라즈베리 파이 5와 Edge TPU 환경에서 전이 학습을 활용한 맞춤형 YOLOv9 모델의 실시간 영상 객체 분할 구현에 관한 연구

박승민[1], 서장원[2]
[1]동서울대학교 컴퓨터소프트웨어학과 학부생
[2]동서울대학교 컴퓨터소프트웨어학과 교수
7195sm@naver.com, jwsuh@du.ac.kr

Research on the implementation of real-time video instance segmentation of customized YOLOv9 using transfer learning in a Raspberry Pi 5 and Edge TPU Accelerator Environment

Seung-Min Park[1]
Jang-Won Suh[1]
[1]Dept. of Computer software, Dong seoul University

요 약

본 연구는 YOLOv9의 세그멘테이션 전용 모델인 YOLOv9c-seg을 교육용 임베디드 시스템인 라즈베리 파이 5와 Google Coral Edge TPU 환경에서 실시간 객체 분할 성능을 평가하였다. YOLOv9-seg모델을 커스텀 데이터셋(Customized Dataset)으로 파인튜닝(Full Fine Tuning)하여 TF Edge TPU 포맷으로 변환하여 추론 속도와 메모리 사용량을 크게 개선하였다. 실험 결과, 변환된 모델은 PT(Pytorch)형식의 기존 모델과 유사한 성능을 유지하면서도 평균 추론 시간이 80.14ms 단축되고, FPS가 21.12프레임 증가하여 실시간 성능이 향상되었다. 이는 Edge TPU가 정수 양자화된 모델에 최적화되어 처리 속도와 효율성을 극대화할 수 있음을 보여준다. 본 연구는 엣지 컴퓨팅(Edge Computing) 환경에서 실시간 객체 분할의 가능성을 제시하였다.

1. 서론

라즈베리 파이 5(Raspberry Pi 5)와 TPU 가속기를 활용한 객체 인식 및 분할 기술은 저전력 고성능의 엣지 컴퓨팅(Edge Computing) 환경에서 큰 가능성을 보여준다. 특히, YOLO(You Only Look Once) 시리즈는 실시간 객체 탐지 분야에서 높은 정확도와 빠른 속도로 주목받아 왔으며, 최신 버전인 YOLOv9은 이전 모델에 비해 성능이 향상되어 다양한 응용 분야에 활용될 수 있다. 본 연구는 YOLOv9 모델을 라즈베리 파이 5와 TPU 가속기 환경에 최적화하여 맞춤형 Data 분할 모델을 구현하는 것을 목표로 한다. 이를 통해 엣지 컴퓨팅 환경에서도 높은 정확도와 성능을 유지하면서 실시간으로 동작 가능한 객체 분할 모델의 가능성을 탐색하고자 한다.

2. YOLOv9(You Only Look Once version 9)

YOLOv9은 기존 YOLO 시리즈의 한계를 극복하며 실시간 객체 탐지 성능을 크게 향상시킨 모델로써 Utralytics사에서 개발되었다. YOLOv9 모델은 정보 손실 문제를 해결하기 위해 PGI(Programmable Gradient Information)와 GELAN(Generalized Efficient Layer Aggregation Network)을 도입하여, 경량 모델에서도 학습 신뢰성을 높이고 최적의 성능을 발휘하도록 설계되었다. PGI는 네트워크의 깊은 층에서 발생할 수 있는 정보 손실을 방지하여 모델 학습의 신뢰성을 높이는 기술로, 보조 가역 함수를 활용하여 그래디언트(Gradient) 흐름을 안정적으로 유지함으로써 특히 경량 모델에서도 우수한 학습 성능을 보장한다. GELAN은 다양한 크기의 모델에서 최적의 성능을 발휘할 수 있도록 설계된 효율적인 계층 집계 네트워크로, 최소한의 연산량으로 높은 정확도를 유지한다. YOLOv9는 이전 버전보다 최대 15% 적은 파라미터와 25% 적은 연산량으로도 우수한 성능을 제공하며, 엣지 디바이스 등 다양한 환경에서 활용될 수 있는 유연성과 효율성을 갖추고 있다[1].

