미세조정된 VideoLLaMA2 기반의 멀티모달 보행자 횡단 의도 예측

김성훈[1], 함제석[2*], 문진영[3]
[1]순천향대학교 AI·빅데이터학과 학사과정
[2]한국전자통신연구원 연구원
[3]한국전자통신연구원 책임연구원
ksh0816@sch.ac.kr, jsham@etri.re.kr, jymoon@etri.re.kr

Multi-modal Pedestrian Crossing Intention Prediction based on Finetuned VideoLLaMA2

Sunghun Kim[1], Je-Seok Ham[2*], Jinyoung Moon[2]
[1]Dept. of AI·Bigdata, Soonchunhyang University
[2]Electronics and Telecommunications Research Institute (ETRI)

요 약

급속한 도시화와 교통량의 증가로 인해 보행자 안전이 중요한 사회적 문제로 부각되고 있다. 이에 따라 보행자의 횡단 여부를 예측하는 다양한 연구가 활발히 진행 중이다. 본 연구에서는 보행자 행동 예측에 대표적으로 활용되는 JAAD 데이터셋을 기반으로 QA 셋을 제작하고, 이를 최신의 오픈소스 MLLM에 해당하는 VideoLLaMA2 모델에서 미세조정을 진행하였다. 이 모델을 기반으로 과거 16프레임 동안의 보행자 움직임을 관찰한 후, 30프레임 이후 시점에서 보행자의 횡단 의도(crossing/not-crossing)를 예측하고 그 정확도를 비교·분석한다. 그 결과, 미세조정을 진행한 모델에서 더 높은 예측 정확도를 나타내었으며. 향후 복잡하고 새로운 도로 환경에서도 보행자의 미래 횡단 의도를 예측하여 보행자의 안전성을 높일 수 있다.

1. 서론

현대 도시화와 교통량의 증가로 인해 보행자 안전이 중요한 문제로 언급되고 있다[1]. 도로에서 보행자는 차량과의 충돌에 매우 취약한 존재이며, 이러한 충돌은 치명적인 결과를 초래할 수 있다. 특히, 보행자 사고는 차량과의 직접적인 신체적 충돌로 인해 치명적인 부상으로 이어질 확률이 높다[2]. 따라서, 보행자 안전을 강화하고 교통사고를 예방하기 위한 시스템의 필요성이 대두되고 있다.

보행자 의도 예측은 자율주행차 및 운전자 첨단 지원 시스템에서 중요한 역할을 한다[3,4,5]. 그러나 기존의 보행자 탐지 및 행동 예측 시스템은 주로 데이터 기반의 블랙박스 형태로 동작하며, 이에 따라 예측 과정의 해석이 불가하다[6]. 해석 가능성이 부족한 시스템은 사용자가 모델의 신뢰성을 평가하는 데 어려움을 초래한다. 교통안전과 같은 중요한 분야에서 예측 시스템의 신뢰성을 높이기 위해서는 투명하게 해석 가능성을 확보해야한다. 따라서 해석 가능한 모델을 통한 보행자 의도 예측이 필요하다.

기존의 연구들은 주로 영상 처리 기술을 사용하여 보행자의 움직임을 감지하고 이를 바탕으로 의도를 예측해 왔다[6]. 그러나 이러한 모델은 보행자의 행동 변화를 반영하는 데 한계가 있었으며, 종종 신뢰도가 부족한 결과를 초래했다. 특히, 다양한 환경에서의 보행자 행동을 일관되게 예측하는 것이 어려웠으며, 일부 연구들은 이러한 문제를 해결하기 위해 LLM(Large-Language Model), VLM(Vision-Language Model), MLLM(Mutlimodal-Large-Language Model) 등의 여러 언어 모델들을 사용한다. 최근에는 MLLM을 활용하여 보행자의 의도를 예측하는 방식이 주목받고 있다[7]. 이러한 방식은 다중 모달 데이터를 결합하여 보행자의 행동을 더욱 정밀하게 분석할 수 있게 한다.

