CNN 모델의 경량화 및 On-Device 추론 가속

안재준[1], 이민서[2], 홍석인[3]
[1]성균관대학교 전자전기공학부 학부생
[2]성균관대학교 바이오메카트로닉스학과 학부생
[3]성균관대학교 반도체시스템공학과 교수

ajj8061@naver.com, yleemm@g.skku.edu, seokin@skku.edu

CNN Model Compression and On-Device Inference Acceleration

An Jae Jun[1], Lee Min Seo[2], Hong Seok In[3]
[1]School of Electronic and Electrical Engineering, Sung-Kyun-Kwan University
[2]Department of Biomechatronics Engineering, Sung-Kyun-Kwan University
[3]Dept. of Semiconductor Systems Engineering, Sung-Kyun-Kwan University

요 약

본 연구에서는 CNN 모델의 경량화 및 on-device 추론 가속을 목표로 한다. 경량화 기법으로는 QAT 기법을 사용하며 여러 환경에서의 성능을 비교한다. 이어서 on-device 추론 가속을 위해 Jetson Nano Board 에서 TensorRT 변환을 통해 모델을 최적화한다.

1. 서론

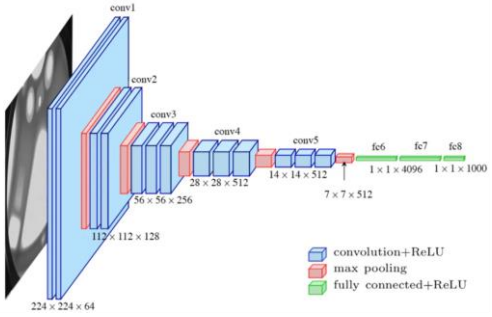
딥러닝 기술의 발전은 컴퓨터 비전, 자연어 처리, 음성 인식 등 다양한 분야에서 혁신적인 결과를 이끌어내고 있으며, 이러한 발전은 대량의 데이터와 컴퓨팅 리소스에 의존하고 있다. 최근의 하드웨어 가속 기술의 발전은 모델의 추론 속도를 향상시키는 데 크게 기여하고 있다.

본 연구는 딥러닝 모델의 경량화와 하드웨어 가속 기술을 결합하여 딥러닝 모델의 추론 속도를 향상하는 방법을 탐구한다. 특히, CNN 모델을 경량화하고 on-device 추론을 가속하는 것을 목표로 한다.

2. 연구 환경 설정

본 연구에서는 대상 모델로 (그림 1)의 구조를 갖는 VGG-16 이미지 분류 모델을 사용한다. VGG-16 모델은 컨벌루션 레이어를 여러 층 쌓은 구조로, 성능에 비해 파라미터 수가 많다고 평가되어 양자화 대상으로 적합하다. 데이터셋으로는 CIFAR10 데이터셋을 사용한다.

상기 모델에 Quantization layer 와 Dequantization layer 를 추가하여 실제 연산은 FP32 로 이루어지지만 주어진 정밀도의 inference 상황을 시뮬레이션하는 Quantization After Training(QAT) 기법을 적용한다



(그림 1) VGG-16 Architecture

추론의 경우 NVIDIA Jetson Nano 에서 TensorRT 변환을 통해 최적화한 후 GPU 환경에서 가속한다.

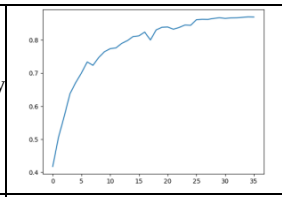
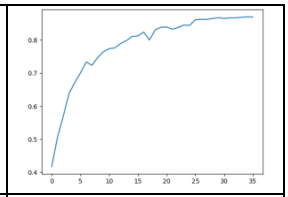
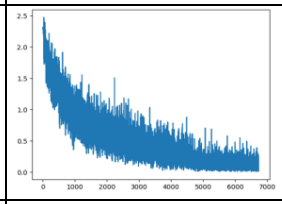
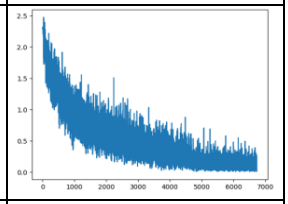
3. VGG 모델의 경량화

연구 환경 설정에서 기술한 QAT 기법을 다양한 환경에서 실험 후 비교한다. 특히 양자화 빈도수, 레이어 퓨전, 각 정밀도에 대한 성능에 대해 실험한다.

3.1. 양자화 빈도수에 따른 성능

(표 1)의 왼쪽 및 오른쪽은 각각 모델의 모든 레이어에 양자화를 적용한 성능과 가중치 및 활성함수에만 8 비트 양자화를 적용한 성능이다.

실험 결과, 양자화 빈도수가 적은 가중치 및 활성함수 양자화 모델이 성능이 뛰어남을 확인할 수 있다.