3. 맞춤형 데이터셋 전이학습과 tflite로 변환

실험에 사용된 모델은 YOLOv9의 세그멘테이션 전용 버전인 YOLOv9c-seg로, 나비 이미지(Butterfly image)와 세그멘테이션 데이터셋 300장을 활용하여 300 에포크(epoch) 동안 전이학습(Transfer Learning)을 진행하였다. 학습이 완료된 후, Google Coral Edge TPU와 같은 하드웨어 가속기에서 최적의 성능을 발휘할 수 있도록 모델의 파일 형식을 기존의 pt(PyTorch) 형식에서 tflite(Tensorflow Lite)형식으로 변환하였다. 이 변환 과정에서는 모델의 모든 연산이 정수화(양자화)되어, 추론 속도를 크게 향상시키고 메모리 사용량을 줄이는 효과를 제공한다. 특히, Edge TPU는 정수 양자화된 모델에 최적화되어 있어, 처리 속도와 효율성을 극대화할 수 있다.[2] 모델의 추론 성능을 나타내는 지표인 mAP(mean Average Precision)50-95는 IoU(Intersection over Union) 임계값이 0.5에서 0.95까지 0.05 단위로 증가하는 총 10개의 임계값에서의 평균 정밀도를 계산한 것으로, 기존 0.907에서 0.869로 0.038만큼 하락하였다.

4. 엣지 컴퓨팅 환경에서의 구현

<표 1>과 <표 2>는 실험에 사용된 라즈베리 파이 5와 Google Coral USB Accelerator의 사양을 나타낸다.

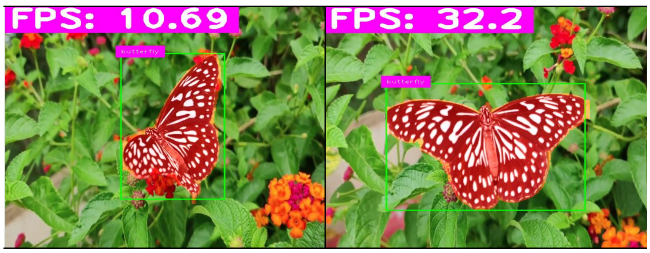
<표 1> 라즈베리 파이 5 사양

H/W	CPU	2.4GHz ARM Cortex-A76 MP4
	GPU	Broadcom VideoCore VII MP12 800 MHz
	메모리	8 GB LPDDR4X-4266 SDRAM
S/W	OS	Debian GNU/Linux 12
	Library	Python == 3.9.12 opencv-python == 4.8.1.78 torch == 2.0.1 torchvision == 0.15.2 edge-tpu-silva == 1.0.5

<표 2> Google Coral USB Accelerator의 사양

H/W	Google Edge TPU coprocessor 4 TOPS (int8)
S/W	Python == 3.9.12 edge-tpu-silva == 1.0.5 libedgetpu == 2.15.0 Pycoral == 2.0.0

동일한 환경에서 TF Edge TPU 포맷으로 변환된 모델은 추론 성능 면에서는 변환 전 모델과 거의 차이가 없었으나, 추론 시간과 추론 결과를 보여주는 영상의 FPS(초당 프레임)에서는 유의미한 차이를 보였다. 200번의 추론을 수행한 결과, 평균 추론 시간은 기존 118.98ms에서 38.84ms로 80.14ms 단축되었고, 평균 FPS는 기존 10.71프레임에서 31.83프레임으로 21.12프레임 증가하였다.



(a) PyTorch (기존) (b) TF Edge TPU

(그림 1) 모델 포맷별 추론 결과.

<표 2> 모델 포맷별 성능 비교

파일형식	mAP 50-95	평균 추론 시간(ms)	평균 FPS (frame)
PyTorch (CPU)	0.907	118.98	10.71
TensorFlow Lite (TPU)	0.869	38.84	31.83

5. 결론

본 논문에서는 맞춤형 데이터 셋인 나비 이미지 데이터셋을 이용하여 YOLOv9c 세그멘트(Segment)모델을 전이학습을 진행하였고 On-Device AI 환경에서 동작하기 위해 파일 형식을 pt 파일 형식에서 tflite 형식으로 변환을 진행하였다. 실험 결과, TF Edge TPU 포맷으로 변환된 모델은 추론 성능에서 약간의 차이가 있었지만, 평균 추론 시간은 기존 118.98ms에서 38.84ms로 단축되었으며, FPS는 10.71에서 31.83으로 증가하였다. 이는 Edge TPU가 정수 양자화된 모델에 최적화되어 있어, 처리 속도와 에너지 효율성을 극대화할 수 있음을 알 수 있다. 본 연구에서는 엣지 컴퓨팅 환경에서도 실시간 객체 분할이 가능한 모델의 구현 가능성을 제시하였다. 향후 연구에서는 YOLOv9 뿐만 아니라 다른 다양한 모델들을 특히 SemiVL 모델을 Edge TPU를 활용하여 On-Device 환경에서 구현하고자 한다.

참고문헌

[1] C.-Y. Wang et al., "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," arXiv preprint, arXiv:2402.13616, 2024

[2] Yipeng Sun, Andreas M Kist, “Deep Learning on Edge TPUs” arXiv preprint, arXiv: 2108.13732, 2021