딥러닝 양자화 기술 기반 보행자 안전을 위한 도로 및 인도 영상 분할 기술

황종철
*큐랩
e-mail:tory0405@cuelab.kr

Road and Sidewalk Image Segmentation Technology for Pedestrian Safety Based on Deep Learning Quantization Technology

Jong-Cheol Hwang*
*CueLab Inc.

요 약

본 논문에서는 차량 지능형 시스템의 보행자 안전 확보를 위해 딥러닝 양자화 기반의 경량화 영상 분할 기법을 제안하며, 정확도와 연산 효율을 동시에 만족하는 도로 및 인도 영상 분할 시스템을 제안한다. 이를 위해 Intel에서 제공하는 road-segmentation-adas-0001모델을 양자화 버전으로 변환하여 비교 분석하였다. 원본 FP32 모델과 이를 양자화한 FP16 및 INT8 모델의 성능을 모델 크기, 그리고 추론 속도(FPS) 측면에서 평가하였다. 표준 CPU 환경에서의 실험 결과, INT8 양자화 모델이 원본 모델 대비 정확도 저하를 최소화하면서도 모델 크기와 추론 시간을 크게 단축하여 성능과 효율성 간의 가장 이상적인 균형점을 확인하였다.

1. 서론

지능형 자동차 기술의 발전과 함께, 카메라를 이용한 주변 환경 인지는 보행자 안전을 위한 핵심 요소로 자리 잡았다. 특히 주행 가능한 도로와 보행자 영역인 인도를 정확히 구분하는 것은 안전한 경로 계획의 foundational step이다 [7]. 초기 컴퓨터 비전 기술은 특정 조건에서는 유효했으나, 도심지 환경의 그림자, 차량으로 인한 가려짐 등 복잡한 외부 요인으로 인해 정확도가 저하되는 한계가 있었다 [5]. 이러한 한계를 극복하기 위해 딥러닝 기반의 영상 분할(Image Segmentation) 기술이 활발히 도입되고 있으나, 높은 정확도를 가진 모델일수록 많은 연산량을 요구한다 [3, 6]. 이는 연산 능력, 메모리, 전력이 제한된 차량용 임베디드 시스템에 고성능 모델을 직접 탑재하기 어렵게 만드는 주요 원인이다[6]. 이러한 문제를 해결하기 위한 핵심 기술로 딥러닝 모델 경량화가 주목받고 있으며, 그중 양자화(Quantization)는 모델의 재학습 없이도 크기와 연산량을 획기적으로 줄일 수 있는 강력한 기법이다 [3, 9].

본 논문은 딥러닝 양자화 기술을 도로 분할 모델에 적용하는 것에 초점을 맞춘다. Intel에서 제공하는 road-segmentation-adas-0001 FP32 모델을 FP16, INT8로 양자화 변환한 후 정밀도로 각각 비교 분석하고, 객체 탐지를 위해 경량 모델인 YOLOv8n을 통합하여 CPU 환경에서의 종합적인 실시간 영상 분할 시스템의 효율성을 평가하고자 한다.

2. 관련 연구

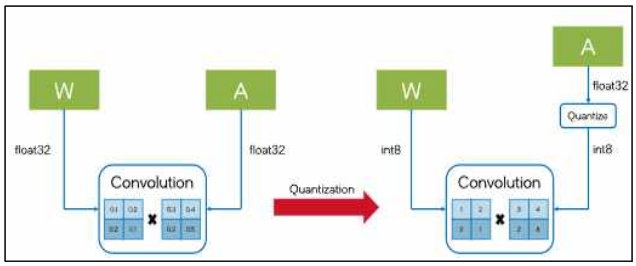
2.1. Semantic Segmentation

Semantic Segmentation은 이미지의 모든 픽셀에 의미론적 레이블을 할당하여, 주행 가능한 도로 영역과 보행자가 이용하는 인도 영역을 정밀하게 구분하는 데 사용된다 [4].

road-segmentation-adas-0001 모델은 주행 영상의 각 픽셀을 배경, 도로, 연석, 차선 표시의 네 가지 클래스로 분류하는 데 특화되어 있으며, DeepLabv3+와 유사한 효율적인 Encoder-Decoder 구조를 가진다 [1, 9].

