임베디드 환경에서 효율적인 동작을 위한 객체검출 모델 변환 및 경량화

*최인규 **송혁
한국전자기술연구원
*cig2982@keti.re.kr **hsong@keti.re.kr

Object detection model conversion and weight reduction for efficient operation in embedded environment

*Choi, In-Kyu **Song, Hyuk
Korea Electronics Technology Institute

요약

최근에는 우수한 성능의 딥러닝 기술을 활용한 장비와 프로그램이 개발되고 있으나 기술의 특성상 모든 환경에서 우수한 성능을 보여주지 못하고 고 사양의 서버와 같은 환경에서의 성능만을 보장하고 있다. 따라서 이에 대한 개선으로 엣지 디바이스 독립적으로 혹은 클라우드 의존과 인터넷 연결을 최소화 할 수 있는 엣지 컴퓨팅 기술이 제안되고 있으며 경량 내장형 시스템에 적합한 인공지능 기술의 개발이 필요하다. 본 논문에서는 객체검출 모델을 적은 연산과 효율적인 구조로 설계하고 생성된 모델을 임베디드 보드에서 원활하게 실행할 수 있도록 중립 모델로 변환하고 경량화 하는 방법에 대해 소개한다. Qualcomm snapdragon 프로세서가 갖춰진 임베디드 보드를 목표로 하였고 편의를 위해 SNPE(snapdragon neural processing engine) SDK를 이용하여 실험을 진행하였다. 실험 결과 변환된 중립모델이 기존 모델과 비교하여 압축된 모델 크기 대비 미미한 성능 저하가 발생함을 확인할 수 있었다.

1. 서론

현재 영상분석 기술은 영상의 정보를 분석하여 자동으로 이상 행위를 탐지하고 그 정보를 전송하는 기술로서 사전에 사고를 예방하고 사고가 발생한 경우에는 신속하게 대응하여 피해를 줄일 수 있게 해준다. 초기에 잦은 오경보로 인해 영상보안 현장에서 외면되었지만 각종 사회적 안전에 이슈가 제기되면서 해결책으로 다시 수요가 증가하고 있으며 다양한 객체검출 및 추적 알고리즘 개발과 카메라 성능의 발전에 따라 영상분석 기술을 이용한 지능형 영상보안 시스템이 개발되고 있다. 다양한 분야에서 인공지능 기술이 혁신적 변화를 가져오고 있으나 대부분 고사양의 서버나 클라우드에 의존하고 있다. 인공지능 기술의 상용화를 위해서는 클라우드의 도움 없이 엣지 디바이스 독립적으로 처리하는 엣지 컴퓨팅 기술이 제안되고 있으며 경량 내장형시스템에 적합한 내장형 인공지능 기술의 개발이 필요하다.

본 논문에서는 간소화된 객체검출 모델을 설계 및 생성하고 이를 임베디드 환경에서 효율적으로 실행시키기 위한 변환 및 경량화에 대해 소개한다. 객체검출 모델을 실제로 삽입해야 하는 qualcomm snapdragon 프로세서 환경에서의 실행을 목표로 하였으며 SNPE SDK를 이용하여 모델 변환 및 양자화를 진행하였다. SNPE에서 기본적으로 지원하는 모델이 아닌 사용자가 새로 정의한 모델을 변환하도록 하였으며 YOLOv3,4[1,2]를 간소화한 구조의 객체검출 모델을 설계하였다. 생성된 모델을 경량화 중립 모델 변환 후 SDK에서 실행시켜 검출 결과를 획득하여 딥러닝 프레임워크로 학습한 기존 모델과 객관적 성능을 비교하였으며 압축된 모델 크기 대비 적은 정확도 저하가 발생함을 확인하였다.

2. 객체검출 모델

그림 1과 같이 YOLOv3,4를 참고하여 크기에 강인한 객체검출을 위해 세 개의 특징지도에서 검출 후보 영역을 추출하였으며 short cut을 이용하여 원본 영상의 context 정보를 유지할 수 있도록 하였다. 또한 K-mean clustering[3]을 이용하여 학습 데이터에 유리한 초기 anchor 크기를 설정하여 보다 정확한 검출 bounding box를 획득하도록 하였다.

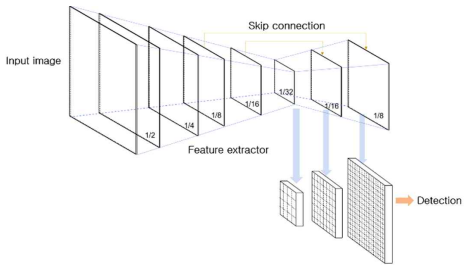


그림 1. 객체검출 구조

3. SNPE SDK를 이용한 모델 변환 및 경량화

아래 그림 2와 같이 학습은 서버에서 기존의 딥러닝 프레임워크를 이용하여 진행하고 임베디드 보드에서 추론을 위해 SNPE SDK를 이용하여 중립 모델(.dlc)로 변환 및 경량화 한다. SNPE는 CPU, GPU, DSP 등 다양한 프로세서에서 딥러닝 모델을 실행가능하고 Caffe, ONNX, Tensorflow 등 다양한 프레임워크의 모델을 변환가능하다. 또한 중립 모델을 8비트 고정 소수점의 양자화 할 수 있다.

