ADWQ: 활성화 분포를 반영한 LLM 가중치 양자화

박서윤, 마예은, 김명준, 안진현*
명지대학교

dhn04100@mju.ac.kr, k41031614@mju.ac.kr, tjmjtjmj2237@mju.ac.kr,
*wlsgus3396@mju.ac.kr

ADWQ: Activation-Distribution-aware Weight Quantization for LLMs

Seoyun Park, Yeeun Ma, Myoungjun Kim, Jin-Hyun Ahn*
Myongji Univ.

요 약

본 논문은 대표적인 INT4 가중치 양자화 방법인 AWQ[1]의 한계를 분석하고, 활성화 분포의 고차 통계량을 활용하여 이를 개선한 ADWQ 를 제안한다. AWQ 는 활성화의 크기 정보만을 사용하여 양자화 스케일을 결정하지만, ADWQ 는 각 트랜스포머 블록의 특성에 맞춰 활성화 분포의 평균과 분산을 동적으로 혼합하는 새로운 통계량을 도입한다. 또한, 블록 및 레이어 단위의 계층적 탐색을 통해 최적의 스케일링 파라미터 (λ, α) 를 결정하여 재구성 오차를 최소화한다. TinyLlama-1.1B 모델을 사용한 실험에서, 제안하는 ADWQ 는 Wikitext 및 C4 데이터셋에서 AWQ 대비 Perplexity (PPL)를 유의미하게 개선했으며, 원본 모델의 지식 보존 및 예측 정확도 지표에서도 전반적인 성능 향상을 달성하였다. 본 연구는 활성화 분포의 통계적 특성을 정교하게 반영하는 것이 양자화 성능 향상에 기여함을 입증하였다.

Ⅰ. 서 론

대규모 언어모델 (LLM)의 메모리 및 연산 비용 문제를 해결하기 위해, 가중치만 4 비트로 양자화하는 Post-Training Quantization(PTQ)이 널리 사용되고 있다. 대표적인 PTQ 기법인 GPTQ[2]는 소규모 캘리브레이션 데이터에서 얻은 2 차 정보(second-order)를 이용해 가중치를 재구성함으로써 높은 정확도를 달성한다. 다만, 캘리브레이션 데이터 분포에 민감하다는 한계가 있으며, 이를 보완한 AWQ[1]는 채널별 활성화의 크기를 기반으로 한 스케일링을 통해 안정적인 성능을 달성했다.

하지만 AWQ의 스케일링 방식은 활성화 분포의 단순 크기 정보에만 의존하므로, 레이어마다 상이한 분포의 형태를 충분히 반영하지 못하는 문제를 가진다. 이러한 문제를 해결하기 위해, 본 논문은 ADWQ(Activation-Distribution-aware Weight Quantization)를 제안한다. ADWQ는 활성화 분포의 평균과 분산을 모두 고려하는 새로운 통계량을 정의하고, 각 트랜스포머 블록의 특성에 맞춰 이들의 중요도를 동적으로 조절하는 계층적 탐색 프레임워크를 도입한다. 이를 통해 기존 방식보다 더 정교하고 적응적인 양자화를 수행하여 모델의 성능 저하를 효과적으로 최소화함을 보인다.

Ⅱ. 제안 기법: ADWQ

대칭 균일 양자화에서 가중치 W 는 $Q(W) = \Delta \cdot Round\left(\frac{W}{\Delta}\right)$, $\Delta = \frac{\max(|W|)}{2^{N-1}}$로 변환된다. 여기서 Δ는 양자화 스케일러이다. AWQ는 일부 채널의 활성화 값이 다른 채널보다 훨씬 큰 문제에 주목한다. 이 경우, 해당 채널의 가중치 W는 양자화 오차에 더 민감해진다. 이 문제를 완화하기 위해 AWQ는 $W' = W \cdot diag(s)$와 $X' = diag(s)^{-1} \cdot X$와 같이 가중치와 활성화를 동시에 재조정하는 채널별 스케일링 계수 s를 도입한다.

$$Q(w \cdot s) \cdot \frac{x}{s} = \Delta' \cdot Round\left(\frac{ws}{\Delta'}\right) \cdot x \cdot \frac{1}{s} \; (w \in W)$$

이 변환은 $W'X' = WX$를 만족하여 수학적으로 동치이지만, s를 통해 가중치의 동적 범위를 조정하여 양자화 오차를 최소화한다.

