모달리티 불균형 완화를 위한 개별 헤드 기반 예측 융합

윤영우, 최원준, 이동규

경북대학교

dbsduddn@knu.ac.kr, sangju@knu.ac.kr, dglee@knu.ac..kr

Per-Modality Heads with Logit Fusion for Mitigating Modality Imbalance

Young-Woo Youn, Wonjun Choi, Dong-Gyu Lee

Kyungpook National University.

요 약

본 논문은 멀티모달 인간 행동 인식(MHAR)에서 발생하는 모달리티 간 최적화 불균형 문제를 해결하기 위한 새로운 학습 전략을 제안한다. 제안 방법은 학습 초기에는 대조 학습(contrastive learning)을 통해 모달리티 간 표현 차이를 줄이고, 이후 단계에서는 개별 손실과 융합 손실을 병행하는 지도 학습을 수행한다. 이를 위해 각 모달리티별 독립적인 classifier head를 두고 logit 수준에서 융합을 수행함으로써 균형 있는 학습을 유도하였다. 실험 결과, 제안한 방법은 기존 융합 방식 대비 안정적이고 일반화 성능이 높은 모델을 확보할 수 있음을 보여준다.

Ⅰ. 서 론

멀티모달 인간 행동 인식 (Multimodal Human Action Recognition, MHAR)은 인간의 다양한 활동을 인식하기 위해 RGB, skeleton, IMU 등 이질적인 센서 데이터를 결합하는 대표적인 응용 분야이다. 여러 모달리티를 동시에 활용함으로써 단일 모달리티 기반 인식보다 풍부한 표현력을 얻을 수 있다는 장점이 있다[1]. 그러나 멀티모달 학습에서는 모달리티 간 표현력과 학습 난이도의 차이로 인해 최적화 불균형이 발생할 수 있다. 이 경우 모델은 상대적으로 빠르게 최적화되는 모달리티에 의존하게 되며, 다른 정보는 충분히 활용되지 못한다. 이러한 현상은 멀티모달 감정 분야에서 이미 보고된 바 있으며, 특정 모달리티에 대한 과도한 의존이 전체 성능 저하로 이어질 수 있음이 지적되었다[2,3,4]. 또한, 기존 MHAR 연구의 융합 접근 방식은 크게 두 가지로 구분된다. (1) feature-level concatenation 방식, (2) decision-level logit aggregation 방식이다. 그러나 두 방법 모두 모달리티별 최적화 상태를 고려하지 못해 특정 모달리티에 치중되는 문제가 존재한다[5,6,7,8].

본 연구에서도 해당 문제가 인간 행동 인식 분야에서도 발생함을 확인하였다. 특히 MMAct[1] 데이터셋의 RGB와 IMU 모달리티를 실험적으로 분석한 결과, IMU 모달리티가 상대적으로 빠르게 최적화되어 융합 과정에서 불균형이 발생하는 경향이 관찰되었다. 이는 다른 도메인에서 제기된 문제와 유사한 현상으로, MHAR에서도 모달리티 불균형이 중요한 도전 과제임을 보여준다.

위 문제를 해결하기 위해 모달리티별 개별 손실과 통합 손실을 병행하는 logit aggregation 기반 학습 방법을 제안한다. 학습 초기에 대조 학습 기반 커리큘럼 단계를 도입하여 모달리티 간 표현 간극을 줄이고, 후반에는 지도 학습으로 전환하여 개별 손실과 통합 손실을 함께 최적화하는 전략을 적용하였다. 이를 통해 두 모달리티가 균형적으로 학습되며, 최종적으로 안정적이고 일반화 성능이 높은 모델을 확보할 수 있음을 보인다. 더 나아가, 대표적인 벤치마크 데이터셋을 활용한 비교 실험에서 제안하는 방법이 기존 방법들 대비 우수한 성능을 달성하였으며, 이는 멀티모달 학습에서 모달리티 불균형 문제를 효과적으로 완화할 수 있음을 시사한다.

