효율적인 질의 처리를 위한 BitNet-MCP 하이브리드 시스템*

김이든[1], 안효빈[1], 이동재[2]
[1]강원대학교 컴퓨터공학과 학부생
[2]강원대학교 융합보안학과 교수
rladlems1031@kangwon.ac.kr, gyqlsdks@kangwon.ac.kr, dongjae.lee@kangwon.ac.kr

A BitNet-MCP Hybrid System for Efficient Query Processing

Yi-Deun Kim[1], Hyo-Bin Ahn[1], Dong-Jae Lee[2]
[1]Dept. of Computer Engineering, Kangwon National University
[2]Dept. of Convergence Security, Kangwon National University

요 약

최근 대규모 언어 모델(LLM)이 상용화됨에 따라, 자원이 제한된 컴퓨팅 환경에서도 효율적으로 동작할 수 있는 경량 LLM에 대한 수요가 증가하고 있다. 본 논문은 경량화된 1.58비트 모델인 BitNet과 Model Context Protocol(MCP)을 결합한 하이브리드 시스템을 제안한다. 시스템은 질의 복잡도에 따라 동적으로 처리 경로를 결정하는 방식으로 동작한다. 단순한 질의는 메모리 효율적인 BitNet이 직접 처리하여 에너지 소비와 지연 시간을 최소화하고, 실시간 데이터 검색이나 외부 도구 사용이 필요한 복잡한 질의는 BitNet이 MCP 호스트로서 적절한 MCP 서버와 통신한다. 모든 결과는 BitNet 내부에서 통합되어 처리된 후 일관된 응답을 생성한다. 이러한 접근 방식은 엣지 컴퓨팅, IoT 기기, 모바일 애플리케이션과 같이 제한된 컴퓨팅 환경에서 AI 활용성을 높이는데 기여할 수 있을 것으로 기대된다.

1. 서론

최근 LLM의 발전은 AI 시스템의 능력을 크게 향상시켰지만, 이러한 모델들은 여전히 상당한 하드웨어 자원을 요구한다는 한계가 있다. 제한된 컴퓨팅 환경에서 BitNet과 같은 1.58비트 양자화 모델이 효율적인 대안으로 개발되었으나, 실시간 데이터 접근이나 외부 도구 활용에 제한이 있다.

한편, Model Context Protocol(MCP)은 AI 모델과 외부 시스템 간의 상호작용을 표준화하지만, 더 많은 계산 자원이 필요하다. 본 논문은 BitNet의 계산 효율성과 MCP의 도구 상호작용 능력을 결합한 하이브리드 시스템을 제안하여, 제한된 컴퓨팅 환경에서도 확장 가능한 기능을 제공한다.

2. 배경지식

2.1 BitNet

BitNet은 Microsoft Research에서 개발한 메모리 사용량을 획기적으로 줄일 수 있도록 설계된 경량 LLM이다[1]. BitNet-b1.58-2B-4T는 20억 개 파라미터 규모를 가진 최초의 오픈소스 1비트 LLM으로, 모델 내부의 가중치를 {-1, 0, +1}의 삼진값으로 제한하여, 실행 시 메모리 사용량을 최대 0.4GB로 대폭 줄인다.

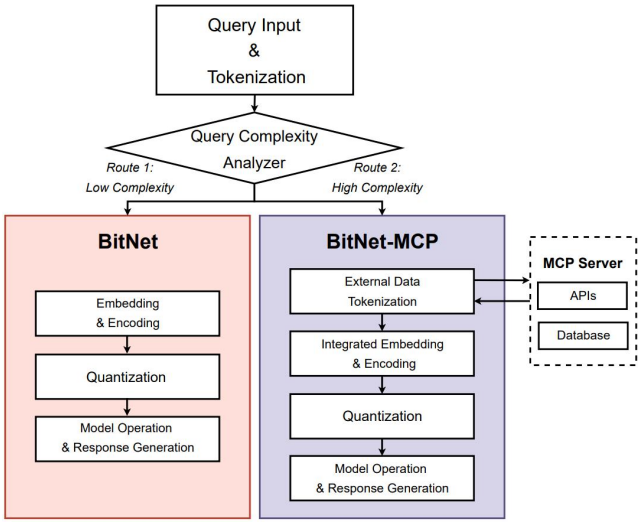
BitNet은 기존 트랜스포머 모델을 기반으로 선형 계층을 BitLinear라는 특수한 계층으로 대체하여 가중치(1.58비트)와 활성화(8비트 정수)를 양자화한다. BitNet은 유사한 규모와 성능의 최신 경량 LLM들보다 메모리 사용량이 크게 낮으며, 특히 MiniCPM 2B 모델과 비교 시 약 91.7% 감소하여 제한된 컴퓨팅 환경에서 효율적인 대안으로 주목받고 있다.