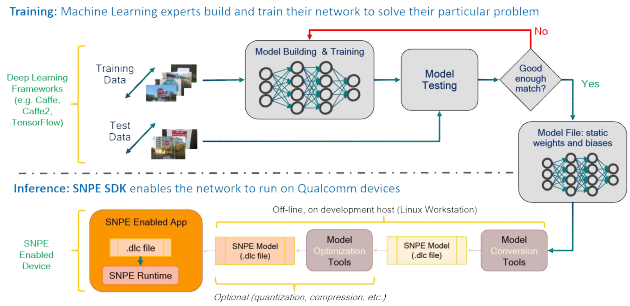


그림 2. Model workflow

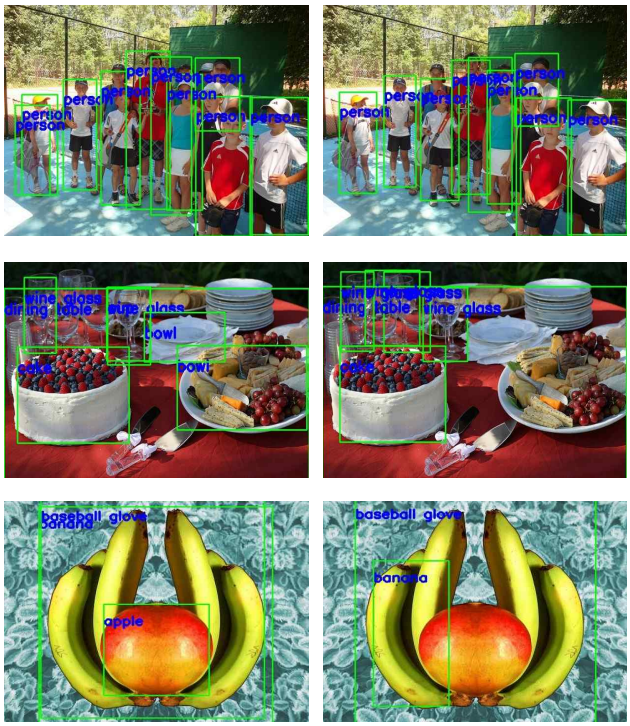
4. 실험 결과

우분투 18.04, tensorflow 1.15의 환경에서 설치한 SNPE SDK 1.52 버전을 이용하였으며 tensorflow를 이용하여 YOLOv3,4를 기반으로 변형한 두 가지의 객체검출 모델을 생성하였다. 각각 중립 모델로 변환 및 양자화 하였으며 COCO validation 데이터를 이용하여 객관적 성능 및 모델 크기를 측정하였다. 표 1은 YOLOv3,4 기반의 변형 모델의 tensorflow와 DLC 버전에 대해 mAP 및 모델크기를 측정한 결과이다. YOLOv3,4 두 가지 경우에 대해서 모델 크기가 4배가량 압축된 데에 비해서 객관적 성능 지표인 mAP는 약 2.5%로 미세하게 감소함을 확인할 수 있다.

	Yolov3-based		Yolov4-based	
	TF	DLC	TF	DLC
mAP (%)	36.57	34.05	38.14	35.49
Model size (MB)	35.5	9.0	24.3	6.2

표 1. 객체검출 모델의 객관적 성능 및 모델 크기

그림 3은 Yolov3 변형모델의 tensorflow와 DLC을 검출 결과를 가시화한 결과로써 DLC의 검출 결과에서 미 검출 대상이 보이지만 유사한 결과를 보여줌을 확인할 수 있다.



Tensorflow DLC

그림 3. Tensorflow와 DLC를 이용한 객체 검출 결과 비교

5. 결론

본 논문에서는 딥러닝 기반의 객체검출 모델을 qualcomm snapdragon 프로세서 임베디드 환경에서 원활하게 동작하도록 중립모델로 변환 및 경량화한 방법에 대해 소개한다. 기존의 객체검출 모델을 개선하여 설계하고 SNPE SDK를 이용하여 생성 모델을 임베디드 보드에서 활용할 수 있도록 중립 모델로 변환 및 양자화를 진행하였다. 실험을 통해 변환한 중립모델이 압축 대비 미미한 성능저하가 있음을 확인할 수 있었다. 추후 연구에서 해당 임베디드 보드에 삽입하여 수행 속도 및 검출 결과를 분석할 예정이다.

Acknowledgement

이 논문은 2022년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [S2977538].

참고문헌

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[2] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020

[3] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, Vol. 63, No. 2, pp.503-527, 2007.