논문 21-46-03-21 The Journal of Korean Institute of Communications and Information Sciences '21-03 Vol.46 No.03
https://doi.org/10.7840/kics.2021.46.3.583

SWNQ: 스케일링된 가중치 정규화 기반 사후학습 양자화 기법

반 근 우◆, 유 준 혁*

SWNQ: Scaled Weight Normalization Based Post-Training Quantization Method

Geun-Woo Ban◆, Joonhyuk Yoo*

요 약

사후 학습 양자화는 학습 과정이 요구되지 않아 학습데이터 의존성이 없다는 장점이 있지만 저정밀도에서 성능 저하가 크게 발생한다는 단점이 있다. 이러한 문제를 해결하기 위해 본 논문에서는 기존의 양자화에 사용되는 가중치 정규화 기법에 스케일링 계수를 도입하여 긴 꼬리를 가진 가중치 분포로부터 발생하는 양자화 오류를 줄이는 스케일링된 가중치 정규화 기반 사후학습 양자화 기법 SWNQ를 제안한다. 실험 결과는 SWNQ가 추가적인 학습이나 미세 조정 없이 기존 가중치 정규화 기반 양자화에 비해 양자화 성능을 향상시키면서 즉각적인 양자화가 가능함을 입증한다. 또한 SWNQ는 4비트 기반의 혼합 정밀도 양자화에서 완전정밀도 모델과 단 1.2%의 성능차이로 양자화가 가능한 것을 보여줌으로써 사후학습 양자화의 성능저하 문제를 효과적으로 해결할 수 있음을 입증한다.

Key Words : DNN(Deep Neural networks), Post-Training Quantization, Outlier, Model Compression, Quantization Sensitivity

ABSTRACT

Post-training quantization has an advantage that it does not require any training process and thus does not depend on the training data, but has a disadvantage in that its performance severely degrades especially at low precision. To solve aforementioned problem, in this paper, we propose a novel SWNQ(Scaled Weight Normalization based post-training Quantization) method that reduces quantization errors arising from the long-tailed weight distribution by introducing a scaling factor to the weight normalization technique used for the existing quantization methods. Experimental results demonstrate that the SWNQ can perform an immediate quantization while increasing quantization performance compared to the state-of-the-art weight normalization-based quantization with no further training or fine-tuning. Moreover, the SWNQ proves that it can effectively solve the performance degradation problem of post-training quantization by showing that the proposed method can be quantized by the performance gap of only 1.2% compared with a full-precision model in the 4-bit-based mixed-precision quantization.

※이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020R1A2C1014768).
◆ First Author : Daegu University Department of Information and Communication, smilebjh@daegu.ac.kr, 정회원
° Corresponding Author : Daegu University Department of Artificial Intelligence, joonhyuk@daegu.ac.kr, 정회원
논문번호 : 202012-337-B-RE, Received December 30, 2020; Revised January 7, 2021; Accepted January 7, 2021

Ⅰ. 서 론

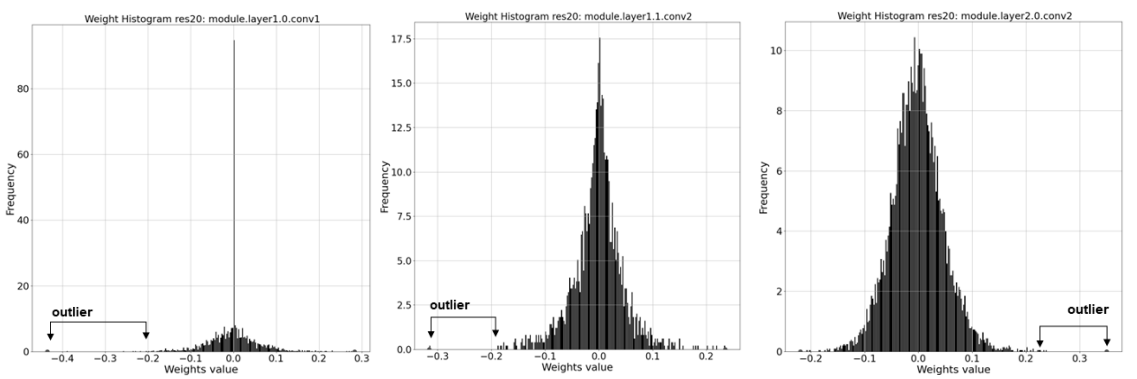
DNN의 방대한 메모리 사용량 및 계산 비용은 컴퓨팅 자원이 제한된 임베디드 장치에 학습된 모델을 실질적으로 배포하고 사용하는데 어려움을 준다. 이러한 문제를 해결하기 위해 양자화(Quantization) 기반의 모델 압축 기법들이 연구되고 있다[1]. 양자화 기법은 가중치를 표현하는 정밀도를 32비트에서 최대 1비트까지 줄임으로써 DNN 모델의 크기 감소와 연산 속도 향상 및 메모리 사용량을 감소시켜 에너지 소모량을 줄일 수 있다는 장점이 있다.