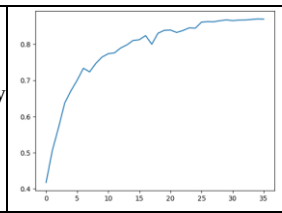
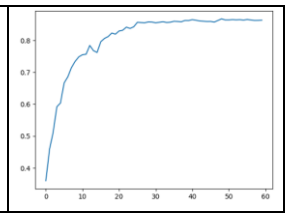
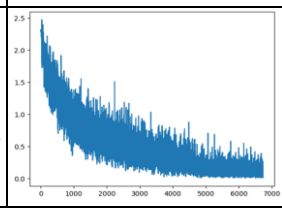
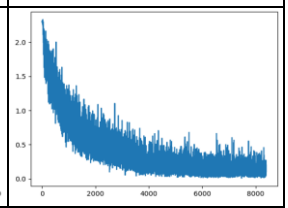
Valid Accuracy / Epoch		
Loss / Iteration		
Accuracy	85.28%	85.47%

(표 1) 모든 레이어 양자화 모델과
가중치 및 활성함수 양자화 모델의 성능 비교

3.2. 레이어 퓨전에 따른 성능

(표 2)의 왼쪽 및 오른쪽은 각각 기존 모델에 배치 정규화를 적용한 성능과 컨벌루션 레이어와 배치 정규화 레이어를 합친 8 비트 양자화 모델의 성능이다.

실험 결과, 컨벌루션 레이어와 배치 정규화 레이어를 합쳤을 때 성능이 뛰어남을 확인할 수 있다.

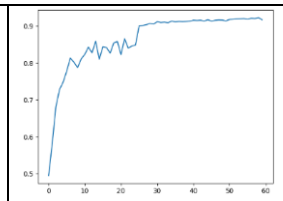
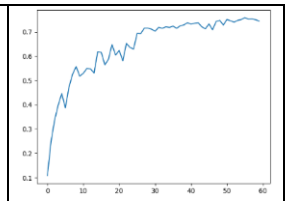
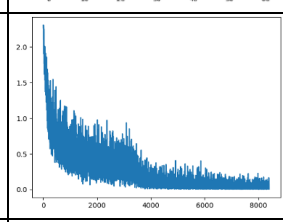
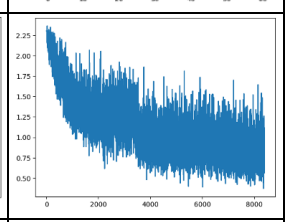
Valid Accuracy / Epoch		
Loss / Iteration		
Accuracy	91.54%	91.09%

(표 2) 정규화 적용 모델과
모델 퓨전 적용 양자화 모델의 성능 비교

3.3. 정밀도에 따른 성능

(표 3)의 왼쪽 및 오른쪽은 각각 4 비트 양자화 모델의 성능과 2 비트 양자화 모델의 성능이다.

실험 결과, 4 비트 양자화까지는 기존 성능이 유지되지만 그 이하의 비트에서는 성능이 크게 떨어지는 것을 확인할 수 있다.

Valid Accuracy / Epoch		
Loss / Iteration		
Accuracy	90.95%	74.45%

(표 3) 4 비트 양자화 모델과
2 비트 양자화 모델의 성능 비교

4. On-Device 추론 가속 성능

VGG-16 모델을 TensorRT 변환을 통해 FP16 양자화 및 그래프 최적화를 통해 추론을 최적화한 후, Jetson Nano 보드에서 추론 성능을 측정한다.

실험 결과, 기존 성능을 유지하며 메모리 사용량 및 추론 속도가 향상됨을 확인할 수 있다.

	기존	Tensorrt (FP16)
Inference Time	3.6 ms	0.647 ms
Memory	463.81 MB	274.17 MB

(표 4) Jetson Nano 에서의 TensorRT 추론 성능

5. 결론

본 연구에서는 다양한 설정으로 QAT 양자화 기법을 VGG-16 모델에 적용해 기존 성능을 유지한다. 실험 결과 양자화 빈도수 최소화, 레이어 퓨전 적용 시 성능이 뛰어남을 확인할 수 있다. 또한, 4 비트 양자화까지는 기존 성능이 유지됨을 확인할 수 있었다.

TensorRT 변환을 통해 Jetson Nano 에서 추론 성능을 실험한 결과 기존 성능을 유지하며 메모리 사용량 및 추론 속도의 향상을 확인할 수 있었다.

ACKNOWLEDGMENT

이 논문은 정부(교육부-산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0022098, 2024 년 미래형자동차 기술융합혁신인재양성사업)

참고문헌

[1] Amir Cholami, et al. "A survey of quantization methods for efficient neural network inference." Low-Power Computer Vision. Chapman and Hall/CRC, 2022. 291-326.