LLM의 온디바이스 추론을 위한 양자화에 관한 연구

정성훈, 이영재
한국전자통신연구원
tamtam211@etri.re.kr, lyj4295@etri.re.kr

A Study on the quantization for on-device LLM inference

Jung Sung Hoon, Lee Young Jae
ETRI

요 약

대규모 언어 모델의 발전과 함께 증가하는 모델 크기와 연산량은 고성능 클라우드 컴퓨팅 환경에 대한 의존성을 심화시킨다. 이 때문에 다양한 LLM 응용 솔루션은 데이터 프라이버시 및 서비스 지연 시간과 같은 다양한 한계를 맞고 있다. 이러한 문제를 해결하기 위해, LLM을 스마트폰 데스크톱 등 여러 엣지 디바이스에서 직접 추론하기 위한 온디바이스 추론 최적화 기술이 필수적으로 요구된다. 본 논문은 추론 최적화 기술 중 하나인 PTQ(Post-Training-Quantization)을 중심으로, 다양한 LLM에 대한 비트 정밀도별 메모리 절감과 성능 저하 사이의 상충 관계를 정량적으로 분석하고 실질적인 가이드라인을 제시하는 것을 목표로 한다. 본 연구는 엣지 디바이스의 하드웨어 제약 조건에 맞는 양자화 전략을 선정할 수 있는 통찰력을 제공하여 LLM 기반 솔루션의 온디바이스 배포에 기여한다.

Ⅰ. 서 론

대규모 언어 모델(Large Language Model, LLM)은 급격한 발전을 바탕으로 전문적인 작업 수행부터 개인화된 일상 활동에 이르기까지 광범위한 영역에 보급되었다. 이러한 LLM은 기술 발전과 함께, 모델 크기와 요구 연산량이 점점 커졌다. 이에 따라, 고성능 GPU는 LLM기반 솔루션 운용에 필수 요소가 되었으며, 결과적으로 LLM 기반 솔루션 운용의 클라우드 컴퓨팅 환경에 대한 의존성이 심화되었다. 서비스의 적용 분야가 다양해짐에 따라, 클라우드 컴퓨팅 환경은 데이터 프라이버시, 서비스 지연 시간 그리고 운영 비용 측면에서 한계를 야기한다. 이러한 문제를 해결하기 위해, 스마트폰, 데스크탑 같은 다양한 디바이스 내에서만 LLM을 추론시키는 연구 및 시도가 활발하게 이루어지고 있다.

온디바이스 LLM 추론은 각 디바이스의 연산 자원 및 메모리 용량에 제약을 받는다. 이러한 제약조건을 만족하며, 기존 LLM의 성능을 유지하기 위해 추론 시스템과 모델 차원의 최적화 기술 개발이 이루어지고 있다, [2] 그중 모델 압축(Model Compression)은 모델의 구조나 파라미터를 압축시켜 모델의 크기를 줄이는 방식으로, 크게 양자화(Quantization), 가지치기(pruning) 그리고 지식 증류(Knowledge Distillation, KD) 3가지로 분류된다.

양자화는 네트워크의 가중치 혹은 활성화의 데이터 타입을 고-정밀도(high-precision)에서 저-정밀도(low-precision)로 나타내는 기술이다. 또한, 가지치기는 모델의 불필요한 파라미터를 제거하는 기술로서, 네트워크의 구성요소(예: neurons, attention heads, layer etc) 전체를 제거하는 structured와 일부를 0으로 처리하는 unstructured 방식이 있다. 지식 증류는 선생이라 부르는 네트워크의 정보를 더 작은 학생이라는 네트워크에 학습을 통해 전달하는 모델 압축 기술이다. 특히, 양자화는 다른 모델 압축 기법과 비교하여 모델의 크기 절감 효과가 직접적이기 때문에, 엣지 디바이스의 제한된 메모리 용량에 적합하며, on-device 추론의 표준기법으로 자리매김했다. [2] 그러나, 다양한 모델 아키텍처와 대표 양자화 기법 간의 성능 및 메모리 효율성을 종합적으로 비교한 분석은 부족한 실정이다.

따라서, 이러한 배경 아래, 본 연구는 다양한 디바이스 내 LLM 추론을 위해 대표적인 양자화 기법을 적용한 모델들의 크기와 성능을 제공하는 것을 목표로 한다. 이를 통해 엣지 디바이스의 하드웨어 제약 조건에 맞게 최적의 양자화 방식과 모델을 선정할 수 있는 실질적인 가이드라인을 제공하고자 하며, 궁극적으로 향후 온디바이스 LLM 최적화 연구를 위한 통찰력을 제공하는데 기여할 것이다.

Ⅱ. 양자화

양자화는 학습 여부에 따라 Quantization-aware Training(QAT)와 Post-training Quantization(PTQ)로 분류된다. QAT는 미리 학습된(pre-trained) 고-정밀도의 LLM을 활용해, 저-정밀도 모델을 재학습시키는 방법으로, 양자화로 인해 발생하는 성능 저하를 효과적으로 완화 시킨다. 그러나 QAT는 현대 LLM의 규모에서는 엄청난 학습 비용을 요구하는 실질적인 한계가 존재한다.

반면에 PTQ는 모델을 재훈련하는 과정 없이 pre-trained LLM을 즉시 양자화하는 방법으로, 소량의 데이터셋(calibration dataset)혹은 데이터 없이 진행된다. 이러한 실용성 덕분에 PTQ는 LLM 추론을 위해 널리 채택되는 경량화 방식으로 자리매김했다.