양자화 기법은 수를 표현하는 비트수에 따라 각 가중치를 특정 양자화 포인트에 맵핑하는 라운딩(Rounding) 과정을 거치는데, 양자화 포인트의 간격에 따라 균일 양자화와 비균일 양자화 기법으로 나뉜다[2-4]. 균일 양자화 기법은 모든 양자화 포인트의 간격이 동일하며 비균일 양자화 기법은 비선형 맵핑 함수에 따라 포인트의 위치를 다르게 생성한다. 양자화 포인트의 간격이 좁을수록 양자화 해상도가 높다고 할 수 있는데, 표현 비트수가 작아질수록 해상도가 낮아져 양자화로 인한 오류가 커지게 된다. 이러한 성능 저하 문제는 4비트 이하의 저정밀도로 갈수록 더 크게 발생하는 문제가 있다.

또한 실제 학습된 가중치의 분포는 그림 1과 같이 양 끝단에 분포의 평균에 비해 매우 큰 이탈값(outlier)들이 존재하며 이는 각 계층별로 상이한 형태로 발생한다. 이로 인해 양자화의 동적 범위(dynamic range)가 넓어지고 양자화 해상도가 낮아져 양자화 오류가 증가하게 된다. 이를 해결하기 위해 이탈값 처리를 위한 가중치 정규화(weight normalization) 또는 클리핑(clipping) 기법을 사용하는 양자화 기법이 연구되고 있다[5,6]. 기존에 사용되는 가중치 정규화와 배치 정규화(batch normalization)와의 차이점은 정규화가 컨볼루션 계층의 가중치와 입력(또는 이전 계층의 활성화 값)에 각각 적용된다는 점에서 다르다.

이러한 이탈값 처리 기반의 양자화 기법은 정규화 요소값 또는 클리핑 임계값을 찾기 위해 가중치와 양자화 파라미터의 공동 학습(jointly training) 또는 양자화 후의 미세조정(fine-tuning) 과정이 요구된다[7-9]. 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[10-12]. 사후 학습 양자화 기법의 경우 양자화를 적용한 후 손실된 성능의 보상을 위한 학습과정이 필요가 없어 학습에 시간을 소요하거나 학습데이터를 요구하지 않는다는 장점이 있다. 이는 사용자의 데이터 프라이버시(privacy)로 인해 데이터 수집이 불가하거나 데이터셋이 공개되지 않을 경우 학습이 불가능한 문제를 해결할 수 있다는 측면에서 매우 큰 장점이다. 그러나 저정밀도 양자화 시 성능 손실을 보상할 방법이 없다는 어려움이 있다.

DNN의 모든 계층을 동일한 저정밀도로 양자화할 경우, 성능저하가 현저하게 발생하는 것을 해결하기 위해 각 계층별로 다른 정밀도로 양자화를 수행하는 혼합 정밀도(mixed precision) 양자화 방법이 연구되고 있다[13,14]. DNN의 모든 계층이 양자화에 대해 동일한 민감도를 가지지 않기 때문에 민감도가 낮은 계층의 경우 양자화로 인한 성능저하가 적게 발생하고, 반대의 경우에는 성능저하가 크게 발생한다. 이러한



(a) Residual Block_1_layer_1 (b) Residual Block_2_layer_2 (c) Residual Block_7_layer_2

그림 1. DNN 가중치 분포의 이탈값 사례들
Fig. 1. Outliers in DNN Weight Distribution

관측을 기반으로 계층별 민감도에 따라 양자화 정밀도를 달리하여 성능을 향상시킬 수 있다. 양자화에 대한 민감도를 측정하는 방법으로는 쿨백-라이블러 발산(Kullback-Leibler divergence)이나 헤시안(Hessian) 기반 방법이 사용되고 있다[15-17].

