MISAKA-NETWORK-H: 실시간 긴급 탐지를 위한 Hailo-8 NPU 최적화 비디오 트랜스포머

안용호*, 이성민

숭실대학교, 서경대학교

*nanhosoft@soongsil.ac.kr, zoq2039@skuniv.ac.kr

MISAKA-NETWORK-H: An Edge-Deployable Video Transformer for Real-Time Emergency Detection on Hailo-8 NPU

Ahn Yong Ho*, Lee Seong Min

*Soongsil Univ., Seokyeong Univ.

요 약

본 논문은 Hailo-8 NPU에 최적화된 Transformer 모델을 통해 비디오 기반 긴급 상황 탐지의 실시간성과 효율성을 개선하였다. 제안된 모델은 PyTorch 구현 대비 유사한 정확도와 재현율을 유지하면서도 짧은 추론 시간 내에 엣지 디바이스에서 실행 가능하도록 설계되었으며, 비교적 우수한 성능을 보여 긴급 상황 감지의 신뢰성을 향상시켰다.

1. 서 론

최근 고령화된 사회, 재난 대응, 사회 및 산업현장 안전 등 다양한 응용 분야(예: CCTV 등)는 실시간성이 중요한 환경으로[1], 중앙 서버에 의존하지 않고 현장에서 빠르게 판단 가능한 엣지 AI 기술의 중요성이 보다 강조되고 있다[2]. 한편, CNN, LSTM, Transformer 등의 모델을 단독 혹은 결합하여 사용하는 사례가 우수한 성능을 보이며 주목받고 있으나[3][4], 대부분 고성능 GPU 환경을 전제로 설계되어 있어 전력·연산 자원이 제한된 엣지 디바이스에는 직접 적용하기 어려운 현실이다.

본 논문은 실제 배포를 위한 하드웨어 최적화에 초점을 맞춘 연구로, 현재 자체적으로 연구 중에 있는 시공간 및 모션 특성 기반 처리모델 "MISAKA(Motion Inference with Spatio-temporal Attention for Kinematics Analysis)" 네트워크 구조를 바탕으로 엣지 디바이스에서의 실행을 위해 구조를 단순화하였다.

2.1. 엣지 디바이스에서의 신경망 연산 처리

엣지 디바이스는 통상적으로 서버/하이엔드급 GPU에 비해 VRAM 용량 제한, 특정 명령어 및 동적 연산의 미지원, 파라미터 양자화를 통한 최적화 요구, 전력 및 발열량 제한 등의 제약을 갖는다.

본 논문에서는 엣지 디바이스 중에서 비교적 널리 혼용되며, 저렴한 Hailo-8 NPU를 대상으로 선정하였다. Hailo-8 NPU는 최대 26TOPs의 성능을 발휘하며, 대부분의 CNN 관련 연산자를 지원한다. 하지만 현재까지도 복잡한 텐서 연산은 지원하지 않거나, 제한적으로만 지원하고 있다.[5]

본 논문에서는 Hailo-8의 연산자 호환성 및 추론 파이프라인을 고려하여, 미지원 연산자(예: Transpose, LayerNorm, Expand 등)를 제거하고 모든 모듈을 고정 연산 기반 구조로 재설계하였다. 이를 통해 모델 전체를 Hailo 상에서 컴파일 및 추론 가능한 형태로 최적화하고, 실제 엣지 디바이스 환경에서 실시간 동작이 가능한 경량 구조를 구현하였다.

2.2. MISAKA-Network-H 모델 설계

MISAKA-Network-H는 실시간 영상 기반 긴급 탐지를 위해 설계된 시공간 인식 파이프라인이다. 그림 1은 전체 모델 파이프라인 구조를 나타낸다. 영상 시퀀스 $[B, T, C, H, W]$는 YOLOv11 기반 백본을 통해 사람 객체의 바운딩 박스와 특징 벡터를 추출하고, 크롭된 사람 이미지를 시계열에 따라 CNN 기반 모션 인코더에 주입한다. 모션 인코더가 시계열에 따른 움직임 정보를 요약하고, 이 정보는 앞서 추출된 YOLOv11의 특징 벡터와 게이트 방식으로 융합된다. 이후 4D 입력을 처리할 수 있도록 구성된 Transformer 모델을 통해 최종적으로 시공간 패턴을 인식한다.