2.2. 모델 양자화 기술

양자화(Quantization)는 모델의 가중치를 표현하는 데이터 타입을 낮은 비트의 정밀도로 변환하는 기술이다. 예를 들어, 32비트 부동소수점(FP32)을 8비트 정수(INT8)로 변환하면, 모델의 크기는 이론적으로 4배 감소하고 추론 처리량은 2배에서 4배까지 향상될 수 있다. 이러한 최적화는 일반적으로 1% 미만의 미미한 정확도 손실을 동반한다 [9]. 본 논문에서는 [그림 1]과 같이 FP32 파라미터를 저 정밀도로 매핑하는 후처리 양자화(PTQ) 방식을 적용했다.



[그림 1] 모델 양자화 개념도 (FP32 to INT8)

2.3. 제안 시스템 구성

제안하는 시스템은 비디오 프레임 입력, 양자화된 모델을 이용한 도로 분할, YOLOv8n을 이용한 객체 탐지, 그리고 결과 시각화의 4단계로 구성된 순차적 파이프라인 구조를 가진다. 도로 분할에는 Intel의 road-segmentation-adas-0001 FP32 모델과 양자화로 변환한 FP16, INT8 버전을 사용하였으며, 객체 탐지에는 YOLOv8n 모델을 통합하여 도로 상황에 대한 포괄적인 인지를 수행하도록 설계했다 [2].

3. 구현 및 실험 결과

3.1 실험 환경

양자화 모델의 성능 평가는 [표 1]에 명시된 일반 PC 환경에서 원본 이미지(4000x2252)를 (720x480)으로 축소하여 수행되었다. GPU를 사용하지 않고 오직 CPU 연산만으로 성능을 측정하여, 저비용 임베디드 시스템에서의 동작 가능성을 평가하였다.

[표 1] 실험 환경

구분	사양
CPU	Intel(R) Core(TM) i5-12400F
RAM	64GB
OS	Window 11
SW	Python 3.12, OpenCV 4.10, OpenVINO 2024.5.0

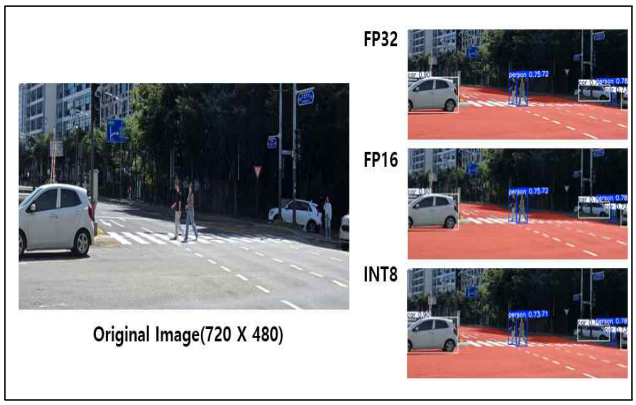
3.2 양자화 모델 성능 평가

road-segmentation-adas-0001 모델의 정밀도에 따른 성능 변화를 측정한 결과는 [표 2]와 같다. 이 표는 각 모델의 크기, 그리고 CPU에서의 분할 작업 평균 처리 속도(FPS)를 비교한다.

