JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 7, pp. 11-20, July 2024
https://doi.org/10.9708/jksci.2024.29.07.011

Bit-width Aware Generator and Intermediate Layer Knowledge Distillation using Channel-wise Attention for Generative Data-Free Quantization

Jae-Yong Baek*, Du-Hwan Hur*, Deok-Woong Kim*, Yong-Sang Yoo*, Hyuk-Jin Shin*, Dae-Hyeon Park*, Seung-Hwan Bae**

*Ph.D. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
*M. S. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
*M. S. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
*Ph.D. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
*M. S. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
*Ph.D. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
**Associate Professor, Vision & Learning Lab, Dept. of Computer Engineering, Inha University, Incheon, Korea

[Abstract]

In this paper, we propose the BAG (Bit-width Aware Generator) and the Intermediate Layer Knowledge Distillation using Channel-wise Attention to reduce the knowledge gap between a quantized network, a full-precision network, and a generator in GDFQ (Generative Data-Free Quantization). Since the generator in GDFQ is only trained by the feedback from the full-precision network, the gap resulting in decreased capability due to low bit-width of the quantized network has no effect on training the generator. To alleviate this problem, BAG is quantized with same bit-width of the quantized network, and it can generate synthetic images, which are effectively used for training the quantized network. Typically, the knowledge gap between the quantized network and the full-precision network is also important. To resolve this, we compute channel-wise attention of outputs of convolutional layers, and minimize the loss function as the distance of them. As the result, the quantized network can learn which channels to focus on more from mimicking the full-precision network. To prove the efficiency of proposed methods, we quantize the network trained on CIFAR-100 with 3 bit-width weights and activations, and train it and the generator with our method. As the result, we achieve 56.14% Top-1 Accuracy and increase 3.4% higher accuracy compared to our baseline AdaDFQ.

▸ **Key words**: Neural Network Quantization, Data-free Quantization, Generative model, Knowledge Distillation, Attention Mechanism

• First Author: Jae-Yong Baek, Du-Hwan Hur, Deok-Woong Kim, Yong-Sang Yoo, Hyuk-Jin Shin, Dae-Hyeon Park, Corresponding Author: Seung-Hwan Bae
*Jae-Yong Baek (jy1213@inha.edu), Vision & Learning Lab, Inha University
*Du-Hwan Hur (gjenghks@inha.edu), Vision & Learning Lab, Inha University
*Deok-Woong Kim (k5000plus@inha.edu), Vision & Learning Lab, Inha University
*Yong-Sang Yoo (cpp0094@inha.edu), Vision & Learning Lab, Inha University
*Hyuk-Jin Shin (shin0528@inha.edu), Vision & Learning Lab, Inha University
*Dae-Hyeon Park (saintPalite2221@inha.edu), Vision & Learning Lab, Inha University
**Seung-Hwan Bae (shbae@inha.ac.kr), Vision & Learning Lab, Dept. of Computer Engineering, Inha University
• Received: 2024. 05. 23, Revised: 2024. 06. 25, Accepted: 2024. 06. 26.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문에서는 생성 모델을 이용한 데이터 프리 양자화에서 발생할 수 있는 지식 격차를 줄이기 위하여 BAG (Bit-width Aware Generator)와 채널 어텐션 기반 중간 레이어 지식 증류를 제안한다. 생성 모델을 이용한 데이터 프리 양자화의 생성자는 오직 원본 네트워크의 피드백에만 의존하여 학습하기 때문에, 양자화된 네트워크의 낮은 bit-width로 인한 감소된 수용 능력 차이를 학습에 반영하지 못한다. 제안한 BAG는 양자화된 네트워크와 동일한 bit-width로 양자화하여, 양자화된 네트워크에 맞는 합성 이미지를 생성하여 이러한 문제를 완화한다. 또한, 양자화된 네트워크와 원본 모델 간의 지식 격차를 줄이는 것 역시 양자화에서 매우 중요한 문제이다. 이를 완화하기 위해 제안한 채널 어텐션 기반 중간 레이어 지식 증류는 학생 모델이 교사 모델로부터 어떤 채널에 더 집중해서 학습해야 하는지를 가르친다. 제안한 기법의 효율성을 보이기 위해, CIFAR-100에서 학습한 원본 네트워크를 가중치와 활성값을 각각 3-bit로 양자화하여 학습을 수행하였다. 그 결과 56.14%의 Top-1 Accuracy를 달성하였으며, 베이스라인 모델인 AdaDFQ 대비 3.4% 정확도를 향상했다.

▸ **주제어**: 뉴럴 네트워크 양자화, 데이터 프리 양자화, 생성 모델, 지식 증류, 어텐션 매커니즘

I. Introduction

최근 심층 신경망 (Deep Neural Network)이 비약적인 발전을 보임에 따라, 자율주행, 보안, 헬스케어 등 다양한 응용 분야에 활용하려는 시도가 연구되고 있다. 하지만 이러한 미션 크리티컬 시스템 (mission critical system)들은 빠른 응답속도를 필수적으로 요구하나, 심층 신경망의 높은 연산량으로 인하여 즉각적인 응답속도를 충족시키는 것은 큰 문제로 남아있다.