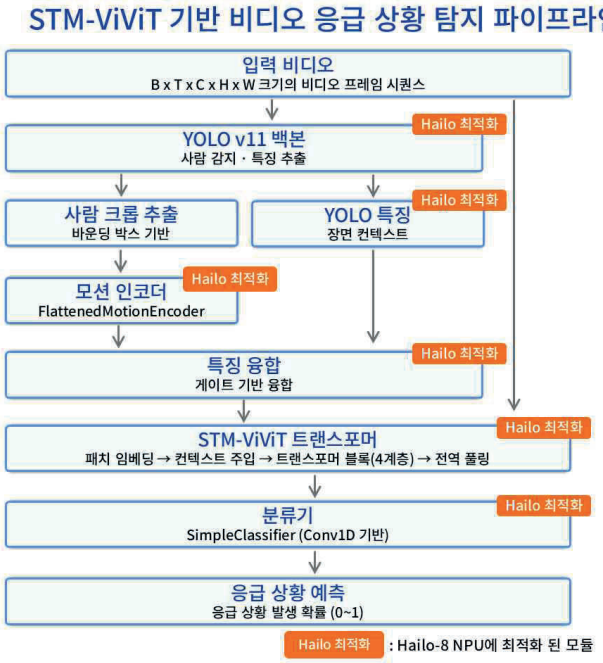

그림 1 MISAKA-Network-H 모델 전체 구조

3. MISAKA-Network-H 모델의 구현 및 변환

MISAKA-Network-H는 앞서 제시한 모델 구조에서 ConvLSTM, LayerNorm, Transpose, Expand 등 Hailo-8 하드웨어가 지원하지 않거나 변환 불가능한 구조의 연산자들을 제거하고, 단순 CNN 기반 시계열 요약 모델 및 BatchNorm 기반 트랜스포머 블록으로 대체함으로써 모든 연산자가 Hailo 변환에 적합하도록 변형되었다. 보다 자세히는, 모션 인코더는 각 프레임에 대해 2-layer CNN과 FC Layer로 특징을 요약하며, 이 특징을 YOLO 모델에서 추출된 특징과 함께 게이트 방식으로 융합된 후, 4D 구조의 Transformer를 통해 시공간 특징을 인식하고 Sigmoid 기반 분류기를 거쳐 긴급 여부를 예측한다. 이 과정에서 고차원 텐서 연산이 필요한 Motion Encoder 출력과 YOLO Feature Map의 Gated Fusion 처리는 호스트 CPU에서 처리되도록 파이프라인을 구성하였다. 최종 출력은 GlobalAveragePooling 기반 분류기를 통해 영상 전체 프레임에 대한 긴급 여부 확률로 예측한다.

4.1. MISAKA-Network-H 학습 및 실험 설계

MISAKA-Network-H의 성능 평가를 위해 총 여섯 개의 낙상 관련 데이터셋 - 공개 데이터셋 5종(GMDCSA24, Le2i Fall, CAUCAFall, Cutup and Detect, UR Fall Detection)[6][7][8][9][10]과 자체 제작 데이터셋 - 을 활용했다. 이 중 CCTV 환경과 유사한 데이터만 선별하여 5 FPS, 총 20프레임으로 표준화하고, YOLOv11로 인물 검출 후 112×112 크기로 크롭했다. 최종적으로 1,218개 영상으로 학습, 405개로 검증, 191개로 테스트를 진행했다. 모델 학습은 BCEWithLogitsLoss와 AdamW를 사용했으며, Hailo-8 SDK로 HEF 컴파일 후 추론 속도(FPS), 평균 지연시간(ms), OPS를 측정했다. 성능 평가는 정확도, 정밀도, 재현율, F1 점수 등을 기준으로 하였으며, 전체 모델은 실제 Hailo-8 NPU를 탑재한 x86 시스템 상에서 실시간으로 구동하였다.

4.2. MISAKA-Network-H 실험 결과

본 논문에서는 응급 상황 감지를 위한 딥러닝 모델의 성능을 평가하고, 임베디드 환경 배포를 위한 모델 최적화 효과를 분석하였다. 원본 PyTorch 모델과 Hailo 플랫폼으로 최적화된 모델의 성능을 비교한 결과는 <표 1>과 같다.

모델	Accuracy	Precision	Recall	F1 Score	Specificity
PyTorch	0.7481	0.7234	0.7312	0.7273	0.7626
Hailo	0.7644	0.8037	0.7818	0.7926	0.7407

표 1 PyTorch baseline 모델과 Hailo 변환 모델 성능 측정표
*PyTorch 모델의 Specificity는 혼동 행렬 [[167 52] [50 136]]에서 계산됨
**Hailo 모델의 Specificity는 혼동 행렬 [[60, 21], [24, 86]]에서 계산됨

변환된 모델은 보다 정제된 테스트셋을 사용하였음을 감안하더라도 전반적인 정확도가 76%로 높았으며, 특히 재현율(Recall)이 78%인 점은 주목할 만하다. F1 점수 역시 79% 수준으로, 정밀도와 재현율 간 더 나은 균형을 달성했다. 이러한 점에서 본 모델은 응급 상황 감지 시스템에서 중요한 실제 응급 상황의 누락을 최소화하는 데 기여한다.