Ⅱ. 본론

본 논문에서는 멀티모달 인간 행동 인식에서 발생하는 모달리티 간 최적화 불균형 문제를 해결하기 위하여 커리큘럼 기반 학습 프레임워크를 제안한다. 제안 방법은 크게 두 단계로 구성된다. 초기 단계에서는 대조 학습을 통해 모달리티 간 표현 간극을 줄이고, 후기 단계에서는 cross-entropy 기반 지도 학습을 통해 개별 모달리티와 융합 출력을 동시에 최적화한다. 전체 프레임워크는 그림 1과 같이 구성된다.

제안한 프레임워크의 입력은 두 가지 모달리티로 구성되어 있다. RGB 데이터는 ResNet-50을 backbone으로 사용하여 특징을 추출하며, 최종 fully-connected layer를 제거한 후 projection head를 통해 embedding 공간으로 투영한다. IMU 데이터는 시계열 신호의 장기 의존성을 포착하기 위해 temporal convolutional network 인코더를 사용한다. 모든 인코더의 출력은 동일한 차원의 embedding으로 변환되어 모달리티 간 비교가 가능하다.

학습 초기에는 각 모달리티의 표현 차이가 크기 때문에 직접적인 지도 학습을 수행하면 특정 모달리티로 최적화가 집중될 수 있다. 이를 방지하기 위해 대조학습을 먼저 적용한다. 동일한 클래스에 속하는 모달리티 표현은 embedding 공간에서 서로 가깝게, 서로 다른 클래스에 속하는 표현은 멀어지도록 학습한다. 대조 학습 손실은 다음과 같이 정리된다

$$\mathcal{L}_{con} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathrm{sim}(z_i, z_i^{+})/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}(z_i, z_j)/\tau\right)}, \quad (1)$$

본 단계는 비지도 학습 형태로 진행되지만 클래스 라벨 정보를 토대로 positive/negative 쌍을 구성함으로써 실제 task와 밀접하게 연관된 특징 정렬을 유도한다. 이를 통해 서로 다른 모달리티가 공통된 표현 공간에서 일정 수준의 정렬을 확보할 수 있다.

대조 학습을 마친 후에는 cross-entropy 기반 지도 학습 단계로 전환된다. 이 단계에서는 각 모달리티별 classifier head와 융합 classifier를 동시에 학습한다. 모달리티별 head는 RGB, IMU 각각의 logit 출력을 산출하

며, 여기에 개별 cross-entropy 손실을 부여한다.

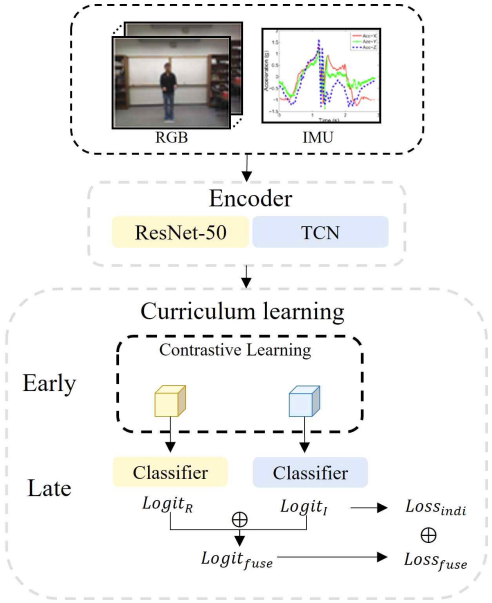

그림 1 제안하는 프레임워크의 전체 구조

동시에 두 모달리티의 logit을 평균하여 융합 결과를 얻고, 이에 대해서도 추가적인 손실을 계산한다. 최종 손실은 다음과 같이 정의된다.

$$\mathcal{L}_{total} = \sum_{m \in \{RGB, IMU\}} \mathcal{L}_{CE}^{(m)} + \mathcal{L}_{CE}^{fuse}, \quad (2)$$

최종 손실은 모달리티별 독립적인 성능을 보존하면서도 융합 결과의 정확도를 동시에 높일 수 있도록 설계되어 있다.

실험은 MMAct[1] 벤치마크 데이터셋을 사용하여 수행되었으며, 평가 척도로는 F1-score를 사용하였고 두 개의 cross-validation인 subject와 session에 대하여 평가하였다.

