경량형 에지 DNN을 위한 스케일링된 가중치 정규화 기반 사후학습 양자화 기법

반종희, 권경필, 유준혁*

대구대학교

smilebjh@daegu.ac.kr, kyungpil08@gmail.com, joonhyuk@deagu.ac.kr

Post-training Quantization Technique Based on Scaled Weight Normalization for Lightweight Edge DNN

Jong-Hee Ban, Kyung-Pil Gwon, Joonhyuk Yoo*

Daegu University

요 약

기존 양자화 방법은 양자화 오류를 줄이기 위해 학습된 모델의 양자화를 거친 후에 학습 파라미터의 미세 조정 과정이 요구된다. 특히 메모리나 에너지 자원이 제한된 임베디드 IoT 및 모바일 장치에서는 복잡한 DNN 모델의 방대한 연산 처리가 어렵고, 저장 공간이 부족하거나 개인정보 보호 및 보안 등의 문제로 인해 학습 데이터에 대한 접근자체가 어려울 수 있다. 본 논문에서는 이러한 문제를 해결하기 위해 기존 가중치 분포의 동적 범위를 조정하는 가중치 정규화 기법에 스케일링 계수를 도입하여 긴 꼬리를 가진 가중치 분포로부터 발생하는 양자화 오류의 영향을 줄이는 스케일링된 가중치 정규화 기반 사후학습 양자화 기법을 제안한다. 실험을 통해 추가적인 학습이나 미세 조정 없이 기존 가중치 정규화 기법에 비해 성능을 향상시키고 4비트 양자화의 경우 완전정밀도 대비 단 1~2%의 성능차이로 즉각적인 양자화가 가능함을 입증한다.

Ⅰ. 서 론

DNN의 방대한 메모리 사용량 및 계산 비용은 컴퓨팅 자원이 제한된 임베디드 장치에 학습된 모델을 실질적으로 배포하고 사용하는데 어려움을 준다. 이러한 문제를 해결하기 위해 양자화(Quantization) 기반의 모델 압축 기법들이 연구되고 있다. 양자화 기법은 가중치를 표현하는 정밀도를 32비트에서 최대 1비트까지 줄임으로써 DNN 모델의 크기 감소와 연산 속도 향상 및 메모리 사용량을 감소시켜 에너지 소모량을 줄일 수 있다는 장점이 있다.

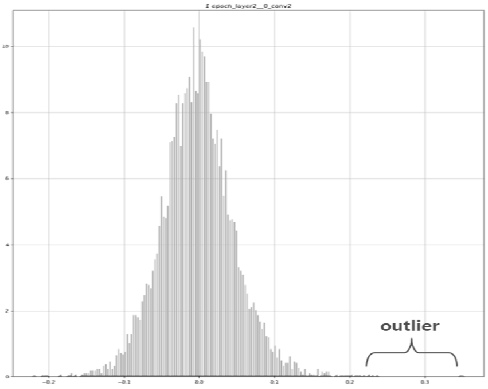

그림 1. DNN 가중치 분포의 아웃라이어
Fig. 1. Outlier of DNN weight distribution

그러나 학습된 DNN 가중치의 분포는 그림 1과 같이 양 끝단에 분포의 평균에 비해 매우 큰 아웃라이어(outlier) 값들이 존재하며 이는 각 계층별로 상이한 형태로 발생한다. 이로 인해 전체 가중치 분포를 균일한 간격의 양자화 포인트에 맵핑하는 균일 양자화 방식에서 양자화의 동적 범위(dynamic range)가 넓어지고 양자화 해상도가 낮아져 양자화 오류가 증가하는 문제가 발생한다. 이를 해결하기 위해 아웃라이어 처리를 위한 가중치 정규화(weight normalization) 혹은 클리핑(clipping) 기법을 사용하는 양자화 기법이 연구되고 있다[1,2].

그러나 이와 같은 양자화 기법들은 정규화 요소값 또는 클리핑 임계값의 최적화를 위해 학습 과정이 요구된다. 그러나 임베디드 장치에서 이러한 학습과정은 오랜 시간을 소요하며 방대한 연산량으로 인해 학습이 불가능한 문제가 발생하여 학습이나 미세조정 없이 바로 적용 가능한 사후학습(post-training) 기반의 양자화 기법이 연구되고 있다[3]. 그러나 사후 학습 양자화 역시 저정밀도 양자화에서 성능 저하가 발생하며 이를 보상할 방법이 없다는 측면에서 어려움이 있다.

