자원 제약 환경에서의 온디바이스 AI 시스템을 위한 모델 경량화 분석 및 성능 평가

권기웅, 주영민, 김재하, 박승현
한국전자기술연구원
{kiwoong.kwon, yjoo, jaeha8888, park.seunghyeon}@keti.re.kr

Analysis and Performance Evaluation of Model Compression for Resource-Constrained On-Device AI Systems

Kiwoong Kwon, Youngmin Joo, Jaeha Kim, Seunghyeon Park
Korea Electronics Technology Institute

요 약

본 논문에서는 자원 제약 환경에서 동작하는 온디바이스 AI 시스템의 실시간 추론 성능 최적화를 위해 프루닝(Pruning) 및 양자화(Quantization) 기반의 모델 경량화 기법을 분석하고 성능을 평가한다. 고성능 딥러닝 모델을 엣지 디바이스에 탑재할 경우 발생하는 연산 지연과 전력 소모 문제를 해결하기 위해, 다중 퍼셉트론(MLP) 기반 회귀 예측 모델을 대상으로 경량화를 수행하였다. 실험 결과, 프루닝과 FP16 양자화를 결합한 모델은 원본 대비 메모리 사용량을 크게 절감하고 추론 속도를 현저히 개선하면서도, 예측 정확도 측면에서는 원본과 유사한 성능을 유지하여 전반적으로 우수한 정확도 - 효율 균형을 보였다.

Ⅰ. 서 론

자원 제약 환경에서 동작하는 온디바이스 AI 시스템은 실시간 환경 인식과 즉각적인 제어 수행을 위해 다양한 설비로부터 데이터를 초 단위로 수집·분석해야 한다. 그러나 이러한 기능을 수행하는 엣지 게이트웨이는 제한된 연산 성능, 메모리 용량, 전력 소비 한계를 가지며, 클라우드 환경에서 사용되는 고성능 AI 모델을 그대로 적용할 경우 응답 지연, 발열 증가, 전력 소모 가중 등의 문제가 발생할 수 있다[1].

특히 실시간 제어가 요구되는 시스템에서는 추론 지연이 제어 안정성에 영향을 미치며, 온디바이스 환경의 특성상 장시간 운용을 고려한 자원 효율 또한 중요한 성능 요구사항으로 작용한다. 이에 따라 온디바이스 환경에 적합한 경량 AI 모델 설계는 필수적이며, 모델 크기·메모리·추론 속도·정확도 간의 균형을 체계적으로 고려할 필요가 있다.

한편, 기존의 모델 경량화 연구들은 이미지 분류나 객체 인식과 같은 분류 문제를 중심으로 수행된 경우가 대부분이며, 회귀 문제를 대상으로 한 체계적인 분석은 상대적으로 부족하다. 회귀 문제는 출력 값이 연속적인 실수 영역에 분포하므로, 양자화 과정에서 발생하는 근사 오차가 예측 정확도에 미치는 영향이 상대적으로 크게 나타날 수 있다. 따라서 회귀 기반 모델에 경량화 기법을 적용할 경우, 정확도 변화 양상을 정량적으로 분석하고 적용 가능성을 검토하는 과정이 필요하다.

본 연구에서는 MLP 기반 회귀 예측 모델을 대상으로, 프루닝(Pruning)과 양자화(Quantization) 중심의 모델 경량화 기법을 적용하고 성능을 평가하였다. TensorFlow/Keras 및 TensorFlow Model Optimization 기반으로 모델 경량화 기법을 구현하였고, FP16 및 INT8 연산을 지원하는 온디바이스 AI 게이트웨이 환경을 고려하여, TFLite 변환 이후 동일한 실행 환경에서 모델 간 상대적인 성능 변화 양상을 비교·분석하였다.

Ⅱ. AI 모델 경량화 기법

본 논문에서는 자원 제약 환경에서 동작하는 온디바이스 AI 시스템을 대상으로, 대표적인 모델 경량화 기법인 프루닝과 양자화를 적용하였다.