본 논문에서는 양자화의 성능저하를 유발하는 이탈값을 처리하는 스케일링된 가중치 정규화 기반의 사후 학습 양자화 기법(SWNQ: Scaled Weight Normalization based Quantization)을 제안한다. 제안하는 방법은 기존 양자화에 사용되는 가중치 정규화 방법에 스케일링 계수를 도입하여 이탈값이 양자화에 미치는 영향을 줄여 좁은 동적범위를 만들어줌으로써 사후학습 양자화의 성능을 향상시킨다. 또한 SWNQ 기법이 양자화 민감도를 낮추는 효과가 있음을 실험을 통해 입증하고, 양자화 민감도에 따라 혼합정밀도를 사용하고 계층별 스케일링 계수를 조절하여 4비트 기반의 혼합 정밀도 양자화에서 32비트 완전정밀도 모델과 약 1.2%의 성능차이로 양자화를 할 수 있음을 보여준다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존 양자화에 사용되는 가중치 정규화 기법과 제안하는 스케일링된 가중치 정규화 기법, 그리고 혼합 정밀도 양자화를 위한 민감도 측정 방법을 설명한다. 3장에서는 제안하는 가중치 정규화 방법과 기존 가중치 정규화 방법 기반의 사후학습 양자화의 실험 결과를 비교한다. 4장에서는 결론 및 향후 연구를 제시한다.

Ⅱ. 제안하는 가중치 정규화 기반 양자화 방법

2.1 균일 양자화의 가중치 정규화 기법

기존 양자화에 사용되는 가중치 정규화 기법(WNQ: Weight Normalization based Quantization)[5]은 표 1의 좌측과 같은 3단계로 진행되며 각 단계는 다음과 같이 수행된다. 첫 번째로 가중치 값을 [-1, 1] 사이의 범위를 가지는 값으로 정규화하기 위해 각 가중치 값을 전체 가중치의 최대 절대값으로 나눈다. 이때 계산은 각 계층에 존재하는 모든 가중치의 요소별(element-wise)로 적용된다. 다음은 정규화된 가중치 $\hat{W}$를 균일한 간격의 양자화 값으로 라운딩하는 과정을 거친다. 여기서 $\Pi(\bullet)$는 라운딩 함수이며, n_q는 $2^{bitwidth}-1$값을 가지는 양자화 포인트의 개수이다. 라운딩은 정규화 된 가중치 $\hat{W}$를 $\frac{1}{n_q}$ 간격을 가지는 균일한 양자화 포인트로 맵핑한다. 마지막으로 양자화된 가중치 $\hat{W}^Q$를 원래의 가중치 값의 범위로 되돌리는 과정을 수행한다.

학습 기반의 양자화에서 양자화는 학습의 순전파 과정에만 적용되며 W^Q값으로 계산한 손실함수 값을 기반으로 완전 정밀도 가중치 W를 학습한다. 가중치 정규화 또는 클리핑 과정에서 발생하는 가중치 정보의 손실은 가중치를 학습함으로써 보상될 수 있지만 학습된 가중치에 양자화를 적용하는 사후 학습 기반의 양자화 방식에서는 이러한 학습 기반의 보상이 불가능하다. 또한 각 계층별 가중치 분포는 모두 다르며, 특정 계층의 이탈값 이 분포 평균값보다 월등히 클 경우 1단계 정규화 과정에서 대부분의 가중치가 매우 작은 값으로 정규화된다. 반면, 이탈값으로 인해 넓은 동적범위에서 양자화 포인트가 생성되기 때문에 2단계 라운딩 과정에서 0 근처의 양자화 포인트에 집중적으로 라운딩되어 성능 저하가 발생하는 원인이 된다.