본 논문에서는 사후 학습 양자화에 기존 가중치 정규화 기법을 적용할 경우에 발생하는 아웃라이어로 인한 성능 손실을 줄이기 위해 기존 가중치 정규화 기법에 스케일링 계수를 적용한 후에 클리핑을 수행하여 아웃라이어를 처리함으로써 양자화 성능을 향상시키는 스케일링된 가중치 정규화 기반의 사후 학습 양자화 기법(SWNQ: Scaled Weight Normalization based Quantization)을 제안하고 실험을 통해 우수성을 입증한다.

Ⅱ. 본론

학습된 모델의 각 계층별 가중치 분포는 모두 다르며, 특정 계층의 아웃라이어 값이 분포 평균값보다 월등히 클 경우 기존의 최대 절대값 기준으로 가중치를 정규화하는 WNQ 방법[1]에서는 0 근처로 많은 가중치가 이동하게 되어 손실되는 정보가 많아지게 된다. 클리핑 역시 임계값이 적절하지 않을 경우 높은 손실을 유발할 수 있다. 본 논문에서 제안된 아이디어는 가중치 정규화 과정에서 드물게 존재하는 아웃라이어로 인한 양자화 오류를 줄이기 위해 기존 가중치 정규화 과정에 스케일링 계수를 추가하

표 1. 제안된 SWNQ 방법의 양자화 과정
Table 1. Quantization process of proposed SWNQ method

	SWNQ(The Proposed Method)
step 1: normalizing	$\widehat{W} = \frac{W}{\max(\lvert W \rvert) \cdot \gamma}$
step 2: clipping	$clip(\hat{w}, 1) = \begin{cases} \hat{w} & \text{if } \lvert\hat{w}\rvert \le 1 \\ sign(\hat{w}) \cdot 1 & \text{if } \lvert\hat{w}\rvert \ge 1 \end{cases}$
step 3: rounding	$\widehat{W}^{Q} = \Pi(\widehat{W} / n_q)$
step 4: dequantizing	$W^{Q} = (\widehat{W}^{Q} \cdot n_q) \cdot \max(\lvert W \rvert) \cdot \gamma$

여 아웃라이어의 영향을 줄여주는 스케일링된 가중치 정규화 방식(SWNQ)을 제안한다.

표 1의 제안하는 SWNQ의 양자화 과정을 살펴보면 1단계 정규화 과정에서 γ는 각 계층의 최대 절대값의 크기를 줄이기 위한 스케일링 계수로써 [0, 1] 사이의 범위를 가진다. γ값에 따라 정규화 항의 크기가 줄어들면서 정규화된 가중치 $\widehat{W}$의 가중치 값은 [-1, 1]의 범위보다 작거나 클 수 있기 때문에 라운딩 범위를 벗어나게 된다. 이러한 현상을 막기 위해 라운딩과정 전에 2단계에서 클리핑을 먼저 수행하여 가중치 $\widehat{W}$가 [-1, 1]사이의 값을 가지도록 한다. 마지막으로 4단계 dequantization 과정에서도 γ값을 추가하여 보정한다.

클리핑 과정에서 정보손실이 발생할 수 있지만 아웃라이어로 인한 영향을 감소시킴으로써 가중치 정규화 과정에서 발생하는 양자화 오류를 줄이고 사후 학습 기반의 균일 양자화에 적합한 가중치 정규화를 수행할 수 있다.

제안된 SWNQ의 우수성을 입증하기 위해 CIFAR10, MNIST 데이터셋과 ResNet20 모델을 기반으로 정규화 기법에 따른 양자화 성능 비교 실험을 수행하였다. 기존 양자화 기법과 같이 첫 번째 레이어와 마지막 레이어는 양자화를 수행하지 않는다. 제안하는 스케일링된 가중치 정규화 기반의 사후 학습 방식 균일 양자화 기법의 스케일링 계수 γ 값은 0.5를 기준으로 값을 3, 4비트의 경우 ±0.05, 2비트의 경우 ±0.01씩 변경하며 실험을 수행하였다.