본 논문은 최신의 오픈소스 MLLM에 해당하는 VideoLLaMA2[9] 모델을 미세 조정하여 보행자 횡단 의도를 예측하는 해석 가능한 멀티모달 접근 방식을 제안한다. 기존의 VideoLLaMA2 모델은 도로 환경뿐만 아니라 다양한 종류의 비디오들에 사전학습되어 있다. 본 연구에서는 보행자 행동 예측에서 대표적으로 활용되는 JAAD 데이터셋에 대하여 질문과

* 교신저자 (Corresponding Author)

답변 형태로 구성된 QA 형태로 데이터셋을 변환하여 제작하고, 이 셋을 기반으로 VideoLLaMA2 사전 학습 모델에서 미세조정을 수행하였다. 그리고 추론 과정에서 4가지 입력 특징과 문자 프롬프트를 결합하여 멀티모달 추론 과정을 수행하였다. 그 결과, 미세조정하지 않은 모델에 비해 예측 정확도는 2% 높은 60%를 나타냈으며, 입력 특징 소거 실험을 수행하여 4가지 입력 특징의 유효성을 입증하였다. 본 논문은 해석 가능한 예측 모델을 도입함으로써 교통 안전성을 높이고, 자율주행차나 운전자 첨단 지원 시스템에서 보행자와 탑승자의 안전을 확보함으로써 시스템의 신뢰도를 향상하는 데 기여할 것이다.

2. 제안하는 방법

2.1. 데이터셋

본 연구에서는 JAAD(Joint Attention in Autonomous Driving)[8] 데이터셋을 활용하여 미세조정과 예측 성능 평가를 진행한다. JAAD 데이터셋은 차량과 보행자 간의 상호작용을 파악하기 위해 사용된다. JAAD 데이터셋은 총 346개의 비디오 클립으로 구성되어 있으며, 이 중 학습 데이터는 188개, 검증 데이터셋은 32개, 테스트 데이터셋은 126개이다.

<표 1> JAAD 데이터셋 정보

JAAD 데이터셋	
총 프레임 수	82,032
총 보행자 수	2,786
초당 프레임 수 (FPS)	30
클립당 길이 (초)	5-10
행동 주석이 있는 보행자 수	686
도로를 횡단한 보행자 수	495
도로를 횡단하지 않은 보행자 수	191

각 보행자에 대한 바운딩박스 좌표, 횡단 여부, 보행자의 행동(걷기, 멈춤, 주시) 등에 대한 주석이 제공된다. 주석을 기반으로 아래와 같이 모든 프레임에 대해 JAAD QA셋을 제작하였다. 이 형식은 VideoLLaMA2에서 학습 가능한 셋의 형태이다.

```
"conversations": [
    {
        "from": "human",
        "value": "<image>\nIs the pedestrian in the red box crossing the road?"
    },
    {
        "from": "gpt",
        "value": "not-crossing"
    }
]
```

Q: Is the pedestrian crossing the road? (보행자가 도로를 건너고 있는가?)

A: Crossing/Not-Crossing (횡단/비횡단)

본 연구에서는 $JAAD_{train}$과 $JAAD_{test}$로 나누어 실험을 진행하였다. $JAAD_{train}$은 VideoLLaMA2을 미세조정 하기 위해 사용되었다. 반면, $JAAD_{test}$는 학습된 모델의 성능을 평가하는 데 사용되었으며, 모델이 보행자의 횡단 여부를 얼마나 예측 성능을 확인하고자 한다.

2.2. 베이스라인 모델

본 연구에서는 비디오 데이터의 복잡한 공간적 및 시간적 역학을 효율적으로 포착하기 위해 공개된 VideoLLaMA2[9]의 STC(Spatial-Temporal Convolution) connector를 활용한다. STC 커넥터는 풀링(Pooling)과 컨볼루션(Convolution) 연산을 통해 입력과 출력 간의 공간-시간 순서를 유지하며, 이를 통해 시간적 일관성을 보장한다. 이러한 방식은 데이터의 과거 다중 프레임에서도 시간적 흐름을 왜곡하지 않으면서도 효과적으로 패턴을 추출할 수 있는 장점이 있다. 이러한 장점으로 본 연구의 베이스라인 모델을 VideoLLaMA2로 채택하였다.

2.3. 보행자 횡단 의도 예측 방법

본 연구에서는 연속적인 과거 16프레임 이미지에 해당되는 Scene Context(도시 주행 환경), Local Context(보행자의 국소적 행동), Bounding Box Coordinates(이미지 내 보행자의 위치), Speed of Ego-Vehicle(자율주행 차량 속도 정보)과 보행자 횡단 및 데이터를 설명하는 프롬프트를 활용하여 30프레임 이후에서의 보행자 횡단 여부를 예측하고자 한다.