[표 2] 도로 분할 모델의 양자화 수준에 따른 성능 비교

양자화	모델 크기(MB)	분할 속도(FPS)
INT8	0.88	30.40
FP16	1.08	21.95
FP32	1.21	22.99

실험 결과, INT8 모델은 원본 FP32 모델 대비 모델 크기는 약 27.3% 감소하고 처리 속도는 약 32.2% 향상되어, 양자화가 성능 최적화에 매우 효과적임을 보여주었다. 이는 최신 CPU가 INT8 연산에 대한 하드웨어 가속을 지원하기 때문이다. 반면, FP16 모델은 모델 크기는 약 10.7% 감소했지만, 처리 속도는 오히려 소폭 하락했다. 이는 실험에 사용된 CPU가 FP16 연산을 네이티브로 지원하지 않아 내부적으로 FP16로 처리하는 과정에서 추가 오버헤드가 발생했기 때문으로 분석된다. [그림 2]는 각 모델의 도로 분할 결과를 시각적으로 비교한 것이다.



[그림 2] 양자화 모델의 도로 분할 결과 비교 (FP32, FP16 vs. INT8)

4. 결론

본 논문에서는 딥러닝 양자화 기술을 중심으로 보행자 안전을 위한 실시간 도로 및 인도 영상 분할 시스템을 제안하고 구현하였다. 도로 분할을 위해 road-segmentation-adas- 0001 FP32 모델과 양자화 모델 FP16, INT8 버전을 비교 분석하여 모델 양자화가 시스템 효율성에 미치는 영향을 정량적으로 평가했다.

실험 결과, INT8 양자화가 원본 모델의 정확도를 거의 유지하면서도 모델의 크기와 추론 시간을 대폭 개선하여 실제 차량 환경에 적용하기에 가장 적합한 균형점을 제공한다는 결론을 도출하였다. 이는 후처리 양자화(PTQ)가 복잡한 재학습 과정 없이도 딥러닝 모델을 효과적으로 경량화할 수 있는 실용적이고 강력한 기술임을 보여준다.

향후, INT8 정밀도에서의 정확도 저하를 완화하기 위해 소량의 데이터로 모델을 미세 조정하는 QAT(Quantization Aware Training) 기법의 적용 가능성을 탐색하고 [9], 다양한 도로 및 기상 조건에 대한 모델의 강건성을 평가할 계획이다.

참고문헌

[1] 박세진, 한정훈, 문영식, "효율적인 비정형 도로영역 인식을 위한 Semantic segmentation 기반 심층 신경망 구조", 한국정보통신학회논문지, 제 24권 11호, pp. 1437-1444, 11월, 2020년.

[2] 서정희, "YOLOv8 알고리즘 기반의 주행 가능한 도로 영역 인식과 실시간 추적 기법에 관한 연구", 한국전자통신학회논문지, 제 19권 3호, pp. 563-570, 6월, 2024년.

[3] 김은희, 이경하, 성원경, "딥 러닝 모델의 경량화 기술 동향", 정보과학회지, 2020년 8월.

[4] B. Sophia and D. Chitra, "Segmentation Based Real Time Anomaly Detection and Tracking Model for

Pedestrian Walkways," Intelligent Automation & Soft Computing, vol. 36, no. 3, pp. 2491-2507, 2023.
[5] 박승준, 한상용, 박상배, 김정하, "자율주행을 위한 딥러닝 기반의 차선 검출 방법에 관한 연구", 한국산업융합학회 논문집, 제 23권 6호, pp. 979-987, 2020년.
[6] 이용주, 문용혁, 박준용, 민옥기, "경량 딥러닝 기술 동향", Electronics and Telecommunications Trends, 2019년.
[7] G. Weld, E. Jang, A. Li, A. Zeng, K. Heimerl, and J. E. Froehlich, "Deep Learning for Automatically Detecting Sidewalk Accessibility Problems Using Streetscape Imagery," In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19), 2019.
[8] 전효진, 조수선, "딥러닝 기반의 주행가능 영역 추출 모델에 관한 연구", 한국인터넷정보학회논문지, 제 20권 5호, pp. 105-111, 10월, 2019년.
[9] OpenVINO NNCF를 활용한 road-segmentation-adas-0001 모델의 INT8 후처리 양자화 종합 분석, 2024년.