2.2 Model Context Protocol

MCP는 AI 모델과 외부 시스템 간 상호작용을 위한 표준화된 프로토콜이다[2]. MCP는 호스트(AI 모델), 클라이언트(중개자), 서버(외부 시스템 접근 지원 환경)로 구성된다. 클라이언트는 호스트와 서버 간의 통신을 관리하며, 서버는 도구(외부 서비스 및 API 호출)와 자원(구조화된 데이터 접근)을 통해 AI 모델의 외부 시스템 접근을 지원한다.

* 본 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2022-II221196, 50%)과 교육부와 한국연구재단의 재원으로 첨단분야 혁신융합대학사업의 지원(50%)을 받아 수행된 연구임

3. BitNet-MCP 하이브리드 시스템



(그림 1) BitNet-MCP WORK FLOW

본 논문에서 제안하는 BitNet-MCP 하이브리드 시스템은 다음의 단계에 따라 질의를 처리하며 전체적인 워크 플로우는 (그림 1)과 같다. 이 시스템은 외부 네트워크가 연결되어 있다는 것을 전제한다.

- **Step 1: 질의 입력 및 토큰화**

사용자가 자연어 형태로 BitNet에 질의를 입력하면 LLaMA3 토크나이저로 텍스트를 토큰화한다.

- **Step 2: 복잡도 분석 및 처리 경로 선택**

복잡도 분석에 따라 경로 중 하나가 선택된다.

- **Route 1 (BitNet 직접 처리):** 낮은 복잡도의 질의는 BitNet에서 직접 처리한다.
- **Route 2 (BitNet-MCP):** 높은 복잡도의 질의는 MCP 호스트인 BitNet이 적합한 MCP 서버와 통신한다. MCP 서버로부터 받은 직렬화된 JSON 형식의 외부 데이터는 BitNet 내부에서 역직렬화되어 토큰화된 후, 원본 질의와 통합된다.

- **Step 3: 임베딩 및 인코딩**

임베딩 및 인코딩 과정은 각 토큰을 숫자 벡터로 변환하고, 문장 내 단어 위치 정보와 저장하여 의미와 순서를 컴퓨터가 이해할 수 있게 한다[1].

- **Step 4: 양자화**

양자화 단계에서는 BitNet 아키텍처의 핵심 요소로서 다음 두 가지 양자화 방식이 적용된다. 이때 임베딩 및 인코딩된 토큰은 활성화 값이 된다[1].

- **활성화 값 양자화**

“Absmax Quantization” 방식을 활용해, 각 토큰의 특성을 보존하며 부동소수점 활성화 값을 8비트 정수로 양자화한다. 이는 각 토큰의 최대 절대값을 기준으로 모든 활성화 값을 스케일링하여 -128~127 범위의 8비트 정수로 변환함으로써, 토큰별 동적 범위를 효율적으로 보존한다[1].

- **가중치 양자화**

가중치 양자화란 모든 가중치를 삼진값으로 제한하는 양자화 방식이다. 이는 가중치당 약 1.58비트만을 필요로 하여 메모리 사용량을 크게 줄인다. 이 양자화는 학습 전 과정에 적용되어 모델이 제한된 표현 내에서 최적의 성능을 달성하도록 한다[1].

- **Step 5: 모델 연산 및 응답 전달**

최종 단계에서 BitNet은 양자화된 활성화 값에 대해 BitNet에 특화된 행렬 곱셈을 수행한다. TL 및 I2_S 커널을 활용해 계산 효율성을 높이고, 문맥 관계 분석과 패턴 인식 처리를 통해 다음 토큰을 예측하여 응답을 생성한 후 사용자에게 반환한다[1].

4. 기대 효과 및 기술적 제약

- **제한된 컴퓨팅 환경에서의 AI 접근성 향상**

BitNet-MCP 시스템은 0.4~0.9GB 메모리로도 외부 도구와 최신 데이터에 접근할 수 있는 확장성을 제공한다. 이는 경량 모델의 한계를 보완하고 다양한 도메인의 전문 지식을 실시간으로 활용하도록 한다.

- **데이터 손실 가능성**

MCP 서버에서 BitNet으로 전송된 JSON 데이터를 양자화하는 과정에서 고정밀 부동소수점 값이 삼진값으로 압축된다. 이로 인해 숫자 정보의 정확도가 감소하고 세밀한 값의 손실이 발생할 수 있다.

5. 결론

본 논문은 제한된 컴퓨팅 환경에서 사용자 질의 복잡도에 따른 효율적인 질의 처리를 위한 BitNet-MCP 하이브리드 시스템 방법론을 제시하였다. 향후 연구에서는 BitNet-MCP의 구체적인 구현 방법과 성능 평가 지표를 제시할 예정이다.

참고문헌

[1] [1] S. Ma, H. Wang, S. Huang, X. Zhang, Y. Hu, T. Song, Y. Xia, and F. Wei, "BitNet b1.58 2B4T Technical Report," arXiv:2504.12285v2 [cs.CL], pp. 1-14, 2025.

[2] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions," arXiv:2503.23278 [cs.CR], pp. 1-19, 2025.