먼저, Visual Encoder(V_E)는 16개의 프레임 이미지에 해당하는 Scene Context(F_s)와 Local Context (F_l)에서 입력된 비디오 시퀀스의 시각적 특징을 추출하고, 각 프레임에서 보행자와 주변 환경의 시각 정보를 학습할 수 있는 형태로 변환한다.

$$F_s^t, F_l^t = V_E(X_s^t, X_l^t)\ ,\ t = 0,1,...,15$$

추출된 시각적 정보는 STC(Spatial-Temporal Convolution) connector에 입력되어, 공간-시간 상호작용을 효과적으로 포착한다.

$$F_{stc}^t = STC(F_s^t, F_l^t)\ ,\ t = 0,1,...,15$$

Bounding Box Coordinates(B)는 이미지 내에서 보행자의 위치를 정확히 파악하는 데 사용되며, 이를 통해 모델은 보행자의 이동 경로와 위치 변화를 분석할 수 있다. Speed of Ego-Vehicle(S)는 차량의 가속, 감속 등 속도 변화를 나타내며, 벡터로 변환된

후 모델에 입력된다.

$$B_{vec}^{t} = f_{bbox}(B^{t}),\ S_{vec}^{t} = f_{speed}(S^{t}),\ t = 0,1,...,15$$

이러한 벡터화된 정보와 시각적 특징을 결합하여 30프레임 이후에서의 보행자 횡단 여부를 예측하도록 한다.

$$P_{cross} = f_{pred}(F_{stc}^{(t)}, B_{vec}^{(t)}, S_{vec}^{(t)}),\ t = 0,1,...,15$$

3. 실험 및 결과

3.1. 평가 지표

본 연구에서 정확도(ACC), AUC, F1-Score(F1), 정밀도(P), 재현율(R)과 같은 평가 지표를 사용하여 성능을 측정한다. 정확도는 모델이 예측한 값이 실제 값과 얼마나 일치하는지를 나타내며, 다음 공식으로 계산된다: $ACC = \frac{TP + TN}{TP + FP + TN + FN}$

정밀도(Precision)는 모델이 횡단으로 예측한 사례 중 실제로 횡단한 보행자의 비율을 나타내며, 다음 공식으로 계산된다: $\mathrm{Precision} = \frac{TP}{TP + FP}$

재현율(Recall)은 실제로 도로를 횡단 보행자 중 모델이 얼마나 많이 정확하게 예측했는지를 나타내며, 다음 공식으로 계산된다: $Recall = \frac{TP}{TP + FN}$

F1-score는 정밀도와 재현율 간의 균형을 평가 하기 위해 사용되며, 이는 두 지표의 조화 평균으로 계산된다: $F1 = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall}$

AUC(Area Under the ROC Curve)는 모델의 횡단 및 비횡단 예측 성능을 이진 분류 문제로 평가하며, ROC 곡선 하의 면적을 나타낸다. x_i는 ROC곡선의 거짓 긍정률(FPR), y_i는 참 긍정률(TPR)을 의미한다. AUC는 모델이 보행자의 횡단 여부를 얼마나 정확하게 판별하는지를 측정하는 지표이다.

$$AUC = \sum_{i}^{n-1} \frac{1}{2}(x_{i+1} - x_i) \times (y_i + y_{i+1})$$

3.2. 정량적 평가

본 실험은 VideoLLaMA2와 JAAD 데이터셋을 사용하여 미세 조정된 VideoLLaMA2 모델의 성능을 비교하고자 한다. <표 2>을 통해 미세조정이 적용된 VideoLLaMA2의 성능이 미세조정을 하지 않은 경우보다 정확도가 2% 상승한 것을 확인할 수 있다. 이외에도 재현율(R) (0.52 → 0.59), F1 Score (0.59 → 0.63), AUC (0.59 → 0.61) 에서 보행자 횡단 예측 성능이 향상되었으며, 이는 미세조정이 모델의 전반적인 성능 향상에 긍정적인 영향을 주었음을 시사한다. 반면, 정밀도(P)에서는 성능 차이가 나타나지 않아, 이 측면에서 미세조정의 영향이 크지 않음을 알 수 있다.

<표 2> Finetuned VideoLLaMA2 성능 비교

Models	ACC	P	R	F1	AUC
VideoLLaMA2	0.58	0.67	0.52	0.59	0.59
Finetuned VideoLLaMA2	0.60	0.67	0.59	0.63	0.61

3.3. Ablation Study

본 실험은 모델이 보행자 횡단 의도를 예측하는데 사용되는 각 입력 특징이 모델 성능에 미치는 영향을 분석하기 위해 진행한다.

<표 3>의 결과에서 확인할 수 있듯이, 모든 특징을 사용한 경우가 가장 높은 성능을 보이며, 정확도는 0.603, F1 Score는 0.632, AUC는 0.605로 나타났다. 이는 모든 특징이 모델의 예측 성능을 향상하는데 중요한 역할을 하는 것을 시사한다. 반면, 일부 특징을 제거했을 때 성능 저하가 발생한다. Local Context를 제거하면 정밀도는 0.686으로 유지되었으나 재현율은 0.473으로 감소, 정확도는 0.572로 감소하여 모델이 실제 횡단하는 보행자를 충분히 인식하지 못하는 문제가 발생한다.

