대규모 언어 모델 기반 서비스의 네트워크 효율적 컨텍스트 관리 기법 연구

윤남경, 김황남*
고려대학교
nkyoon93@korea.ac.kr, hnkim@korea.ac.kr

Network-Efficient Context Management for LLM-Based Conversational Services without Model Modification

Namkyung Yoon, Hwangnam Kim*
Korea Univ.

요 약

최근 대형 언어모델(LLM) 기반의 서비스는 클라우드 인프라을 통해 인공지능과 사용자가 대화를 나누는 새로운 유형의 통신 집약적 어플리케이션으로 부상되고 있다. 기존의 request-response 기반의 챗봇 서비스와 달리, LLM 어플리케이션 내의 서비스는 누적된 상호 대화 컨텍스트 기반의 상호작용에 기반하여, 세션이 길어질수록 LLM 어플리케이션이 부담하는 트래픽이 급격히 증가한다. 이러한 현상은 클라우드 인프라를 통한 중앙집약적인 LLM 서비스의 컨텍스트 길이 제약 하에서 종단 간 지연 시간 증가, 과도한 계산 오버헤드, 서비스 품질 저하를 초래할 수 있다. 따라서 본 논문에서는 서비스 트래픽 최적화를 위한 언어모델의 컨텍스트 관리 기법을 제안한다. 제안하는 기법은 사용자의 쿼리와의 의미적 관련성을 바탕으로 유효성이 낮은 대화 컨텍스트를 억제하는 프레임워크이며, 최근의 대화 연속성을 유지하면서도 높은 관련성을 가진 컨텍스트 세크먼트를 유지하도록 한다. 이를통해 언어모델의 아키텍처를 수정하거나 추가적인 훈련을 하지 않고도 서비스에 불필요한 네트워크 트래픽을 줄이는 것을 목표로 한다. 우리는 제안하는 기법을 검증하기 위해 채팅 기반 대규모 언어 모델 오픈소스 OpenAssistant Conversations Dataset (OASST1)에 대해 3가지의 언어모델을 사용하여 동일한 페이로드 하에서 제안된 컨텍스트 인식 방법이 전체 컨텍스트 전송에 비해 응답 품질을 최대 96% 향상시키면서 종단 간 지연 시간의 부담을 줄여주는 것을 확인했다.

Ⅰ. 서 론

대형 언어모델 (Large Language Model, LLM) 서비스는 답변의 신뢰도를 높이며 더 정확한 품질을 제공하기 위해 연구 및 개발이 진행되고 있다 [1]. 하지만 LLM 서비스는 사용자와 상호 대화 세션이 진행됨에 따라 이전 대화 맥락 전체를 참조하기 위해 어플리케이션 계층에서 페이로드가 지속적으로 증가하여 트래픽이 증가하고 처리 시간과 지연 시간이 길어지며 네트워크 품질이 저하되는 한계가 나타난다 [2].

기존 LLM 서비스는 이러한 문제에 직면하였을 때, 일반적으로 최대 컨텍스트 길이에 수용 가능한 만큼의 과거 컨텍스트를 전송한 후, 잘라내기를 적용한다 [3]. 그러나 이러한 방식은 현재 사용자의 질의와 밀접하게 연관된 과거의 중요한 정보를 손실시킬 위험이 있으며, 불필요한 맥락까지 전송하게 되어 네트워크 대역폭 낭비와 지연 시간 문제를 근본적으로 해결하지 못한다는 한계가 있다.

따라서 본 논문에서는 대화형 LLM 서비스 환경에서 발생하는 이러한 문제를 네트워크 트래픽 관점에서 분석하고, 불필요한 맥락 전송을 억제함으로써 서비스 지연을 완화할 수 있는 컨텍스트 관리 기반의 트래픽 최적화 기법을 제안한다.

우리가 제안하는 기법은 LLM과의 상호 대화 과정에서 현재 사용자의 질의와 과거 대화 맥락 간의 의미적 유사성에 기반하여, 응답 생성에 기여도가 높은 맥락만을 선택적으로 유지하고, 상대적으로 기여도가 낮은 맥락의 전송을 억제함으로써 어플리케이션 계층의 불필요한 트래픽을 감소시키는 방식이다.