본 연구는 이러한 PTQ 기법 중 LLM 추론 분야에 성능 보존 효과가 입증되며, 널리 배포된 GPTQ [9], AWQ(Activation-aware Weight Quantization) [6]를 연구 대상으로 선정한다. 또한, 대표 LLM인 Llama families [6, 7], Qwen 3 에 적용된 결과를 분석한다. [8] 특히, 엣지 디바이스의 일반적인 메모리 제약을 고려해, 파라미터가 100조 개(10B) 이하 규모의 모델들을 분석 대상으로 선정하였다.

모델	bit precision	ppl (↓)	크기(GB)
Llama-7B	16	6.14	13.5
	4	7.36	3.38
Llama2-7B	16	5.47	13.5
	4	5.83	3.38
Llama3-8B	16	5.68	16
	4	5.97	4
Qwen3-0.6B	16	12.7	1.52
	8	12.7	0.76
	4	14.9	0.38
	3	85.9	0.29
	2	64200000	0.19
Qwen3-1.7B	16	9.39	4.08
	8	9.38	2.04
	4	9.99	1.02
	3	41.8	0.77
	2	11300000	0.51
Qwen3-4B	16	7.9	8.06
	8	7.89	4.03
	4	8.19	2.02
	3	26.3	1.51
	2	113	1.01
Qwen3-8B	16	6.99	16.4
	8	6.99	8.2
	4	7.22	4.1
	3	22.6	3.08
	2	53.1	2.05

표 1

Ⅲ. 실험 결과

표 1은 Llama 및 Qwen3 등 주요 LLM에 비트 정밀도(bit precision)가 다른 PTQ를 적용한 후, 토큰 생성 능력 지표인 퍼플렉시티(Perplexity, ppl)와 모델 크기 변화를 비교 분석한 결과이다. 해당 결과는 양자화를 통한 메모리 절감과 성능 저하 사이의 상충-관계(trade-off)를 보여준다.

표 1을 통해, 4 비트 양자화는 메모리 효율성과 성능 안정성 측면에서 가장 실용적인 전략으로 확인된다. 4비트 양자화는 모델 크기를 75% 절감하며, 퍼플렉시티가 평균 8.89% 증가했으며, 특히 8B급 모델의 4비트 양자화는 그 이하의 정밀도 양자화에서 발생하는 성능 저하와 비교할 때 안정성을 보인다. 이를 통해 적은 성능 저하로 최대의 메모리 이득을 취할 수 있음을 입증한다. 이러한 전략은 10B 이하의 LLM을 최대 약 4GB로 압축해 스마트폰과 같은 엣지 디바이스 환경에서 추론을 가능하게 하는 핵심 기술이다.

반면, 8비트 양자화는 4비트 대비 낮은 메모리 절감률(50%)을 갖지만, 모든 모델 크기에서 퍼플렉시티 변화율이 0에 가까워 성능 민감도에 가장 안전한 표준 옵션임을 시사한다. 3비트 및 2비트와 같은 극단적인 양자화 결과는 퍼플렉시티 값이 수백만 단위로 폭증하여 치명적인 성능 저하를 보인다. 이러한 극단적인 양자화의 성능 저하를 극복하기 위해 새로운 양자화 방식을 도입하거나, 상대적으로 크기가 작은 LLM을 선택하는 것이 효율적이다.

Ⅳ. 결론

본 논문은 온디바이스 LLM 추론을 위해 디바이스 제약 극복을 목표로, 양자화된 다양한 LLM에 대한 메모리 절감 및 성능 저하 사이의 상충 관계를 정량적으로 분석하고 실질적인 가이드라인을 제시하였다. 그 결과, 4비트 양자화는 가장 실용적인 최적화 지점으로 확인되었다. 4비트 양자화는 평균적으로 모델 크기를 75% 절감하는 동시에 8B급 모델에서 극단적인 저 정밀도 양자화 대비 현저히 우수한 안정성을 보이며, 적은 성능 손실로 최대의 메모리 이득을 취할 수 있음을 입증하였다.

이러한 결과는 LLM 기반 솔루션 운용의 클라우드 컴퓨팅 환경에 대한 의존성을 낮추고, 데이터 프라이버시 및 서비스 지연 시간 문제 해결의 실마리를 제공한다. 특히, 4비트 양자화 전략은 10B 이하의 LLM을 4GB로 압축하여 스마트폰과 같은 엣지 디바이스 내에서 추론을 가능하게 하는 핵심 기술임을 입증하였다.

향후 연구에서는 현재 극단적으로 낮은 정밀도의 양자화된 모델에서 나타나는 성능 저하를 극복하는 방법에 관해 탐구할 필요가 있다. 이를 통해 보다 다양한 자원 제한을 갖는 엣지 디바이스에서 고성능 LLM 추론 기반 솔루션을 개발할 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 IITP 사업의 일환으로 수행되었음 [25HT1210, 온디바이스 엣지 AI용 eFlash 기반 아날로그 PIM 반도체 개발]

참 고 문 헌

[1] Xu, Jiajun, et al. "On-device language models: A comprehensive review." arXiv preprint arXiv:2409.00088 (2024).

[2] Park, Youngsuk, et al. "Inference optimization of foundation models on ai accelerators." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024.

[3] Zhou, Zixuan, et al. "A survey on efficient inference for large language models." arXiv preprint arXiv:2404.14294 (2024).

[4] Pope, Reiner, et al. "Efficiently scaling transformer inference." Proceedings of machine learning and systems 5 (2023): 606-624.

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

[6] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[7] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[8] Zheng, Xingyu, et al. "An empirical study of qwen3 quantization." arXiv preprint arXiv:2505.02214 (2025).

[9] Frantar, Elias, et al. "Gptq: Accurate post-training quantization for generative pre-trained transformers." arXiv preprint arXiv:2210.17323 (2022).