엣지 디바이스를 위한 LLM 추론 가속화: 라즈베리 파이 5에서의 양자화 기법 비교 연구

강서진[1]
[1]성결대학교 컴퓨터공학과 학부생
kseojin0205@sungkyul.ac.kr

Accelerating LLM Inference for Edge Devices: A Comparative Study of Quantization Techniques on Raspberry Pi 5

Seo-Jin Kang[1]
[1]Dept. of Computer Engineering, Sungkyul University

요 약

본 연구는 라즈베리 파이 5에서 대규모 언어 모델(LLM)의 추론 성능을 개선하기 위한 양자화 기법을 비교 분석하였다. 균일 양자화(8비트, 4비트)와 레이어 중요도 기반 적응형 양자화의 성능을 평가한 결과, 균일한 8비트 양자화가 추론 속도를 2.29배 향상시키면서도 출력 품질을 우수하게 유지하는 최적의 방법임을 확인하였다. 이는 ARM 아키텍처의 하드웨어 특성과 메모리 접근 패턴의 최적화 측면에서 설명할 수 있다. 본 연구는 제한된 자원을 가진 엣지 디바이스에서 LLM을 효율적으로 구동하기 위한 실용적인 가이드라인을 제시한다.

1. 서론

대규모 언어 모델(LLM)은 자연어 처리 분야에서 혁신적인 성능을 보이고 있으나, 그 크기와 복잡성으로 인해 저전력 엣지 디바이스에서의 실행이 어렵다는 한계가 있다[1]. 특히 라즈베리 파이와 같은 저비용, 저전력 컴퓨팅 플랫폼에서 LLM을 효율적으로 실행하는 것은 에너지 효율성, 개인 정보 보호, 오프라인 접근성 측면에서 중요한 연구 과제로 여겨지고 있다. 본 연구에서는 라즈베리 파이 5(8GB RAM)에서 LLM 추론 성능을 개선하기 위한 다양한 양자화 기법을 실험적으로 비교한다. 양자화는 모델의 가중치를 낮은 비트 정밀도로 표현하여 메모리 사용량을 줄이고 연산 효율성을 높이는 기법이다[2]. 특히 모델의 모든 레이어에 동일한 양자화를 적용하는 균일 양자화와 레이어의 중요도에 따라 다른 비트 정밀도를 적용하는 적응형 양자화의 효과를 비교 분석하였다.

2. 실험 방법

본 연구는 2.4GHz 쿼드코어 Arm Cortex-A76 프로세서와 8GB RAM을 탑재한 라즈베리 파이 5에서 진행되었으며, 기본 모델로는 DistilGPT-2[3]를 사용하였다. 다음 네 가지 모델 구성을 비교하였다: (1) 원본 32비트 부동소수점(FP32) 모델, (2) 모든 가중치를 8비트로 양자화한 모델[4], (3) 모든 가중치를 4비트로 양자화한 모델, (4) 레이어 중요도에 따라 상위 30%는 FP32, 중간 30%는 8비트, 하위 40%는 4비트로 양자화한 적응형 양자화 모델[5]. 각 모델은 추론 속도와 출력 품질 측면에서 평가되었으며, 동일한 프롬프트 "The future of artificial intelligence is"에 대한 응답을 분석하였다.