Bounding Box와 Local Context가 모두 제거된 경우, AUC는 0.550으로 가장 크게 감소하였며, 모델의 분류 능력이 크게 저하되었음을 보여주었다. 이는 Bounding Box와 Local Context가 보행자의 의도 예측에서 중요한 역할을 한다는 점을 시사하며, 이들 특징을 제거했을 때 모델이 보행자의 횡단 여부를 정확히 구분하지 못함을 알 수 있다.

모든 특징은 모델의 성능을 결정하는 중요한 요소로 작용하며, 이는 모델 성능을 최적화하는 데 크게 기여한다는 결론을 도출할 수 있다.

4. 결론 및 향후 연구

VideoLLaMA2 모델을 JAAD QA셋으로 미세조정하여 과거 16프레임을 관찰하고 이후 30프레임에서의 보행자의 횡단 의도(crossing/not-crossing)를 예측한 결과를 평가 및 분석하였다. 시각적인 요소와 보행자 위치 및 차량의 속도를 나타내는 텍스트를

<표 3> Ablation Study 성능 비교

Scene Context	Bounding Box	Ego-vehicle speed	Local Context	ACC	P	R	F1	AUC
✓	✓	✓	✓	0.603	0.676	0.593	0.632	0.605
✓	✓	✓	-	0.572	0.686	0.473	0.560	0.590
✓	✓	-	✓	0.581	0.663	0.551	0.601	0.586
✓	-	✓	✓	0.573	0.657	0.537	0.591	0.579
✓	-	✓	-	0.573	0.687	0.473	0.560	0.591
✓	✓	-	-	0.560	0.676	0.451	0.541	0.579
-	✓	✓	✓	0.576	0.611	0.722	0.662	0.550
-	-	-	✓	0.570	0.649	0.548	0.594	0.574

동시에 학습시켜 60%의 정확도로 횡단 의도를 예측하였다. 또한, 모델의 성능에 가장 큰 영향을 미치는 특징을 찾기 위해 특징들을 단계적으로 제거하여 성능 비교를 진행하였다. 향후에는 보행자의 자세 정보 등을 포함한 더 많은 데이터로 모델 성능을 개선하고 실시간 처리 기능을 추가하여 보행자 안정성 향상을 높이기 위해 연구할 계획이다.

사 사

이 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발).

참고문헌

[1] Sanganaikar, R.S., Mulangi, R.H. “Pedestrian Safety Studies on Urban Infrastructure: A Review”, Sustainable Infrastructure: Innovation, Opportunities and Challenges, Singapore, 2024, 183-188.

[2] Namatovu S, Balugaba BE, Muni K, Ningwa A, Nsabagwa L, Oporia F, et al. “Interventions to reduce pedestrian road traffic injuries: A systematic review of randomized controlled trials, cluster randomized controlled trials, interrupted time-series, and controlled before-after studies.”, PLOS ONE, 17, 1, ,DOI: https://doi.org/10.1371/journal.pone.0262681, 2022

[3] Job RFS, “Policies and Interventions to Provide Safety for Pedestrians and Overcome the Systematic Biases Underlying the Failures”, Frontiers in Sustainable Cities, 2, 30, DOI: https://doi.org/10.3389/frsc.2020.00030, 2020

[4] Ham, Je-Seok, et al., “CIPF: Crossing Intention Prediction Network based on Feature Fusion Modules for Improving Pedestrian Safety”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023, pp.3666-3675

[5] Ham, Je-Seok, et al., “MCIP: Multi-Stream Network for Pedestrian Crossing Intention Prediction”, European Conference on Computer Vision, Tel-Aviv, 2022, pp.663-679

[6] Soroori, Emad, et al., "Spatial association between urban neighbourhood characteristics and child pedestrian–motor vehicle collisions.", Applied Spatial Analysis and Policy, 16, 4, 1443-1462, 2023

[7] Adinarayana, Badveeti, and Mohammad Shafi Mir. "Development of pedestrian safety index models for safety of pedestrian flow at un-signalized junctions on urban roads under mixed traffic conditions using MLR." Innovative Infrastructure Solutions, 6, 54, 1-9, 2021

[8] Rasouli, Amir, Iuliia Kotseruba, and John K. Tsotsos. "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior." Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, 2017, 206-213.

[9] Cheng, Zesen, et al. "VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs." arXiv preprint arXiv:2406.07476, 2024.