2.2 스케일링된 가중치 정규화 기반의 사후 학습 양자화 기법

본 논문에서 제안된 아이디어는 가중치 정규화 과정에서 드물게 존재하는 이탈값으로 인한 양자화 오류를 줄이기 위해 기존 가중치 정규화 과정에 스케일

표 1. 기존 양자화 기법과 제안된 방법의 양자화 방법 비교
Table 1. Comparison of the existing quantization method and the proposed SWNQ

	WNQ[5]		SWNQ(The Proposed Method)
step 1: normalizing	$\hat{W}=\frac{W}{\max(\lvert W\rvert)}$	step 1: normalizing	$\hat{W}=\frac{W}{\max(\lvert W\rvert)\bullet\gamma}$
		step 2: clipping	$clip(\hat{w},1)=\begin{cases}\hat{w} & \text{if } \lvert\hat{w}\rvert\le 1\\ sign(\hat{w})\bullet 1 & \text{if } \lvert\hat{w}\rvert\ge 1\end{cases}$
step 2: rounding	$\hat{W}^Q=\Pi(\hat{W}/n_q)$	step 3: rounding	$\hat{W}^Q=\Pi(\hat{W}/n_q)$
step 3: denormalizing	$W^Q=(\hat{W}^Q\bullet n_q)\bullet\max(\lvert W\rvert)$	step 4: denormalizing	$W^Q=(\hat{W}^Q\bullet n_q)\bullet\max(\lvert W\rvert)\bullet\gamma$

링 계수를 추가하여 이탈값의 영향을 줄여주는 스케일링된 가중치 정규화 방식(SWNQ)을 제안한다.

표 1의 우측 1단계 정규화 과정에서 γ는 최대 절대값의 크기를 줄이기 위한 스케일링 계수로써 [0, 1] 사이의 범위를 가진다. γ값에 따라 정규화 항의 크기가 줄어들면서 정규화된 가중치 $\hat{w}$의 가중치 값은 [-1, 1]의 범위보다 작거나 클 수 있기 때문에 라운딩 범위를 벗어나게 된다. 이러한 현상을 막기 위해 WNQ와 달리 라운딩과정 전에 2단계에서 클리핑을 먼저 수행하여 가중치 $\hat{w}$가 [-1, 1]사이의 값을 가지도록 한다. 마지막으로 4단계 역정규화(denormalizing) 과정에서도 γ값을 추가하여 보정한다.

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

2.3 쿨백-라이블러 발산 기반의 양자화 민감도 측정 방법

아래 식은 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)을 사용하여 완전정밀도 모델과 양자화 된 모델간의 양자화 민감도를 측정하는 방법을 나타낸다.

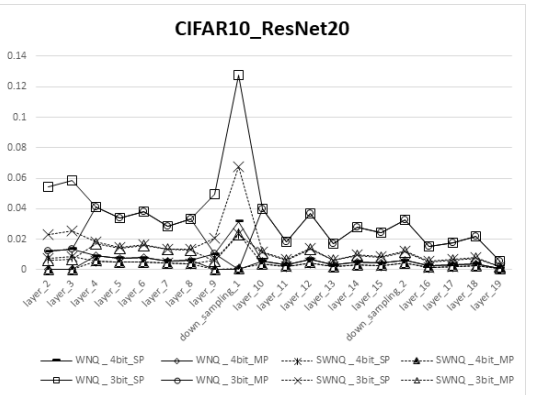
$$\Omega_i(k) = \frac{1}{N}\sum_{j=1}^{N} KLD(M(W_i;x_j), M(W_i^Q(k);x_j)) \quad (1)$$

식 (1)에서 $\Omega_i(k)$는 k비트로 양자화된 i번째 계층의 민감도이며 N개의 학습데이터에 대한 완전정밀도 모델과 양자화된 모델간의 쿨백-라이블러 발산값으로 계산된다. $W_i^Q(k)$는 k비트로 양자화된 i번째 계층의 가중치를 뜻한다. $\Omega_i(k)$값이 작은 계층의 경우 k비트 양자화에 대한 결과가 완전정밀도 모델과 유사하다고 볼 수 있기 때문에 더 낮은 비트로 양자화가 가능하다고 볼 수 있다. 반대의 경우에는 양자화로 인한 손실이 크다고 볼 수 있기 때문에 더 높은 비트로 양자화를 할 경우 성능 저하를 감소시킬 수 있다.