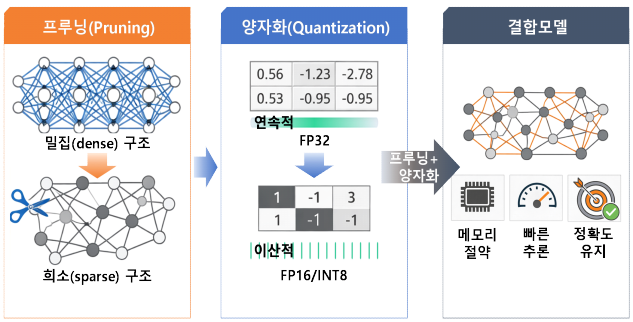

그림 1. AI 모델 경량화 기법 개요

프루닝은 신경망 내부에서 중요도가 낮은 가중치를 제거하여 모델의 파라미터 수와 연산량을 감소시키는 기법으로, 적용 전에는 밀집(dense) 구조를 가지던 모델이 적용 후에는 희소(sparse) 구조를 형성하게 된다. 본 연구에서는 TensorFlow Model Optimization 프레임워크를 활용하여 점진적 희소화 기반의 프루닝을 적용하였으며, 학습 단계에 따라 희소도를 점진적으로 증가시키는 Polynomial Decay 스케줄을 사용함으로써 급격한 성능 저하 없이 모델 경량화를 수행하였다. 또한 프루닝 적용 이후 재학습과 strip_pruning 과정을 통해 추론에 최적화된 모델을 생성하였다.

프루닝 비율을 일괄적으로 적용할 경우 모델의 수렴 안정성이 저하될 수 있으며, 특히 얕은 구조의 MLP 기반 회귀 모델에서는 과도한 프루닝이 출력 민감도를 증가시킬 위험이 있다. 이에 본 연구에서는 학습 초기에는 모델의 표현력을 유지하고, 학습이 안정화되는 후반부에 프루닝 비율을 약 10%에서 최대 50% 수준까지 점진적으로 증가시키는 전략을 채택함으로써, 불필요한 가중치를 단계적으로 제거하고 정확도 저하를 최소화하고자 하였다.

양자화는 모델의 가중치와 활성값을 부동소수점(FP32) 표현에서 저정밀 수치 표현으로 변환하여 저장 공간과 연산 비용을 줄이는 기법이다. FP16

표 1. 실험 결과

모델 구분	크기 (MB)	속도 (ms)	메모리 (MB)	정확도		
				MAE	RMSE	MAPE
원본	0.0668	52.994	3.7305	0.1682	0.2688	10.12
프루닝	0.0287	54.765	0.1	0.1696	0.2697	10.22
프루닝 +FP16	0.0086	0.0036	0.0129	0.1695	0.2697	10.22
프루닝 +INT8	0.0054	0.0151	0.0081	0.2425	0.5261	14.89

표 2. 실험 결과 분석

모델	모델 크기	추론 속도	메모리 사용량	정확도 손실
원본	100%	1.00x	100%	기준
프루닝	41.7%	0.97x	2.7%	**0.1%**
프루닝 +FP16	12.5%	**14,700x**	0.35%	**0.1%**
프루닝 +INT8	**7.9%**	3,510x	**0.22%**	4.77%

양자화는 반정밀 부동소수점 연산을 통해 모델 크기 감소와 연산 효율 향상을 동시에 달성할 수 있는 반면, INT8 양자화는 캘리브레이션 과정을 거쳐 정수 연산 기반의 추론을 가능하게 하여 가장 높은 연산 효율을 제공한다. 다만 INT8 양자화는 스케일 및 영점 기반 근사로 인해 연속적인 출력 값을 가지는 회귀 문제에서는 오차 민감도가 증가할 수 있다.

본 연구에서는 이러한 특성을 고려하여 FP16 및 INT8 양자화를 모두 적용하고, 프루닝 단독 모델과 프루닝 - 양자화 결합 모델을 함께 구성하여 비교하였다. 이를 통해 각 경량화 기법의 단독 및 결합 적용 시 모델 크기, 연산 효율, 예측 정확도 변화 양상을 종합적으로 분석하고, 자원 제약 환경의 온디바이스 AI 시스템에 적합한 정확도 - 효율 균형점을 도출하고자 하였다. 특히 회귀 기반 모델에 대한 경량화 적용 시 정확도 저하 양상과 자원 절감 효과 간의 관계를 정량적으로 비교함으로써, 실제 온디바이스 배포 관점에서의 적용 가능성을 함께 검토하였다.

Ⅲ. 성능평가

경량화 모델의 성능 평가는 자원 제약 환경에서의 온디바이스 AI 시스템 적용 가능성을 정량적으로 검증하기 위해 수행되었다. 실험 대상은 기 보유한 산업용 보일러 운전 데이터를 기반으로 학습된 MLP 예측 모델로, 특정 운영 조건에서 산업용 보일러의 스팀량을 예측하는 단일 모델을 기준으로 하였다. 해당 모델은 2단계 연속 구조(Sequential Dual MLP)로 구성되며, 히든 레이어 구조는 (15, 10)과 (50, 50)으로 설계되었고, ReLU 활성화 함수와 Adam Optimizer를 적용하였다. 학습에는 약 9개월간 수집된 실측 운전 데이터를 활용하여, 다양한 부하 조건과 운전 상태가 충분히 반영되도록 하였다.

