LLM Quantization 연구 동향과 탄소 배출 절감 효과에 대한 연구

하예영, 박수현
숙명여자대학교

haye4031@sookmyung.ac.kr, soohyun.park@sookmyung.ac.kr

Research Trends in LLM Quantization and Its Effects on Carbon Emission Reduction

Yeyoung Ha, Soohyun Park
Sookmyung Women's University

요 약

Large language model(LLM) 양자화(Quantization)은 LLM의 확산으로 증가하는 연산 자원 소모와 탄소 배출 문제를 완화하기 위한 핵심 기법이다. 본 논문에서는 양자화 기반 경량화 기술의 연구 동향과 효과를 분석하고, 이를 통해 LLM 양자화가 미래 자율통신 환경에서의 에너지 효율 향상과 탄소 배출 저감에 기여할 잠재력을 제시하며, 한계점과 향후 도전 과제에 대해 논의한다.

Ⅰ. 서 론

LLM 은 현재 전 세계 다양한 응용 서비스에 깊이 통합되어 있으며, 최근에는 자율주행, 군집 UAV, 이동형 로봇 네트워크 등 미래 자율통신 시스템에도 활용이 확산되고 있다. 이러한 이동체 환경에서는 전력 제약이 존재하며, 실시간 추론 요구로 인해 연산 효율의 중요성이 더욱 부각된다. 한 연구에서는 LLM 의 추론 단계에서 탄소 배출이 무시할 수 없는 수준임을 보였으며 [1], BLOOM 과 같은 대규모 모델의 학습 과정에서는, 하드웨어 제조, 인프라 운영 등을 모두 포함하면 최대 약 50.5 톤의 이산화탄소가 배출될 수 있다는 분석이 보고되었다 [2]. 특히 엣지 디바이스나 UAV 온보드 시스템과 같이 제한된 전력 환경에서의 LLM 운용은 양자화 기술 없이는 비효율적임이 지적되고 있다 [3].

이처럼 LLM 보급 확대와 에너지 부담의 상관관계가 명확해지는 가운데, 단순히 성능만을 추구하는 기존의 접근은 지속 가능하지 않다. 이러한 배경에서 LLM 경량화 기술은 성능 저하를 최소화하면서 연산 효율을 극대화하기 위한 핵심 연구 분야로 주목받고 있다. 이에 본 논문에서는 LLM 경량화 기법 중 하나인 양자화의 개발 동향을 분석하고, 이동체 자율제어용 LLM 모델의 효율적 운용 측면에서 고찰하며, 이를 통해 얻은 인사이트와 향후 연구 방향을 제시하고자 한다.

Ⅱ. 본론

LLM 경량화 기법에는 양자화, 가지치기, 지식증류, LoRA, MoE 등이 존재한다. 특히 양자화는 메모리 사용량과 계산량을 크게 줄이면서도, 비교적 적은 성능 저하를 보여주며, 구현이 간단하고 기존 하드웨어와의 호환성이 뛰어나 다른 경량화 기법들보다 많이 사용되고 있다 [3]. 양자화란 신경망의 가중치나 활성값을 낮은 비트 정밀도(예: 8bits, 4bit 등)로 표현하여 모델의 메모리 사용량과 연산량을 줄이는 기술이다. 그림 1 은 본 논문에서 언급되는 양자화 기법들을 분류한 내용이다.

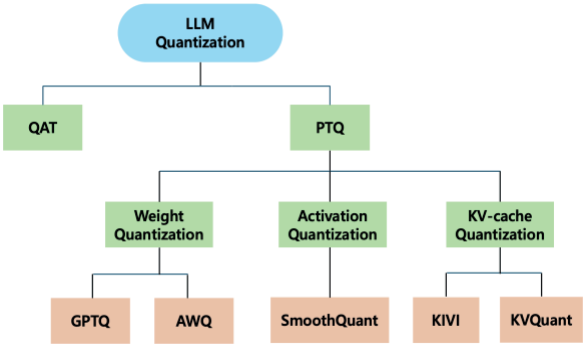

그림 1 양자화 기법

양자화는 모델의 효율성을 높이기 이해 정밀도를 낮추는 과정으로, 크게 두 가지 접근 방식이 있다. 첫째, Quantization-Aware Training(QAT)는 훈련 과정에서 양자화 효과를 모사하여 모델이 양자화로 인한 노이즈에 적응하도록 학습하는 방법이다. 둘째, Post-Training Quantization(PTQ)는 모델이 훈련이 완료된 후 별도의 추가 학습 없이 양자화를 적용하는 방식으로, 구현이 간단하고 빠르다는 장점이 있다. QAT 는 낮은 비트 표현에서도 메모리 소비를 줄이고 정확도 손실을 최소화하는 해결책을 제공하지만, 막대한 훈련 자원 소모로 인해 실제로는 적용이 어렵다 [4]. 따라서 최근의 연구들은 PTQ 기반의 양자화에 기반하여 이루어지고 있다. PTQ 기법 안에서도 학습된 가중치를 저비트로 표현하는 가중치 양자화, 활성값의 범위를 스케일링하거나 이상치를 이동시켜 저비트 표현에서도 정확도 손실을 최소화하는 활성값 양자화, 추론 중 저장되는 Key/Value 캐시를 양자화하는 KV 캐시 양자화가 존재한다.