Ⅲ. SWNQ 성능 평가 결과

제안된 SWNQ의 우수성을 입증하기 위해 CIFAR10과 CIFAR100 데이터셋에 대해 ResNet20 모델을 기반으로 정규화 기법에 따른 양자화 성능 비교 실험을 수행하였다. 완전정밀도 ResNet20 모델의 경우 미리 학습된 가중치를 사용하였다. 또한 기존 양자화 기법과 같이 첫 번째 레이어와 마지막 레이어는 양자화를 수행하지 않았다[11]. 제안된 스케일링된 가중치 정규화 기반의 사후 학습 방식 균일 양자화 기법의 스케일링 계수 γ 값은 0.5를 기준으로 값을 ±0.05씩 변경하며 최적의 γ값을 휴리스틱에 따라 결정하였다. 양자화 민감도는 양자화 전 단계에서 각 데이터셋의 학습데이터 5만장을 입력으로 사용하여 양자화 민감도를 계산하였고 민감도 추정에 소요되는 시간은 약 6.7초가 소요되었다. 또한 테스트데이터 1만장에 대한 추론에 소요되는 시간은 사후학습 양자화 과정을 포함하여 CIFAR10 데이터셋의 경우 약 1.5초, CIFAR100 데이터셋의 경우 약 3.7초가 소요되었다. 학습 기반 양자화가 CIFAR10 기준 약 2~3시간이 소요되는 것을 감안하면 사후학습 양자화가 매우 빠르다는 것을 알 수 있다.

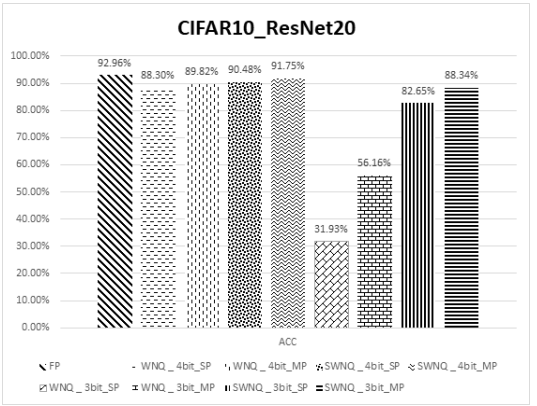
그림 2는 CIFAR10에 대한 양자화된 ResNet20 모델의 계층별 양자화 민감도를 나타낸다. 여기서 SP는 Single Precision으로 모든 계층이 동일한 비트로 양자화가 되었음을 나타내며 MP는 Mixed Precision으로 혼합 정밀도 양자화를 뜻한다. 먼저 단일 정밀도 SP 기반의 양자화 결과를 보았을 때, 3비트 및 4비트 양자화 모두 민감도를 오름차순으로 정리했을 때 (2, 3, 9, down_sampling_1) 4개 계층이 높은 민감도를 보이고 있으며 해당 계층에 8비트 양자화를 적용하여 혼합정밀도(MP) 양자화를 수행하였다. 또한 제안하는 SWNQ가 적용된 결과의 민감도를 살펴보면 각 양자



* SP : Single Precision / MP : Mixed Precision

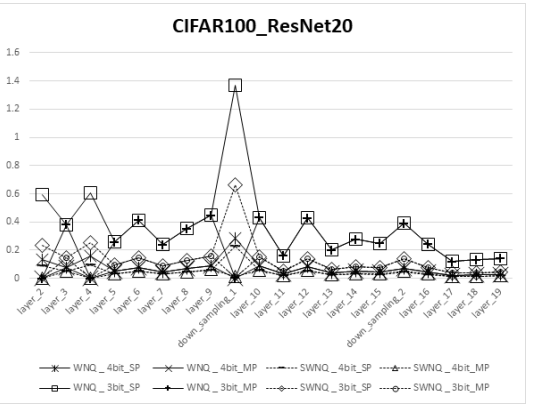
그림 2. CIFAR10에 대한 양자화된 ResNet20의 양자화 민감도 비교

Fig. 2. Sensitivity comparison of the quantized ResNet20 on CIFAR10



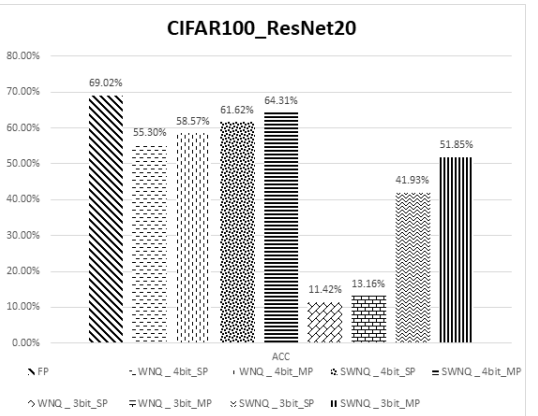
* SP : Single Precision / MP : Mixed Precision

