JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 8, pp. 31-40, August 2026
https://doi.org/10.9708/jksci.2026.31.08.031

SADA-DFQ: Self-Attention Correlation Distillation and Difficulty-Aware Sample Generation for Generative Data-Free Quantization

Du-Hwan Hur*, Deok-Woong Kim**, Seung-Hwan Bae***

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

[Abstract]

In this paper, we propose SADA-DFQ, a generative data-free quantization (DFQ) framework that restores a low-bit network without any access to the original training data, as required in privacy-sensitive domains such as medical imaging and security. Although generator-based DFQ offers high sample diversity, its accuracy still lags behind. We attribute this gap to two factors: an unreliable intermediate-layer transfer signal at low bit-width, where rounding and clipping errors prevent the quantized model from reproducing the full-precision (FP) model's activation magnitudes, and the limited informativeness of the synthetic dataset used for fine-tuning. To address the first, Self-Attention Correlation Distillation (SACD) reformulates each intermediate feature into multi-head self-attention and transfers the relational structure among channel tokens, which tends to be less sensitive to quantization error and thus provides a more stable signal. To address the second, Difficulty-Aware Sample Generation (DASG) re-weights the synthetic dataset by sample difficulty to emphasize examples near the decision boundary. On CIFAR-100, SADA-DFQ attains the best accuracy among the compared generator-based DFQ methods at both 3-bit and 4-bit, improving Top-1 accuracy by about 2.6 percentage points over AdaDFQ at 3-bit.

▸ **Key words**: Model Quantization, Model Compression, Data-free Quantization, Generative model, Knowledge Distillation, Self-Attention, Image Classification

• First Author: Du-Hwan Hur, Deok-Woong Kim, Corresponding Author: Seung-Hwan Bae
*Du-Hwan Hur (gjenghks@inha.edu), Vision & Learning Lab, Inha University
**Deok-Woong Kim (k5000plus@inha.edu), Vision & Learning Lab, Inha University
***Seung-Hwan Bae (shbae@inha.ac.kr), Vision & Learning Lab, Dept. of Electrical and Computer Engineering, Inha University
• Received: 2026. 06. 30, Revised: 2026. 07. 14, Accepted: 2026. 07. 22.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

[요 약]

저전력 엣지 환경에 심층 신경망을 배포하려는 수요가 늘면서 양자화의 중요성이 커지고 있으나, 의료·보안처럼 데이터 공개가 제한되는 분야에서는 원본 데이터에 접근할 수 없어 원본 없이 저비트 네트워크를 복원하는 데이터 프리 양자화(Data-Free Quantization, DFQ)가 요구된다. 생성자 기반 DFQ는 표본 다양성은 높지만 정확도가 뒤처지는데, 본 논문은 그 원인을 두 가지로 본다. 하나는 저비트에서 양자화 모델이 FP (Full-Precision) 모델의 활성값 크기를 재현하지 못해 생기는 중간 레이어 전이 신호의 불안정성이고, 다른 하나는 미세조정에 쓰이는 합성 데이터셋의 정보량 한계이다. 전자를 위해 셀프 어텐션 상관 증류(SACD)는 중간 특징을 멀티헤드 셀프 어텐션의 채널 토큰 간 상관 구조로 재구성하여, 양자화 오차에 덜 민감한 관계 구조를 전이한다. 후자를 위해 난이도 인지 샘플 생성(DASG)은 합성 데이터셋을 난이도로 재가중하여 결정 경계 부근의 표본을 강조한다. CIFAR-100에서 SADA-DFQ는 3비트와 4비트 모두에서 생성자 기반 비교 기법 중 최고 정확도를 달성했으며, 3비트에서 AdaDFQ 대비 Top-1 정확도를 약 2.6%p 향상시켰다.

▸ **주제어**: 모델 양자화, 모델 압축, 데이터 프리 양자화, 생성 모델, 지식 증류, 셀프 어텐션, 이미지 분류

I. Introduction

최근 저전력·실시간 환경에 심층 신경망을 올리기 위한 모델 압축에서, 양자화(quantization)는 가중치와 활성값의 비트 폭을 줄여 연산량과 메모리를 동시에 절감하는 핵심 수단이다[1, 2]. 비트 폭을 낮추면 정확도가 떨어지므로 통상적으로는 원본 학습 데이터의 일부 또는 전부를 사용해 보정이나 재학습을 수행한다[3, 4]. 그러나 의료·보안과 같이 데이터 공개가 제한되는 영역에서는 원본 데이터에 접근할 수 없는 경우가 많아 이러한 복원 절차를 그대로 적용하기 어렵다.

데이터 프리 양자화(DFQ)는 이 문제를 우회하기 위해 원본 데이터를 합성 데이터로 대체한다[5-8]. 입력 자체를 그래디언트로 최적화하는 노이즈 최적화 방식[9]은 합성 표본의 표현력이 제한적인 반면, 사전 분포로부터 표본을 생성하도록 생성자를 학습하는 생성자 기반 방식[6-8]은 합성 데이터셋의 다양성과 사실성을 높인다. 그럼에도 생성자 기반 방식의 보고 정확도는 노이즈 최적화 방식에 미치지 못하는 역설적 상황이 이어지고 있다.

본 논문은 이 격차를 두 가지 원인으로 설명한다. 첫째는 중간 레이어 전이 신호의 신뢰도 문제이고, 둘째는 미세조정에 쓰이는 합성 데이터셋의 정보량 한계이다. 먼저 전이 신호 측면에서, 양자화 모델은 정수 연산과 좁은 표현 범위로 인해 반올림·클리핑 오차를 누적하므로 FP(원본) 모델의 활성값 크기를 정밀하게 재현하기 어렵다. 이때 활성값을 직접 정렬하거나 채널별 통계를 맞추는 중간 레이어 증류는 양자화 모델이 도달하기 어려운 목표를 부과하여 지도 신호가 약해질 수 있다. 반면 한 레이어 안에서 특징들이 형성하는 상대적 관계 구조는 활성값의 절대 크기 변화에 상대적으로 덜 민감할 것으로 기대되어, 보다 안정적인 전이 대상이 될 수 있다. 다음으로 표본 측면에서, 합성 데이터셋이 양자화 모델 관점에서 충분히 어려운(정보량이 높은) 표본을 포함하지 못하면 미세조정 신호가 제한된다.