본 연구는 바로 이 스케일링 계수 s를 찾는 더 정교한 방법인 ADWQ를 제안한다. AWQ가 활성화의 크기 정보에만 의존하는 한계를 극복하기 위해, ADWQ는 채널별 평균과 분산을 모두 반영하여 스케일링 계수 s를 탐색함으로써 양자화로 인한 성능 저하를 최소화한다.

2.1 활성화 분포의 고차 통계량 도입

AWQ는 활성화 텐서 X의 입력 채널별 절댓값 평균에 의존하여 스케일링 계수를 탐색한다. 이 방식은 분포의 크기 외에 중심 위치나 퍼짐 정도와 같은 다른 중요한 통계적 특성을 충분히 활용하지 못하는 한계를 가진다. 이를 극복하고자, ADWQ는 활성화의 채널별 평균과 분산을 모두 반영하는 새로운 통계량 $stat_\lambda(X)$을 정의한다.

$$stat_\lambda(X) = \sqrt{\mu(X)^2 + \lambda \cdot Var(X) + \epsilon},$$

$$s = (stat_\lambda(X))^\alpha$$

여기서 하이퍼파라미터 λ는 평균과 분산의 상대적 중요도를 조절하며, 이 통계량은 활성화 분포를 더 종합적으로 표현하는 기반이 된다.

2.2 스케일링 계수 매개화 및 목적 함수

ADWQ는 $stat_\lambda(X)$를 기반으로 스케일링 계수 s를 매개화하며, 여기서 지수 α는 채널별 민감도를 제어하는

역할을 한다. 탐색의 목표는 아래와 같이 정의된 양자화 재구성 오차 $L(s)$를 최소화하는 것이다.

$$L(s) = \left\| Q\left(W \cdot diag(s)\right)\left(diag(s)^{-1} \cdot X\right) - WX \right\|^2$$

2.3 계층적 탐색

최적의 스케일링 계수를 찾기 위해, ADWQ는 계층적 탐색 방식을 도입한다. 탐색 과정은 다음과 같은 2 단계로 구성된다.

2.3.1 레이어 단위 α 탐색

각 트랜스포머 블록 b내의 기능적 레이어 그룹 g마다, 주어진 λ값에 대해 양자화 오차 $L_{b,g}$를 최소화하는 최적의 지수 $\alpha^*_{b,g}(\lambda)$를 독립적으로 탐색한다.

$$\alpha^*_{b,g} = \arg min_{\alpha \in [0,1]} L_{b,g}(\lambda, \alpha)$$

2.3.2 블록 단위 λ 탐색

후보 집합 Λ의 모든 λ값에 대해 1 단계 과정을 수행하고, 각 레이어에서 계산된 최소 오차들의 총합이 가장 작아지는 블록 최적 λ^*_b를 최종적으로 선택한다.

$$\lambda^*_b = \arg min_{\lambda \in \Lambda} \sum_{g \in Layers} L_{b,g}(\lambda, \alpha^*_{b,g}(\lambda))$$

이러한 계층적 접근을 통해, ADWQ는 블록의 거시적 특성(λ)과 레이어의 미시적 분포(α)를 모두 고려하여, 기존 방식보다 더 정교하고 적응적인 양자화를 수행한다.

Ⅲ. 실험 및 결과

3.1 실험 환경

모든 실험은 NVIDIA A100 GPU 40GB 환경에서 수행하였다. 베이스 모델로는 TinyLlama-1.1B[3]를 사용하였으며, 모든 양자화는 4-bit 정밀도와 128 의 그룹 크기를 적용하였다. 평가는 기존 AWQ 기법을 baseline으로 설정하고, ADWQ를 비교 분석하였다.

3.2 데이터셋 및 평가 지표

모델 평가는 Wikitext-103, Wikitext-2[4], C4[5]데이터 셋의 초기 32,768 개 토큰을 사용하여 진행하였다. 성능은 PPL, Teacher-Student Metrics(KL Divergence, Top-k Overlap). Ground Truth Metrics(NTA(Next Token Accuracy), Top-k Recall) 지표를 통해 다각적으로 측정하였다.

