센서 데이터를 활용한 인간 활동 인식을 위한 버트 기반 언어 모델과 하이브리드 딥러닝 접근법의 비교 연구

황준경[1], 김지인[2], 황의석[2]
[1]서울과학기술대학교, [2]광주과학기술원

wnsghkdqk0802@seoultech.ac.kr, jeeinkim@gm.gist.ac.kr, euiseokh@gist.ac.kr

Comparative Study of BERT-Based Language Models and Hybrid Deep Learning Approaches for Human Activity Recognition Using Sensor Data

Hwang Junkyung[1], Kim Jeein[2], Hwang Euiseok[2]

[1]Seoul Nat'l Univ. of Science and Technology, [2]Gwangju Inst. of Science and Technology

요 약

본 연구는 스마트 홈 센서 데이터를 활용한 인간 활동 인식에서 BERT 기반 언어 모델과 심층학습 기법을 결합한 하이브리드 분류 방식을 제안한다. 원시 센서 데이터는 BERT 계열 사전학습 언어 모델(all-MiniLM-L6-v2)과 다양한 심층학습 분류기에 입력되어 임베딩 벡터를 생성하고, 이를 완전연결 계층을 통해 최종 활동을 예측한다. 실험 결과, 제안된 하이브리드 방식은 심층학습 단독 모델보다 높은 분류 정확도를 달성하였다. 이는 언어 모델이 센서 데이터의 문맥적 특성을 반영하여 심층학습 기법이 간과할 수 있는 의미적 정보를 보완했기 때문이다. 또한 실험은 임베딩 벡터의 의미적 공간에서의 정렬 과정이 정교하게 이루어지지 않을 경우 기대한 성능을 충분히 발휘하지 못할 수 있음을 시사하였다. 따라서 의미적 공간에서의 정합성을 강화할 경우 더 높은 분류 정확도를 달성할 수 있음을 보여준다.

Ⅰ. 서 론

센서 데이터를 기반으로 한 인간 활동 인식(Human Activity Recognition, HAR)은 최근 스마트 홈, 헬스케어, 웨어러블 기기 등 다양한 분야에서 그 활용성이 확대되며 주목받고 있다. 사물인터넷(Internet of Things, IoT) 기술의 발달로 다양한 센서들이 인간의 행동을 감지하고, 이를 바탕으로 상황 인지형 서비스를 지원할 수 있게 되었다. 동시에, 인공지능의 발전 특히 ChatGPT 와 같은 대규모 언어 모델의 등장은 언어 처리 영역을 넘어 맥락적 추론 능력이 다양한 응용 분야에서 활용될 가능성을 보여주고 있다.

기존의 센서 기반 활동 인식 연구들은 대체로 특정 공간을 구분하여 그 구역 내에서 발생하는 활동만을 제한적으로 측정하는 방식에 의존해왔다.[1] 예를 들어 식사나 요리와 같은 활동은 부엌에서만 일어난다고 가정하고, 해당 구역의 센서 데이터를 통해 분석하는 것이다. 그러나 실제 인간의 행동은 특정 구역에만 국한되지 않으며, 서로 다른 공간에서 유사한 활동이 나타나거나 여러 활동이 복합적으로 발생하는 경우가 많다. 이러한 점은 기존 연구의 한계로 작용하며, 실제 환경에서의 정확한 활동 인식을 어렵게 한다.

따라서 단순히 센서 신호에 기반한 전통적인 심층학습 접근만으로는 활동 인식의 맥락적 복잡성을 충분히 반영하기 어렵다. 특히 데이터의 의미적 관계와 맥락을 해석할 수 있는 언어 모델과의 결합은 기존 한계를 보완할 수 있는 새로운 가능성을 제시한다. 언어 모델의 맥락적 해석 능력과 심층학습의 특징 추출 능력을 융합함으로써, 보다 정교하고 실질적인 활동 인식 성능 향상을 기대할 수 있다.

이에 본 연구는 스마트 홈 환경에서 센서 데이터를 활용한 인간 활동 인식에 있어 언어 모델과 심층학습의 결합 방법을 제시한다. 이를 통해 실제 생활 환경의 복잡성을 반영하고, 기존 방법의 한계를 극복할 수 있는 접근의 필요성과 가능성을 확인하고자 한다.