이를 해결하기 위해 인공 신경망 양자화 (Neural Network Quantization)[1-4]가 제안되었다. 양자화는 원본 네트워크 (full-precision network)의 활성값 (activation)과 가중치 (weight)의 bit-width를 낮춘 양자화된 네트워크 (quantized network)를 생성하여, 연산량과 에너지 소모를 급격히 감소시킬 수 있는 가장 대표적인 모델 압축 (model compression) 방법이다. 낮아진 bit-width로 인한 인공 신경망 모델의 정확도를 복원하기 위해, 원본 네트워크의 학습에 사용된 데이터의 전체 (fine-tuning)[5-7] 혹은 일부 (calibration)[2, 8, 9]의 이용이 필수적으로 요구된다. 불행히도, 보안과 헬스케어 등의 분야에서는 개인정보 혹은 윤리적인 이유로 이러한 원본 데이터에 접근할 수 없는 경우가 발생할 수 있다. 생성 모델을 이용한 양자화 (Generative Data-Free Quantization)[10-12]는 접근할 수 없는 원본 데이터를 대체하기 위해 생성자 (generator)를 원본 네트워크를 이용하여 학습하고, 이를 통해 생성한 합성 데이터로 양자화된 네트워크를 학습한다. 이때, 원본 네트워크와 양자화된 네트워크 간의 지식 격차를 줄이기 위해 지식 증류 (Knowledge Distillation)[13-15]를 사용할 수 있다. 인공 신경망 양자화에서 지식 증류는 원본 네트워크를 교사 모델로, 양자화된 네트워크를 학생 모델로 하여 특정 활성값 간의 거리를 줄이는 과정을 통해, 학생 모델이 교사 모델을 모방하여, 지식을 전이할 수 있다. 추가적인 레이어 (layer) 혹은 데이터 없이 학생 모델의 성능을 큰 폭으로 향상할 수 있다는 장점 때문에, 다양한 분야에서 이를 적용하려는 시도가 활발히 연구되고 있으나, 생성 모델을 이용한 양자화에 지식 증류를 활용하는 연구는 아직 제한적이다.

GAN (Generative Adversarial Networks)[16-18]에서 생성자와 구별자 (discriminator) 간의 큰 수용 능력 차이는 성능 감소로 이어진다[19, 20]. 생성 모델을 이용한 양자화에서 역시 양자화된 네트워크의 낮은 bit-width와 양자화 과정에서 발생하는 rounding error와 clipping error로 인하여 수용 능력이 낮아지므로, 이는 같은 문제가 발생할 수 있다.

이를 해결하기 위해, 본 논문에서는 BAG와 채널 어텐션 (channel-wise attention) 기반 중간 레이어 (intermediate layer) 지식 증류 기법을 제안한다. BAG (Bit-width Aware Generator)는 생성자를 양자화된 네트워크와 동일한 낮은 bit-width로 양자화하여, 같은 rounding error와 clipping error가 발생함으로써, 양자화된 네트워크의 수준에 맞는 적합한 합성 이미지들을 생

성할 수 있다. 또한, 양자화된 네트워크와 원본 네트워크 간의 지식 격차 역시 인공 신경망 양자화에서 매우 중요한 문제이다. 중간 레이어 지식 증류는 이를 해결하기 위한 효과적인 기법이지만, 자료형(정수형과 부동 소수점)과 bit-width의 차이로 기존 연구에서는 소프트 라벨 (soft label)을 주로 이용하여 지식 증류를 활용하였다[10, 12]. 이를 해결하기 위해, 우리는 우선 양자화된 네트워크의 중간 활성값들을 양자화 복원 (de-quantization) 과정 후 지식 증류를 수행하였다. 이때, 양자화 과정에서 발생한 rounding error와 clipping error로 학생 모델이 교사 모델을 완벽히 모사하는 학습 하는 것보다, 중요한 채널에 더 집중할 수 있도록 채널 어텐션을 사용하여 거리를 줄이도록 지식 증류를 수행한다.

제안한 기법들의 효과를 보이기 위해, ResNet-20[21] 모델을 CIFAR-100[22] 데이터셋으로 학습 후 3비트와 4비트로 양자화하였다[1]. 그 결과 우리가 제안한 기법은 56.14%와 67.47%의 Top-1 Accuracy를 달성하였으며, 베이스라인 모델인 AdaDFQ (Adaptive Data-Free Quantization)[12] 대비 3.40%와 0.66%의 성능 향상을 보였다.

본 논문의 주요 기여 사항은 다음과 같다. i) 양자화된 네트워크에 더 적합한 합성 이미지를 생성할 수 있는 Bit-width Aware Generator를 제안한다. ii) 생성 모델을 이용한 양자화에서 효과적인 채널 어텐션 기반 중간 레이어 지식 증류를 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존의 인공 신경망 양자화에 관한 연구를 설명한다. 3장에서는 본 논문에서 제안한 기법들에 관해 설명한다. 4장에서는 CIFAR-100 데이터셋에서 평가 및 최근 데이터 프리 양자화 (data-free quantization) 방법들과 비교를 수행한다. 5장에서는 본 논문의 결론을 도출한다.

II. Related Works

이 장에서는 본 연구와 연관된 기존 연구에 대해 기술한다.

1. Generative Data-Free Quantization

인공 신경망 양자화는 효율적인 deployment를 위한 가장 대표적인 모델 압축 기법 중 하나로 양자화된 네트워크의 정확도 복원을 위하여, fine-tuning이나 calibration 과정을 필수적으로 요구한다. 이때, 개인정보나 윤리적인 문제로 인하여 원본 데이터에 접근할 수 없는 문제를 해결하기 위해 데이터 프리 양자화 [10, 12, 23, 24]가 제안되었다. 데이터 프리 양자화에는 노이즈 최적화 (noise optimization) 데이터 프리 양자화[23, 24]와 생성 모델을 이용한 양자화[10, 12]가 있다. 노이즈 최적화 데이터 프리 양자화의 경우 노이즈를 그레디언트 (gradient)에 의해 업데이트하여 그럴듯한 이미지를 만들 수 있으나, 노이즈 수가 제한적이므로 원본 데이터의 분포를 캡처하는 데 어려움이 있다.