이 관찰에 근거하여 SADA-DFQ를 제안한다. Figure 1은 제안 기법의 전체 구조를 보인다. 셀프 어텐션 상관 증류(SACD)는 각 중간 레이어의 특징맵을 채널 단위 토큰으로 바꾸고 멀티헤드 셀프 어텐션을 적용하여 채널 토큰 간 상관 구조를 부호화한 셀프 어텐션 상관 텐서를 얻는다. FP 모델과 양자화 모델의 이 텐서를 일치시킴으로써, 크기 정보가 아니라 양자화 오차에 덜 민감한 관계 구조를 전이한다. 난이도 인지 샘플 생성(DASG)은 원본 모델이 어렵게 분류하는 표본에 더 큰 비중을 두어 생성을 유도함으로써, 결정 경계 인근의 정보량 높은 표본으로 합성 데이터셋을 보강한다.

제안 기법은 생성자 기반 베이스라인 위에서 동작하며, CIFAR-100[12]에서 대표적인 생성자 기반 DFQ 기법과 비교한 결과 3비트와 4비트 모두에서 생성자 기반 방법 중 가장 높은 정확도를 보였고, 특히 3비트에서 AdaDFQ[8]를 Top-1 정확도 기준 약 2.6%p를 앞섰다. 같은 조건에서 기존 중간 레이어 증류 방식인 FitNet[11]과 AT[15]보다 우수하였으며, ImageNet[20]의 3비트 설정에서도

www.kci.go.kr

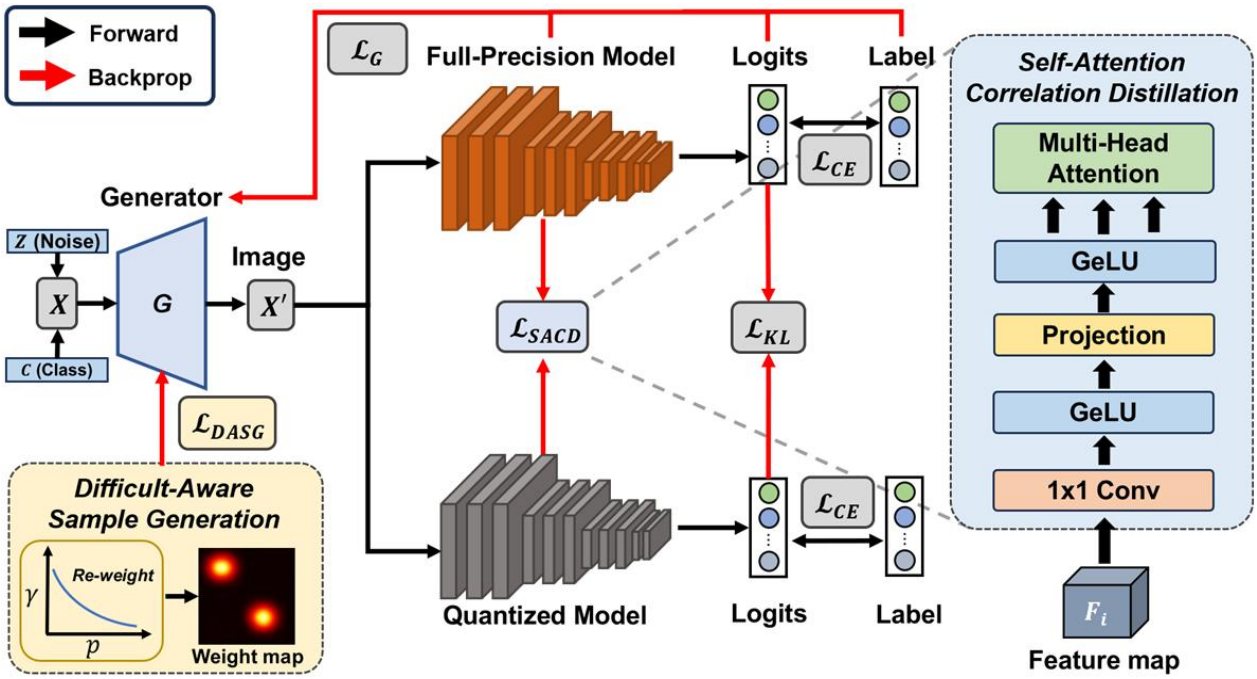

Fig. 1. Overall framework of the proposed SADA-DFQ.

AdaDFQ 대비 +3.65%p의 향상을 확인하여 데이터셋과 백본에 걸친 일반화를 보였다. 본 논문의 기여는 다음과 같다. i) 중간 레이어 특징을 멀티헤드 셀프 어텐션의 관계 구조로 변환해 증류하여 저비트 양자화에서 안정적인 지식 전이를 달성하는 SACD를 제안한다. ii) 원본(FP) 모델 기준 난이도로 생성을 재가중하여 합성 표본의 정보량을 높이는 DASG를 제안한다. iii) 기존 중간 레이어 증류 방식(FitNet, AT)과의 동일 조건 직접 비교와 반복 실험을 통해, 제안 기법이 저비트에서 효과적임을 실증한다. 이어지는 구성은 2장 관련 연구, 3장 제안 기법, 4장 실험, 5장 결론 순이다.

II. Related Works

본 장에서는 제안 기법과 연관된 두 흐름, 즉 데이터 프리 양자화의 표본 생성 전략과 중간 레이어 지식 증류를 정리한다.

1. Data-Free Quantization

DFQ는 원본 데이터 없이 저비트 네트워크의 정확도를 복원한다. 노이즈 최적화 계열[9]은 무작위 입력을 그래디언트로 갱신해 합성 표본을 만들지만 표현력이 제한적이며, 생성자 기반 계열은 생성 모델로 합성 데이터셋의 다양성과 사실성을 높인다. 생성자 기반 기법의 핵심 차이는 '어떤 표본을 생성할 것인가'에 있다. GDFQ[6]는 배치 정규화 통계 정합으로 합성 분포를 원본에 근접시키고, DSG[7]는 표본 다양성을, Qimera[13]는 클래스 경계를 지지하는 표본을, AIT[14]는 손실 구성을 단순화해 학습 안정성을 높였으며, AdaDFQ[8]는 FP 모델과 양자화 모델의 합치 정도를 기준으로 표본의 정보량을 적응적으로 조절한다. 그러나 이들은 모두 생성 분포 자체의 품질을 높이는 데 초점을 두며, 원본 모델이 현재 무엇으로 분류할지 어려워하는 표본 생성에 직접 반영하는 난이도 기반 재가중은 생성자 기반 DFQ에서 다루어지지 않았다. 본 논문의 DASG는 이 점을 처음으로 도입한다.

