지 파라미터로 동작하며, 지식의 구조를 결정하는 은닉 차원인 k, sparse coding으로 인코딩된 지식의 구조를 제어하는 regularization coefficient인 μ, 목적함수의 regularization을 위한 λ, logistics regression의 파라미터인 ridge 값을 그리드탐색으로 최적 값을 찾는다. 차원이 많이 큰 관계로 은닉 차원 k는 ELLA에서 권장하는 최대인 10으로 고정한다. λ는 {-5, -2, 1, 4}, μ는 {-40, ..., 0}, logistics regressions의 ridge는 {-3, -1, 0, 1, 3, 5}로 그리드를 지정한다.

4.3 결과

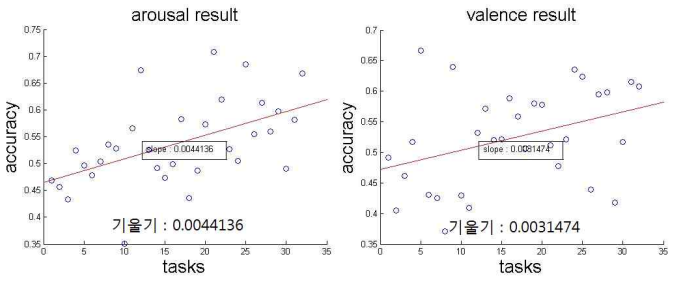

그림 3 감성기반 컴퓨팅 데이터를 ELLA를 통해 학습하여 검증한 결과 중, task 진행에 따른 정확도의 least square line 기울기가 가장 높은 파라미터 집합 중 하나

그리드검색으로 정확도의 증가 기울기가 높은 파라미터의 집합을 찾아 그 결과를 나타낸 것이 [그림 3]이다. 각 점들은 하나의 task와 그 정확도를 나타내며, 실선은 각 점들 사이의 거리를 최소로 하는 직선인 least square line을 그리고 있다. 첫 task의 성능은 50% 이하로 매우 저조하지만 task가 진행됨에 따라 점차 성능이 향상되어 마지막에는 65% 가까이 도달하는 것을 볼 수 있다. 또한 각각의 task 안에서 instance를 무작위로 섞어가며 반복 시행을 통한 검증을 한 결과, 전체 task에 대한 평균 정확도는 53.73%이다. least square line의 기울기로 보면 성능은 상승하는 경향성을 보인다.

기준으로 삼았던 배치 학습은 10-fold cross validation으로 logistics regression을 weka 3.7으로 배치 학습하여 검증했으며, 그 정확도가 arousal 예측 결과의 경우 59.51%, valence 예측 결과의 경우 59.45% 인 것과 비교해보면 Lifelong 학습의 결과와 큰 차이가 없다.

배치 학습으로 한 결과와 비교했을 때 큰 차이를 보이지 않으면서, Lifelong 학습 측면에서 봤을 때 데이터가 추가적으로 주어지면 지속적으로 성능이 향상될 것으로 기대된다.

5. 결론

본 논문에서는 사용자의 개입을 최대한 배제한 웨어러블 센서 기반 감정 예측 프레임웍으로서 멀티모달 Lifelong 감정 예측 방법과 실용적 프로토콜을 제안했다.

이러한 Lifelong 학습은 실용적 측면에서 매우 유용하게 쓰일 수 있다. 비디오 콘텐츠는 끊임없이 생산되는데 기업의 입장에서는 배치 학습을 할 경우에 학습된 이후에 나오는 콘텐츠에 대한 대처를 할 수 없게 되는 반면, Lifelong 러닝을 통해 비디오 콘텐츠를 학습할 수 있게 된다면 지속적으로 대처가 가능하게 된다.

기존의 Lifelong 학습에서 더 나아가 멀티모달에 대응하는 초석으로서 사전 퓨전을 시도하여 위와 같은 결과를 얻었다. 지금은 다중 모달리티를 하나의 데이터로 합쳤지만, 이후 각각의 모달리티를 따로 학습하여 합치는 사후 퓨전 기반 Lifelong 러닝이 가능할 것으로 보인다.

감사의 글

*박경화, 김병희 2명은 공동 1저자임. 본 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(NRF-2010-0017734-Videome)과 정보통신기술진흥센터의 지원(R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI)을 받아 수행된 연구임.

참고문헌

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: A survey,” *Multimed. Syst.*, 16:345-379, 2010.

[2] R. W. Picard, “Response to Sloman’ s Review of Affective Computing,” *AI Mag.*, 20(1):134-137, 1999.

[3] Y. Baveye, E. Dellandrea, C. Chamaret, L. Chen, Y. Baveye, and E. Dellandr, “LIRIS-ACCEDE: A Video Database for Affective Content Analysis,” *IEEE Trans. Affect. Comput.*, 6(1):43-55, 2015.

[4] R. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *IEEE Trans. Affect. Comput.*, 1(1):18-37, 2010.

[5] P. Ruvolo and E. Eaton, “ELLA: An efficient Lifelong learning algorithm,” in *Proc. ICML*, 28(1):507-515, 2013.

[6] A. Kumar, H. Daum, and H. D. Iii, “Learning Task Grouping and Overlap in Multi-task Learning,” in *Proc. ICML* pp. 1383-1390, 2012.

[7] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cogn. Emot.*, 24(7):1153-1172, 2010.

[8] 온경운, 김병희, 김경민, 곽동현, 박태서, 장병탁, “CogDIEM: 인지컴퓨팅 연구를 위한 멀티미디어 시청자의 암묵적 감정 반응 데이터베이스”, *한국정보과학회 동계학술발표회 논문집*, pp. 571-573, 2014.

[9] J. Fleureau, P. Guillotel, and I. Orlac, “Affective Benchmarking of Movies Based on the Physiological Responses of a Real Audience,” in *Proc. ACII*, 2013.

[10] S. Koelstra, M. Soleymani, J. Lee, A. Yazdani, T. Pun, A. Nijholt, and I. Patras, “DEAP : A Database for Emotion Analysis Using Physiological Signals,” *IEEE Trans. Affect. Comput.*, 3(1):18-31, 2012.