sLLM 양자화의 성능-효율성 트레이드오프에 대한 실증적 분석 연구

김강우, 함승재, 박상재, 성민규, 강재모*

경북대학교

{kkwo1013, hamsj99, psj1069, alsrb0351, *jmkang}@knu.ac.kr

sLLM Quantization : An Empirical Study of the Performance-Efficiency Trade-off

Kim Kang woo, Ham Seung Jae, Park Sang Jae, Sung Min Gyu Kang Jae-Mo*

Kyungpook National Univ.

요 약

최근 인공지능 분야는 LLM(Large Language Model)의 등장으로 다방면에 걸쳐 꾸준한 발전을 거듭하고 있다. LLM은 수천억 개에 달하는 파라미터를 기반으로 인간의 언어를 이해하고 생성하는 LLM의 능력은 다양한 산업 분야에서 새로운 가능성을 제시했다. 하지만 막대한 연산과 높은 비용, 느린 추론 속도라는 제약이 있다. 이러한 한계는 실시간 서비스나 On-device AI와 같이 빠른 응답 속도를 요구하는 환경에서 LLM의 도입을 가로막는 주요한 장벽이 되고 있다. 이러한 문제의 대안으로 적은 파라미터 수와 빠른 추론 속도를 기대할 수 있는 sLLM(small Large Language Model)이 강력한 대안으로 부상하고 있다. 본 논문은 sLLM의 효율성을 극대화하기 위한 양자화(Quantization) 알고리즘을 적용했을 때 발생하는 성능 저하와 그로 인해 얻을 수 있는 이점 사이의 상충 관계를 규명한다. 또한 향후 발전할 sLLM과 같은 언어 모델 연구를 위한 기반이 되는 자료를 제시한다.

Ⅰ. 서론

최근 인공지능 분야는 LLM의 놀라운 기술의 성장을 보여주고 있다. LLM은 인간의 언어를 이해하고 생성하는 능력을 바탕으로 AI 챗봇 상담, 소프트웨어 코드 생성, 콘텐츠 창작 등 산업 전반에 걸쳐 산업의 성장을 촉진하고 있다. 특히, Google, META, xAI와 같은 글로벌 플랫폼 기업들은 대규모 데이터 센터와 막대한 컴퓨팅 자원을 기반으로 하는 중앙 집중식 클라우드 모델을 통해 사용자 경험을 극대화하고 있다. 중앙 집중식 클라우드 방식을 기반으로 개인화된 AI 서비스를 제공하고 있으나, 막대한 자원과 높은 운용비용, 네트워크 지연 시간이라는 본질적인 한계가 존재한다. 사용자들은 서비스 이용을 위해 중앙 서버에 매번 요청을 보내고 받는다. 이러한 방식은 실시간 상호작용이 필수적인 응용 분야에서는 물리적인 제약으로 인해 불필요한 지연이 발생하며, 사용자 데이터를 중앙 서버로 전송해야 하는 과정에서 데이터 프라이버시와 보안에 대한 우려 또한 증대된다. 이러한 한계를 극복하기 위해서는 클라우드 의존성을 줄이고, E2E(End-to-End)에서 직접 AI 연산을 수행하는 On-device AI 패러다임의 필요성이 대두된다.

Ⅱ. 관련 연구

LLM : 최근 LLM의 파라미터 수가 증가하면서 막대한 자원을 소모하게 되자 LLM의 비용에 대한 중요성이 커지고 있다. Google의 Gemma[2]와 같이 QAT(Quantization Aware Training)[1]을 사용하여 메모리 사용량을 대폭 줄인 결과를 보여주거나 OpenAI의 gpt-oss 20b과 같은 모델과 같이 MoE(Mixture-of-Experts)[2] 아키텍처를 사용하여 추론 시 비용을 크게 절감한 LLM들이 등장하고 있다. 이는 고비용의 모델이 아닌 저비용 고성능의 모델이 필요하다는 것을 시사한다.

sLLM: 최근 LLM의 발전과 함께 sLLM의 성장은 AI 기술의 개인 보급화를 앞당기는 핵심이다. Microsoft, Google, Meta 등은 거대 모델의 막대한 비용과 자원 한계를 극복하고자, 고도로 정제된 데이터와 최적화된 아키텍처를 통해 적은 파라미터와 비용으로도 놀라운 성능을 내는 sLLM 개발에 집중하고 있다. 이러한 연구는 중앙집중형 방식에서 개인의 단만 장비로 가져오는 On-device AI를 현실화하고 있다. sLLM의 발전은 인터넷 연결 없이도 빠르고 안전하게 AI를 사용하는 환경을 조성하며, 기술의 실용성과 접근성을 극대화하는 환경을 만들 수 있다.