2. Intermediate-Layer Knowledge Distillation

지식 증류[10, 11]는 교사(teacher) 모델의 출력을 학생(student) 모델이 모사하도록 지식을 전이하는 일반적인 기법이다. 출력 소프트 라벨만으로는 지식 전이가 제한적이므로 FitNet[11]은 중간 레이어 활성값을, AT[15]는 공간 어텐션 맵을 정렬해 더 풍부한 단서를 제공하였다. 양자화에서는 FP 원본 네트워크를 교사로, 양자화 네트워크를 학생으로 두는 증류가 일반적이며, 중간 레이어 정렬은 양자화 인지 학습(Quantization-Aware Training)을 포함한 양자화 전반에서 쓰이는 손실이다. 다만 저비트에서는 자료형과 표현 범위의 간극 때문에 활성값 크기 기반 정렬이 약화되기 쉽고, 특히 생성자 기반 DFQ에서는 GDFQ·AdaDFQ가 대체로 소프트 라벨 증류에 머물러 중

www.kci.go.kr

간 레이어의 관계 구조를 직접 전이하려는 시도가 드물었다. 본 논문의 SACD는 이 비어 있는 축을 겨냥하여, 중간 특징을 멀티헤드 셀프 어텐션[16]으로 재구성해 채널 토큰 간 상관 구조를 정렬 대상으로 삼는다. 이 관계 구조는 활성값의 절대 크기 변화에 상대적으로 덜 민감할 것으로 기대되어 저비트에서도 비교적 안정적인 전이 신호를 제공하며, 표본 생성 개선에 집중한 기존 연구와 구분된다. 이와 관련하여, 4장에서는 FitNet과 AT를 동일한 생성자 기반 베이스라인에 결합한 직접 비교를 통해 이 차이를 정량적으로 검증한다.

3. Generator-based Data-Free Quantization

생성자 기반 DFQ 방법론들은 사전 학습된 모델로부터 추출한 정보를 활용해, 생성자가 실제 데이터의 분포에 근접한 합성 데이터셋을 만들어내고, 이를 통해 양자화된 모델의 성능을 회복시킨다. GDFQ는 생성자를 활용한 DFQ 방법론의 첫 시도에 해당한다. 생성자를 사용하지 않던 기존 DFQ 기법들이 Batch Normalization Statistics (BNS) 손실을 핵심으로 삼아온 가운데, GDFQ의 차별점은 생성자의 입력에 클래스 정보를 부여하고 교차 엔트로피 (CE) 손실을 함께 최적화한 데 있다. 즉 BNS 손실과 클래스 조건부 생성을 결합함으로써, 실제 데이터의 분포 특성과 클래스 속성을 동시에 반영하는 의미 있는 합성 데이터를 생성하고자 했다. Qimera는 클래스 조건부 생성자가 CE 손실을 통해 샘플들을 서로 잘 분리되도록 학습하기 때문에 생성된 샘플이 클래스 중심에 몰리고 결과적으로 클러스터 사이의 경계 영역이 비게 된다는 점을 지적하고, superposed latent embedding을 활용해 결정 경계 부근의 boundary supporting sample을 집중적으로 생성한다. AdaSG[17]는 DFQ를 생성자와 양자화 모델 사이의 제로섬 게임으로 재정의한다. 생성자 학습에 오로지 원본 모델의 정보만을 활용하던 기존 방식과 달리, 양자화 모델의 정보까지 끌어들여 생성된 샘플이 양자화 모델의 학습에 이로운지를 나타내는 샘플 적응성(sample adaptability)을 함께 최적화한다. AdaDFQ는 적응성을 단순히 극대화하면 과적합이, 반대로 너무 낮으면 과소적합이 발생한다고 지적하며, 사전 학습된 모델과 양자화 모델의 예측을 바탕으로 불합의(disagreement)와 합의(agreement) 샘플로 두 경계를 정의하고, 그 사이의 마진을 최적화하여 생성 샘플의 적응성을 적절한 수준으로 조절한다. 한편 HAST[18]는 노이즈 최적화 기반 방법으로, 모델이 어려워하는 hard sample을 강조하여 합성하고 학습에 활용함으로써 복원 성능을 높인다. 그러나 이러한 전략은 입력 노이즈를 최적화하는 방식에서만 다루어졌을 뿐, 생성자 기반 DFQ에서는 충분히 탐구되지 않았다. 본 연구는 이를 생성자 학습으로 확장하여, 사전 학습 모델 기준의 난이도로 생성자의 분류 손실을 재가중하는 기법을 제안한다.

III. Methodology

본 장에서는 먼저 생성자 기반 양자화의 기본 구성을 정의한 뒤, 셀프 어텐션 상관 증류로 중간 레이어의 관계 구조를 전이하는 방법과, 난이도 인지 샘플 생성으로 합성 표본의 정보량을 높이는 방법을 기술한다.