3.3 실험 결과

Table 1: PPL (Perplexity)

Method	Wikitext-103	Wikitext-2	C4
FP16 (Baseline)	8.732	7.888	10.004
AWQ	9.132	8.220	10.544
ADWQ	**9.076**	**8.204**	**10.465**

Table 2: KL Divergence

Method	Wikitext-103	Wikitext-2	C4
AWQ	0.04482	0.04190	0.05162
ADWQ	**0.04376**	**0.04172**	**0.04091**

Table 3: Top-k Overlap(%) ($k \in (3,5)$)

Method ($k=3$ / $k=5$)	Wikitext-103	Wikitext-2	C4
AWQ	87.38/87.11	87.32/87.04	87.44/87.28
ADWQ	**87.42/87.20**	**87.40/87.93**	**87.51/87.48**

Table 4: NTA (%)(Next Token Accuracy)

Method	Wikitext-103	Wikitext-2	C4
FP16 (baseline)	53.35	55.67	52.04
AWQ	52.68	55.08	50.98
ADWQ	**52.76**	**55.18**	**51.13**

Table 5: Top-k Recall (%) ($k \in (3,5)$)

Method ($k=3$ / $k=5$)	Wikitext-103	Wikitext-2	C4
FP16 (baseline)	70.15/76.13	71.84/77.46	68.20/74.29
AWQ	69.41/75.70	71.17/76.79	**67.50**/73.67
ADWQ	**69.52/75.75**	**71.18/77.01**	67.49/**73.75**

본 연구에서 제안하는 ADWQ는 기존 양자화 기법인 AWQ와 비교하여 모든 평가 지표에서 일관된 성능 향상을 보였다. Table 1 에서 확인할 수 있듯이, Wikitext-103, Wikitext-2, C4 데이터 셋에서 모두 AWQ의 PPL을 능가하는 결과를 달성하며 제안 방법론의 효용성을 입증했다. 또한 Table 2 와 Table 3 은 ADWQ가 KL Divergence와 Top-k Overlap 지표에서 더 나은 성능을 보임을 보여주는데, 이는 원본 모델의 언어적 지식을 더 효과적으로 보존한다는 것을 의미한다. 마지막으로 Table 4 와 Table 5 의 NTA 및 Top-k Recall 결과는 이러한 지식 보존 능력이 실제 정답 예측 정확도의 향상으로 이어진다는 것을 명확히 보여준다.

Ⅳ. 결론

본 연구는 최신 양자화 기법인 AWQ가 활성화의 1 차원적 통계량에 의존하는 한계를 극복하고자, 분포의 평균과 분산을 종합적으로 고려하는 ADWQ를 제안한다. 제안 기법은 트랜스포머 블록과 레이어의 고유한 특성에 맞춰 스케일링 파라미터(λ, α)를 정교한 계층적 탐색 메커니즘으로 최적화하여 재구성 오차를 최소화한다.

실험을 통해, ADWQ는 PPL을 포함한 거의 모든 평가 지표에서 기존 AWQ의 성능을 상회함을 실증적으로 입증하였으며, 이는 활성화의 고차 모멘트를 활용하는 것이 양자화 효율을 극대화하는 핵심 전략임을 시사한다. 향후 본 연구에서 제시한 ADWQ를 LLM에 확장 적용함으로써, 효율적인 압축 기술 발전에 기여할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00212836)

참 고 문 헌

[1] Lin, Ji, et al, "Awq: Activation-awqre weight quantization for on-device llm compression and acceleration." *Proceedings of machine learning and systems* 6 (2024): 87-100.

[2] Frantar, Elias, et al. "Gptq: Accurate post-training quantization for generative pre-trained transformers." *arXiv preprint arXiv:2210.12323* (2022).

[3] Zhang, Peiyhuan, et al. "Tinyllama: An open-source small language model." *arXiv preprint arXiv:2401.02385* (2024).

[4] Merity, Stephen et al. "Pointer sentinel mixture models." *arXiv preprint arXiv:1609.07843* (2016).

[5] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020):1-67.