최적화된 모델의 추론 성능 또한 주목할 만하다. 인코더 단계에서 0.021초, 트랜스포머 단계에서 0.066초로 빠른 추론 시간을 보였으며, 전체 파이프라인은 파일당 평균 4.55초의 처리 시간을 기록했다. 입력 영상 파일을 비교적 레이턴시가 높은 USB 드라이브에서 로드했음을 감안할 때, 실제로는 더 빠른 처리 속도를 기대할 수 있다. 이러한 결과는 본 연구가 실시간 응급 상황 모니터링 시스템에 적합한 수준의 성능을 달성했음을 나타낸다.

5. 결론

본 논문에서는 Hailo-8 NPU를 활용한 경량 Transformer 모델을 설계하여 영상 기반 긴급 상황 탐지의 실시간성 및 에너지 효율성을 개선하고자 하였다. 최적화된 모델은 경량화에도 불구하고 응급 상황 식별 능력이 유지되었으며 파일당 평균 4.55초의 처리 시간을 기록하여 엣지 디바이스 상에서 실시간 응급 상황 모니터링 시스템에 적합한 수준의 성능을 달성하였다.

향후 연구에서는 실제 임베디드 시스템에서의 현장 평가, 사례 분석을 통한 개선, 그리고 양자화 관련 실험 및 후처리 기법에 대한 추가 연구가 필요하다. 본 연구는 하드웨어 제약이 있는 환경에서 Transformer 기반 모델의 실용성을 입증하며, 엣지 디바이스를 활용한 실시간 응급 상황 감지 시스템의 상용화 가능성을 높일 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 한이음 드림업 프로젝트 결과물입니다.

참 고 문 헌

[1] 박용석(Yong-Suk Park), 김현식(Hyun-Sik Kim), 최규만(Kyuman Choi), "영상기반 독거노인 낙상감지 시스템의 구현," 한국통신학회 학술대회논문집, vol. 2019.11, pp. 304-305, Nov. 2019.

[2] B. Charyyev, E. Arslan, and M. H. Gunes, "Latency Comparison of Cloud Datacenters and Edge Servers," in IEEE Global Communications Conference (GLOBECOM), 2020, pp. 1-6.

[3] 정승수(Seung Su Jeong), 김남호(Nam Ho Kim), 유윤섭(Yun Seop Yu), "Yolo-pose를 이용한 장단기 메모리의 낙상감지 시스템 연구," 한국정보통신학회 종합학술대회 논문집, vol. 26.2, pp. 123-125, 2022.

[4] Y. Sanjalawe, S. Fraihat, M. Abualhaj, S. R. Al-E'Mari, and E. Alzubi, "Hybrid Deep Learning for Human Fall Detection: A Synergistic Approach Using YOLOv8 and Time-Space Transformers," IEEE Access, vol. 13, pp. 41336-41366, 2025, doi: 10.1109/ACCESS.2025.3547914.

[5] M. N. Achmadiah, N. Setyawan, A. A. Bryantono, C.-C. Sun, and W.-K. Kuo, "Fast Person Detection Using YOLOX With AI Accelerator For Train Station Safety," in Proc. 2024 International Electronics Symposium (IES), Denpasar, Indonesia, 2024, pp. 504-509, doi: 10.1109/IES63037.2024.10665874.

[6] E. Alam, A. Sufian, P. Dutta, M. Leo, and I. A. Hameed, "GMDCSA24: A Dataset for Human Fall Detection in Videos," Zenodo, 2024. [Online]. Available: https://doi.org/10.5281/zenodo.12921216

[7] I. Charfi, J. Mitran, J. Dubois, M. Atri, and R. Tourki, "Optimised spatio-temporal descriptors for real-time fall detection: comparison of SVM and Adaboost based classification," Journal of Electronic Imaging (JEI), vol. 22, no. 4, pp. 17, Oct. 2013.

[8] Jose Camilo Eraso, Elena Muñoz, Mariela Muñoz, and Jesus Pinto, "Dataset CAUCAFall," Mendeley Data, V4, 2022. doi: 10.17632/7w7fccy7ky.4

[9] T. Grutschus, O. Karrar, E. Esenov, and E. Vats, "Cutup and Detect: Human Fall Detection on Cutup Untrimmed Videos Using a Large Foundational Video Understanding Model," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2401.16280

[10] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," Computer Methods and Programs in Biomedicine, vol. 117, no. 3, pp. 489-501, Dec. 2014. doi: 10.1016/j.cmpb.2014.09.003