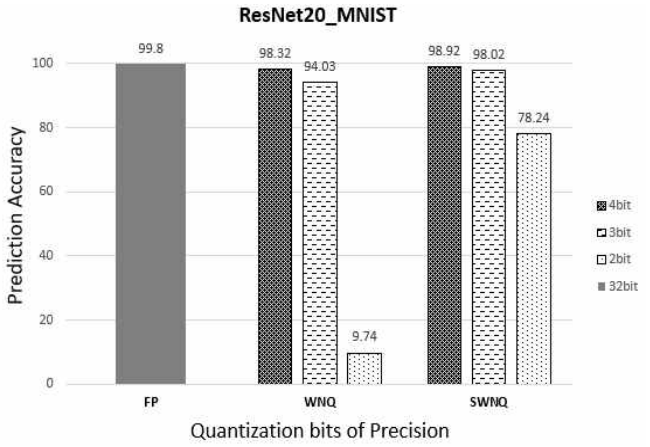

그림 2. MNIST 데이터셋에서 양자화된 ResNet20의 성능 비교
Fig. 2. Accuracy comparison of the quantized ResNet20 model on MNIST

먼저 그림 2의 MNIST에 대한 실험 결과를 살펴보면 SWNQ의 경우 γ가 0.9일 때 완전정밀도에 비해 약 0.9%의 정확도 감소만으로도 4비트 양자화가 즉시 가능한 것을 볼 수 있다. WNQ 방법의 경우 별도의 학습없이 양자화를 적용할 경우 4비트 정밀도에서는 약 1.5%정도의 미세한 손실을 보이지만 2비트 정밀도의 경우 9.74%로 약 90%의 성능 저하가 발생한다.

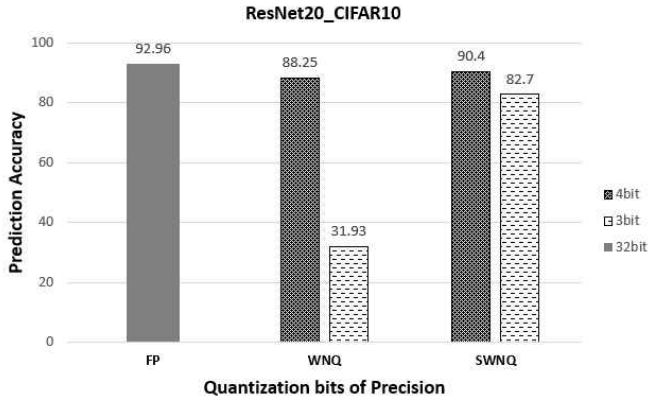

그림 3. CIFAR10 데이터셋에서 양자화된 ResNet20의 성능 비교
Fig. 3. Accuracy comparison of the quantized ResNet20 model on CIFAR10

그러나 SWNQ의 경우 2비트 정밀도에서 γ가 0.47일 때 78.24%로 WNQ에 비해 약 69% 향상된 성능을 보인다.

CIFAR10에 대한 실험 결과는 그림 3에 보이는 것과 같이 MNIST 데이터셋에 비해 성능 저하가 비교적 크게 일어난다. 하지만 4비트 정밀도에서 SWNQ는 CIFAR10의 경우 γ가 0.8일 때 WNQ에 비해 약 2.2% 성능이 향상되며 완전정밀도에 비해 약 2.5%의 성능 저하만으로도 양자화가 가능한 것을 보여준다. 하지만 3비트 저정밀도에서 WNQ 기반 양자화 방법의 성능 저하 문제가 극명하게 드러나는데, CIFAR10에서 WNQ의 경우 3비트 정밀도에서 61%의 성능 저하가 발생하는 반면, SWNQ의 경우 γ가 0.6일 때 약 10%의 성능 저하로 양자화가 가능한 것을 보여준다.

Ⅲ. 결론

본 논문에서 제안하는 SWNQ는 4비트 정밀도에서 완전정밀도 모델에 비해 약 1~2%의 성능손실만으로도 2시간에 가까운 학습 시간을 소요하지 않고 양자화가 가능한 우수한 성능을 보여준다. 향후 연구에서는 계층별 γ값을 적용하여 가중치 정규화를 수행하는 방법을 연구할 계획이다.

ACKNOWLEDGMENT

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020R1A2C1014768)

참 고 문 헌

[1] W. Cai, and W. Li, "Weight normalization based quantization for deep neural network compression," arXiv preprint arXiv:1907.00593, 2019.

[2] S. Jung, C. Son, S. Lee, J. Son, J. Han, Y. Kwak, and C. Choi, "Learning to quantize deep networks by optimizing quantization intervals with task loss," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4350-4359, 2019.

[3] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," In Advances in Neural Information Processing Systems, pp. 7950-7958, 2019.