Method	F1-score	
	Cross-subject	Cross-session
SAKDN'21 [9]	77.23	82.77
MuMu'22 [5]	76.28	87.50
MAWKDN'23 [6]	78.77	85.26
SMTDKD'24 [7]	79.48	86.16
MSMFT'25 [8]	81.07	93.06
Ours	**84.88**	**93.40**

표 1 벤치마크 데이터셋에 대한 실험 결과

표 1의 실험결과를 통해 제안하는 방법이 기존 최신 방법들보다 우수한 성능을 보임을 확인 할 수 있다. 이는 멀티모달 학습에서 단순 융합 방식에 의존하기 보다는, 각 모달리티의 정보를 충분히 반영하여 통합 모델이 균형이 있게 학습되도록 유도하는 것이 중요함을 시사한다.

Ⅲ. 결론

본 논문에서는 멀티모달 인간 행동 인식에서 발생하는 모달리티 간 최적화 불균형 문제를 해결하기 위해 커리큘럼 기반 학습 전략을 제안하였다. 초기 단계에서는 클래스 정보를 활용한 대조 학습을 통해 모달리티 간 표현 간극을 줄였으며, 후반부에는 cross-entropy 기반 지도 학습으로 전환하여 개별 손실과 통합 손실을 동시에 최적화하였다. 이를 통해 서로 다른 모달리티가 균형적으로 학습될 수 있었으며, 최종적으로 안정적이고 일반화 성능이 높은 모델을 확보할 수 있었다.

MMAct 데이터셋을 대상으로 한 실험 결과, 제안 방법은 기존 융합 방식 대비 우수한 성능을 보였으며, 이는 멀티모달 학습에서 융합 전략뿐만 아니라 개별 모달리티의 학습을 병행적으로 고려하는 것이 중요함을 보여준다. 특히 모달리티 간 불균형 문제가 실제 데이터셋에서도 관찰되었음을 확인하였고, 제안하는 학습 전략이 이를 효과적으로 완화할 수 있음을 입증하였다.

향후 연구에서는 depth, skeleton과 같은 추가 모달리티를 포함하여 더 복잡한 환경에서의 확장 가능성을 검증하고, 다양한 커리큘럼 설계 방식을 탐구할 예정이다. 또한 본 방법을 대규모 데이터셋과 실시간 추론 환경에 적용하여 효율성과 범용성을 동시에 향상시키는 방향으로 발전시킬 계획이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터의 지원(IITP-2025-RS-2020-II201808, 50%)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2025-02214941, 50%)

참 고 문 헌

[1] Kong, Quan, et al. "Mmact: A large-scale dataset for cross modal human action understanding." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Yang, Yang, et al. "Facilitating multimodal classification via dynamically learning modality gap." Advances in Neural Information Processing Systems 37 (2024): 62108-62122.

[3] Qian, Chengxuan, et al. "Dyncim: Dynamic curriculum for imbalanced multimodal learning." arXiv preprint arXiv:2503.06456 (2025).

[4] Wei, Yake, and Di Hu. "MMPareto: boosting multimodal learning with innocent unimodal assistance." Proceedings of the 41st International Conference on Machine Learning. 2024.

[5] Islam, Md Mofijul, and Tariq Iqbal. "Mumu: Cooperative multitask learning-based guided multimodal fusion." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 1. 2022.

[6] Quan, Zhenzhen, et al. "MAWKDN: A multimodal fusion wavelet knowledge distillation approach based on cross-view attention for action recognition." IEEE Transactions on Circuits and Systems for Video Technology 33.10 (2023): 5734-5749.

[7] Quan, Zhenzhen, et al. "SMTDKD: A semantic-aware multimodal transformer fusion decoupled knowledge distillation method for action recognition." IEEE Sensors Journal 24.2 (2023): 2289-2304.

[8] Zhou, Xianfa, et al. "MSMFT: Multi-Stream Multimodal Factorised Transformer for Human Activity Recognition." IEEE Sensors Journal (2025).

[9] Liu, Yang, et al. "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition." IEEE Transactions on Image Processing 30 (2021): 5573-5588.