Ⅱ. 본론

본 논문에서는 대화형 LLM 서비스에서 발생하는 불필요한 네트워크 트래픽을 줄이기 위해, 현재 사용자 질의와 과거 대화 맥락 간의 의미적 유사성에 기반한 컨텍스트 인식 트래픽 최적화 기법을 제안한다.

그림 1과 같이 제안하는 기법은 응답 생성에 기여도가 높은 맥락만을 선택적으로 전송함으로써, 동일한 컨텍스트 예산 하에서 네트워크 효율성과 응답 품질을 동시에 개선하는 것을 목표로 한다.

이에, 현재 질의 ut와 과거 대화 맥락 간의 관련성을 정량화하기 위해, 각 발화를 의미 공간 상의 임베딩 벡터로 변환한다. 계산된 의미적 유사도를 기반으로, 과거 대화 맥락 집합에서 유사도가 높은 일부 맥락만을 선택한다.

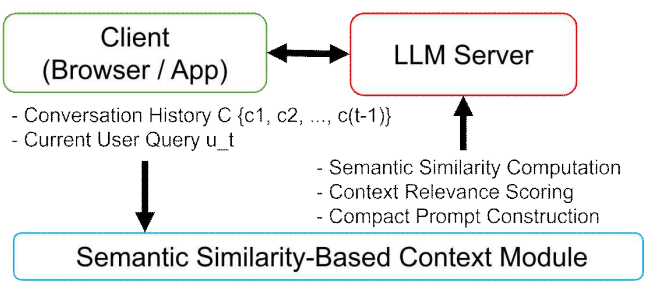

그림 1. 제안된 기술의 전반적인 시스템 구조도.

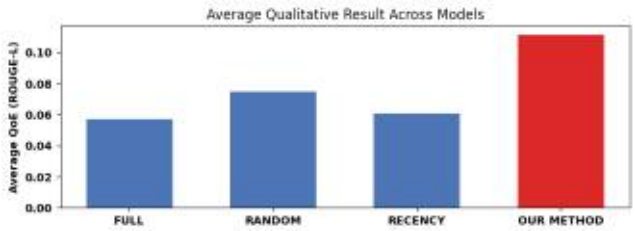

그림 2. 다양한 응답 생성 기법에 대한 정성 평가 비교 결과

최종적으로 서버로 전송되는 프롬프트는 선택된 컨텍스트와 현재 사용자 질의를 포함하여 구성된다. 이를 통해 응답 생성에 기여도가 낮은 맥락의 전송을 억제할 수 있으며, 업링크 페이로드 크기를 감소시켜 네트워크 지연 시간을 효과적으로 완화할 수 있다.

제안하는 기법은 모델의 내부 어텐션 구조를 수정하지 않고, 추론 이전 단계에서 의미적 유사도 계산을 수행함으로써 기존 LLM 서비스 구조와의 호환성을 유지한다.

Ⅲ. 결론

실험에는 채팅 기반 대규모 언어 모델 오픈소스 데이터셋 OpenAssistant Conversations Dataset(OASST1)을 사용하였다 [4]. 해당 데이터셋은 실제 사용자 - 모델 간의 다중 턴 대화로 구성되어 있어, 대화 누적에 따른 트래픽 증가 현상을 분석하기에 적합하다.

언어 모델로는 경량 모델부터 상대적으로 큰 모델까지 서로 다른 크기를 가지는 세 가지 사전학습 언어 모델을 사용하였다. 사용된 모델은 distilgpt2, tiny-gpt2, gpt2이며, 모든 실험은 동일한 하드웨어 및 소프트웨어 환경에서 수행되었다. 각 질의에 대해 동일한 최대 페이로드 예산과 생성 토큰 수를 적용하여 공정한 비교가 이루어지도록 하였다.