그림 3. CIFAR10에 대한 양자화된 ResNet20의 성능 비교
Fig. 3. Accuracy comparison of the quantized ResNet20 on CIFAR10



* SP : Single Precision / MP : Mixed Precision

그림 4. CIFAR100에 대한 양자화된 ResNet20의 양자화 민감도 비교
Fig. 4. Sensitivity comparison of the quantized ResNet20 on CIFAR100



* SP : Single Precision / MP : Mixed Precision

그림 5. CIFAR100에 대한 양자화된 ResNet20의 성능 비교
Fig. 5. Accuracy comparison of the quantized ResNet20 on CIFAR100

화 결과에서 모두 기존 가중치 정규화 기반 방식 WNQ보다 전반적으로 민감도가 낮아지는 것을 볼 수 있는데, 이는 제안하는 방법이 이탈값을 처리함으로써 양자화로 인한 성능 저하를 해결함으로써 민감도를 감소시키는 효과가 있다고 볼 수 있다.

그림 3은 단일 정밀도 방식과 혼합 정밀도 방식 양자화에서 WNQ 기반 균일 양자화와 제안하는 SWNQ가 적용된 균일 양자화 모델의 성능을 비교하는 그래프이다. 먼저 4비트 단일 정밀도 양자화에서 제안하는 SWNQ는 γ값을 0.8로 적용했을 때 WNQ 기반의 양자화 대비 약 2.2%가량 성능이 향상되는 것을 볼 수 있다. 이때, 완전정밀도 대비 약 2.5%의 성능하락이 발생하지만 혼합 정밀도 기반의 SWNQ의 경우 민감도가 높은 계층을 8비트로 양자화 함으로써 완전정밀도 대비 1.2%의 성능 차이로 91.75%의 정확도를 달성하는 것을 볼 수 있다. 하지만 3비트 저정밀도에서 사후학습 양자화 방법의 성능 저하 문제가 극명하게 드러나는데, WNQ의 경우 3비트 단일정밀도에서 완전정밀도 대비 61%의 성능 저하가 발생하는 반면, SWNQ의 경우 γ가 0.6일 때 약 10%의 성능 저하로 양자화가 가능한 것을 보여준다. 이러한 성능저하는 혼합정밀도를 적용하였을 때 WNQ의 경우 약 24% 정도 성능이 향상되긴 하지만 여전히 완전정밀도에 비해 약 37%의 성능저하가 발생한다. 반면에 제안하는 SWNQ의 경우 혼합정밀도를 적용하였을 때 약 6%가량 성능이 향상되면서 완전정밀도와 단 4%의 차이로 3비트 기반의 혼합정밀도 양자화가 가능한 것을 보여준다.

그림 4와 5는 CIFAR100 데이터셋 기반의 실험 결과를 나타낸다. 양자화 민감도의 경우 CIFAR10 실험 결과와 전반적으로 유사한 결과를 보이고 있으며 CIFAR100이 클래스 개수가 CIFAR10의 10배인 것을 감안했을 때 민감도가 더 높게 나타나고 성능저하 역시 더 크게 나타나는 것을 볼 수 있다. CIFAR100의 혼합정밀도 양자화는 민감도가 높게 나타나는 (2, 4, down_sampling_1) 3개 계층에 대하여 8비트 양자화를 적용하였다.

우선 4비트 단일정밀도 양자화에서 비교적 완전정밀도와 유사한 성능을 보였던 CIFAR10 결과와 달리 CIFAR100에서는 WNQ의 경우 약 14%의 성능저하가 발생한다. 그러나 SWNQ의 경우 γ값이 0.81로 적용되었을 때 WNQ대비 약 6%의 성능 향상을 보이며,

4비트 기반 혼합정밀도의 경우에는 완전정밀도와 약 4.7%의 차이로 성능이 향상된 것을 보인다. 3비트 단일정밀도 양자화에서는 CIFAR10에 비해 성능이 매우 현저하게 떨어지는 것을 볼 수 있는데 WNQ의 경우에는 혼합정밀도를 적용하였음에도 불구하고 완전정밀도 대비 약 56%가량의 성능저하가 발생한다. 반면, 제안하는 SWNQ의 경우에는 γ가 0.55일 때 3비트 단일정밀도에서는 완전정밀도 대비 약 27% 성능저하가 발생하지만 WNQ 대비 약 30%가량 성능이 향상되는 것을 알 수 있고, 혼합정밀도를 적용하였을 때 약 10% 성능이 더 향상되는 것을 볼 수 있다. 이는 제안하는 방법이 사후학습 양자화의 성능저하 문제를 해결함과 동시에 민감도 기반의 혼합정밀도 양자화의 성능 역시 향상시킬 수 있음을 나타낸다.

