엣지 컴퓨터 환경(Raspberry Pi5)에서 BitNet-b1.58 모델과 Gemma3 모델의 구현 및 성능비교에 관한 연구

이승환[1], 박성준[1], 김정훈[1]
[1]동서울대학교 컴퓨터 소프트웨어과 학부생
seung_hwan10@naver.com, syaa123@naver.com, thdnfrhfem@gmail.com

A Study on the Implementation and Performance Comparison of the BitNet-b1.58 and Gemma3 Models in an Edge Computing Environment Using Raspberry Pi 5

Seung-Hwan Lee[1], Seong-Jun Park[1], Jung-Hun Kim[1],
[1] Dept. of Computer Software, Dong-Seoul University

요 약

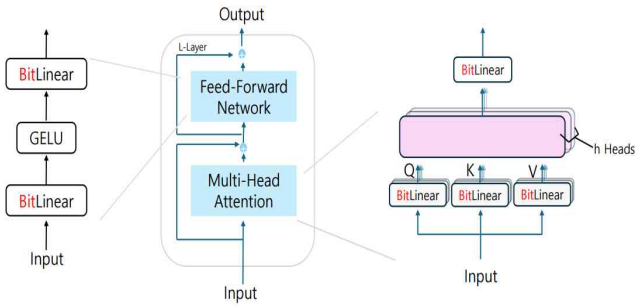
본 연구는 마이크로소프트(Microsoft)에서 공개한 경량화 LLM(Large Lanuage Model) BitNet b1.58 모델과 구글(Google)에서 공개한 Gemma3 모델을 엣지 컴퓨팅 환경인 Raspberry Pi 5에서 설치, 운영하여 각 성능을 비교 평가한 것이다. 본 연구에서는 초당 토큰(Token) 생성 속도, CPU 효율, 메모리(Memory)효율 그리고 생성된 문장에 대한 품질성능을 정량적 측정지표 ROUGE-1, BLEU, METEOR를 사용하여 평가하였다. 실험 결과, BitNet b1.58은 Gemma3 대비 약 4.4배 빠른 토큰 생성 속도, 약 4.5배 높은 CPU효율, 약 15배 높은 메모리 효율을 보였으며, 생성 문장의 품질 성능평가는 Gemma3가 모든 분야에서 좋은 성능을 보였다.

1. 서론

최근 온 디바이스(On-Device) AI의 중요성이 높아짐에 따라 한정된 자원을 사용하면서 높은 성능을 유지할 수 있는 경량화 LLM모델에 대한 연구가 활발히 진행되고 있다. 특히 대규모 언어 모델(Large Language Model)의 큰 문제점은 성능이 좋아질수록 연산량과 메모리 사용량 등 자원 요구가 기하급수적으로 증가한다는 점이다. 따라서 최근 LLM의 연구 동향은 성능을 최대한 유지하면서 사용하는 자원량을 최소화 하는 것을 목표로 많은 연구가 이루어지는 추세이다[1]. BitNet-b1.58 역시 이러한 흐름 속에서 MS(Microsoft)사가 제안한 경량화 LLM 모델 중 하나이다. 본 연구에서는 BitNet-b1.58 모델에 대해 분석하고, Raspberry Pi 5와 같은 엣지 컴퓨터 환경에서 구글에서 발표한 경량화 LLM 모델인 Gemma3와 성능비교평가를 수행하였다.

2. BitNet b1.58

BitNet-b1.58[2] 모델은 BitNet 프레임워크(Framework)를 활용하여, QAT(Quantization-Aware Training) 방식으로 처음부터 1.58-Bit 가중치로 학습된 네이티브 경량화 모델이다. BitNet 프레임워크는 기존 LLM 구조를 유지하며 피드포워드 네트워크(Feed-Forward Network)와 멀티 헤드 어텐션(Multi-Head Attention) 부분에 리니어 레이어(Linear Layer) 부분을 비트리니어 레이어(BitLinear Layer)로 단순히 교체한 형태이다. 그림 1은 BitNet 프레임워크를 나타낸 구조도 이다.

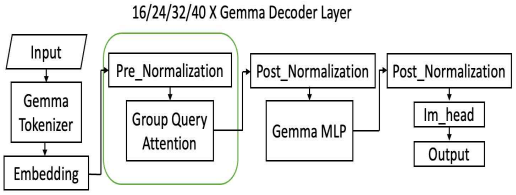


[그림 1] BitNet 프레임워크 구조도

3. Gemma3

Gemma3는 디코더 기반(Decoder-only) 트랜스포머 구조로 설계되었으며, 단일 GPU 또는 TPU 환경에서도 학습 및 추론이 가능하다. GQA(Group Query Attention) 방식을 이용하여 Query만 그룹화하여 연산량을 줄이는 방식으로 메모리 사용을 효율화하였다. FAv2(Flash Attention v2) 고속 연산과 메모리 저소비를 동시에 달성하는 어텐션(Attention) 알고리즘을 도입하였다. MLP 구조 개선은 정적 학습과 성능 향상을 위해 시그모이드

(Sigmoid) 활성화 함수를 채택하였다. 트리플 프로젝션 구조를 통해 표현 능력을 강화하면서도 연산 효율성을 유지하였다. 그림 2는 Gemma3의 구조도 이다.



