VA-QKD: 가상 앵커를 이용한 효율적인 One-Shot W2A2 양자화

최범영, 조성환, 정의림, 김영동[1]

국립한밭대학교

zerotiger0930@gmail.com, josseong1227@gmail.com, erjeong@hanbat.ac.kr, [1]ydkim1293@hanbat.ac.kr

Efficient One-Shot W2A2 Quantization using Virtual Anchors

Choi Bum Young, Jo Seong Hwan, Jeong Eui Rim, Kim Young Dong[1]

Hanbat National University

요 약

최근 엣지 AI 환경에서 연산량과 메모리 효율성을 개선하기 위한 방법으로, 가중치와 활성값을 2비트 정밀도로 제한하는 W2A2 양자화 기법이 연구되고 있다. 이러한 초저비트 모델의 성능 저하는 주로 고정밀도 교사(Teacher) 모델의 지식을 전이하는 양자화 지식 증류(Quantization Knowledge Distillation, QKD)를 통해 완화된다. 그러나 FP32 교사 모델과 W2A2 학생(Student) 모델 간의 표현력 차이로 인해 증류 과정에서 정보 손실이 발생하며, 이는 성능 저하로 이어진다. 기존 연구들은 중간 정밀도의 보조 모델을 추가하여 이 문제를 완화했으나, 이는 추가적인 학습 단계로 인해 학습 비용과 시간이 증가한다. 본 연구는 추가적인 학습 비용 없이 가상 앵커를 활용하여 교사-학생 모델간 표현력 격차를 완화하는 VA-QKD(Virtual Anchor Quantization Knowledge Distillation)를 제안한다. 제안 기법은 교사 모델의 출력값을 학습 과정에서 즉시 양자화하여 해당 결과를 학생 모델의 증류 타겟으로 사용한다. 이를 통해 학생 모델은 FP32 교사 출력과 해당 출력의 양자화 결과들을 손실 계산에 사용하여 학습한다. CIFAR-100 데이터셋 실험 결과, 제안 기법은 기존 QKD 대비 정확도를 향상시키면서 초저비트 양자화 모델의 경량화 효율을 유지하면서 성능 저하를 완화하였다.

I. 서론

최근 모바일 및 IoT(Internet of Things) 기반 엣지 디바이스에서 실시간 추론 요구가 증가함에 따라, 심층 신경망(DNN)의 연산량과 메모리 사용량을 줄이기 위한 경량화가 주요 과제로 다루어지고 있다 [1]. 이러한 접근 중 하나로, 가중치(Weight)와 활성값(Activation)을 모두 2비트 정밀도로 제한하는 W2A2가 초저비트 양자화 기법의 한 형태로서 연구되고 있다 [2]. W2A2는 FP32 모델 대비 메모리 사용량을 최대 16배 감소시키며, 곱셈 연산을 비트 연산으로 대체함으로써 연산 복잡도를 줄일 수 있다.

그러나 초저비트 양자화는 모델의 표현력을 크게 제한하며, 이로 인한 추론 성능 저하가 발생한다. 이러한 성능 저하를 완화하기 위한 방법으로, 교사(Teacher) 모델의 출력을 학생(Student) 모델이 학습하도록 하는 지식 증류(Knowledge Distillation, KD)가 주로 사용된다 [3]. 특히 양자화 환경을 고려하여 설계된 양자화 지식 증류(Quantization Knowledge Distillation, QKD)는 비트 정밀도가 제한된 학생 모델이 FP32 교사 모델의 출력을 학습하도록 함으로써, 양자화로 인한 정확도 감소를 줄이는 데 활용된다 [4].