그림 6과 7은 각 양자화 결과에 따른 가중치 히스토그램을 나타낸다. 그림 6의 WNQ 결과의 경우 음수 영역에서 매우 작은 값의 이탈값이 존재하고 있으며 이로 인해 넓은 양자화 동적범위가 설정되고 해당 영역에서 라운딩이 되지 않은 양자화 포인트가 낭비되는 것을 볼 수 있다. 결과적으로 0부근에서 높은 가중치 빈도수를 볼 수 있다. 반면에 그림 7의 SWNQ 결과를 살펴보면 동일 계층에서 음수 영역에 있던 이탈값이 줄어들게 되면서 양자화 동적범위가 좁아지고 WNQ에 비해 각 양자화 포인트에 고르게 가중치가 맵핑되는 것을 볼 수 있다. 이러한 현상은 이탈값이 존재하는 모든 계층에서 발생하고 있으며 이러한 이탈값 처리 과정에서 좁은 동적범위에서 가중치가 골고루 맵핑되면서 DNN 모델의 표현 능력이 향상된다고 볼 수 있다.

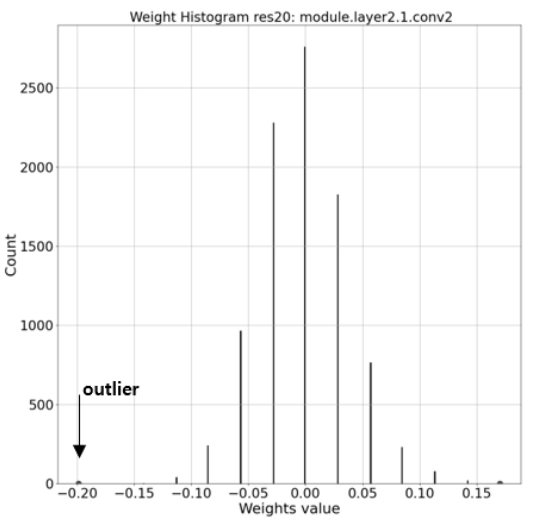


그림 6. WNQ 기반의 양자화된 가중치 히스토그램
Fig. 6. Quantized weight histogram based on WNQ

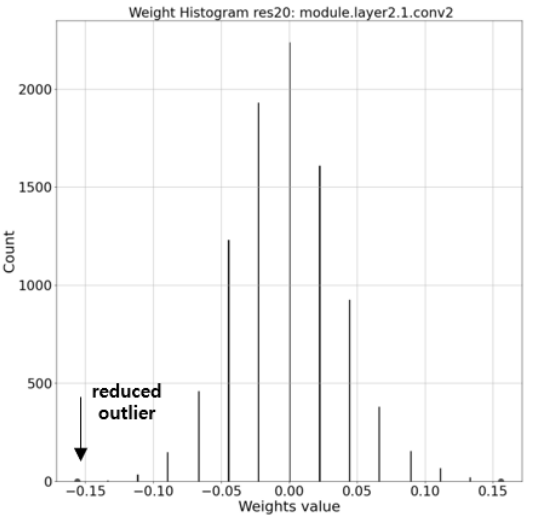


그림 7. SWNQ 기반의 양자화된 가중치 히스토그램
Fig. 7. Quantized weight histogram based on SWNQ

Ⅳ. 결 론

본 논문에서 제안하는 SWNQ는 4비트 정밀도에서 완전정밀도 모델에 비해 약 1.2%의 성능 손실만으로도 추가적인 학습 시간을 전혀 소요하지 않고 양자화가 가능한 우수한 성능을 보여준다. 제안된 SWNQ의 경우 하이퍼-패러미터(hyper-parameter)인 γ값에 따라 성능 차이가 컸으며 이는 이탈값이 양자화에 미치는 영향이 크다는 것으로 해석된다. 특히 정밀도가 줄어들 때, γ값이 작아질수록 성능이 향상되는 것을 관찰할 수 있었으며, 이는 저정밀도 양자화일수록 좁은 양자화 동적범위를 요구한다는 것을 알 수 있다. 향후 연구에서는 혼합정밀도의 압축률을 최소화 할 수 있도록 민감도에 따른 계층별 정밀도를 최적화하는 방안을 연구할 계획이다.

