임베디드 시스템 환경에서의 INT8 및 FP32 기반 Mixed Precision 의 정확도 실험 및 분석

장경빈 [1], 이종은 [1], 임승호 [2]
[1] 한국외국어대학교 컴퓨터전자시스템공학부
[2] 한국외국어대학교 컴퓨터공학부

jkb12377@naver.com, whddms1208@gmail.com, slim@hufs.ac.kr

Accuracy Experiment and Analysis of INT8 and FP32 based Mixed Precision Layer in Embedded System Environments

Kyung-Bin Jang[1], Jong-Eun Lee[1], Seung-Ho Lim[2]
[1]Division of Computer Electronic System Engineering, Hankuk University of Foreign Studies
[2] Division of Computer Engineering, Hankuk University of Foreign Studies

요 약

최근 CNN 기반 객체인식 시스템은 고정밀도 모델을 기반으로 정확도를 높이고 있다. 하지만 고정밀도 모델일수록 모델의 크기가 늘어나고 더 많은 하드웨어 자원을 필요로 한다. 따라서 모델 경량화 기술이 많이 연구되고 있으며, 그 중에 대표적인 경량화 기술이 양자화 기술이다. 양자화 기술은 파라미터의 크기와 연산 오버헤드를 줄이지만, 정확도 역시 줄어들게 된다. 영자화와 정확도의 상관관계를 분석하기 위해서 본 논문에서는 INT8 과 FP32 을 이용한 Mixed precision CNN 을 실행시키기 위한 프레임워크를 구성하고, 임베디드 시스템 환경에서의 INT8 연산에 기반하여 맞추어 각 layer 별 Mixed Precision 연산을 수행하여 보고, 모델의 정확도를 측정하여 분석하여 보았다.

1. 서론

최근 AI 분야에서는 고정밀도 모델들이 더욱 많이 개발되어 제품화되고 있으며, 글로벌 기업들도 이에 박차를 가하고 있다. 하지만 딥러닝 신경망에서 모델의 정확도와 성능이 높아질수록 필요한 피라미터 수도 증가하며, 이는 모델의 크기를 늘려주고, 처리에 필요한 하드웨어 자원을 더욱 필요로 한다[1]. 게다가, 모바일 기기, IoT 시스템, 자율 주행차와 같은 분야에서는 제한된 하드웨어 자원에서도 실시간 추론이 필요하기 때문에 거대한 신경망 모델을 실행하는 것은 효율적이지 않다[2].

따라서, 최근에는 모델 경량화 기술이 많이 연구되고 있다. 이는 가지치기, 양자화, 지식 증류, 경량 네트워크 설계 등을 활용하여 모델의 파라미터 수를 줄이면서도 높은 정확도와 성능을 유지하는 방식이다[3]. 그 중에서 양자화의 경우, 일반적으로 네트워크의 파라미터를 FP32 가 아닌 INT8 을 사용함으로써 파라미터의 크기를 줄이고 연산의 오버헤드를 줄일 수 있다. 그러나 INT8 의 데이터 범위의 한계로 인해서 모델의 정확도를 감소시킨다.

모델의 양자화에 따른 정확도 감소의 주요 요인으로는 INT8 을 이용해 합성곱 연산을 하게 되면 오버플로우가 일어나 필연적으로 학습 중 오차가 발생할 수 있으며 학습이 진행되지 않는 문제가 발생할 수 있다. 이러한 문제를 Mixed Precision 으로 해결할 수 있다. 본 논문에서는 임베디드 시스템 환경에서의 INT8 연산을 기반으로 FP32 연산을 혼용해서 사용하는 Mixed Precision 에 대한 분석과 실험을 진행하고 모델의 정확도와 연산 시간을 분석하여 보았다. 먼저, ARM 기반 임베디드 시스템에서 mixed precision 을 위한 c/tensorflow 프레임워크 구성[4]을 참고하여 프레임워크를 구성하고 높은 정밀도를 요구하는 layer 의 위치와 특성을 파악하고 그것을 바탕으로 네트워크의 경량화 수준과 정확도 수준을 분석해보았다.

2. INT8 overflow 이슈

먼저, INT8 양자화의 문제점을 파악하기 위해서 본 연구에서는 INT8 inference 중 overflow 발생에 대해서 분석을 진행하였다. 표 1 은 yolov3-tiny CNN 의 layer 1~10 에서, 단일 픽셀에 대해 weight 와 input 을 곱했을 때 INT8 범위에서 overflow 가 발생하는 비율을 나

타낸 것이다. 표 1 을 보면 전체 layer 중에서 2, 3, 4, 10 번 convolutional layer 에서 average overflow 가 상대적으로 높은 비율을 보인다. 전체 layer 중에서 overflow 비율이 30%가 넘어가는 경우도 존재하였다. 우리는 이를 해결하기 위해 단일 픽셀 연산 결과를 INT16 data type 에 저장하도록 고안하였다. INT16 범위에서 위와 같이 overflow 발생 비율을 분석했을 때 단 1 건의 overflow 도 발생하지 않는 것을 확인하였다.

