# 비구조적 가지치기와 양자화를 통한 인공지능 모델 경량화 및 추론 성능 분석

성주형, 박경원
한국전자기술연구원

{jh.sung, kwpark}@keti.re.kr

# Lightweight AI Model through Unstructured Pruning and Quantization with Inference Performance Analysis

Juhyoung Sung, Kyoungwon Park
Korea Electronics Technology Institute (KETI)

요 약

외부에서 훈련된 인공지능(Artificial intelligence, AI) 모델을 다양한 플랫폼에서 의도하는 대로 동작 시키기 위해 AI 모델의 경량화는 많은 관심을 받고 있다. 본 논문에서는 합성곱 신경망(Convolutional neural network, CNN) 모델 중 하나인 ResNet-50 기반의 의료 이미지 분류 모델에 대해 가지치기(Pruning) 및 양자화(Quantization)를 적용하여 모델의 경량화를 수행한다. 그 이후, 경량화 된 모델에 대하여 기존 모델 대비 이미지 분류 예측 성능이 어떻게 변화하는지 분석한다.

## Ⅰ. 서 론

인공지능(Artificial intelligence, AI) 기술이 급속도로 발전함에 따라 여러 분야에서 AI에 대한 활용도가 높아지며, 성능도 크게 향상되고 있다. 이와 함께, 개별 기기 내부에서 AI 기능을 자체적으로 실행하는 온 디바이스(On-device) AI 기술에 대한 수요가 증가함에 따라 개발된 AI 모델이 특정 환경에서만 동작하는 것이 아니라 배포를 통하여 다양한 플랫폼에서 AI 기능이 동작하도록 하는 것 또한 매우 중요한 과제이다. 이를 위하여 플랫폼 고유의 하드웨어(Hardware, HW) 구성이나 성능에 따라 AI 모델을 최적화시키는 것이 필요하다. 일반적으로 배포 환경에서는 개발 환경과 비교하였을 때 전력량 등 가용 자원 측면에서 제한이 있으며, 운영체제(Operating system, OS) 등의 제약과 함께 그래픽 처리 장치(Graphics processing unit, GPU)의 유무 및 호환성 등 고려해야 되는 요소가 매우 다양하다. 특히, 플랫폼의 연산 자원 성능에 따라 동일한 AI 모델을 실행하더라도 속도에서 크게 차이가 발생할 수 있고, AI 모델의 규모에 따라서 모델의 탑재 자체가 불가능할 수도 있다. 여기에 더해, GPU 가속을 사용하지 못하는 플랫폼 환경에서는 CPU(Central processing unit)만 이용해서 AI 모델을 동작시켜야 하므로 다양한 연산 자원 환경을 고려해야 AI 모델을 배포해야 한다[1]. 이를 위하여, AI 모델을 성능을 유지하면서도 경량화를 통하여 모델을 압축하는 것이 필요하다.

본 논문에서는 혈액 세포 이미지로 구성된 데이터셋인 BloodMNIST[2]를 분류하도록 훈련된 합성곱 신경망(Convolutional neural network, CNN) 계열의 ResNet-50 모델을 이용하여 비구조적 가지치기(Pruning) 및 양자화(Quantization)를 이용하여 모델의 경량화를 적용한다. 모델에 대해 가지치기를 적용하는 경우, 미세 조정(fine-tuning)을 통하여 모델을 다시 학습시킨다. 양자화를 적용하는 경우에는 추가적인 재학습 없이 모델의 가중치만 양자의 비트(bit) 레벨에 따라 변경한다. 다양한 경량화 설정을 통하여 AI 모델의 경량화를 적용하고, 경량화 설정 방법에 따른 모델의 추론 성능을 비교 및 평가한다.