Generative Pre-trained Transformer Quantization(GPTQ)는 가중치 양자화를 통해 PTQ 만으로도 기존 대비 3-8 배 메모리 절감 및 4 배 이상의 추론 속도 상승을 보이면서도, 원본 대비 99%

이상의 정확도를 달성하였다 [5]. Activation-aware Weight Quantization(AWQ)는 기존 GPTQ 의 calibration 데이터셋에 과적합되어 도메인 일반화 성능이 저하될 수 있는 문제를 해결하기 위해 활성화 분포를 기준으로 중요한 가중치를 보호하고 Per-channel scaling 을 적용해 양자화 오차를 줄였다. 이에 GPTQ 대비 일반화 성능이 우수하고, 다양한 모델/도메인/멀티모달 모델에도 적용이 가능해졌다 [6]. SmoothQuant 는 가중치는 잘 양자화 되지만 활성값은 outlier 때문에 양자화 시 성능이 급락하는 문제를 해결하기 위해서 이상치를 가중치로 옮겨 활성화 분포를 평탄하게 만들어 INT8 가중치+활성값 양자화가 가능하도록 하였다 [7]. KIVI 는 LLM 에서 KV 캐시가 배치 크기 증가와 긴 컨텍스트 처리 시 메모리 소모가 폭발적으로 늘어나고, 이로 인해 속도와 메모리 병목 현상이 발생하는 문제를 해결하기 위해 key 캐시는 채널 단위로, value 캐시는 토큰 단위로 양자화하여 Llama-2, Falcon, Mistral 모델에서 품질을 유지하며 최대 2.6 배 메모리 사용 절감, 최대 4 배 배치크기 지원, 2.35~3.47 배 추론 속도 개선을 달성하였다 [8]. KVQuant 는 긴 문맥에서도 메모리 사용을 줄이면서 정확도 저하를 최소화하는 KV-cache 양자화 기법을 위해 Per-Channel Key 양자화, Pre-RoPE Key 양자화, NUQ, Per-Vector-Dense-and-Sparse-Quantization 기법을 적용하였다. 그 결과 3bit 정도의 낮은 정밀도로 KV 캐시를 양자화 해도 원본 대비 0.1 미만의 perplextity 손실만 발생시키며, 긴 문맥(최대 1 백만~1 천만 토큰)에서 빠른 추론을 가능하였다 [9].

Ⅲ. 결론

본 논문에서는 LLM 의 확산에 따른 에너지 소비와 탄소 배출 문제를 완화하기 위한 주요 접근으로 LLM 양자화 기술의 연구 동향을 살펴보았다. 기존 연구들은 주로 가중치 양자화를 중심으로 이루어졌으며, GPTQ, AWQ 등의 기법을 통해 높은 정확도와 효율을 달성하였다. 최근에는 UAV swarm 등 이동체 간 협력 통신 환경에서 LLM 이 경로 최적화, 협력 제어 의사결정에 활용되기 시작했다. 이러한 시스템은 통신 지연과 전력 소모에 민감하므로, PTQ 기반의 양자화 모델이 온보드 GPU 나 AI 가속기 환경에 적합한 솔루션으로 주목받고 있다 [11]. 특히, UAV 의 실시간 상황 인식·제어를 위한 경량 LLM 은 KV 캐시 양자화를 통해 메모리와 전력 소모를 2~3 배 절감하면서도, 명령 해석 및 통신 응답 지연을 최소화할 수 있다. 한편, 활성값과 KV 캐시는 모두 추론 시 동적으로 생성되는 활성값이라는 공통점을 가지므로, 두 영역을 동시에 최적화 (활성값+KV 캐시 결합 양자화)하는 통합적 접근이 새로운 연구 방향으로 제시될 수 있다.

나아가, 이러한 양자화 연구는 클라우드 중심에서 엣지, 이동체 중심의 분산형 자율통신 구조로 확장될 것으로 예상된다. 특히 UAV 나 자율주행 차량 내 탑재형 LLM 의 효율적 운용은 탄소 저감형 자율통신 시스템 구현의 핵심요소로 자리잡을 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터사업의 연구결과로 수행되었음(IITP-2024-RS-2024-00436887).

참 고 문 헌

[1] Z. Fu, F. Chen, S. Zhou, H. Li, and L. Jiang, "LLMCO$_2$: Advancing accurate carbon footprint prediction for LLM inferences," *arXiv preprint arXiv:2405.18972*, 2024.

[2] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," *arXiv preprint arXiv:2211.02001*, 2022.

[3] S. Zhang et al., "EdgeLLM: Efficient deployment of large language models on autonomous edge systems," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11523- 11535, 2025.

[4] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 1556- 1577, 2024.

[5] M. Chen, Y. Li, S. Wang, and H. Zhang, "EfficientQAT: Efficient quantization-aware training for large language models," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2025, pp. 6123- 6135.

[6] E. Frantar, T. Lin, B. Ginsburg, and J. Leutgeb, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, 2023.

[7] J. Lin *et al.*, "AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration," in *Proc. 7th Conf. Mach. Learn. Syst. (MLSys)*, 2024, pp. 3883- 3900.

[8] G. Xiao *et al.*, "SmoothQuant: Accurate and efficient post-training quantization for large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, PMLR, 2023, pp. 34707- 34733.

[9] Z. Liu *et al.*, "KIVI: A tuning-free asymmetric 2-bit quantization for KV cache," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, vol. 235, PMLR, 2024, pp. 32332- 32344.

[10] C. Hooper *et al.*, "KVQuant: Towards 10 million context length LLM inference with KV cache quantization," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, vol. 240, PMLR, 2024, pp. 32332- 32344.

[11] H. Lee, J. Kim, and S. Park, "Lightweight transformer adaptation for UAV swarm intelligence," *IEEE Commun. Lett.*, vol. 29, no. 4, pp. 2014- 2018, 2024.