부정 샘플 선택 전략을 통한 비지도 대조 학습

윤영우, 이동규
경북대학교
dbsduddn@knu.ac.kr, dglee@knu.ac.kr

Unsupervised Contrastive Learning with Negative Selection Strategy

Young-woo Youn, Dong-Gyu Lee
Kyungpook National University.

요 약

본 논문은 부정 샘플 선택 전략을 이용한 비지도 대조 학습에 관한 연구이다. 기존의 대조 학습 모델들은 부정 샘플을 선택 후, 학습에서 선택된 부정 샘플들을 제외하는 방식의 연구가 보편적이다. 하지만 이러한 방식은 임시 레이블을 최대한으로 활용하지 못하는 문제점이 존재한다. 본 논문에서는 클러스터링 기법을 사용하여 각 모달리티의 임시 레이블들을 형성하고 그 조합을 통해 정확한 임시 레이블들을 생성한다. 생성한 임시 레이블들을 토대로 대조학습을 진행하여 향상된 학습결과를 보여준다. 제안한 방법은 동일한 데이터셋 실험 결과, 기존의 방법과 비교하여 우수한 성능을 보인다.

Ⅰ. 서 론

인공지능과 기계 학습의 발전에 따라, 인간 활동 인식(Human Activity Recogntion, HAR)은 다양한 응용 분야에서 중요한 역할을 하고 있다. HAR 시스템은 헬스케어, 스포츠 분석과 같은 여러 분야에 사용되며, 인간의 활동을 자동으로 인식하고 분석함으로써 효율성을 향상시킬 수 있다 [1]. 기존의 HAR 연구는 주로 컴퓨터 비젼 기술들을 활용하여 활동 인식을 집중해왔다. 그러나 단일 모달리티 데이터는 제한된 정보만을 제공하므로, 복잡한 활동을 인식하는데 한계가 존재한다. 이러한 한계를 극복하기 위해 최근에는 멀티모달 데이터(Multimodal data)를 활용한 HAR 연구가 활발히 진행되고 있다 [2].

멀티모달 데이터는 여러 유형의 센서 데이터를 결합하여 풍부한 정보를 제공하며, 이를 통해 HAR 시스템의 정확도와 범용성을 크게 향상시킬 수 있다. 최근의 멀티모달 HAR 연구에서는 다양한 방법론이 제안되었다. 대표적으로, 대조 학습(Contrastive learning) 기법을 사용하여 레이블되지 않은 데이터로부터 개별 모달리티의 특징을 추출하고 유용한 표현을 학습하는데 있어 많은 주목을 받고 있다. 이 기법은 데이터의 유사성과 차이점을 학습하고 최근의 연구들에서는 SimCLR [3]의 대조 학습 모델을 주로 사용한다.

그러나 이러한 방법들에도 여전히 몇 가지 문제점이 존재한다. 첫째, 대조 학습 방법은 부정 샘플(Negative samples)의 선택 전략에 크게 의존한다. 잘못된 부정 샘플의 선택은 모델이 잘못된 학습을 하게 만들고, 과도한 군집화(Over-clustering)가 발생하여 결과적으로는 모델의 성능저하를 유발한다 [4]. 둘째, 양성 샘플(Positive samples)의 수를 고려하지 않는 문제 [5]가 있으며 모델의 일반화 성능에 부정적인 영향을 준다.

본 논문은 이러한 문제점을 해결하는 새로운 멀티모달 HAR 모델을 제안한다. 우리는 다양한 모달리티 데이터를 활용한 Contrastive Multiview Coding(CMC) [6] 기법을 도입하여, 레이블되지 않은 데이터로부터 효과적으로 유용한 표현을 학습하고자 한다. 또한, K-Means 알고리즘과 조합 기반 재정렬 방법을 결합한 새로운 부정 샘플 선택 전략을 제안하여 모델의 성능을 개선하고자 한다. 이러한 방법들을 통해 우리는 기존의 문제점들을 극복하고, 최종적으로 HAR 시스템의 성능을 향상시킨다.