GDFQ (Generative Low-bitwidth Data Free Quantization)[10]는 생성자를 이용하여, 생성자가 원본 데이터와 같은 분포를 가진 합성 이미지들을 생성하기 위해, 합성곱 신경망의 활성값 분포가 원본 데이터의 분포와 일치하도록 BNS (Batch Normalization Statistics) 손실함수[10]를 제안하였다. BNS 손실함수는 인상적인 성능을 보였으나, 여전히 원본 데이터를 이용한 fine-tuning과는 큰 성능 차를 보인다. Qimera[25]는 latent embeddings를 사용하여 원본 네트워크의 분포를 반영하는 boundary supporting 샘플을 생성하고, 양자화된 네트워크가 구별하기 어려운 confusing 샘플을 생성하여 정확도를 높이는 학습 방법을 제안한다. DSG (Diverse Sample Generation)[26]는 BNS 손실함수로 인하여 합성 이미지의 다양성이 훼손되어 성능이 하락함을 주장하고 이를 해결하기 위해 데이터에 따라 특정 레이어를 강조하는 방법을 제안하였다. AIT (All In the Teacher)[27]는 다양한 손실함수를 사용한 최적화로 인한 학습의 불안정성과 합성 이미지로 인한 일반화 능력 저하 문제를 지적하며, 이를 해결하기 위해 크로스 엔트로피 (cross entropy) 없이 KL (Kullback-Leibler) 발산만을 이용하여 손실함수를 구성하고, 임계값을 이용하여 가중치를 제한하는 방법을 제안하였다. TexQ[35]는 원본 데이터 및 생성자에 의해 생성된 합성 이미지 간 텍스쳐 (texture) 분포 비율에 유의미한 차이가 있다는 점을 지적하였고, 이러한 텍스쳐 특성을 원본 데이터와 유사하게 개선할 수 있는 손실함수를 제안하였다. ACQ[36]는 합성 데이터에 어텐션 (attention) 기법을 적용하여 특성을 분석했을 때, 원본 데이터에 비해 동일 클래스 내 샘플 다양성이 부족하다는 점과 합성 데이터로 학생 모델을 학습할 경우, 원본 데이터와 달리 모델 네트워크 설정에 따라 모델 정확도 및 BNS 오류에 큰 차이가 있다는 점을 지적하여, 이러한 문제점들을 개선할 수

1) 가중치와 activation은 같은 비트로 양자화하였다.

있는 손실함수를 제안한다. AdaDFQ[12]는 생성자가 오로지 원본 네트워크에 의해 학습되어 과적합 문제가 발생할 수 있음을 지적하여, 교사와 학생 모델의 과적합과 과소적합 경계에 위치할 수 있는 손실함수를 제안하였다. 다중 구별자는 생성자의 mode collapse 문제 완화에 도움이 되므로[37], 교사와 학생 모델 양쪽의 예측 결과를 활용해 생성자의 강건성뿐만 아니라 mode collapse 문제도 완화한다. 기존 연구들은 BNS 손실함수 개선을 통한 생성자의 다양성과 과적합 문제만을 주로 다루며, bit-width 차이로 인한 양자화된 네트워크와 생성자 간의 수용 능력 차이를 줄이려는 연구는 부족한 상황이다.

2. Knowledge Distillation for Generative Data-Free Quantization

지식 증류[13-15]는 고성능의 교사 모델의 지식을 경량화된 학생 모델로 전이할 수 있는 효율적인 모델 압축 방법이다. Hinton et al.[13]는 학생 모델이 교사 모델이 예측한 소프트 라벨을 모사하는 방법을 제안하였다. 하지만 이러한 방법에도 불구하고 여전히 교사 모델과 학생 모델 간 큰 성능 격차가 존재함으로, 더 풍부한 지식을 전이하기 위해 중간 레이어를 활용하는 방법들이 제안되었다. FitNet[14]는 중간 레이어의 활성값을 최소화하는 방법을 제안하였다. 중간 레이어의 지식 증류의 경우 과적합 문제가 발생할 수 있으며, 교사 모델과 학생 모델 간의 수용 능력 차이가 큰 경우 성능이 저하되는 경우가 발생할 수 있다. 이를 해결하기 위해 AT[28]는 spatial attention map 간의 거리를 줄이는 방법을 제안하였다.

인공 신경망 양자화에서도 양자화된 네트워크의 성능을 복원하기 위해, 원본 네트워크를 교사 모델로, 양자화된 네트워크를 학생 모델로 지식 증류를 수행한다. 앞서 언급한 지식 증류 방법의 발전에도 불구하고, [10, 12]는 소프트 라벨을 이용한 지식 증류만을 사용하였다. HAST (Hard sample Synthesizing and Training)[24]는 노이즈 최적화 데이터 프리 양자화 방법 중 하나로, 양자화 복원 과정 후 중간 레이어의 채널 어텐션을 구한 후 지식 증류하는 방법을 제안하였다. 하지만 HAST는 노이즈 최적화 데이터 프리 양자화로 제한된 노이즈만 사용하기 때문에, 생성 모델을 이용한 데이터 프리 양자화에서는 해당 지식 증류 방법의 효과는 아직 연구되지 않았다.

III. Methodology

이 장에서는 기존의 인공 신경망 양자화와 본 논문에서 제안하는 BAG와 채널 어텐션 지식 증류 방법에 대해 설명한다.

1. Generative Data-Free Quantization

생성 모델을 이용한 양자화는 먼저 원본 네트워크를 이용하여 양자화된 네트워크를 생성한다. 이때, 각 레이어의 가중치 혹은 활성값을 n 비트로 양자화하며, 이는 수식 (1)로 정의된다.

$$x_q = \psi(x) \tag{1}$$

수식 (1)에서 입력 x는 원본 네트워크의 가중치 혹은 활성값을 나타내며, x_q는 양자화된 결과를 의미한다. $clamp(\cdot)$ 함수는 입력 값 v가 a,c를 상한값과 하한값으로 갖도록 범위를 제한하는 함수로 다음과 같이 정의된다.

$$clamp(v;a,c) = \begin{cases} a, v < a \\ v, a \le v \le c \\ c, v > c \end{cases} \tag{2}$$

함수 $\psi(\cdot)$는 quantizer로 다음과 같이 정의된다.

$$\psi(x) = clamp(Round(\frac{x}{s} + b); 0, 2^n - 1) \tag{3}$$

$$\text{where } s = \frac{\theta_{\max} - \theta_{\min}}{2^n - 1},\ b = -\frac{\theta_{\min}}{s}$$

$Round(\cdot)$함수는 반올림, s는 scaling factor, b는 zero point, $\theta_{\max}, \theta_{\min}$는 x가 속한 가중치 혹은 활성값의 최댓값과 최솟값이며, n은 양자화 비트 정밀도이다.