1. Preliminaries

FP 원본 네트워크 M_{FP}로부터 양자화 네트워크 M_q를 얻을 때, 각 가중치 및 활성값 x는 비트 폭 n에 대해 균일 양자화기로 변환된다. 스케일 s와 영점 Z를 이용하면 양자화 값은 수식 (1)과 (2)로 표현된다.

$$x_q = \lfloor s \cdot clip(x; x_{\min}, x_{\max}) - Z \rceil, \quad (1)$$

$$s = \frac{2^n - 1}{x_{\max} - x_{\min}} \quad (2)$$

생성자 G는 잡음 z와 라벨 y를 입력받아 합성 표본 $\hat{x} = G(z, y)$를 생성하며, 합성 표본이 목표 클래스를 따르도록 하는 분류 손실과, 각 배치 정규화 레이어 l의 실행 통계 (평균 μ_l, 분산 σ_l)를 합성 배치 통계와 맞추는 BNS 손실은 아래 수식 (3)으로 학습된다.

$$L_{BNS} = \sum_l (\| \mu_l(\hat{x}) - \overline{\mu_l}(\hat{x}) \|^2 + \| \sigma_l(\hat{x}) - \overline{\sigma_l}(\hat{x}) \|_2) \quad (3)$$

양자화 네트워크는 합성 데이터셋에 대한 교차 엔트로피와, FP 모델과의 출력 분포를 맞추는 KL 발산 기반 소프트 라벨 증류로 미세조정된다. 이 기본 손실을 수식 (4)인 $\mathcal{L}_{base}$ 로 둔다.

$$L_{base} = CE(M_q(\hat{x}), y) + KL(M_{FP}(\hat{x}) \| M_q(\hat{x})) \quad (4)$$

본 절 이후의 모든 구성요소는 이 생성자 기반 베이스라인 위에 추가된다.

www.kci.go.kr

2. Self-Attention Correlation Distillation (SACD)

SACD의 출발점은, 저비트 양자화에서 신뢰할 수 있는 전이 대상이 활성값의 크기가 아니라 특징 간 관계라는 점이다. 반올림·클리핑 오차는 개별 활성값의 절댓값을 교란하지만, 채널들이 서로 이루는 상대적 패턴은 상대적으로 덜 교란되는 경향이 있다. 따라서 우리는 크기를 직접 맞추는 대신 이 관계 구조를 명시적으로 부호화하여 전이한다. 증류 지점으로는 각 스테이지의 대표 특징맵을 사용한다. Figure 1 우측 SACD 모듈과 같이, 스테이지 l의 특징맵 $F^l \in \mathbb{R}^{(C\times H\times W)}$에 대해 토크나이저 T는 특징맵을 채널 단위의 C개 토큰으로 변환한다. 구체적으로 T는 1×1 합성곱과 GeLU로 각 채널의 표현을 추출한 뒤, 선형 투영(Projection)과 GeLU를 거쳐 토큰 임베딩 $T(F^l) \in \mathbb{R}^{(K\times D)}$을 생성한다. 이 과정에서 높이 H × 너비 W의 공간 정보는 채널별로 요약되고, 각 채널이 하나의 토큰이 되어 총 K개의 토큰이 D차원 임베딩으로 표현된다. 채널을 토큰으로 두는 까닭은, 전이 대상이 개별 활성값의 크기가 아니라 채널들 사이의 상대적 관계이기 때문이다. 또한, 토큰 임베딩 $T(F^l)$을 선형 투영해 질의 Q, 키 K, 값 V를 구성하고, h개의 헤드로 분할한 멀티헤드 셀프 어텐션을 적용한다. 각 헤드에서 질의와 키의 내적을 정규화한 어텐션 가중치가 채널 토큰 간 상관을 부호화하며, 이를 헤드에 대해 모은 것이 셀프 어텐션 상관 텐서 $A^l \in \mathbb{R}^{(K\times K)}$이다.(수식 5.) SACD는 값 V로 가중합한 어텐션 출력이 아니라 이 상관 텐서 A^l 자체를 증류 대상으로 삼아, 채널 간 관계 구조를 FP 모델로부터 양자화 모델로 전이한다.

$$A^l = MHSA(T(F^l)) = Concat(head_1, \cdots, head_h)W_O, \quad (5)$$
$$head_j = softmax(\frac{Q_j K_j^T}{\sqrt{D/h}})V_j$$

FP 모델과 양자화 네트워크에서 각각 산출한 A_f^l, A_q^l에 대해, SACD 손실(수식 6.)은 두 셀프 어텐션 텐서 간 거리로 정의된다. 헤드 별로 산출된 상관 패턴을 모두 일치시킴으로써, 양자화 모델은 FP 모델이 레이어 내부에서 형성하는 관계 구조를 학습한다.

$$\mathcal{L}_{SACD} = \frac{1}{s}\Sigma_{l\in s}\ MSE\|A_f^l - A_q^l\|_2 \qquad (6)$$

$\mathcal{S}$는 증류에 사용한 지점의 집합이며, $\|\cdot\|_2$는 L2 노름이다. 양자화 네트워크의 최종 학습 손실(수식 7.)은 기본 손실에 SACD 손실을 가중 결합하여 정의된다.

$$\mathcal{L}_q = \mathcal{L}_{base} + \lambda\cdot\mathcal{L}_{SACD} \qquad (7)$$

λ는 SACD의 반영 비율을 조절하는 하이퍼 파라미터이다. 어텐션 상관 맵은 채널 간 상대적 관계를 부호화하므로 개별 활성값의 절대 크기보다 채널 사이의 패턴에 의존한다. 따라서 FP 모델과 양자화 모델의 표현 범위가 다르더라도 비교적 안정적인 지도 신호를 제공한다. 이는 활성값을 직접 정렬하는 FitNet 계열이나 채널 통계만 맞추는 채널 어텐션 정렬이 저비트에서 겪는 신호 약화를 구조적으로 완화하며, 이 효과는 양자화 오차가 큰 저비트에서 두드러진다. SACD의 토크나이저와 멀티헤드 셀프 어텐션은 학습 시 손실 계산에만 사용되며, 미세조정이 끝난 양자화 모델의 추론 그래프에는 포함되지 않는다. 따라서 SACD는 추론 시 추가 연산이나 메모리 비용을 발생시키지 않는다.

3. Difficulty-Aware Sample Generation (DASG)

SACD가 전이 신호의 질을 높이는 동안, 합성 표본의 정보량을 함께 끌어올리면 추가 이득을 기대할 수 있다. 생성자 기반 DFQ에서 합성 표본은 양자화 모델을 미세조정하는 학습 신호의 원천이므로, 원본 모델이 이미 쉽게 맞히는 표본보다 어려워하는 표본이 더 큰 학습 신호를 제공한다. DASG는 이 점에 착안하여 표본의 난이도를 생성 단계에 직접 반영한다.