양자화 : 기존 양자화는 FP4나 FP8과 같은 양자화는 성능의 감소를 감수하더라도 빠른 추론 속도와 낮은 비용으로 추론을 할 수 있도록 실용적 목표에 집중했다. 최근 양자화 기술은 GPTQ(Generative Pre-trained Transformer Quantization)[3]와 AWQ(Activation-aware Weight Quantization)[4]와 같은 방식들은 모델의 중요한 가중치를 보호하는 것과 같이 양자화 과정에서 성능 저하를 최소화하면서도 비용 효율성을 확보하여, 원본 모델의 성능을 거의 그대로 유지를 중심으로 발전하고 있다. 이는 모델의 성능과 비용의 균형이 중요한 것을 알 수 있다.

Ⅲ. 본론

본 논문에서는 최신 양자화 기법들을 적용한 sLLM의 성능을 변화를 측정한다. 이를 통해, 제한된 연산 자원을 가진 엣지 디바이스 환경에서도 충분한 가능성을 입증하고자 한다.

1.실험환경

본 논문에서 진행한 실험 환경은 표 1과 같다.

	Content
GPU	Nvidia RTX A5000 * 4
Framework	Pytorch, vLLM
models	EXAONE-3.5-2.4B-Instruct
	Gemma-3-1B-it
	Llama-3.2-1B

표 1. 실험 환경

	EXAONE-3.5-2.4B-Instruct			Gemma-3-1B-it			Llama-3.2-1B		
	vanilla	GPTQ	AWQ	vanilla	GPTQ	AWQ	vanilla	GPTQ	AWQ
MMLU	59.29	34.31	57.24	39.87	34.41	42.52	31.10	40.45	30.76
HellaSwag	54.04	42.88	53.01	42.45	40.88	42.54	48.34	41.86	36.91
Inference time	26.15 t/sec	8.72 t/sec	13.90 t/sec	16.34 t/sec	11.67 t/sec	13.03 t/sec	43.66 t/sec	34.29 t/sec	31.48 t/sec
Vram Usage	4.61GB	0.94GB	2.07GB	0.96GB	0.93GB	0.97GB	1.03GB	0.99GB	0.99GB

표 2. sLLm 정량적 성능 평가표

2.실험방법

표 2와 같이 성능을 평가하기 위해 57개 분야의 전문 지식을 평가하는 MMLU와 문맥 기반 상식 추론 능력을 평가하는 HellaSwag를 활용하여 기존 모델과 성능을 비교했다. 또한 양자화된 모델이 long-context를 처리하는 강건성을 검증하기 위해 LongBench 벤치마크를 활용하여 성능을 검증했다.

실험으로 활용한 sLLM 모델은 각기 다른 학습 방식을 통해 학습된 모델을 활용했다. EXAONE 3.5[5]는 사전학습 단계에서 학습 데이터 정제와 SFT(Supervised Fine-tuning)과 DPO(Direct Preference Optimization)을 통해 학습되었다. Gemma 3[6]는 QAT를 활용하여 비용을 감소시켰다. Llama 3.2[7]는 Pruning을 통해 성능은 높이고 비용은 감소시켰다. 이와 같이 현재 모델들은 각기 다른 학습 방식을 통해 성능을 높이거나 연산 비용을 낮추는 형식으로 학습되고 있다.

본 논문의 양자화는 autogptq, autoawq라이브러리를 사용했다. 양자화는 PTQ(Post-Training Quantization) 방식으로 GPTQ와 AWQ를 통해 이뤄졌다. GPTQ는 입력층부터 출력층까지 순차적으로 처리하는 Layer-wise Quantization을 사용하여 양자화를 했으며, wikitext2를 보정 데이터셋으로 사용했다. AWQ는 중요 가중치를 보존하는 Per-channel scaling방식을 사용하며, The Pile 데이터셋의 검증 데이터셋을 보정 데이터셋으로 이용했다.