양자화된 네트워크의 학습을 위한 생성자는 다음과 같이 정의된다[16].

$$\hat{\boldsymbol{x}} = G(\boldsymbol{z}|y), \boldsymbol{z} \sim N(0,1) \tag{4}$$

G는 생성자, $\boldsymbol{z}$는 latent vector, y는 클래스 라벨, N은 가우시안 분포, $(0,1)$은 각각 가우시안 분포의 평균과 분산, $\hat{\boldsymbol{x}}$는 생성된 이미지를 의미한다. 이때, 생성자가 원하는 클래스 이미지를 생성하기 위한 classification 손실함수는 다음과 같이 정의된다.

$$\ell_{CE}^{G} = \mathbb{E}_{\boldsymbol{z},y}[CE(M(\boldsymbol{z}|y), y)] \tag{5}$$

$CE(\cdot)$는 크로스 엔트로피 손실함수, M은 원본 네트워크이다. 생성자는 생성한 이미지를 주어진 라벨에 맞게 분류할 수 있도록 합성 이미지를 생성해야 하는 것뿐만 아니라, 원본 이미지와 같은 분포를 갖는 이미지를 생성해야 한다. 이를 위해 BNS 손실함수[10]는 다음과 같이 정의된다.

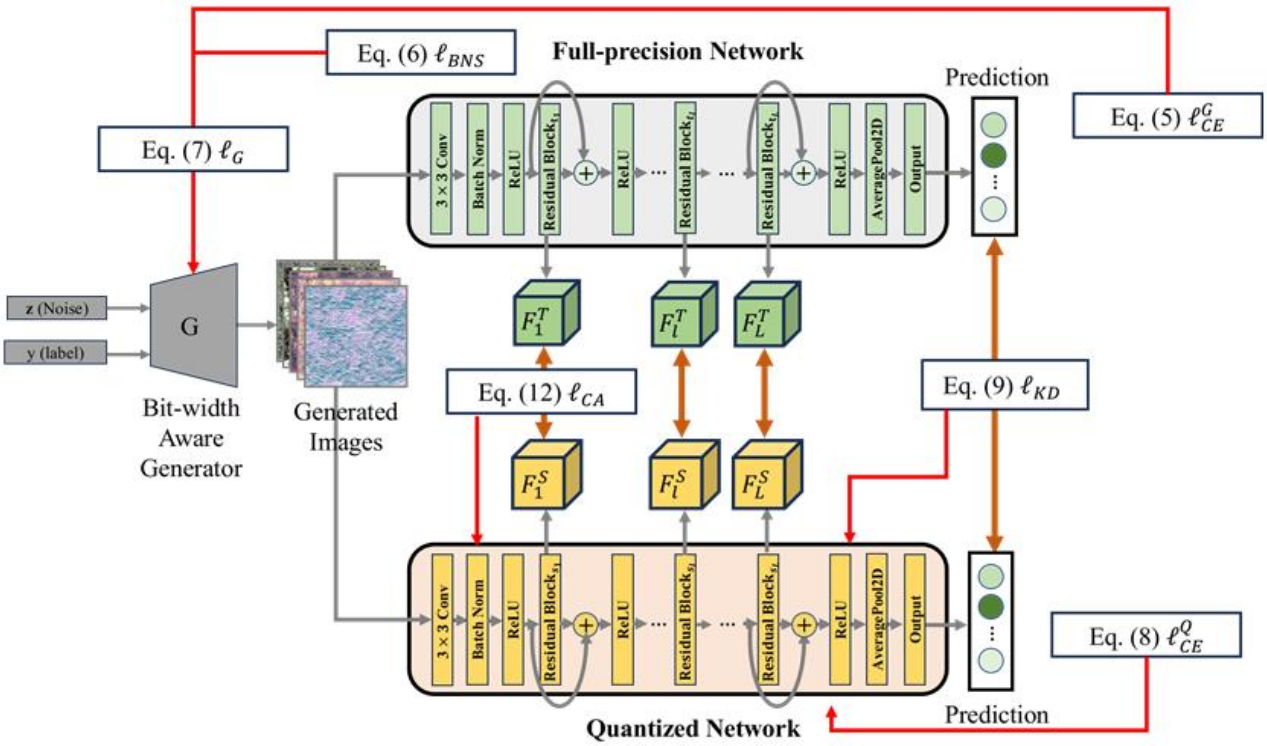

Fig. 1. The overall framework of proposed methods.

$$\ell_{BNS} = \frac{1}{L}\sum_{l=1}^{L} \| \mu_l^r - \mu_l \|_2^2 + \| \sigma_l^r - \sigma_l \|_2^2 \quad (6)$$

L은 Batch Normalization[29] 레이어의 수, μ_l^r는 EMA (Exponential Moving Average)에 의해 계산된 l번째 Batch Normalization 레이어의 원본 데이터의 평균, μ_l는 생성한 이미지에 대한 평균, σ_l는 각각의 분산을 의미한다. 위 두 손실함수에 의해 생성자의 손실함수는 다음과 같이 정의된다.

$$\ell_G = \ell_{CE}^{G} + \lambda \cdot \ell_{BNS} \quad (7)$$

λ는 BNS 손실함수의 비율을 결정하는 하이퍼 파라미터(hyper parameter)이다.

양자화된 네트워크를 Q로 정의할 때, 크로스 엔트로피 손실함수 ℓ_{CE}^{Q}와 지식 증류 손실함수 ℓ_{KD}는 다음과 같이 정의된다[10].

$$\ell_{CE}^{Q} = \mathbb{E}_{\hat{x},y}[CE(Q(\hat{x}), y)] \quad (8)$$

$$\ell_{KD} = \mathbb{E}_{\hat{x}}[KL(Q(\hat{x}), M(\hat{x}))] \quad (9)$$

KL은 Kullback-Leibler divergence 손실함수를 의미한다. ℓ_{KD}는 학생 모델의 학습을 위한 일반적인 지식 증류 과정을 표현한 수식으로, 각 훈련 데이터 $\hat{x}$에 대한 선생 모델의 결괏값 $M(\hat{x})$와 학생 모델의 결괏값 $Q(\hat{x})$의 KL 결괏값의 평균치를 최소화하여 선생 모델의 지식을 학생 모델이 습득할 수 있도록 유도하는 손실함수이다.