구체적으로, 사전 학습된 원본(FP) 모델 M_f가 합성 표본 $\hat{x}_i$를 정답 y_i로 분류할 확률을 $p_i = M_f(\hat{x}_i)[y_i]$로 둘 때, 표본 난이도를 $d_i = 1 - p_i$로 정의한다. 즉 원본 모델이 확신을 갖고 맞히는 표본일수록 d_i는 0에 가깝고, 어려워하는 표본일수록 1에 가깝다. 이 난이도를 가중치로 사용해 생성자의 분류 손실을 수식 (8)과 같이 재가중한다.

$$\mathcal{L}_{DASG} = \beta \cdot \frac{1}{N}\Sigma_i(1-p_i)^{\tau}\cdot CE(M_f(\hat{x}_i), y_i \qquad (8)$$

수식 (8)에서 N은 배치 내 표본 수, CE는 교차 엔트로피 손실이다. τ는 난이도 강조 지수로 클수록 어려운 표본에 가중치가 더 집중되며, β는 DASG 손실 항의 반영 비율을 조절하는 계수이다. 난이도 가중치 d_i와 분류 손실 $CE(M_f(\hat{x}_i), y_i)$는 모두 사전 학습된 원본(FP) 모델 M_f

www.kci.go.kr

의 예측을 기준으로 계산된다. 따라서 재가중은 원본 모델이 낮은 확신으로 분류하는, 결정 경계 인근의 표본을 더 많이 생성하도록 유도한다. 이는 베이스라인의 적응적 생성이 표본의 정보량을 일정 구간 안으로 제한하는 것과 달리, 난이도를 생성 손실에 직접 결합하는 방식이다. 또한 DASG는 생성자 손실의 분류 항에만 개입하므로 BNS 손실이나 양자화 모델의 미세조정 손실(L_{base}) 구성을 바꾸지 않으며, SACD와 독립적으로 결합된다. 또한, SACD와 DASG는 모두 학습 시 손실에만 관여하며, DASG는 생성자 학습에, SACD는 양자화 모델 미세조정에 쓰인 뒤 추론에는 영향을 주지 않는다.

IV. Experiments

본 장에서는 실험 설정을 기술한 뒤, 최신 DFQ 기법과의 정확도 비교, 기존 중간 레이어 증류 방식과의 직접 비교 및 각 구성요소의 기여를 분리한 분해 실험, ImageNet으로의 일반화 검증, 그리고 두 구성 요소의 작동 방식에 대한 심층 분석을 통해 제안 기법의 효과를 검증한다.

1. Experiment Settings

본 연구에서는 100개의 클래스와 학습 이미지 50K개, 테스트 이미지 10K개로 구성된 CIFAR-100[12] 데이터셋에서 ResNet-20[19] 모델을 사용하여 제안 기법을 평가하였다. 평가 지표로는 3비트(W3A3)와 4비트(W4A4) 양자화를 적용한 모델의 Top-1 정확도(Top-1 Accuracy)를 보고하며, 여기서 W는 가중치(Weight), A는 활성화(Activation)를 의미하고 각 기호 뒤의 숫자는 해당 대상의 양자화 비트 수를 나타낸다. 비교를 위한 베이스라인으로는 생성자 기반의 대표적인 기법인 GDFQ와 AdaDFQ를 채택하였다. 데이터 생성에는 모든 실험에서 ACGAN 기반 생성자를 사용하였으며, Optimizer, Momentum, Learning rate 등 생성자 관련 하이퍼파라미터는 각 베이스라인의 원 설정을 그대로 유지하였다. 양자화 모델의 미세조정은 Epoch 400, 배치 사이즈 16, SGD Optimizer로 수행하였다. 제안하는 DASG의 보조 가중 β와 강조 지수 τ는 3비트와 4비트 실험에 대해 각각 {5e-3, 2.0}과 {5e-3, 0.5}를 적용하였으며, 이는 GDFQ와 AdaDFQ 두 베이스라인에 동일하게 사용하였다. 3비트에 더 큰 τ(=2.0)를 사용한 것은 비트 폭이 낮을수록 양자화 오차가 커져 어려운 표본을 더 강하게 부각하는 편이 유리하기 때문이며, 오차가 상대적으로 작은 4비트에서는 완만한 강조(τ=0.5)가 더 좋은 결과를 보였다.

SACD의 반영 비율 가중치 λ와 증류 지점 수 $\mathcal{S}$, 헤드 수 h, 임베딩 차원 D는 베이스라인에 따라 다르게 설정하였다. AdaDFQ에서는 모든 실험에서 λ=100, $\mathcal{S}$=4, h=4, D=64를 적용하였고, GDFQ에서는 λ를 3비트와 4비트에 대해 각각 100과 1000으로 두고 $\mathcal{S}$=4, h=4, D=512를 적용하였다. 증류 지점 수 $\mathcal{S}$는 백본 구조를 따라 ResNet-20에서는 4개, ResNet-18에서는 5개를 사용하였다. 모든 실험은 Intel Xeon 2.10GHz CPU와 NVIDIA RTX A6000 GPU 환경에서 수행하였으며, PyTorch를 비롯한 파이썬 라이브러리 버전 등 세부 환경은 각 베이스라인의 설정에 맞추어 구성하였다. 제안 기법이 포함된 실험은 서로 다른 무작위 시드(seed)로 3회 반복하여 best-epoch Top-1 정확도의 평균과 표준편차(mean ± std)로 표기한다. FitNet[11]과 AT[15]를 동일한 GDFQ 베이스라인에 결합하였으며, 공정한 비교를 위해 두 기법 모두 원 저자 구현의 기본 설정을 사용하였다. 일반화 검증을 위해 ImageNet[20]에서 ResNet-18을 백본으로 AdaDFQ 베이스라인에 제안 기법을 결합해 W3A3 설정으로 평가하였으며, β와 τ는 각각 {1e-1, 5e-1}를 적용하고 그 외 학습 구성은 AdaDFQ 기본 설정과 동일한 원칙을 따랐다.