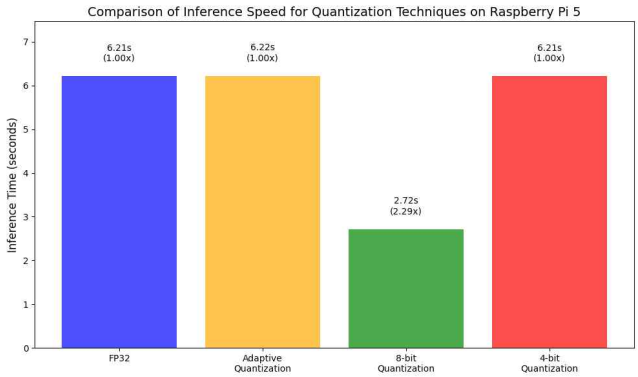
3. 실험 결과

실험 결과, 8비트 균일 양자화 모델이 가장 우수한 성능을 보였다. FP32 모델의 추론 시간은 평균 6.2150초였으며, 적응형 양자화 모델은 6.2204초로 거의 동일한 성능을 보였다. 8비트 모델은 2.7178초로 원본 대비 2.29배 빠른 추론 속도를 달성하였다. 반면, 4비트 모델은 6.2120초로 예상과 달리 속도 향상을 보이지 않았다. 이러한 결과는 표 1에 정리되어 있으며, 그림 1은 각 모델의 추론 시간과 속도 향상을 시각적으로 비교하여 보여준다. 출력 품질 측면에서도 8비트 양자화 모델이 원본과 가장 유사한 결과를 생성하였다. FP32 모델은 "The

future of artificial intelligence is not yet clear." 라는 출력을 생성하였으며, 8비트 양자화 모델은 "The future of artificial intelligence is not yet clear. But it is possible that artificial intelligence could be developed in the future."로 원본과 매우 유사하면서도 추가 정보를 제공하는 출력을 생성하였다. 반면, 적응형 양자화 모델은 "The future of artificial intelligence is:::::::::::::::::::::" 와 같이 의미 없는 반복 패턴이 포함된 출력을 생성하였으며, 4비트 모델은 "The future of artificial intelligence is in in in in in in in in..."와 같이 단어 반복이 심한 저품질 출력을 생성하였다. 이러한 출력 품질의 차이는 양자화 과정에서 발생하는 정보 손실과 관련이 있는 것으로 보인다.

<표 1> 양자화 기법별 성능 비교

모델 구성	추론 시간 (초)	속도 향상 (배)	출력 품질
FP32	6.2150	1.00	우수
적응형 양자화	6.2204	1.00	불량
8비트 양자화	2.7178	2.29	우수
4비트 양자화	6.2120	1.00	불량



(그림 1) 양자화 기법별 추론 시간 및 속도 향상 비교

4. 논의 및 결론

라즈베리 파이 5에서는 균일한 8비트 양자화가 가장 효과적인 LLM 추론 가속 방법임을 확인하였다. 이는 다음과 같은 이유로 설명할 수 있다. 첫째, ARM Cortex-A76의 SIMD 명령어 세트가 8비트 연산에 최적화되어 있어 하드웨어 가속의 이점을 최대로 활용할 수 있다. 둘째, 균일한 데이터 형식은 메모리 접근 패턴을 단순화하여 캐시 효율성을 높인다. 셋째, 적응형 양자화에서 서로 다른 비트 정밀도를 처리하는 오버헤드가 성능 이점을 상쇄한다[5].

4비트 양자화의 경우, ARM 아키텍처의 4비트 연산 처리 제한으로 인해 속도 향상을 보이지 않았으며, 출력 품질도 크게 저하되었다. 적응형 양자화의 성능이 기대에 미치지 못한 것은 이론적 중요도와 실제 추론 과정에서의 중요도 간 차이가 있을 수 있음을 시사한다.

향후 연구에서는 ARM 아키텍처에 최적화된 커스텀 GEMM 연산자 구현, 더 큰 모델에 대한 양자화 효과 검증, 양자화 과정에서의 정보 손실을 최소화하기 위한 교정 기법 연구, 동적 양자화 및 양자화 인식 훈련 기법의 적용 가능성을 탐색할 계획이다. 이러한 연구는 엣지 컴퓨팅 환경에서 LLM의 활용 가능성을 크게 향상시키고, 개인 정보 보호 및 오프라인 접근성이 중요한 응용 분야에서의 활용을 촉진할 것으로 기대된다.

참고문헌

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al., "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vol. 33, pp. 1877-1901, 2020.

[2] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D., "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 2704-2713.

[3] Sanh, V., Debut, L., Chaumond, J., Wolf, T., "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter", arXiv preprint arXiv:1910.01108, 2019.

[4] Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L., "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale", Advances in Neural Information Processing Systems, Vol. 35, pp. 27423-27436, 2022.

[5] Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D., "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers", Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 2023, pp. 1-17.