2. Bit-width Aware Generator and Intermediate Layer Distillation using Channel-wise Attention

Figure 1은 제안한 방법의 전체 프레임워크이다. 우선 생성자를 BAG로 대체하며, BAG로 생성한 이미지를 원본 네트워크와 양자화된 네트워크의 입력으로 하여, 중간 레이어의 활성 값들의 채널 기반 어텐션을 구하고 이를 지식 증류한다. 또한, 소프트 라벨을 이용한 지식 증류와 양자화된 네트워크의 classification 손실함수를 통해 학습한다. 각각 BAG와 채널 기반 중간 레이어 지식 증류는 다음 장에서 설명한다.

3. Bit-width Aware Generator

생성자와 구별자 간의 큰 수용 능력 차이는 성능 저하로 이어질 수 있다[19, 20]. 생성 모델을 이용한 양자화 역시 생성자와 양자화된 네트워크 간의 bit-width와 자료형(정수형과 부동 소수점)의 차이로 인하여 같은 문제가 발생할 수 있다. 일반적으로, 생성자는 원본 네트워크와 같은 FP32 (Floating Point 32 bits)를 사용한다. 예를 들어 3w3a (3w는 가중치를 3비트로, 3a는 활성값을 3비트로 양자화 함을 의미)로 양자화할 때, 표현 방식을 제외한다면 bit-width는 약 10.6배 차이이다. 이때 3비트는 III-1장의 n값이다. 이러한 bit-width의 차이는 각 모델의 수용 능력 차이와 지식 격차를 유발한다.

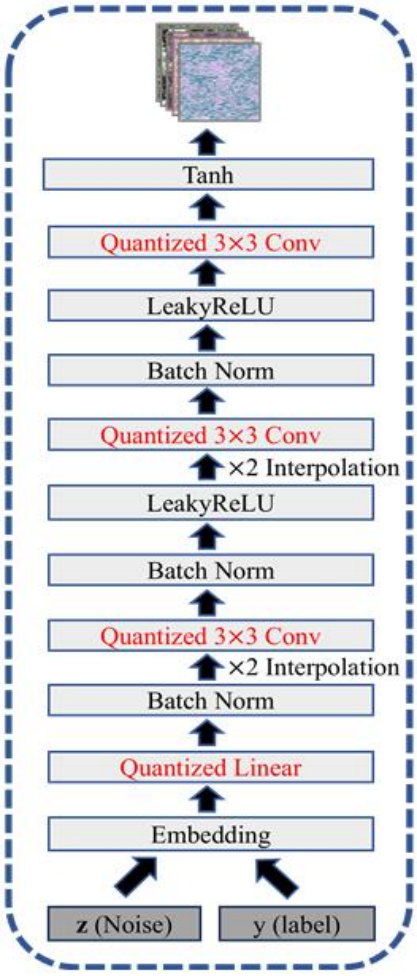

Fig. 2. The architecture of Bit-width Aware Generator. We highlight the main differences from the baseline with the red color.

이를 해결하기 위하여 생성자를 양자화된 네트워크와 동일한 bit-width 및 양자화 방법을 사용하여 양자화한다. Figure 2 는 BAG의 구조이다. 빨간색으로 표시된 부분이 양자화한 레이어이다. 우선 latent vector $\boldsymbol{z}$와 라벨 y를 합쳐 양자화된 선형 레이어 (linear layer)의 입력으로 사용한다. 이후 3개의 양자화된 컨볼루셔널 레이어 (convolutional layer)와 Batch Normalization 레이어, 그리고 Leaky ReLU (Rectified Linear Unit)[30] 활성 함수를 통과한 다음, 마지막으로 Tanh (Hyperbolic Tangent) 활성 함수를 통해 이미지를 생성한다. 각 선형 레이어와 컨볼루셔널 레이어의 가중치와 활성값들은 양자화된 네트워크와 동일한 bit-width와 방법으로 양자화하였다.

4. Intermediate Layer Distillation using Channel-wise Attention

중간 레이어 지식 증류를 위해 2개의 컨볼루셔널 레이어와 2개의 Batch Normalization 레이어, 하나의 ReLU 활성 함수로 구성된 ResNet[21]의 Residual Block의 출력들을 선택하였다.

특징맵 F와 전치행렬 F^T에 대해 채널 기반 어텐션 $Attn(F)$는 다음과 같다.

$$A = F \cdot F^T \tag{10}$$

$$Attn(F) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} A_{(i,j)} \tag{11}$$

W, H ($W = H$) 는 특징맵의 각각 width와 height이다. 채널 기반 어텐션을 이용한 손실함수 ℓ_{CA} 는 다음과 같이 정의된다.

$$\ell_{CA} = \frac{1}{L_B} \sum_{l=1}^{L_B} \| Attn(F_l^{Tea}) - Attn(F_l^{Stu}) \|_2^2 \tag{12}$$

L_B는 Residual Block의 수, F^{Tea}는 원본 네트워크의 특징맵, F^{Stu}는 양자화된 네트워크의 특징맵이다. 이렇게 구한 채널 기반 어텐션 손실함수를 식 (10), (11)과 함께 합산하여 구성한 양자화된 네트워크의 손실함수 ℓ_Q 는 다음과 같다.

$$\ell_Q = \ell_{CE}^Q + \ell_{KD} + \gamma \ell_{CA} \tag{13}$$

γ는 채널 어텐션 손실함수의 반영 비율을 조절하는 하이퍼 파라미터이다.

IV. Experiments

본 논문에서 제안한 방법의 효과를 보이기 위해, CIFAR-100[22] 상에서 3비트와 4비트 양자화 실험 및 비교를 수행하였다.