## Ⅱ. BloodMNIST 데이터셋 및 벤치마크 성능

MedMNIST는 다양한 의학 분야의 시각 데이터로 구성된 AI 모델을 학습 및 평가하기 위해 공개된 벤치마크 데이터셋으로, 훈련용 데이터셋, 교차 검증용 데이터셋, 테스트 데이터셋으로 분리되어 있다. 본 논문에서는 MedMNIST의 하위 항목 중 하나인 혈액 세포에 대한 RGB 이미지로 구성된 BloodMNIST 데이터셋을 이용한다. 데이터셋에 포함된 각각의 이미지에 대한 픽셀의 크기는 28x28 이며, 총 8가지 범주로 구성된다. BloodMNIST의 테스트 데이터셋에 대한 8가지 범주에 대한 분포는 그림 1과 같다. BloodMNIST 데이터셋을 이용하여 훈련된 ResNet-50 모델의 테스트 데이터셋에 포함된 3421개의 이미지에 대한 분류 정확도는 0.956 정도 수준이며, AUC(Area Under the Curve)는 0.997 정도로 매우 높은 성능을 보인다[2]. 본 논문에서는 공개된 벤치마크 성능 기반으로 AI 모델의 경량화 전후 성능 비교를 하기 위해, 먼저 ResNet-50 모델을 구현하였다. 그 이후, 벤치마크 성능과 유사한 결과를 얻을 수 있도록 모델을 훈련시켜 테스트 데이터셋에 분류 정확도와 AUC가 각각 0.9669, 0.9983이 되도록 초기 환경을 구성하였다.

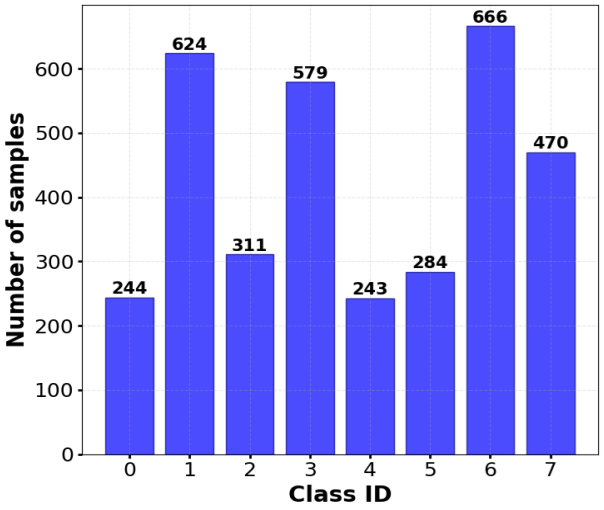


그림 1. BloodMNIST 데이터셋의 8 가지 범주에 대한 데이터 분포

## Ⅲ. AI 모델 경량화 설정에 따른 추론 성능 분석

본 절에서는 II절에서 BloodMNIST를 분류하는 ResNet-50 모델을 이용하여 다양한 AI 모델 경량화 방법 중, 비구조적 가지치기와 양자화 방식을 이용하여 경량화를 적용하고 성능을 분석한다. 비구조적 가지치기를 수행하는 경우, 설정한 가지치기 비율에 따라 모델의 가중치를 0으로 변환시키고, 모델의 자체 크기는 가지치기 전후 거의 변함이 없지만 모델의 희소성(Sparsity)이 증가하여 배포 환경에서 메모리 대역폭이나 과적합(overfitting) 방지, 에너지 효율 측면에서 이득을 얻을 수 있고, 희소 연산을 지원하는 환경에서는 연산 속도 또한 향상된다. 비구조적 가지치기를 적용한 이후에는 AI 모델의 성능 확보를 위하여 작은 epoch 동안 모델을 훈련하여 가중치를 미세 조정하고, 가지치기가 적용된 가중치는 마스킹 처리하여 가중치를 0으로 만든다. 한편, 양자화를 적용하는 경우, bit 레벨을 축소하는 만큼 비례해서 가중치를 축소하여 AI 모델의 크기 자체를 압축할 수 있다. AI 모델의 가중치는 부동 소수점(Floating-point, FP)으로 구성되어 있고, 일반적으로 개발 환경에서는 32-bit 레벨 또는 16-bit 레빌인 FP32 또는 FP16을 기본 설정으로 사용한다. 반면, 배포 환경에서는 메모리 대역폭 및 정수 연산을 위한 8-bit 또는 4-bit 레벨인 INT8, INT4 설정을 요구하는 경우가 있어, 모델의 크기를 축소할 뿐만 아니라 호환성을 위해서도 양자화 적용이 필요하다. 양자화를 수행하는 경우, 일반적으로 부동 소수점의 자리수만 변화하며 값 자체에는 크게 변함이 없기 때문에 추가적인 미세조정은 수행하지 않는다. 정리하면, 비구조적 가지치기와 양자화를 동시에 적용하는 경우, 모델의 경량화는 가지치기→미세조정→양자화 순으로 이루어진다. 본 절에서는 비구조적 가지치기의 비율은 20%, 50%, 80% 중 하나로 설정하고, 양자화 bit 레벨은 INT4 또는 INT8로 적용하여 ResNet-50 모델에 대한 경량화를 수행한다. 경량화 이전의 벤치마크 ResNet-50 모델은 가지치기를 적용하지 않고, FP16의 양자화가 적용된 가중치를 이용한다. 경량화 설정에 따른 ResNet-50 모델의 BloodMNIST 분류 성능을 요약하면 표 1과 같다. 표 1의 결과로부터 8-bit 양자화를 적용한 경우, 가지치기 비율이 50%까지 증가함에도 불구하고 오히려 벤치마크 모델보다 분류 정확도는 소폭 향상되는 것을 확인할 수 있다. 이는 비구조적 가지치기 이후 미세 조정을 통해 과적합 현상을 완화시킨 효과라고 볼 수 있다. 반면, 가지치기 비율이 80%까지 올라간 경우, 모델의 분류 정확도는 크게 감소하는 것을 동시에 확인할 수 있다. 이를 통해, 과도한 비율로 가지치기를 적용한 경우 모델이 과소적합(underfitting)에 빠져 추론 성능이 크게 떨어진다는 사실을 알 수 있다. 따라서, 비구조적 가지치기를 적용할 때, 모델 및 목적 함수 특성에 따라 추론 성능을 평가하여 가지치기 비율을 설정하는 것이 필요하다. 한편, 4-bit 양자화를 적용한 경우에는 8-bit 양자화 대비 성능이 크게 떨어지는 것을 확인할 수 있는데, 이러한 성능 저하의 원인은 4-bit 양자화에서 표현 가능한 부동 소수점의 정밀도에 한계가 있기 때문이라고 볼 수 있다.