하지만 W2A2와 같이 표현력이 극도로 제한된 학생 모델이 FP32 교사 모델의 출력을 직접 모방하는 경우, 두 모델 간 표현력 차이로 인해 증류 과정에서 큰 정보 손실이 발생하여 학습이 불안정해지거나 충분한 성능 향상을 얻지 못하는 문제가 존재한다 [5]. 이를 해결하기 위해 기존 연구들은 INT8 또는 INT4 수준의 중간 정밀도 모델을 교사 보조자로 도입하는 다단계 증류 방식(Teacher Assistant Knowledge Distillation, TAKD)을 제안하였다 [6]. 하지만 이 기법은 학습 비용과 학습 시간이 크게 증가한다는 한계가 존재한다.이러한 한계는 중간 정밀도 모델을 추가로 학습하는 구조에서 기인한다. 중간 모델은 교사 출력과 학생 출력 간의 분포 차이를 완화하는 역할을 수행하지만, 이를 위해 별도의 학습 단계가 요구된다.

이에 본 연구는 중간 정밀도의 모델을 학습하지 않고도 교사–학생 간 표현력 격차를 완화하는 중간 출력을 활용하는 VA-QKD 프레임워크를 제안한다. 구체적으로, FP32 교사 모델의 출력을 학습 과정에서 즉시 양자화하여, FP32 출력보다 단순화된 형태의 중간 출력을 가상 앵커로 생성하고, 이를 학생 모델의 증류 타겟으로 정답 라벨과 함께 학습을 수행한다.

II. 가상 앵커를 활용한 지식 증류 방법

본 연구에서 제안하는 VA-QKD 프레임워크는 가상 앵커 생성 단계와 이를 이용한 지식 증류 단계로 구성된다.

A. 가상 앵커 생성 단계

기존의 Teacher Assistant(TA) 기반 방법과 달리, 제안 기법은 별도의 보조 네트워크를 학습하지 않는다. 대신 FP32 교사 모델의 출력값에 양자화를 직접 적용하여 가상 앵커를 생성한다. 가상 앵커는 다음과 같이 정의된다.

$$\mathbf{z}_{\mathrm{VA}} = Q(\mathbf{z}_T, b) \tag{1}$$

여기서 $\mathbf{z}_T$는 FP32 교사 모델의 출력을 의미하며, b는 가상 앵커의 비트 정밀도를 나타낸다. 이 과정은 학습 과정 중 순전파 단계에서 수행되는 수치 연산으로 구성되며, 역전파를 필요로 하지 않는다. 따라서 추가적인 메모리 사용이나 학습 시간 증가 없이

[1]교신 저자

가상 앵커를 생성할 수 있다.

가상 앵커는 FP32 교사 출력의 상대적인 클래스 순위 정보를 유지하면서, 비트 정밀도를 제한하여 학생 모델에게 단순화된 출력 분포를 제공한다.

B. 지식 증류 단계

학생(Student) 모델은 최종 정답(Hard Target)과 가상 앵커로부터 제공되는 중간 출력(Soft Target)을 함께 사용하여 학습한다. 이를 위해 학생 모델의 출력은 가상 앵커 및 정답 레이블과 각각 비교되며, 이들로부터 계산된 손실을 결합하여 최종 손실을 구성한다.

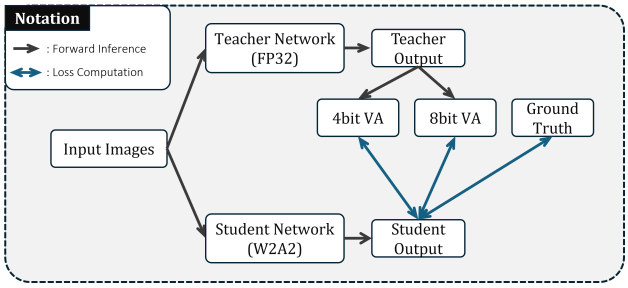

그림 1: VA-QKD의 손실 함수 구성 및 학습 구조

그림 1은 제안하는 손실 함수와 전체 학습 구조를 나타낸다. 교사 모델의 FP32 출력을 양자화한 가상 앵커는 정답 라벨과 함께 학생 모델의 출력과 각각 비교되며, 이 과정에서 계산된 손실 항들은 합산되어 최종 학습 손실로 사용된다.