1. Experiment Settings

CIFAR-100 데이터셋은 100개의 클래스로 이뤄져 있는 이미지 분류 데이터셋이다. 학습 이미지 50,000장과 평가 이미지 10,000장으로 구성되어 있으며, 각 이미지는 32×32의 해상도를 갖는다. 성능 평가를 위해 이미지 분류에서 가장 널리 사용되는 metric인 Top-1 Accuracy를 사용하였다. 우리는 베이스라인 모델을 AdaDFQ[12]로 하여, ResNet-20[21]을 CIFAR-100에서 학습한 후 양자화를 진행하였다. BAG의 구현은 [18]의 생성자 네트워크 구조를 베이스라인으로 채택하였다.

양자화된 네트워크와 생성자의 학습을 위해 순서대로 Stochastic Gradient Descent optimizer와 Adam optimizer[31]를 사용하여 400 epoch 학습하였으며, 학

습률은 각각 10^{-4},10^{-3}이며, 양자화된 네트워크의 경우 100 epoch마다 학습률을 0.1 배씩 감소시켰다. 초기 생성자는 노이즈에 가까운 의미 없는 이미지를 생성하기 때문에 초기 4 epochs는 생성자만 학습하였으며[10], 이후 epoch에서는 두 네트워크를 모두 학습하였다. 그 외 하이퍼 파라미터의 경우 배치 사이즈는 16, 식 (9)의 λ는 0.1, 식 (15)의 γ는 30이다. [10, 12, 24]에 따라 매 epoch마다 테스트 데이터셋을 이용하여 양자화된 네트워크의 평가를 수행하였으며, 학습 중 가장 높은 성능을 최종 성능으로 정하였다. 모든 실험은 NVIDIA TITAN V GPU와 Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.4GHz 상에서 수행하였으며, PyTorch 1.4.0, CUDA 10.2를 사용하였다.

Table 1. Comparison with recent DFQ methods on CIFAR-100. All methods are quantized with 3-bits weights and activations. G and N indicate Generator and Noise optimization methods, respectively.

Method	Data	Publication	Top-1 Accuracy
GDFQ[10]	G	ECCV20	47.16
ZeroQ+IL[23]	N	CVPR20	26.35
DSG+IL[26]	G	CVPR21	43.42
ARC[32]	G	IJCAI21	40.15
Qimera[25]	G	NeurIPS21	46.13
IntraQ[33]	N	CVPR22	48.25
ARC+AIT[27]	G	CVPR22	41.34
AdaSG[34]	G	AAAI23	52.76
AdaDFQ[12]	G	CVPR23	52.74
HAST[24]	N	CVPR23	55.67
Ours	**G**	Proposed	56.14

Table 2. 4-bits quantization comparison with recent DFQ methods on CIFAR-100.

Method	Data	Publication	Top-1 Accuracy
GDFQ[10]	G	ECCV20	63.75
ZeroQ+IL[23]	N	CVPR20	63.97
DSG+IL[26]	G	CVPR21	62.62
ZAQ[11]	G	CVPR21	60.42
ARC[32]	G	IJCAI21	62.76
Qimera[25]	G	NeurIPS21	65.10
IntraQ[33]	N	CVPR22	64.98
ARC+AIT[27]	G	CVPR22	61.05
AdaSG[34]	G	AAAI23	66.42
AdaDFQ[12]	G	CVPR23	66.81
HAST[24]	N	CVPR23	66.91
Ours	**G**	Proposed	67.47

2. Comparison with SOTA Methods

Table 1은 본 논문에서 제안한 방법과 최근 SOTA (State of the art) 데이터 프리 양자화 방법 간의 가중치 및 활성값 3비트 양자화 성능 비교 결과이다. 제안한 방법은 56.14%로 가장 높은 정확도를 달성하였다. 일반적으로 노이즈 최적화 데이터 프리 양자화 방법들이 생성 모델을 이용한 양자화 대비 더 높은 성능을 보이나 제안한 방법은 가장 높은 성능을 보이는 노이즈 최적화 데이터 프리 양자화 모델인 HAST보다도 0.47% 높은 매우 우수한 성능을 보였다. 또한, 베이스라인인 AdaDFQ와 비교하여 3.4%를 향상 시켜, 제안한 방법이 매우 큰 성능 향상 효과가 있음을 보인다. CIFAR-100에서 3비트 양자화는 가장 도전적인 양자화 평가 방법 중 하나로, 제안한 방법이 매우 효과적임을 나타낸다. Table 2는 가중치 및 활성값 4비트 양자화 성능 비교 결과이다. 제안한 방법이 67.47%로 모든 방법 중 가장 높은 성능을 보였다. 베이스라인 대비 약 0.66%의 성능 향상을 보였다. 기존 가장 높은 성능을 보인 HAST와 비교해서도 0.56% 높은 성능을 보였다. 최근 양자화 기술 발달로 4비트 양자화에서는 AdaSG[34], AdaDFQ[12], HAST[24]와 같이 0.39%, 0.1%의 매우 적은 차이만은 보이나, 제안한 방법은 특출나게 높은 성능 향상을 보였다.

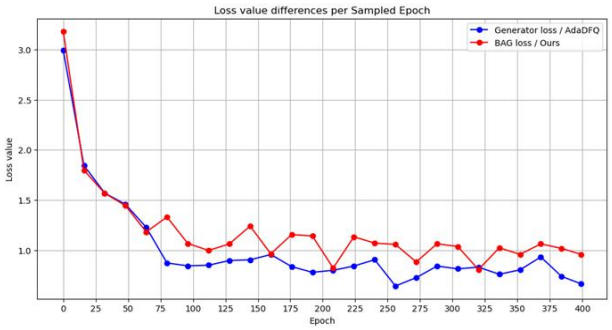

Fig. 3. The comparison of generator model loss trends over 400 epochs on CIFAR-100. All methods are quantized with 3-bits weights and activations.