경량화는 프루닝과 양자화를 중심으로 단계적으로 적용하였다. 점진적 프루닝을 실시한 후, 프루닝이 적용된 모델을 대상으로 FP16 및 INT8 양자화를 각각 적용하였으며, 프루닝과 양자화를 결합한 모델을 생성하여 단일 기법 적용 대비 자원 절감 효과와 정확도 변화를 종합적으로 분석하였다. 성능 평가는 모델 크기, 메모리 사용량, 추론 시간, 예측 정확도를 주요 지표로 수행하였으며, 추론 성능은 실제 온디바이스 환경을 가정한 조건에서 동일한 하드웨어 상에서 동일한 입력 데이터셋에 대해 반복 추론을 수행한 후 평균 추론 시간을 산출하여 평가하였다. 메모리 사용량은 추론 과정에서의 프로세스 RSS(Resident Set Size) 변화를 기준으로 측정함으로써, 모델 크기뿐 아니라 실제 실행 시 메모리 점유 특성을 함께 분석하였고, 모델 크기는 저장 파일 크기를 기준으로 측정하였으며 예측 정확도는 MAE, RMSE, MAPE를 통해 평가하였다.

실험 결과, 프루닝 적용 모델은 원본 대비 모델 크기와 메모리 사용량이 유의미하게 감소하였으나, 추론 속도 측면에서는 큰 차이를 보이지 않아 성능 향상 효과는 제한적인 것으로 나타났다. 이는 프루닝을 통해 파라미터 수는 감소하였지만, 평가 환경에서 희소 연산에 대한 전용 가속이 충분히 활용되지 못했기 때문으로 해석된다. 반면 프루닝과 FP16 양자화를 병행한 모델은 모델 크기와 메모리 사용량을 크게 절감하는 동시에 추론 속도에서도 가장 큰 개선을 보였으며, MAE, RMSE, MAPE 지표에서 원본과 거의 동일한 수준의 예측 정확도를 유지하여 전반적으로 가장 우수한 정확도 - 효율 균형을 나타냈다. 한편, 프루닝과 INT8 양자화를 결합한 모델은 가장 높은 경량화 효과를 달성하였으나, 본 실험 환경에서는 정수 연산에 대한 전용 가속이 제한적이어서 양자화·디양자화 오버헤드가 상대적으로 크게 작용하였고, 이로 인해 FP16 양자화 모델 대비 추론 속도 개선폭이 제한되었다. 또한 회귀 문제의 연속적인 출력 특성으로 인해 INT8 양자화 과정에서 근사 오차가 누적되면서 예측 정확도가 다소 저하되는 경향이 관찰되었다.

본 실험 결과를 바탕으로 회귀 기반 효율 예측 모델의 특성과 실험 조건을 종합적으로 고려할 때, 실시간 추론 성능과 예측 정확도를 동시에 만족하기 위한 경량화 전략으로 프루닝과 FP16 양자화의 조합이 가장 실용적인 선택지임을 확인하였다.

Ⅳ. 결론

본 연구에서는 자원 제약 환경에서의 온디바이스 AI 시스템 적용을 목표로, 기 보유 MLP 모델에 프루닝과 양자화 기반 경량화 기법을 적용하고 성능을 평가하였다. 실험 결과, 점진적 프루닝과 FP16 양자화를 병행한 모델은 원본 대비 메모리 사용량을 절감하고 추론 속도를 개선하면서도 예측 정확도 손실을 0.5% 이하로 유지하여 가장 우수한 정확도 - 효율 균형을 보였다. 반면 INT8 양자화는 추가적인 경량화 효과를 제공하였으나, 일부 정확도 저하가 관찰되었다.

이러한 결과는 본 연구에서 고려한 회귀 모델과 실험 환경 조건 하에서, 프루닝과 FP16 양자화의 조합이 실시간 추론 성능과 예측 정확도를 동시에 만족시키는 데 있어 상대적으로 유효한 경량화 전략임을 시사한다. 향후 연구에서는 적응형 프루닝과 혼합 정밀도 양자화를 결합한 확장 전략을 통해 다양한 운전 조건에 대한 적용 가능성을 검증할 예정이다.

ACKNOWLEDGMENT

본 연구는 기후에너지환경부(MCEE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. RS-2025-02314925)

참 고 문 헌

[1] S. Somvanshi et al., "From Tiny Machine Learning to Tiny Deep Learning: A Survey," ACM Computing Surveys, vol. 58, no. 7, pp. 1-33, 2025.

[2] R. Ding et al., "Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized Object Detector," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16174-16184, 2024.