Ⅱ. 본론

본 논문에서는 효과적인 대조 학습을 위한 새로운 프레임워크를 제안한다. 제안하는 프레임워크는 모달리티마다 클러스터링을 진행하여 클러스터의 중심에 가장 근접하게 위치한 데이터들을 중심으로 다시 클러스터링을 진행한다. 이 과정을 이전 클러스터와 현재 클러스터의 중심거리가 특정 임계점을 벗어나지 않을 때까지 진행한다. 특정 임계점에 도달하면 클러스터의 중심에 가장 근접한 데이터들을 다시 뽑아서 클러스터링을 진행하여도, 이전의 클러스터와 거의 동일한 정도의 결과가 도출되기 때문에 초창기 클러스터 결과보다 더욱 견고한 클러스터를 얻을 수 있다.

Algorithm 1 Finding and Adjusting Skeleton and Inertial Data Clusters

Input: Skeleton data D_s, Inertial data D_i, Number of clusters k
Output: Adjusted data points with valid combinations R
Perform k-means clustering on D_s and D_i to obtain cluster centers C_s and C_i.
Initialize an empty set S to store cluster combinations.
for each data point (d_s, d_i) in (D_s, D_i) **do**
    Find the cluster (c_s, c_i) for (d_s, d_i) in (C_s, C_i).
    Add the tuple (c_s, c_i) to set S.
end for
Count occurrences of each combination in S and sort by frequency to obtain list L.
Define valid combinations M as the top combinations in L.
for each data point (d_s, d_i) in (D_s, D_i) **do**
    Find the cluster (c_s, c_i) for (d_s, d_i) in (C_s, C_i).
    if $(c_s, c_i) \in M$ **then**
        Add (d_s, d_i) to R.
    else
        Compute distances to centers μ_{s_m} and μ_{i_m} for each combination (m_s, m_i) in M.
        Find the combination with minimum total distance d_{total}.
        Adjust (d_s, d_i) to match this best combination.
        Add the adjusted point to R.
    end if
end for
Output: The adjusted data points R.

알고리즘 1. 클러스터 매칭 및 임시 레이블 재정렬

그 이후, 알고리즘 1에 따라서 각 모달리티의 클러스터 결과로 조합을 생성하고 많이 나온 조합 순으로 서로 다른 모달리티의 클러스터들은 같은 임시 레이블로 매칭한다. 매칭되지 않은 임시 레이블들은 클러스터의 중심과 데이터까지의 거리를 계산하여 가장 가까운 클러스터로 임시 레이블

을 재정렬하여 임시 레이블의 노이즈를 줄이도록 설계되었다.

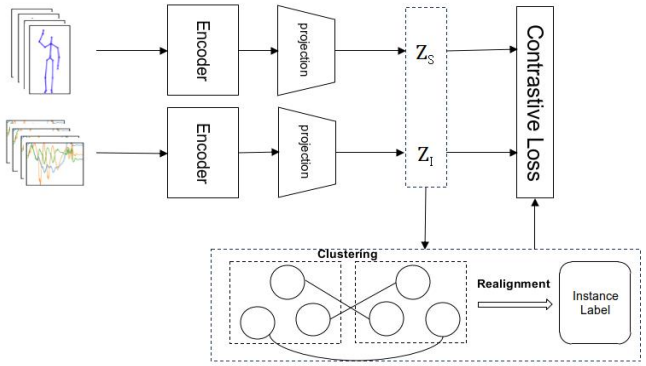

그림 1. 제안한 프레임워크

그림 1을 보면 인코더에서 나온 벡터값을 통해 클러스터링을 하고 그 결과를 조합 기반 재정렬 방법으로 임시 레이블을 생성하여 대조 학습으로 학습한다. 인코더는 미리 학습된 인코더로 Transformer 기반의 인코더, CNN기반의 co-occurrence [7]를 사용하였고 학습 파트에서 제안된 프레임워크 기반으로 추가학습을 진행한다. 이때 사용하는 손실함수는 대조 학습에서 보편적인 InfoNCE loss [8]를 제안된 프레임워크에 맞게 한 개의 고정된 양성 샘플만 사용하는 것이 아닌 임시 레이블을 기반으로 양성 샘플과 음성 샘플을 다시 나누어 손실을 계산하도록하였다. 수정된 손실함수를 수식으로 표현하면 다음과 같다.