2. Comparison with Generator-DFQ Methods

이 실험의 목적은 제안 기법이 기존 생성자 기반 DFQ 대비 어느 정도의 정확도 이득을 주는지를 동일 조건에서 정량화하는 것이다. 비교 대상으로는 노이즈 최적화 계열 IntraQ와 생성자 기반 계열 Qimera·GDFQ·AdaDFQ를 두며, 제안 기법은 대표적인 생성자 기반 베이스라인 GDFQ와 AdaDFQ 위에 각각 SACD와 DASG를 결합한 형태(+Ours)로 평가한다. 모든 기법은 동일한 백본(ResNet-20), 비트 폭(W3A3, W4A4), 평가 프로토콜에서 측정하여, 베이스라인 대비 순수한 이득만 드러나도록 한다. Table 1은 CIFAR-100에서 생성자 기반 DFQ 기법들과 제안 기법

Table 1. Top-1 accuracy on CIFAR-100 (ResNet-20) (%)

Method	W3A3	W4A4
IntraQ	48.25	64.98
Qimera	46.13	65.10
GDFQ	47.20	63.54
GDFQ+(Ours)	**50.13±1.08** (+2.93%)	**64.93±0.46** (+1.39%)
AdaDFQ	52.74	66.81
AdaDFQ+(Ours)	**55.33±0.81** (+2.59%)	**67.27±0.24** (+0.46%)

www.kci.go.kr

의 Top-1 정확도를 비교한 결과이다. 제안 기법은 3비트와 4비트 모두에서 비교 기법 중 가장 높은 정확도를 보였으며, 특히 가장 공격적인 3비트 설정에서 향상 폭이 크다. 구체적으로 3비트에서 AdaDFQ의 52.74%를 55.33±0.81%로 끌어올려 약 +2.59%p를 달성했다. 4비트에서는 66.81%에서 67.27±0.24%로 +0.46%p의 향상을 보였다. 즉 전이 신호를 개선하는 접근은 비트 폭 전반에서 베이스라인 이상을 유지하되, 그 이득은 양자화 오차가 큰 저비트에서 뚜렷하게 커진다. 이 이득이 특정 데이터셋이나 백본에 한정되지 않음을 확인하기 위해, 대규모 벤치마크인 ImageNet[20]에서 ResNet-18을 백본으로 동일한 검증을 수행하였다. 제안 기법의 이득이 저비트에서 두드러진다는 앞선 결과에 따라 W3A3 설정을 대상으로 삼았으며, 그 결과 AdaDFQ의 35.19%를 38.84±1.36%로 +3.65%p 향상시켰다. 클래스 수와 해상도가 크게 늘어난 조건에서도 향상 폭이 유지되는 것은, 채널 간 상관 구조의 전이와 난이도 기반 표본 보강이 데이터셋 규모와 백본에 걸쳐 일반화됨을 보인다.

3. Comparison with Intermediate-Layer Distillation and Component Ablation

본 절의 목적은 두 가지다. 첫째, SACD가 전이하는 채널 간 상관 구조가 활성값 크기를 직접 정렬하는 기존 중간 레이어 증류보다 저비트 양자화에서 실제로 우수한 지도 신호인지를 검증한다. 둘째, 제안 기법을 구성하는 SACD와 DASG가 각각 단독으로 얼마나 기여하며 둘의 결합이 상호 보완적인지를 확인한다. 이를 위해 대표적인 중간 레이어 증류 기법인 FitNet[11]과 AT[15], 그리고 제안 구성요소인 DASG와 SACD를 각각 동일한 GDFQ 베이스라인에 단독으로 결합하고, 동일한 백본·비트 폭·평가 프로토콜에서 비교하였다. Table 2의 각 행은 베이스라인에 해당 기법만을 단독으로 추가한 결과이며 이전 행에 순차적으로 누적한 것이 아니다. 비교군의 설정은 IV.1절에 기술한 바와 같이 원 저자 구현의 기본값을 사용하였다.

Table 2. Component ablation on the GDFQ baseline with CIFAR-100 (%)

Method	W3A3	W4A4
GDFQ	47.20	63.54
+FitNet[11] only	42.42 ±1.84	62.26 ±0.05
+AT[15] only	45.31 ±1.05	64.52 ±0.09
+DASG only	48.90 ±0.77	63.16 ±0.16
+SACD only	48.80 ±1.65	64.65 ±0.42
+Ours(SADA-DFQ)	**50.13 ±1.08** (+2.93%)	**64.93 ±0.46** (+1.39%)

Table 3. Sensitivity to DASG hyperparameters (β, τ) on the AdaDFQ with CIFAR-100 (W4A4) (%)

β	τ	Top-1 Acc.(%)
0.005	0.05	67.05
0.005	0.5	**67.61**
0.005	1.0	67.43
0.005	2.0	67.12
0.01	0.05	66.61
0.01	0.5	66.39

먼저 기존 중간 레이어 증류와 비교하면, W3A3에서 제안 기법은 FitNet 대비 +7.71%p, AT 대비 +4.82%p로 우수하다. 주목할 점은 W3A3에서 FitNet과 AT가 중간 레이어 손실이 없는 베이스라인보다도 낮다는 것이다. 이는 저비트 양자화 모델이 재현할 수 없는 활성값 크기(FitNet)나 양자화로 붕괴된 공간 어텐션 구조(AT)를 정렬 목표로 부과하면, 해당 손실의 그래디언트가 소프트 라벨 증류와 충돌하여 지도 신호가 약해지는 것을 넘어 오히려 해가 됨을 보인다. 반면 AT는 W4A4에서 베이스라인 대비 +0.98%p의 향상을 보이는데, 동일한 가중 계수에서 비트 폭에 따라 성능이 바뀌는 현상은 비트 폭에 따른 정렬 목표의 도달 가능성 차이에서 비롯됨을 시사한다. W4A4에서도 제안 기법이 가장 높은 평균 정확도를 보였으나 AT와의 차이는 표준편차 범위 이내였다. 즉 활성값 크기 기반 정렬(FitNet)과 공간 어텐션 정렬(AT)은 비트 폭이 낮아질수록 지도 신호가 급격히 약화되는 반면, 관계 구조를 전이하는 제안 기법은 저비트에서 그 격차를 크게 벌리며, 특히 W3A3에서는 제안 구성요소를 하나만 결합한 경우조차 두 비교군을 모두 상회한다.

다음으로 구성 요소별 기여를 보면, SACD는 단독으로 3비트에서 +1.60%p, 4비트에서 +1.11%p의 정확도 향상을 보여, 관계 구조 전이가 두 비트 폭 모두에서 유효하다는 것을 보인다. DASG는 3비트에서 단독으로 48.90±0.77%(+1.70%p)의 향상을 더하는 반면, 4비트에서는 63.16±0.16%로 베이스라인과 유사한 수준에 머문다. 이는 난이도 강조의 이득이 양자화 오차가 커 학습 신호가 부족한 저비트에서 발현된다는 본 논문의 관찰과 일치한다. 나아가 4비트 설정에서도 제안 기법들을 모두 적용하면 기법 간 상호작용을 통해 더욱 높은 성능을 달성한다.