Ⅱ. 데이터셋 선택 및 방법론

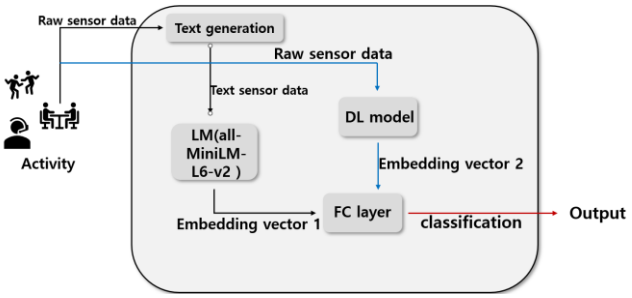

그림 1. 전체 프레임워크

Figure 1. Overall framework

2.1 데이터셋

본 연구에는 DOO-RE 데이터셋을 활용하였다.[2] 해당 데이터셋은 2023 년 KAIST 오피스 환경에서 수집되었으며, 센서는 특성에 따라 환경 기반 센서, 사용자 기반 센서, 액추에이터 기반 센서로 구분된다. 환경 기반 센서는 밝기, 습도, 온도, 소리 등과 같이 주변 환경의 물리적 상태를 측정하며, 사용자 기반 센서는 발표자 감지, 출입, 동작, 착석 등 인간의 존재 및 활동을 인식하는 역할을 한다. 또한 액추에이터 기반 센서는

에어컨, 조명, 프로젝터와 같이 환경을 제어하거나 특정 동작을 유발하는 장치와 연동되어 사용된다. 데이터셋에서 정의된 타겟 활동은 혼자 식사, 독서, 전화, 세미나, 랩 미팅, 대화, 그룹 공부, 기술 토론, 그룹 식사 등 총 10 종으로 구성된다. 데이터셋은 메타데이터와 행동별 센서 데이터로 이루어져 있으며, 메타데이터에는 행동명, 사용 센서, 시작 및 종료 시각, 참여 인원, 지속 시간이 포함되고, 행동별 센서 데이터에는 각 센서 타입별 측정값이 기록된다. 본 연구에서는 해당 데이터셋을 언어 모델 및 심층학습 기반 분류 모델에 적용하기 위해 다음과 같은 전처리 과정을 수행하였다.

2.2 모델 선택 및 임베딩 벡터 결합

전체 프레임워크는 그림 1 과 같고, 연구는 다음과 같이 진행했다. 우선 언어 모델로는 all-MiniLM-L6-v2 를 채택하였다. 본 모델은 BERT 계열 중 경량화 된 구조를 기반으로 연산 효율성이 높고 의미 파악 능력이 보장되므로, 대규모 센서 윈도우 데이터를 처리해야 하는 HAR 과제에 적합하다. 원시 센서 데이터는 언어 모델 입력에 활용하기 위해 텍스트 형태로 변환하였으며, 구체적으로 타겟 활동명, 센서 타입, 측정값을 문장으로 구성하여 언어 모델에 입력한 뒤 임베딩 벡터를 생성하였다.[3]

기계학습 및 심층학습 모델은 LightGBM, MLP, CNN, BiLSTM, DeepConvLSTM 을 선정하였는데, 이는 HAR 연구에서 널리 사용되고 있으며 언어 모델 기반 접근법과의 성능 비교에 적합하다고 판단하였다. 학습 과정에서는 센서 데이터를 윈도우 단위 시계열 데이터로 변환한 후 이를 학습, 검증, 테스트셋으로 분할하여 모델 학습을 수행하였으며, 언어 모델을 통해 생성된 임베딩 벡터는 심층학습 모델의 출력과 함께 FC Layer 에서 결합된 후 최종 분류기를 통해 활동 예측 결과를 도출하였다.