다음으로는 단일 픽셀에 대한 합성곱 연산 결과에 대한 overflow 를 분석하였다. 표 2 는 단일 픽셀에 대해 INT16 범위에서 overflow 가 발생하는 비율이다. 전체적으로 average overflow 가 5%미만의 낮은 비율을 보이며 아예 overflow 가 발생하지 않은 layer 도 있지만 4, 6, 8, 10 번 layer 의 특정 filter 에서 상당히 높은 overflow 비율을 보여 단일 픽셀에 대한 합성곱 연산 결과에 대해 INT16 data type 을 사용할 경우 약간의 정밀도 손실이 있을 것이다. 그렇기 때문에 단일 픽셀 합성곱 연산 결과에 대해서는 INT32 data type 에 저장하도록 고안하였다.

<표 1> layer 1-10 에서 INT8 의 overflow 비율

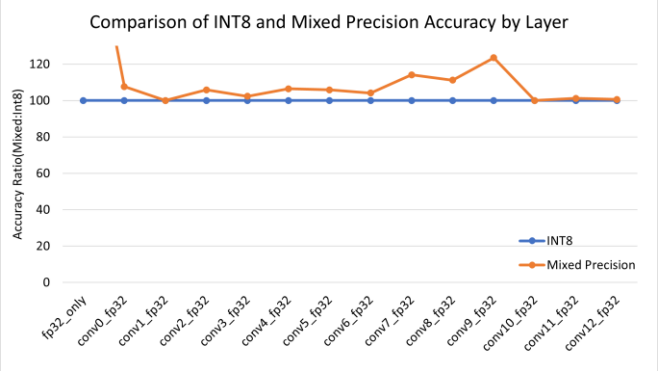
Layer	1	2	3	4	5	6	7	8	9	10
Avg	0.092	0.169	0.203	0.201	0.113	0.069	0.109	0.079	0.092	0.144
Max	0.17	0.296	0.321	0.289	0.186	0.129	0.24	0.163	0.14	0.281
Min	0.008	0.023	0.069	0.113	0.023	0.026	0.018	0.001	0.044	0.017

<표 2> 단일 픽셀에 대해 INT16 범위에서 overflow 가 발생하는 비율

Layer	1	2	3	4	5	6	7	8	9	10
Avg	3E-04	7E-04	0.015	0.253	0.08	0.146	0.005	0.101	0	0.436
Max	0	0	0	0	0	0	0	0	0	0
Min	1E-05	3E-05	0.001	0.017	0.005	0.013	3E-05	0.003	0	0.053

3. 실험 및 결과 분석

양자화된 파라미터를 가지는 CNN 모델에서 layer 별 Mixed Precision 을 적용하는 경우에 대한 정확도 실험을 수행하였다. 우리는 사전 학습된 yolov3-tiny model 을 사용하였으며, 라벨링한 image data 1000 장과 검출된 객체를 비교분석하여 혼합 행렬을 통해 정확도를 계산하였다. 높은 정밀도를 요구하는 layer 의 위치를 특정하기위해 각 0~12 layer 에 대해서 하나의 layer 만 FP32 로 연산을 수행하고 나머지 layer 는 INT8 연산을 수행하도록 진행하였고 환경은 Jetson Nano 와 Raspberry Pi 환경에서 같은 실험을 진행하여 결과를 분석하였다. 그림 1 은 Jetson 과 Pi 환경에서의 정확도 분석이다. FP32 로 연산했을 때 보다 INT8 Mixed Precision 을 사용하여 연산을 했을 때 정확도가 소폭 상승하는 것을 확인하였다.



(그림 1) Jetson Nano 에서 Yolo3-tiny 모델에 대한 레이어별 mixed precision 정확도 실험 결과

또한 0, 2, 4, 5, 7, 8, 9 layer 를 FP32 로 연산하였을 때 총 정확도가 상승하는 것으로 보아 각 layer 가 높은 정밀도를 요구하고 있음을 확인하였다. 반대로 1, 6, 10, 11, 12 layer 는 높은 정밀도로 연산을 수행했음에도 불구하고 오히려 정확도가 유지되거나 떨어지는 것을 확인하였다.

4. 결론

본 논문에서는 임베디드 환경에서의 INT8 연산에 대한 Mixed Precision 실험을 진행하였다. 각 layer 별 높은 정밀도를 요구하는 layer 의 위치와 특성을 분석하고 실제 정확도에 어느 정도 영향을 미치는지 분석하였다. 향후 이번 실험에서 얻은 결과를 토대로 하나의 layer 가 아니라 높은 정밀도를 요구하는 여러 개의 layer 의 정밀도를 변경하여 실험을 진행하고 data set 의 개수를 더 늘려 정확도의 향상에 대해 연구할 예정이다.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (NRF-2021R1F1A1048026).

참고문헌

[1] 이경하, 김은희, 딥러닝 모델 경량화 기술 분석, 대전, 한국과학기술정보연구원, 엠에스기획, 2020

[2] 유승목, 이경희, 박재복, 윤석진, 조창식, 정영준, 조일연, 임베디드 시스템용 딥러닝 추론엔진 기술 동향, 전자통신동향분석, 34 권, 제 4 호, 24 쪽

[3] 이경하, 김은희, 딥러닝 모델 경량화 기술 분석, 대전, 한국과학기술정보연구원, 엠에스기획, 2020

[4] 이종은, 임승호, ARM 기반 임베디드 시스템에서 mixed precision 을 위한 c/tensorflow 프레임워크 구성, AKC 2022 학술발표대회 논문집, 29 권, 제 2 호