4. Sensitivity Analysis of the Proposed Methods

본 절에서는 두 구성 요소의 효과를 분석한다. DASG에 대해서는 난이도 재가중 강도의 영향을, SACD에 대해서

www.kci.go.kr

는 이득이 발생하는 비트 구간, 전이 대상인 채널 간 상관 구조의 변형, 그리고 헤드 수와 증류 지점 구성에 대한 민감도를 확인한다. SACD 분석은 GDFQ 베이스라인(CIFAR-100, ResNet-20)에서 경향 확인을 목적으로 단일 시드로 수행하였다.

DASG는 난이도 강조 지수 τ와 보조 가중 β를 갖는다. Table 3에서 β=0.01 설정들은 모두 67% 미만으로 낮아지고 지나치게 강한 강조(τ=2.0)도 이득을 줄여, 어려운 표본의 보강은 유효하되 과도한 강조는 오히려 학습을 저해함을 보인다. 이는 완만한 재가중을 택한 DASG의 설계와 부합하며, 분해 실험(3절)에서 DASG의 이득이 3비트에서 크고 4비트에서 축소되는 것도 난이도 보강이 학습 신호가 부족한 저비트에서 발현된다는 설계 의도와 일치한다. 이에 본 연구는 4비트에서는 β=0.005, τ=0.5를 사용하며, 양자화 오차가 큰 3비트에서는 더 강한 강조(τ=2.0)를 사용한다. SACD의 이득이 발생하는 구간을 확인하기 위해 KD만 사용하는 베이스라인과 SACD 결합 모델을 W8A8부터 W3A3까지 비교하였다(Table 4). 양자화 오차가 거의 없는 W8A8에서는 이득이 나타나지 않았고 저비트 구간(W5A5~W3A3)에서만 +1.05~+1.30%p의 이득이 관찰되어, SACD가 양자화가 중간 특징을 훼손하는 구간에서 작동함을 시사한다. 그 기전을 확인하기 위해 CIFAR-100 테스트셋을 입력으로 각 증류 지점에서 채널 상관행렬 $C = \hat{X}\hat{X}^{\mathrm{T}}$($\hat{X}$는 Frobenius 정규화한 특징맵)를 계산하고 FP 모델과 비교하였다. 토크나이저와 셀프 어텐션은 학습마다 갱신되는 보조 모듈이므로, 그 기반이 되는 모델 내재적 채널 상관 구조를 측정 대상으로 삼았다. 최심층(stage3) 기준(Table 5), SACD가 없을 때 FP 모델과의 상관행렬 거리는 W4A4의 0.0619에서 W3A3의 0.1630으로 커지고 상관 구조의 복잡도를 나타내는 유효 랭크(effective rank)는 FP 대비 36.79로 팽창하여, 비트 폭이 낮을수록 관계 구조의 왜곡이 심화됨을 보인다. 유효 랭크의 팽창은 채널 간 상관이 약해져 구조가 무질서해졌음을 뜻한다. SACD를 결합하면 W3A3에서 거리는 0.0994로, 유효 랭크는 27.71로 FP 모델에 가까워지며, 이 회복폭은 W4A4보다 크다. 상관행렬에서 채널 간 상관이 차지하는 비중을 나타내는 비대각 에너지 비율(off-diagonal energy)도 같은 방향을 보여, FP 모델의 0.894가 SACD가 없는 W3A3 모델에서는 0.763으로 감소했다가 SACD를 결합하면 0.861로 회복된다. 이러한 회복은 최심층에서 두드러졌고 얕은 층에서는 왜곡과 회복의 폭이 모두 작았다. 총 임베딩 차원을 고정하고 헤드 수 h∈{1, 2, 4, 8}을 비교하면(Table 6) 단일 헤드 대비 h=4가 높아 복수의 헤드로 상관 구조를 정렬하는 설계가 유효함을 시사하며, h=8과의 차이는 변동 범위 이내여서 계산 비용을 고려해 h=4를 기본값으로 사용한다. 증류 지점은 최심층 단일 지점보다 복수 지점 설정들이 대체로 높았으나 복수 지점 간 차이는 변동 범위 이내로, 본 논문은 일관성을 위해 전체 지점을 기본값으로 유지한다.

Table 4. Top-1 accuracy (%) of the KD-only baseline and the SACD-augmented model across bit-widths (GDFQ, ResNet-20, CIFAR-100)

Bit-width	Baseline (KD only)	+SACD	Δ
W8A8	70.33	70.33	0.00
W5A5	67.72	68.98	+1.26
W4A4	63.45	64.75	+1.30
W3A3	48.22	49.27	+1.05

Table 5. Channel correlation statistics at stage3, measured against the full-precision (FP) model on the entire CIFAR-100 test set (batch-averaged).

Model	SACD	$\|C_t - C_s\|_F$ (↓)	Eff. rank	Off-diag. energy
FP	-	-	23.50	0.894
W4A4	w/o	0.0619	27.02	0.869
W4A4	w	0.0517	21.17	0.907
W3A3	w/o	0.1630	36.79	0.763
W3A3	w	0.0994	27.71	0.861

Table 6. Effect of the number of attention heads and distillation points in SACD (GDFQ, W3A3). † denotes the default configuration.

attention heads	Top-1 (%)	distillation points	Top-1 (%)
1	47.12	1 (stage3)	49.00
2	48.55	2 (stage2-3)	50.53
4†	48.91	3 (stage1-3)	49.56
8	49.23	4 (all)†	49.94

V. Conclusions

본 논문은 생성자 기반 DFQ의 정확도 격차를 표본 품질만의 문제가 아니라 중간 레이어 전이 신호의 신뢰도와 합성 표본의 정보량이라는 두 축에서 재해석하였다. SACD는 중간 특징을 멀티헤드 셀프 어텐션의 채널 토큰 간 상관 구조로 재구성하여 활성값 크기 대신 관계 구조를 전이함으로써, 양자화 모델이 도달하기 어려운 목표를 부과하지 않고 비교적 안정적인 지도 신호를 제공한다.

www.kci.go.kr

DASG는 원본(FP) 모델 기준 난이도로 생성을 재가중하여 결정 경계 부근의 표본으로 합성 데이터셋을 보강한다.