III. 실험 결과 및 분석

A. 실험 결과

본 연구는 CIFAR-100 데이터셋을 사용하여 제안하는 VA-QKD 기법의 정량적 성능을 평가한다. 실험에는 ResNet-32 구조를 사용하며, 학생 모델은 가중치와 활성값을 모두 2비트로 양자화한 W2A2 설정으로 학습한다. 비교 실험은 로짓(logit) 기반 증류 손실, 특징 맵(feature map) 정렬 손실, 그리고 대조(contrastive) 손실의 세 가지 설정에 대해 각각 수행되었으며 각 설정마다 동일한 W2A2 환경에서 학습된 QKD 모델과 VA-QKD 모델의 성능을 비교한다.

표 1: CIFAR-100에서 각 모델의 성능 비교

Model	Bit-width (W/A)	Top-1 Acc. (%)	Model Size (MB)	BOPs Reduct.
Teacher (FP32)	32 / 32	70.19	1.78	1×
VA-QKD (Logit)	2 / 2	65.06	0.11	256×
QKD (Logit)	2 / 2	64.97	0.11	256×
VA-QKD (Feature Map)	2 / 2	64.40	0.11	256×
QKD (Feature Map)	2 / 2	64.24	0.11	256×
VA-QKD (Contrastive)	2 / 2	64.64	0.11	256×
QKD (Contrastive)	2 / 2	64.48	0.11	256×

표 1은 W2A2 설정에서 QKD와 VA-QKD의 성능을 비교한 결과를 나타낸다. 실험 결과, VA-QKD는 기존 QKD 모델 대비 Top-1 정확도를 0.09%에서 0.16%까지 향상시킨다. 해당 성능 향상은 추가적인 파라미터 증가나 별도의 보조 모델 학습 없이 달성했다는 것에 의의가 있다.

B. 정성적 분석

모델이 분류 과정에서 활용하는 시각적 단서를 분석하기 위해 Grad-CAM을 적용하였다.

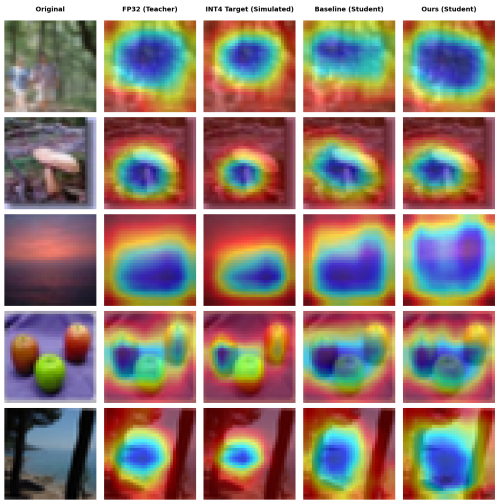

그림 2: Grad-CAM을 통한 기존 QKD 모델과 VA-QKD 모델의 활성화 비교

그림 2은 기존 QKD 모델과 VA-QKD 모델의 활성화 결과를 비교한 예시를 보여준다. 기존 QKD 모델은 초저비트 양자화에 따른 정보 손실로 인해 객체의 전반적인 형상을 포착하지 못하고 활성화 영역이 지엽적인 특징에 국한되는 한계를 보인다. 반면, VA-QKD 모델의 활성화 영역은 객체의 핵심 영역에 강하게 집중하는 패턴을 나타낸다.

이는 가상 앵커가 양자화 과정의 노이즈를 필터링하고 분류에 결정적인 정보만을 선별하여 전달함으로써 초저비트 환경에서도 모델이 분류에 필요한 시각적 단서를 보다 일관되게 활용함을 정성적으로 확인할 수 있다.

참 고 문 헌

[1] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision (ECCV)*, pp. 525–542, 2016.

[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] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] J. Kim, Y. Bhalgat, J. Lee, C. Patel, and N. Kwak, "QKD: Quantization-aware knowledge distillation," *arXiv preprint arXiv:1911.12491*, 2019.

[5] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4794–4802, 2019.

[6] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5191–5198, 2020.