3. Effects of BAG

우리는 양자화된 네트워크와 생성자 간에 지식 격차를 줄이는 것의 중요성을 보이기 위해 베이스라인과 비교를 수행하였다. Figure 3은 생성자의 손실 함수값 비교이다. BAG는 양자화로 인하여 AdaDFQ의 생성자와 비교하여 약간 높은 에러를 보인다. 생성자의 에러가 다소 커졌음에도 불구하고 Table 1, 2와 같이 우리가 제안한 방법은 성능을 향상한다. 이는 생성자의 다양성을 위해 잘 학습시키는 것뿐만 아니라, 원본 네트워크와 양자화된 네트워크 간의 지식 격차가 매우 중요한 문제이며 제안한 BAG가 효과적임을 보인다.

Table 3. Comparison of training time with other DFQ methods on CIFAR-100. All methods are quantized with 3-bits weights and activations.

Method	Data	Publication	Time (ms)
AdaDFQ[12]	G	CVPR23	130.1
HAST[24]	N	CVPR23	184.6
Ours	**G**	Proposed	123.3

4. Discussion for Training Complexity

우리는 학습 시간 측면에서 제안한 방법의 효과를 보이기 위해 베이스라인인 AdaDFQ와 가장 높은 정확도를 가지는 HAST와 비교를 수행하였다. Table 3에서 볼 수 있는 것처럼 우리가 제안한 방법은 베이스라인보다 6.8ms 빠르다. 우리가 제안한 방법은 양자화 복원과정과 채널 어텐션 기반 중간 레이어 지식 증류가 추가되었으나, 양자화된 BAG로 인하여 속도를 향상한다. 또한 가장 높은 성능을 보였던 HAST와 비교했을 때, 우리가 제안한 방법은 61.3ms나 빠르다. 이 결과는 우리가 제안한 방법은 큰 성능 향상을 보이면서, 속도 역시 향상하여 데이터 프리 양자화에서 매우 효과적임을 보인다.

V. Conclusions

생성 모델을 이용한 데이터 프리 양자화에서 생성자와 양자화된 네트워크 간에 bit-width와 자료형 차이로 인한 지식 격차를 해결하기 위하여 BAG를 제안하였다. BAG는 bit-width를 양자화된 네트워크와 동일하게 양자화하여, 양자화된 네트워크 수준에 맞는 이미지를 생성할 수 있다. 또한, 원본 네트워크와 양자화된 네트워크 간의 지식 격차를 줄이기 위해, 채널 어텐션 기반 중간 레이어 지식 증류를 제안하였다. 해당 방법은 Residual block의 출력의 어텐션을 구하여 이를 지식 증류함으로써, 학생 모델은 교사 모델로부터 어떤 채널이 더 중요하고 더 집중해야 하는 지를 배울 수 있다. 제안한 방법의 효과를 보이기 위해 양자화된 네트워크를 각각 3비트, 4비트로 양자화 후 CIFAR-100에서 데이터 프리 양자화 SOTA 모델들과 비교를 수행한 결과, 본 모델의 성능이 가장 높은 56.14%와 67.47%의 특출난 Top-1 Accuracy를 보였다. 또한, 학습 속도 비교를 통하여 속도도 향상함을 보였다. 본 연구를 통하여, 기존 생성 모델을 이용한 데이터 프리 양자화가 생성자의 다양성 향상에만 집중하던 것에서 벗어나 더 다양한 접근법을 기반으로 연구될 수 있기를 기대한다. 또한, 채널 어텐션 기반 중간 레이어 지식 증류를 활용하여, 생성 모델을 이용한 데이터 프리 양자화와 원본 데이터를 사용한 QAT (Quantization Aware Training)의 성능 차이를 줄이는데 기여할 수 있기를 기대한다.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009208) and funded by the Ministry of Education (No.2022R1A6A1A03051705); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2022-0-00448: Deep Total Recall, 30%, No.RS-2022-00155915: Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

REFERENCES

[1] Han, Song, Huizi Mao, and William J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.” arXiv, Oct. 2015.

[2] Jacob, Benoit, et al. “Quantization and training of neural networks for efficient integer-arithmetic-only inference.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2704-2713, June 2018. DOI: 10.1109/CVPR.2018.00286

[3] Gray, Robert M., and David L. Neuhoff. “Quantization.” IEEE transactions on information theory, Vol. 44, No. 6, pp. 7308-7316, 1998. DOI:10.1109/18.720541

[4] Yang, Jiwei, et al. “Quantization networks.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2704-2713, June 2018. DOI: 10.1109/CVPR.2019.00748

[5] Nagel, Markus, et al. "Data-free quantization through weight equalization and bias correction." Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1325-1334, Oct. 2019. DOI: 10.1109/ICCV.2019.00141

[6] Liu, Xingchao, et al. "Post-training quantization with multiple points: Mixed precision without mixed precision." Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 10,

pp. 8697-8705, Feb. 2021. DOI: 10.1609/AAAI.V35I10.17054

[7] Li, Yuhang, et al. "Brecq: Pushing the limit of post-training quantization by block reconstruction." Proceedings of 9th International Conference on Learning Representations, May 2021.

[8] Wang, Kuan, et al. "Haq: Hardware-aware automated quantization with mixed precision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8612-8620, June 2019. DOI: 10.1109/CVPR.2019.00881

[9] Esser, Steven K., et al. "Learned step size quantization." Proceedings of 8th International Conference on Learning Representations, April 2020.

[10] Xu, Shoukai, et al. "Generative low-bitwidth data free quantization." Computer Vision–ECCV 2020: 16th European Conference, Vol. 12357, pp. 1-17, Aug. 2021. DOI: 10.1007/978-3-030-58610-2_1

[11] Liu, Yuang, Wei Zhang, and Jun Wang. "Zero-shot adversarial quantization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1512-1521, June 2021. DOI: 10.1109/CVPR46437.2021.00156

[12] Qian, Biao, et al. "Adaptive data-free quantization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7960-7968, June 2023. DOI: 10.1109/CVPR52729.2023.00769

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv, March 2015. DOI: 10.48550/arXiv.1503.02531

[14] Romero, Adriana et al. “FitNets: Hints for Thin Deep Nets.” Proceedings of 8th International Conference on Learning Representations, May 2015.