Ⅲ. 실험 결과

표 1 과 2 는 각각 심층학습 모델만 사용했을 경우의 정확도와 언어 모델과 심층학습 모델을 같이 사용했을 경우의 정확도를 나타낸다. 각 활동별 F1-score 를 나타내었으며, 각 행의 최고 값은 볼드체, 차순 값은 이탤릭체로 표시하였다. 두 표를 비교해보면 언어 모델을 함께 사용했을 경우 정확도가 오르는 것을 확인할 수 있다. 하지만 기계학습 모델인 LightGBM 의 경우 오히려 언어 모델을 사용한 경우, 정확도가 떨어지는 것을 확인할 수 있는데 두 벡터를 정렬시킬 때 의미적 공간(Semantic Space)에서의 정렬이 제대로 이루어지지 않아 정확도가 낮게 나오는 것으로 추측된다.[4]

표 1. 기계학습/심층학습 모델만 사용 시 정확도

Table 1. Accuracy when using only the deep learning model

Activity	LightGBM	MLP	CNN	BiLSTM	DCLSTM
Eating	**0.9314**	0.1659	0.1408	0.2992	*0.5310*
Eating_together	**0.9315**	0.1957	0.3846	0.0706	*0.4643*
Lab_meeting	**0.9007**	0.7167	0.7290	0.7491	*0.8329*
Phone_call	**0.8333**	0.2398	0.3789	0.4433	*0.5258*
Reading	**0.8893**	0.6588	0.6807	0.7153	*0.7679*
Seminar	**0.9464**	0.7301	0.7267	0.7782	*0.8455*
Small_talk	**0.8576**	0.5724	0.5148	0.6206	*0.7163*
Study_together	**0.9622**	0.6453	0.6466	0.6644	*0.7193*
F1-score	**0.8999**	0.5965	0.6127	0.6413	*0.7324*

표 2. 언어 모델과 기계학습/심층학습 모델 사용 시 정확도

Table 2. Accuracy when using both the language model and the machine learning/deep learning model

Activity	LightGBM	MLP	CNN	BiLSTM	DCLSTM
Eating	**0.7875**	0.2037	0.2667	*0.6277*	0.5819
Eating_together	*0.6071*	0.1818	0.1429	**0.7967**	0.5076
Lab_meeting	**0.8901**	0.6994	0.7083	*0.8577*	0.8396
Phone_call	**0.7346**	0.3929	0.4836	0.5136	*0.5374*
Reading	**0.8907**	0.7198	0.7098	0.7787	*0.7826*
Seminar	0.7346	0.7090	0.7410	**0.8929**	*0.8703*
Small_talk	**0.8907**	0.6268	0.5947	*0.7921*	0.7729
Study_together	**0.9401**	0.6667	0.6438	*0.8814*	0.8220
F1-score	**0.8776**	0.6518	0.6447	*0.7701*	0.7575

Ⅲ. 결론

본 연구는 HAR 분야에서 언어 모델의 활용 가능성을 제시하였다. 센서 데이터를 기반으로 언어 모델과 심층학습 모델에서 각각 임베딩 벡터를 생성하고 이를 결합함으로써, 단독 심층학습 모델에 비해 성능이 개선될 수 있음을 보였다. 따라서 향후 연구에서는 의미적 공간 간 정렬을 통해 정합성을 강화함으로써 보다 높은 정확도를 달성할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2024-00349582).

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2021-II211835)

참 고 문 헌

[1] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, “CASAS: A smart home in a box,” *Computer*, vol. 46, no. 7, pp. 62- 69, Jul. 2012. (doi: 10.1109/MC.2012.328)

[2] H. Kim, G. Kim, T. Lee, K. Kim, and D. Lee, “DOO-RE: A dataset of ambient sensors in a meeting room for activity recognition,” *Scientific Data*, vol. 11, no. 50, pp. 1- 17, Jan. 2024. (doi: 10.1038/s41597-024-01234-5)

[3] G. Civitarese, M. Fiori, P. Choudhary, and C. Bettini, “Large Language Models Are Zero-Shot Recognizers for Activities of Daily Living,” *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 8, no. 4, pp. 1- 26, Dec. 2024. (doi: 10.1145/3678635)

[4] Z. Chen, Y. Zhao, S. Shen, Z. Zhang, and H. Xu, “LanHAR: Language-centered Human Activity Recognition,” *Proc. 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, 2024, pp. 5822- 5840. (arXiv:2402.19468)