Take Home Project Analysis Report (Part4)

Kim, Donghyun

본 Take Home Project에서 CIFAR-10 dataset을 사용해 CNN 모델을 만들고 training 하는 과제를 진행했다. PyTorch Library를 사용해 모델을 구현해 성능을 평가하는 과정을 거쳤다. 이후 학습된 모델에 Dynamic Quantization을 적용해 효율과 정확도, 크기, inference time 등을 지표로 하여 Quantization 적용 이전 모델과의 비교를 통해 발전된 사항, 실제 edge device들에서 나타날 수 있는 유리한 점을 분석하고자 한다.

- The Effect on Quantization: Perspective of Performance and Accuracy

프로젝트에서 진행한 모델 사이 효과 분석인자는 accuracy, file size, inference time 3 가지로 설정했다. Training 이후 모델 사이의 Quantization에 따른 효과를 앞선 요소에 따라 분리해 성능을 계산하고 그 결과를 토대로 분석한다.

Accuracy의 경우 기존 CNN모델에서 0.748의 accuracy를 달성했는데 quantization 이후 FP32에서 Linear Layer를 INT8로 변환한 후에는 0.747의 accuracy를 보이는 것을 확인했다. 감소 정도가 0.001에 불과하기에 본 측정을 통해 quantization에도 불구하고 accuracy의 손실이 거의 발생하지 않음을 알 수 있다. 이에 대해 Dynamic Quantization 이 Convolution이 아니라 Fully-Connected에서만 적용한다는 특성이 feature 추출과정에서 정확도를 유지할 수 있음을 알 수 있다. 따라서, 본 프로젝트에 사용한 quantization 은 분류 능력에 영향을 주지 않는다는 결론을 도출한다.

모델 크기의 경우 앞선 모델이 9148.3KB를 차지하는 한번 quantization 이후 모델의 크기는 2998.2KB로 기존 크기에서 67%가 감소해 32비트에서 8비트로 양자화 하며 4배로 압축한 이론적 결과와 근접하기 때문에 모델 크기의 조절을 통한 저장 공간 절약과 적은 메모리 사용을 기대할 수 있다. 기존 모델의 inference time은 22.7409초에서 양자화 이후 inference time이 22.4910초로 조금 개선된 모습을 보였다.

- Trade-Offs Between Compression and Precision

Quantization을 통해 모델의 크기와 속도가 개선된 모습을 보였다. 이때, 모델 가중치가 32비트에서 8비트로 줄어드는 이유로 모델의 크기가 압축되어 향상된 모습을 보였으나 데이터를 표현하는 precision이 떨어진다. 기존 모델의 경우 float형 32비트인 반면 양자화 이후 모델의 경우 8비트 int형이기에 숫자를 표현하는 precision과 range가 감소한다. 따라서 Compression을 통해 비트 수를 줄이면 memory size가 감소함과 동시에 Precision이 감소하는 Trade-Off가 발생한다.

- Benefits of Deploying Quantized Models on Edge Devices

본 프로젝트를 진행하며 Quantized model이 가지는 이점이 메모리 사용량이 감소하고 inference time이 감소, 추론 속도가 증가함을 알았다. 더불어 줄어든 메모리 사용량은 결국 device의 에너지를 절감해 전력 효율이 개선될 수 있다. Edge Device의 경우 데이터를 메인 서버로 보내기 이전에 edge에서 정보를 처리하는 device로 CPU 성능이 서버보다 능력이 현저히 떨어진다. 모바일 기기와 같이 메모리가 작고 배터리 기반의 전력 소비에 한계를 둔다는 단점을 가진다. 따라서 Quantization을 통한 Model의 메모리 사용량을 절약, 에너지 절감 전력 효율 개선이라는 장점으로 Edge Device의 단점을 보완할 수 있으므로 Quantized Model이 Edge Device에서 사용되는 것이 훨씬 유리하다고 판단할 수 있다.