표 1. 경량화 설정에 따른 ResNet-50 모델의 BloodMNIST 분류 성능 비교

| 가지치기 비율 | 양자화 | ACC | AUC |
|---|---|---|---|
| 0 % | FP16 | 0.9669 | 0.9983 |
| 20 % | INT8 | 0.9701 | 0.9975 |
| | INT4 | 0.8637 | 0.9758 |
| 50 % | INT8 | 0.9725 | 0.9981 |
| | INT4 | 0.3308 | 0.8148 |
| 80 % | INT8 | 0.3256 | 0.7205 |
| | INT4 | 0.1824 | 0.5622 |

## Ⅳ. 결론

본 연구에서는 비구조적 가지치기와 양자화를 적용하여 이미지를 입력으로 받아 분류를 수행하는 ResNet-50 모델의 경량화를 수행하고, BloodMNIST 데이터셋을 이용하여 경량화 방법에 따른 분류 성능을 비교하였다. 비구조적 가지치기를 통하여 훈련된 모델의 가중치에 대하여 희소성을 증가시켜 배포 환경에서의 메모리 효율 증가, 희소 연산 속도 향상 등의 효과를 얻을 수 있다. 양자화를 통해서는 모델의 크기를 양자화 bit 레벨에 따라 감소시킬 수 있고, 정수 처리에 최적화된 다양한 플랫폼에 AI 모델의 배포를 가능하게 한다. 다양한 경량화 설정을 적용하여 경량화를 적용하지 않은 벤치마크 모델과 경량화를 적용한 ResNet-50 모델의 BloodMNIST 데이터셋에 대한 분류 성능을 비교하였을 때, 적절한 크기의 가지치기는 오히려 모델의 과적합을 완화시켜 성능을 약간 개선시킬 수 있다는 사실을 확인하였다.

향후, 이미지 분류를 수행하는 모델 이외에도 다양한 목적 함수를 해결하는 AI 모델에 경량화를 적용한 이후, 배포되는 플랫폼에 경량화가 적용된 AI 모델을 운영하여 처리 속도와 성능을 분석할 예정이다.

## ACKNOWLEGEMENT

이 논문은 2025년 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2025-25442867, SDx 지능형 서비스의 최적 실행을 위한 생성형AI 지원 시스템SW 프레임워크 기술개발)

## 참 고 문 헌

[1] H.-I Liu *et al.*, "Lightweight deep learning for resource-constrained environments: A survey," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1-42, Jun. 2024.

[2] J. Yang, R. Shi and B. Ni, "MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis," 2021, in *arXiv:2010:14925*.