CIFAR-100 데이터셋을 활용한 ResNet-20 모델 실험에서 SADA-DFQ는 3비트와 4비트 모두에서 생성자 기반 DFQ 중 더 나은 정확도를 보였으며, 3비트에서 AdaDFQ 대비 약 2.6%p 향상을 달성하였다. 동일 조건에서 결합한 기존 중간 레이어 증류 방식(FitNet, AT)과 비교해도 일관되게 우수하였고, 이 이득은 ImageNet(ResNet-18) 3비트에서도 +3.65%p로 재현되었다. 주요 비교 실험에서 제안 기법의 이득은 4비트보다 양자화 오차가 큰 3비트에서 크게 나타났다. 심층 분석에서는 SACD의 이득이 양자화 오차가 발생하는 저비트 구간에서 나타나고 저비트에서 크게 왜곡되는 채널 간 상관 구조가 SACD 적용 시 FP 모델에 가깝게 회복되는 것을 관찰하였으며, DASG의 난이도 재가중은 완만한 강도에서 유효함을 확인하였다. 본 연구의 의의는 저비트 양자화에서 무엇을 전이할 것인가를 활성값 크기에서 관계 구조로 옮기고, 난이도 기반 표본 재가중을 생성자 기반 DFQ에 도입한 데 있다. 향후 연구로는 (i) 2비트 이하 극저비트 설정으로의 확장, (ii) 다른 과제·구조로의 일반화 가능성 검토를 계획한다.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No.RS-2022- NR071978) and funded by the Ministry of Education (No.RS-2022-NR070869); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.RS-2022-II220448: Deep Total Recall, 10%, and IITP-2026-RS-2026-25544527: Leading Generative AI Human Resources Development, RS-2026-25548560: Artificial Intelligence Innovation Graduate School (Inha University)).

REFERENCES

[1] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2704-2713, Salt Lake City, USA, Jun. 2018. DOI: 10.1109/CVPR.2018.00286

[2] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2020. DOI: 10.48550/arXiv.1902.08153

[3] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 7950-7958, Vancouver, Canada, Dec. 2019. DOI: 10.48550/arXiv.1810.05723

[4] M. Nagel, M. van Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 1325-1334, Seoul, Korea, Oct. 2019. DOI: 10.1109/ICCV.2019.00141

[5] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer, "ZeroQ: A novel zero shot quantization framework," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 13169-13178, Seattle, USA, Jun. 2020. DOI: 10.1109/CVPR42600.2020.01318

[6] S. Xu, H. Li, B. Zhuang, J. Liu, J. Cao, C. Liang, and M. Tan, "Generative low-bitwidth data free quantization," Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 1-17, Glasgow, UK, Aug. 2020. DOI: 10.1007/978-3-030-58610-2_1

[7] X. Zhang, H. Qin, Y. Ding, R. Gong, Q. Yan, R. Tao, Y. Li, F. Yu, and X. Liu, "Diversifying sample generation for accurate data-free quantization," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15658-15667, Virtual, Jun. 2021. DOI: 10.1109/CVPR46437.2021.01540

[8] B. Qian, Y. Wang, R. Hong, and M. Wang, "Adaptive data-free quantization," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7960-7968, Vancouver, Canada, Jun. 2023. DOI: 10.1109/CVPR52729.2023.00769

[9] Y. Zhong, M. Lin, G. Nan, J. Liu, B. Zhang, Y. Tian, and R. Ji, "IntraQ: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12339-12348, New Orleans, USA, Jun. 2022. DOI: 10.1109/CVPR52688.2022.01202

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, Mar. 2015. DOI: 10.48550/arXiv.1503.02531

[11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, USA, May 2015. DOI: 10.48550/arXiv.1412.6550

[12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Report, Univ. of Toronto, 2009. URL:

www.kci.go.kr

https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf

[13] K. Choi, D. Hong, N. Park, Y. Kim, and J. Lee, "Qimera: Data-free quantization with synthetic boundary supporting samples," Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 14835-14847, Virtual, Dec. 2021. DOI: 10.48550/arXiv.2111.02625

[14] K. Choi, H. Y. Lee, D. Hong, J. Yu, N. Park, Y. Kim, and J. Lee, "It is all in the teacher: Zero-shot quantization brought closer to the teacher," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8311-8321, New Orleans, USA, Jun. 2022. DOI: 10.1109/CVPR52688.2022.00813

[15] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," Proc. Int. Conf. Learn. Represent. (ICLR), Toulon, France, Apr. 2017. DOI: 10.48550/arXiv.1612.03928

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 5998-6008, Long Beach, USA, Dec. 2017. DOI: 10.48550/arXiv.1706.03762

[17] B. Qian, Y. Wang, R. Hong, and M. Wang, "Rethinking data-free quantization as a zero-sum game," Proc. AAAI Conf. Artif. Intell. (AAAI), Vol. 37, No. 8, pp. 9489-9497, Washington, USA, Feb. 2023. DOI: 10.1609/aaai.v37i8.26136

[18] H. Li, X. Wu, F. Lv, D. Liao, T. H. Li, Y. Zhang, B. Han, and M. Tan, "Hard sample matters a lot in zero-shot quantization," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 24417-24426, Vancouver, Canada, Jun. 2023. DOI: 10.1109/CVPR52729.2023.02340

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770-778, Las Vegas, USA, Jun. 2016. DOI: 10.1109/CVPR.2016.90

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 248-255, Miami, USA, Jun. 2009. DOI: 10.1109/CVPR.2009.5206848

Authors

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 Electrical Computer Engineering at Inha University, Korea. His current research interests are object detection, model compression, efficient learning, and on-device AI.

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. course degree with the Department of Electrical Computer Engineering at Inha University, Korea. His research interests are knowledge distillation, quantization, data-free and on-board AI.

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 Electrical and Computer Engineering at Inha University, His research interests include object tracking, object detection, generative model learning, continual learning, on-device ML, etc.

www.kci.go.kr