[그림 2] Gemma3 구조도

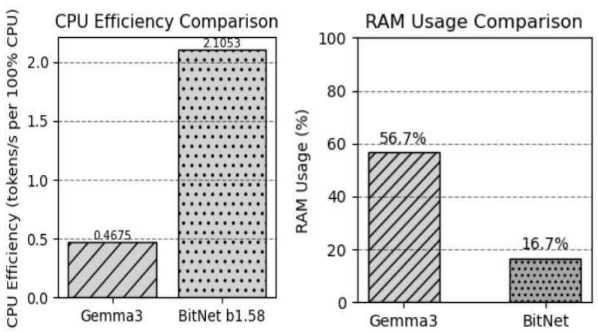
4. 실험 환경 및 구현결과

표 1는 본 연구에서 Bitnet-b1.58-2B 모델과 Gemma3-4B 모델을 설치하여 운영 비교평가 한 Raspberry Pi 5의 주요 규격이다.

<표 1> 실험구현 환경의 주요규격

구분	항목	규격
Raspberry Pi 5	CPU	Broadcom BCM2712
	RAM	8GB LPDDR4X-4267

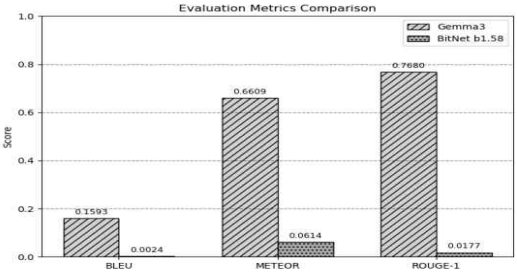
Bitnet-b1.58-2B 모델과 Gemma3-4B 모델을 엣지 컴퓨팅 환경의 대표적인 경우인 Raspberry Pi 5 환경에 설치 후 성능을 비교분석 하였다. 최대 554토큰(Token)길이의 문장을 생성 하였으며 생성하는 과정에서 초당 토큰생성 속도, CPU 효율, 메모리 효율을 비교하였다. 모델의 정량적인 품질성능 측정을 위해 SQuAD v1.1 데이터 중 Super Bowl 50 관련 문단을 기반으로 질문 응답 문제 100개를 통해 성능 평가를 수행하였다. 성능평가의 측정방식은 ROUGE-1, BLEU, METEOR 3가지 방식을 사용하였다. 초당 토큰 생성 개수는 Bitnet-b1.58-2B 모델의 경우 약 8.0[tok/s] 이고 Gemma3-4B 모델은 약 1.80 [tok/s]의 결과가 나왔다. CPU 점유률은 각각 약 385[%] , 약 380[%] 로 측정되었다. 이를 CPU 코어 하나 기준으로 정규화을 수행하여 CPU 효율을 계산하였다. 메모리 효율은 각각 약 16.7[%], 56.7[%]로 측정되었다. 그림 3은 이러한 측정 결과를 나타낸 것이다.



(a) CPU 효율 그래프 (b) 메모리 효율 그래프

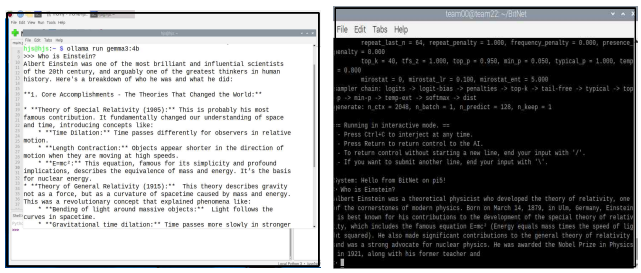
[그림 3] 두 모델의 CPU, 메모리 효율비교 결과

모델별 생성문장에 대한 품질성능 측정 결과는 그림 4 와 같다.



[그림 4] 정량적인 성능 평가 결과

정규화 된 초당 토큰 생성 개수는 Bitnet b1.58-2B는 약 2.1053 [tok/s] 당 [100%CPU] 이고 Gemma3-4B는 약 0.4675 [tok/s] per [100%CPU] 로 BitNet-b1.58-2B이 Gemma3-4B 대비 CPU효율이 약 4.5배 높은 결과를 얻었고, 메모리효율은 Bitnet-b1.58-2B이 Gemma3-4B 대비 15배 높은 결과를 얻었다. ROUGE-1, BLEU, METEOR을 이용한 정량적인 품질 성능평가는 Bitnet-b1.58-2B이 0.0024 ,0.0614 ,0.0177 Gemma3-4B이 0.1593, 0.6609, 0.7680로 Gemma3-4B모델이 모두 더 좋은 결과를 얻었다. 그림 5는 Raspberry Pi 5에서 각 모델을 구현한 동작화면이다.



(a) Gemma3-4B 모델 동작화면 (b) BitNet-b.1.58-2B 모델동작화면

[그림 5] 각 모델의 동작화면

5. 결론

본 연구에서는 MS사의 Bitnet-b1.58-2B모델과 구글의 Gemma3-4B 모델을 엣지 컴퓨팅 환경인 Raspberry Pi 5 에서 구현하고, CPU 효율성, 메모리 효율성 그리고 품질성능 평가를 진행하였다. 본 연구의 실험 결과를 통하여 엣지 컴퓨팅 환경에서 BitNet-b1.58-2B 모델이 에너지 효율성이 더 좋다는 결과를 확인할 수 있었고, 모델의 파라미터를 더 증가시킨다면 생성문장의 품질성능도 더 우수할 것으로 판단된다. 본 연구의 결과는 엣지 컴퓨팅 환경에서 경량 LLM 모델을 설치 및 운영하고자 할 때 활용 될 수 있을 것이다.

참고문헌

[1]Z. Wang, Y. Liang, Z. Xu, T. Sun, Y. Zhang, S. Zhang, Y. Wang, and Y. Chen, “Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models,” arXiv preprint arXiv:2401.00625, 2024.
[2]S. Ma, H. Wang, S. Huang, X. Zhang, Y. Hu, T. Song, Y. Xia, and F. Wei, “BitNet b1.58 2B4T Technical Report,” arXiv preprint arXiv:2504.12285, 2025.