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\left(\frac{\text{sim}(f(z_i),f(z_j))}{\tau}\right)}{\exp\left(\frac{\text{sim}(f(z_i),f(z_j))}{\tau}\right) + \sum_{k=1}^{N-1}\exp\left(\frac{\text{sim}(f(z_i),f(z_k))}{\tau}\right)}$$

수식 1. 손실 계산식

Ⅲ. 실험결과

본 논문에서 제안한 프레임워크의 평가를 위해 동일한 데이터셋으로 기존의 연구들과 비교 실험을 진행한다. 현재 프레임워크의 기반이 되는 SimCLR [3]과 CMC [6]를 같이 비교하였으며, 추가적으로 지도 학습 기반의 모델도 같이 실험을 진행하였다. 사용한 모달리티는 총 두 개로 센서 데이터(Inertial)와 스켈레톤(Skeleton)이다. 성능지표는 UTD-MHAD [9] 데이터의 경우는 Accuracy, MMAct [10] 데이터의 경우는 F1-score를 사용하였다.

Modality	Approach	UTD-MHAD (Accuracy)	MMAct x-subject (F1-score)	MMAct x-scene (F1-score)
Inertial	SimCLR	72.09	52.89	59.23
Inertial	Supervised	76.74	61.22	78.86
Skeleton	SimCLR	95.11	75.82	67.80
Skeleton	Supervised	94.65	82.50	70.58
Multimodal	CMC	96.04	82.05	79.97
Multimodal	Supervised	97.21	84.05	**87.36**
Multimodal	Ours	**97.67**	**84.26**	83.84

표 1. 실험 결과

실험 결과, 본 논문에서 제안한 방법이 기존의 방법들보다 전반적으로 우수한 성능을 보이는 것을 확인할 수 있다. MMAct 데이터셋의 x-scene 실험에서 지도 학습에 비해 낮은 성능을 보이는 것을 제외하고는 UTD-MHAD와 MMAct x-subject 실험에서 지도 학습과 기존 방법과 비교해서 가장 좋은 Accuracy와 F1-score를 보여주고 있다.

Ⅳ. 결론

본 논문은 기존의 대조 학습의 부정 샘플 선택 전략에 존재하는 과도한 군집화 문제, 양성 샘플을 고려하지 않는 문제를 효과적으로 해결하기 위해서 새로운 프레임워크를 제안한다. 제안하는 프레임워크는 각 모달리티의 클러스터링 결과를 토대로 데이터의 임시 레이블 조합을 구하고 이를 토대로 클러스터 매칭 후 임시 레이블 재정렬 과정을 거침으로써, 추가적인 양성 샘플을 획득하고 과도한 군집화 문제를 해결한다. 2가지의 데이터셋에 대한 실험을 통해 제안한 프레임워크가 최신 연구들과 비교해 높은 성능을 보이고 그 방법론이 효과적인 것을 입증한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT연구센터 지원사업의 연구결과로 수행되었음. (IITP-2024-2020-0-01808)

참 고 문 헌

[1] Kaseris, Michail, Ioannis Kostavelis, and Sotiris Malassiotis. "A Comprehensive Survey on Deep Learning Methods in Human Activity Recogntion." Machine Learning and Knowledge Extraction 6.2 (2024): 842-876.

[2] Ni, Jianyuan, et al. "A Survey on Multimodal Wearable Sensor-based Human Action Recognition." arXiv preprint arXiv:2404.15349 (2024).

[3] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[4] Wang, Jinqiang, et al. "Negative selection by clustering for contrastive learning in human activity recogntion." IEEE Internet of Things Journal 10.12 (2023): 10833-10844.

[5] Choi, Hyeongju, Apoorva Beedu, and Irfan Essa. "Multimodal contrastive learning with hard negative sampling for human activity recognition." arXiv preprint arXiv:2309.01262(2023).

[6] Tian, Younglong, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding." Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI 16. Springer International Publishing, 2020.

[7] Brinzea, Razvan, Bulat Khaertdinov, and Stylianos Asteriadis. "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition." 2022 International Joint Conference on Neural Networks(IJCNN). IEEE, 2022.

[8] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

[9] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnava. "UTD-MHAD: A Multimodal dataaset for human action recognition utilizing a depth camera and a wearable inertial sensor." 2015 IEEE International conference on image processing (ICIP). IEEE, 2015.

[10] 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.