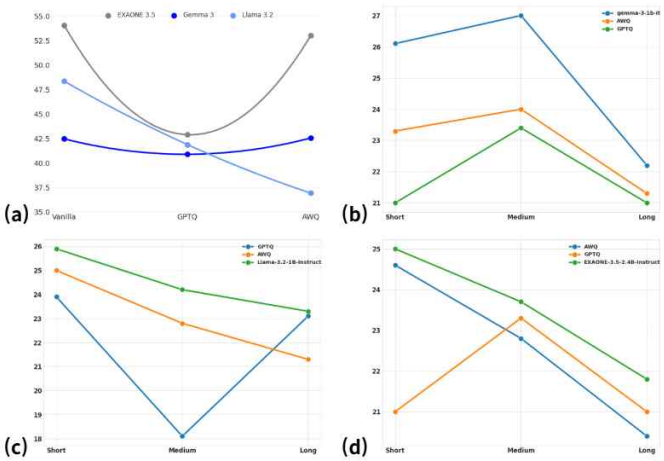

그림 1. (a) 모델별 HellaSwag 시각화, (b, c, d) 모델별 LongBench 문맥 길이에 따른 문맥 이해도 시각화

3.실험결과

그림 1의 (a)는 HellaSwag 벤치마크를 통해 모델의 양자화 방식에 따른 성능을 보여준다. 여기서 GPTQ는 양자화를 적용했을 때 모델의 종류와 관계없이 성능이 유사한 수준으로 수렴하는 경향이 있다. 이는 GPTQ의 성능이 모델 자체의 고유한 특성보다 양자화에 사용된 보정 데이터셋에 크게 의존할 수 있음을 시사한다. 즉, 특정 데이터 유형에서는 높은 성능을 보일 수 있으나, 범용적인 성능은 저하될 가능성이 있다. 반면, AWQ는 Llama모델을 제외한 원본(Vanilla) 모델과 거의 동등한 성능을 유지하는 것을 확인할 수 있다. 이는 AWQ가 추론 속도와 같은 효율성을 개선하면서도 모델의 성능을 잘 보존하는 경향이 있음을 의미한다. 한편, 그림 1(b, c, d)의 문맥 길이별 성능 그래프를 보면, 모든 양자화 적용 여부와 관계없이 긴 시퀀스 입력에 대해 유사한 패턴으로 성능이 감소하는 것을 확인할 수 있다. 이는 양자화 기술 자체가 모델의 장단기 기억 능력에 미치는 영향이 낮은 것을 알 수 있다.

표 2의 Vram 사용량을 보면 알 수 있듯 모델 크기에 따른 압축률 차이를 명확히 보여준다. 2.4B 파라미터의 EXAONE 모델은 높은 압축률을 달성했지만, 1B 규모의 다른 모델들은 그 효과가 미미했다. 이러한 결과는 현재 sLLM은 기본 파라미터가 작을수록 압축 잠재력 또한 낮아진다는 점을 암시한다.

Ⅳ. 결론

본 논문은 서로 다른 방식으로 학습된 sLLM에 최신 양자화 기법을 적용하여 그 성능을 비교 분석했다. 실험 결과로 모델의 학습 방식이 장단기 시퀀스 처리 능력에 미치는 영향은 미미한 것으로 나타났다. 결론적으로, sLLM에 양자화를 적용하면 추론 속도 향상이라는 이점은 분명히 존재하지만, 높은 수준의 모델 압축률을 기대하기는 어렵다는 한계를 확인했다.

이러한 한계점을 통해 특정 sLLM의 내부 가중치 분포와 활성화 패턴을 사전에 분석하여, 이에 최적화된 맞춤형 양자화 연구를 추후 진행할 계획이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2020-II201808)

참 고 문 헌

[1] Mengzhao Chen, Wenqi Shao, Peng Xu “EfficientQAT: Efficient Quantization-Aware Training for Large Language Models“

[2] William Fedus, Barret Zoph, Noam Shazeer “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity“

[3]Elias Frantar, Saleh Ashkboos, Torsten Hoefler “GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers “

[4]Ji Lin, Jiaming Tang, Haotian Tang “AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration”

[5]LG AI Research “EXAONE 3.5: Series of Large Language Models for Real-world Use Cases”

[6]Gemma Team “Gemma 3 Technical Report”

[7]Llama Team “The Llama 3 Herd of Models”