비교 대상 기법으로는 다음 네 가지를 고려하였다 [5]:

- FULL: 모든 과거 대화 맥락을 전송
- RECENCY: 최근 대화 일부만 전송
- RANDOM: 과거 맥락을 무작위로 선택
- UTILITY(제안 기법): 의미적 유사성 기반 선택

응답 품질을 평가하기 위해, 생성된 응답과 기준 응답 간의 의미적 유사도를 기반의 ROUGE-L 기법을 사용하여 비교하였다 [6]. 세 모델의 평균ROUGE-L 지수를 비교한 그림 2과 같이, 동일한 페이로드 조건 하에서 제안 기법은 ROUGE-L 지수가 FULL 방식 대비하여 응답 품질을 최대 96%까지 향상시키는 결과를 보였으며, 이는 불필요한 맥락 제거가 오히려 응답 생성에 긍정적인 영향을 줄 수 있음을 시사한다.

추가적으로, 우리는 네트워크 지연 성능에 대한 정량적인 평가를 위해 다양한 컨텍스트 관리 기법들에 대해 종단 간 지연 성능을 평가하였다. FULL 방식은 대화 전환이 누적됨에 따라 업링크 페이로드가 지속적으로 증가하기 때문에 모든 모델에서 가장 높은 지연 시간을 나타낸다. 반면에, 제안된 컨텍스트 인식 방법은 의미적으로 관련된 과거 컨텍스트만 선택적으로 포함함으로써 불필요한 컨텍스트 전송을 줄여 지연 성능을 다른 기법들에 비해 개선한다.

결과적으로, 제안 기법은 동일한 페이로드 예산 조건 하에서 FULL 방식 대비 평균 업링크 트래픽을 유의미하게 감소시켰으며, 그 결과 종단 간 지연 시간이 모델에 따라 0.93%부터 12%까지 개선되었다.

이에, 본 논문의 기여는 다음과 같이 요약할 수 있다.

우선, 우리는 대화형 LLM 서비스를 어플리케이션 계층의 네트워크 트래픽 문제를 대상으로, 사용자와의 상호 대화의 누적에 따라 발생하는 페이로드 증가와 지연 시간 저하 현상을 체계적으로 분석하였다.

이에 대하여, 의미적 유사성에 기반한 컨텍스트 관리 기법을 통해, 응답 품질에 기여도가 낮은 맥락의 전송을 억제함으로써 불필요한 네트워크 트래픽을 감소시키는 방법을 제시하였다.

또한 OASST1과 다수의 언어 모델을 활용한 실험을 통해, 제안된 본 기법이 기존 전체 컨텍스트 전송 방식 대비 종단 간 지연 시간을 감소시키고 사용자 체감 응답 품질을 유의미하게 향상시킴을 검증하였다.

본 연구는 대화형 LLM 서비스의 성능 저하를 단순한 모델 추론 문제로 보지 않고, 네트워크 트래픽 관점에서 접근함으로써 새로운 최적화 방향을 제시한다. 향후 연구에서는 보다 다양한 네트워크 환경과 실제 서비스 시나리오를 고려한 확장 실험을 통해, 제안 기법의 실용성을 더욱 검증할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Climate, Energy \& Environment(MCEE) of the Republic of Korea (No. RS-2022-KP002860) and also supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2021-II211835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

참 고 문 헌

[1] Gordon Owusu Boateng, Hani Sami, Ahmed Alagha, Hanae Elmekki, Ahmad Hammoud, Rabeb Mizouni, Azzam Mourad, Hadi Otrok, Jamal Bentahar, Sami Muhaidat, et al. A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. IEEE Communications Surveys & Tutorials, 2025.

[2] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. arXiv preprint arXiv:2406.16937, 2024.

[3] Jiachen Liu, Jae-Won Chung, Zhiyu Wu, Fan Lai, Myungjin Lee, and Mosharaf Chowdhury. Andes: Defining and enhancing quality-of-experience in llm-based text streaming services. arXiv preprint arXiv:2404.16283, 2024.

[4] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. Advances in neural information processing systems, 36:47669 - 47681, 2023.

[5] Boheng Sheng, Jiacheng Yao, Meicong Zhang, and Guoxiu He. Dynamic chunking and selection for reading comprehension of ultra-long context in large language models. arXiv preprint arXiv:2506.00773, 2025.

[6] Taojun Hu and Xiao-Hua Zhou. Unveiling llm evaluation focused on metrics: Challenges and solutions. arXiv preprint arXiv:2404.09135, 2024.