References

[1] D. Zhang, J. Yang, D. Ye, and G. Hua, "Lq-nets: Learned quantization for highly accurate and compact deep neural networks," in *Proc. ECCV*, pp. 365-382, 2018.

[2] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth

convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv: 1606.06160*, 2016.

[3] J. Choi and J. Yoo, "Exploiting inverse power of two non-uniform quantization method to increase energy efficiency in deep neural networks," *J. KIISE*, vol. 47, no. 1, pp. 27-35, 2020.

[4] C. Baskin, E. Schwartz, E. Zheltonozhskii, N. Liss, R. Giryes, A. Bronstein, and A. Mendelson, "Uniq: Uniform noise injection for non-uniform quantization of neural networks," *arXiv preprint arXiv:1804.10969*, 2018.

[5] W. Cai and W. Li, "Weight normalization based quantization for deep neural network compression," *arXiv preprint arXiv: 1907. 00593*, 2019.

[6] 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 *Proc. IEEE Conf. CVPR*, pp. 4350-4359, 2019.

[7] J. Choi, Z. Wang, S. Venkataramani, P. Chuang, V. Srinivasan, and K. Gopalakrishnan, "Pact: Parameterized clipping activation for quantized neural networks," *arXiv preprint arXiv:1805.06085*, 2018.

[8] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan, "Differentiable soft quantization: Bridging full-precision and low-bit neural networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019.

[9] Y. Li, X. Dong, and W. Wang, "Additive powers-of-two quantization: an efficient non-uniform discretization for neural networks," *Int. Conf. Learn. Representations*, 2020.

[10] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," in *Advances in NIPS*, pp. 7950-7958, 2019.

[11] M. Nagel, M. V. Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019.

[12] J. Ban, G. Kwon, and J. Yoo, "Post-training quantization technique based on scaled weight normalization for lightweight edge DNN," in *Proc. Symp. KICS*, pp. 279-280, Aug. 2020.

[13] K. Wan, Z. Liu, Y. Lin, J. Lin, and S. Han, "Haq: Hardware-aware automated quantization with mixed precision," in *Proc. IEEE Conf. CVPR*, 2019.

[14] C. Gong, Z. Jiang, D. Wang, Y. Lin, Q. Liu, and D. Z. Pan, "Mixed precision neural architecture search for energy efficient deep learning," in *Proc. IEEE/ACM ICCAD*, Westminster, CO, USA, 2019.

[15] Y. Cai, Z. Yao, and Z. Dong, "Zeroq: A novel zero shot quantization framework," in *Proc. IEEE/CVF Conf. CVPR*, 2020.

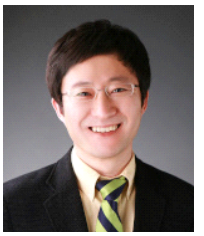
[16] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019.

[17] Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ-V2: Hessian Aware trace-Weighted Quantization of neural networks," *arXiv preprint arXiv:1911.03852*, 2019.

반 근 우 (Geun-Woo Ban)

2016년 2월 : 대구대학교 임베디드공학과 졸업
2018년 2월 : 대구대학교 정보통신공학과 석사
2018년 2월~현재 : 대구대학교 정보통신공학과 박사과정
<관심분야> 딥러닝, 컴퓨터비전, 임베디드 AI
[ORCID:0000-0002-6649-4934]

유 준 혁 (Joonhyuk Yoo)



1993년 2월 : 포항공과대학교 전자전기공학과 졸업
1995년 2월 : 포항공과대학교 전자전기공학과 석사
2007년 12월 : 메릴랜드대학교 컴퓨터공학과 박사
2009년~현재 : 대구대학교 정보통신대학 ICT융합학부 인공지능전공 교수.
<관심분야> 기계학습, 컴퓨터비전, 임베디드 딥러닝, 사이버물리시스템.
[ORCID:0000-0002-8311-5342]