[15] Chen, Defang, et al. "Cross-layer distillation with semantic calibration." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 8. pp. 7028-7036, Feb. 2021. DOI: 10.1609/aaai.v35i8.16865

[16] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in Neural Information Processing Systems. pp. 2672-2680, Dec. 2014.

[17] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." Proceedings of 4th International Conference on Learning Representations . May 2016.

[18] Odena, Augustus, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier gans." International conference on machine learning. Vol. 70. pp. 2642-2651, Aug. 2017.

[19] Li, Shaojie, et al. "Revisiting discriminator in gan compression: A generator-discriminator cooperative compression scheme." Advances in Neural Information Processing Systems. pp. 28560-28572. Dec. 2021.

[20] Metz, Luke, et al. "Unrolled generative adversarial networks." Proceedings of 5th International Conference on Learning Representations. April 2017.

[21] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770-778, June 2016. DOI: 10.1109/CVPR.2016.90

[22] Alex Krizhevsky and Geoffrey Hinton. “Learning multiple layers of features from tiny images.” Technical Report, pp. 1-60, 2009

[23] Cai, Yaohui, et al. "Zeroq: A novel zero shot quantization framework." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13166-13175, June 2020. DOI: 10.1109/CVPR42600.2020.01318

[24] Li, Huantong, et al. "Hard sample matters a lot in zero-shot quantization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24417-24426, June 2023. DOI: 10.1109/CVPR52729.2023.02339

[25] Choi, Kanghyun, et al. "Qimera: Data-free quantization with synthetic boundary supporting samples." Advances in Neural Information Processing Systems. pp. 14835-14847, Dec. 2021.

[26] Zhang, Xiangguo, et al. "Diversifying sample generation for accurate data-free quantization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15658-15667, June 2021. DOI: 10.1109/CVPR46437.2021.01540

[27] Choi, Kanghyun, et al. "It's all in the teacher: Zero-shot quantization brought closer to the teacher." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8301-8311, June 2022. DOI: 10.1109/CVPR52688.2022.00813

[28] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer.” Proceedings of 5th International Conference on Learning Representations. April 2017.

[29] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. Vol. 37, pp. 448-456. July 2015.

[30] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." International conference on machine learning. Vol. 30. pp. 3-8, June 2013.

[31] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." Proceeding of 3rd International Conference on Learning Representations. May 2015.

[32] Zhu, Baozhou, et al. "Autorecon: Neural architecture search-based reconstruction for data-free compression." Proceedings of the AAAI conference on artificial intelligence. pp. 3470-3476. Feb. 2021. DOI: 10.24963/IJCAI.2021/478

[33] Zhong, Yunshan, et al. "Intraq: Learning synthetic images with

intra-class heterogeneity for zero-shot network quantization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12329-12338 June 2022. DOI: 10.1109/CVPR52688.2022.01202

[34] Qian, Biao, et al. "Rethinking data-free quantization as a zero-sum game." Proceedings of the AAAI conference on artificial intelligence. Vol. 37. No. 8. pp.9489-9497, Feb. 2023. DOI: 10.1609/AAAI.V37I8.26136

[35] Chen, Xinrui, et al. "TexQ: Zero-shot Network Quantization with Texture Feature Distribution Calibration." Advances in Neural Information Processing Systems. Dec. 2023.

[36] Li, Jixing, et al. "ACQ: Improving generative data-free quantization via attention correction." Pattern Recognition, Vol. 152. pp.110444, Aug. 2024. DOI: 10.1016/J.PATCOG.2024.110444

[37] I. P. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks." Proceedings of 5th International Conference on Learning Representations, April 2017.

Authors

Jae-Yong Baek received the M.S degree in Computer Science and Engineering from Incheon Nation University, and worked as a software engineer that develop the software for object detection, lane detection and semantic segmentation in autonomous vehicle startup. He is currently pursuing the Ph.D. degree with the Department of Electronic Computer Engineering at Inha University, Korea. His current research interests are object detection, generative model, quantization and optimization for edge AI.

Du-Hwan Hur received the B.S. degree in Computer Engineering from Hanbat National University in 2021, and is currently pursuing the M.S. - Ph.D. integrated course degree with the Department of Electronic Computer Engineering at Inha University, Korea. His current research interest is objection detection, multi-object tracking, and deep learning.

Deok-Woong Kim received the B.S. degree with Department of Industrial engineering, Public administration, Inha University, South Korea. He is currently pursuing the M.S. - Ph.D. integrated course degree. His research interests are knowledge distillation, quantization, data-free and on-board AI.

Yong-Sang Yoo received the BS degree in Computer Science and Engineering from Incheon National University in 2021, and is currently pursuing the MS and PhD integrated degree with the Department of Electronic Computer Engineering at Inha University, Korea. His current research interests include object detection, machine learning, continual learning, on-board AI and deep learning.

Hyuk-Jin Shin received the B.S. degree with Department of Computer Science, Chung-Buk National University, South Korea. He is currently pursuing the M.S. - Ph.D. integrated course degree. His research interests are object detection and continual learning.

Dae-Hyeon Park received the B.S degree in Computer Engineering from Inha University in 2020 and the M.S degree with the Department of Electronic Computer Engineering at Inha University in 2023. He is currently pursuing the Ph.D. degree with the Department of Electronic Computer Engineering at Inha University, Korea and His current research interests are single/multi-object tracking, multi-modal learning, real-time system and self-attention mechanism.

Seung-Hwan Bae received the BS degree in information and communication engineering from Chungbuk National University, in 2009 and the MS and PhD degrees in information and communications from the Gwangju Institute of Science and Technology (GIST), in 2010 and 2015, respectively. He was a senior researcher at Electronics and Telecommunications Research Institute (ETRI) in Korea from 2015 to 2017. He was an assistant professor in the Department of Computer Science and Engineering at Incheon National University, Korea from 2017 to 2020. He is currently an Associate Professor with the Department of Computer Engineering at Inha University, His research interests include object tracking, object detection, generative model learning, continual learning, on-device ML, etc.