모바일 환경에서 대규모 언어 모델 추론의 전력 및 열 효율성 분석에 관한 연구

조용화, 이경한*
서울대학교, *서울대학교

yonghwacho@snu.ac.kr, * kyunghanlee@snu.ac.kr

A Study on the analysis of Power and Thermal Efficiency of Large Language Model Inference on Mobile systems

Yonghwa Cho, Kyunghan Lee*
Seoul National Univ., * Seoul National Univ.

요 약

본 논문에서는 LLM 추론의 실행 특성이 입력 토큰 수와 문맥 길이에 따라 동적으로 변화하며, 주변 온도와 같은 열 환경에 의해서도 성능 및 전력 소모 양상이 달라진다는 점을 관찰한다. 이러한 특성으로 인해 동일한 지연 시간 요구사항을 만족하는 주파수 설정이라 하더라도 성능 정체와 에너지 효율은 크게 달라질 수 있다. 이는 정적인 규칙에 기반한 기존의 동적 전압 및 주파수 조절(DVFS) 기법이 모바일 LLM 추론 환경에 충분히 효과적이지 않음을 시사한다. 이에 본 논문에서는 입력 특성과 열 환경 변화에 적응적으로 대응할 수 있는 학습 기반 주파수 제어 방식의 필요성을 논의한다.

Ⅰ. 서 론

대규모 언어 모델(LLM)은 최근 자연어 처리 분야에서 뛰어난 성능을 보이며 다양한 모바일 서비스에 활용되고 있다. 기존의 LLM 기반 서비스는 주로 원격 데이터 센터에서 추론을 수행하는 방식에 의존해 왔으나, 네트워크 지연과 개인정보 보호 문제로 인해 모바일 기기에서 직접 LLM 을 실행하려는 요구가 증가하고 있다. 이에 따라 주요 모바일 제조사들은 스마트폰 상에서 실시간 번역, 요약, 개인 비서 기능과 같은 온디바이스 LLM 서비스를 적극적으로 도입하고 있다.

그러나 모바일 환경에서의 LLM 추론은 높은 전력 소모를 동반한다. 선행 연구에 따르면 비교적 작은 입력 길이를 갖는 양자화 LLM 조차도 반복적인 질의 수행 시 배터리를 급격히 소모할 수 있음이 보고되었다. 이러한 전력 소모는 단순한 배터리 소진 문제를 넘어, 발열 증가로 인한 열 쓰로틀링을 유발하여 응답 지연 증가 및 사용자 경험 저하로 이어진다.[1]

모바일 시스템은 이러한 문제를 완화하기 위해 DVFS 를 활용하여 CPU 와 메모리의 동작 주파수를 조절한다. DVFS 는 워크로드에 따라 성능과 전력 소모를 균형 있게 조절할 수 있는 기법으로 널리 사용되고 있다. 그러나 실제 모바일 환경에서 사용되는 DVFS governor 들은 성능 극대화를 위해 높은 주파수를 유지하거나, 단순히 주파수를 낮추는 방식으로 전력을 절감하는 경향이 있다. 이러한 접근은 연산 특성이 빠르게 변화하는 LLM 추론에서는 비효율적인 결과를 초래할 수 있다.[2]

본 논문에서는 모바일 LLM 추론의 에너지 비효율성이 단순한 연산량 증가가 아니라, 연산과 메모리 접근 간 불균형으로 인해 발생하는 성능 정체에서 비롯된다는 점에 주목한다. 이를 통해 기존 DVFS 기법의 한계를 분석하고, 보다 세밀한 제어의 필요성을 논의한다

Ⅱ. 본론

LLM 추론의 실행 특성은 입력 토큰 수와 문맥 길이에 따라 시간적으로 지속적으로 변화한다. 초기 추론 단계나 짧은 입력을 처리하는 경우에는 비교적 계산 중심의 연산이 주를 이루는 반면, 문맥 길이가 증가하고 반복적인 추론이 수행될수록 메모리 접근 빈도가 증가하여 메모리 대역폭에 대한 의존도가 점차 커지는 경향을 보인다. 이러한 변화는 추론 과정 전반에 걸쳐 일정하지 않으며, 동일한 질의 처리 중에도 단계에 따라 상이한 성능 특성이 나타난다. 이로 인해 동일한 LLM 모델과 동일한 하드웨어 환경에서도 입력 조건에 따라 성능 병목의 원인은 크게 달라질 수 있다. 특정 구간에서는 계산 자원이 충분히 활용되지 못한 채 메모리 접근을 대기하는 상황이 발생하며, 이러한 유휴 시간은 성능 정체(stall)로 이어진다. 특히 모바일 환경에서는 이러한 성능 정체가 단순한 지연 시간 증가에 그치지 않고, 에너지 효율 저하로 직결된다는 점에서 더욱 중요한 문제가 된다.

본 연구의 관찰 결과, 동일한 지연 시간 요구사항을 만족하는 서로 다른 주파수 조합들 사이에서도 성능 정체 비율은 크게 달라질 수 있으며, 성능 정체가 최소화되는 지점에서 에너지 소모 또한 최소화되는 경향을 보였다. 이는 LLM 추론에서의 에너지 낭비가 단순히 연산 수행에 필요한 에너지보다, 실제로 유효한 작업을 수행하지 못하고 대기하는 시간 동안 소모되는 에너지에서 상당 부분 발생함을 시사한다. 따라서 에너지

효율을 향상시키기 위해서는 평균적인 성능 지표나 단일 자원의 활용도만을 기준으로 주파수를 조절하는 기존 방식에서 벗어날 필요가 있다.

기존의 모바일 DVFS 기법들은 주로 CPU 중심의 일반적인 워크로드를 가정하여 설계되었다. 이러한 기법들은 입력 특성이 비교적 안정적인 애플리케이션에서는 일정 수준의 효과를 보일 수 있으나, LLM 추론과 같이 실행 특성이 입력 및 시간에 따라 급격히 변화하는 워크로드에서는 한계를 드러낸다. 예를 들어, 특정 입력 길이에서 적절한 주파수 설정이 다른 입력 조건에서는 과도한 에너지 소모나 성능 저하를 유발할 수 있으며, 이를 사전에 정의된 규칙이나 고정된 임계값 기반 정책으로 포괄하기는 어렵다.[3]

더 나아가, 모바일 기기는 주변 온도, 사용 환경, 누적 발열 상태에 따라 동일한 주파수 설정에서도 서로 다른 성능 및 전력 소모 양상을 보인다. 초기 온도가 높은 상태에서 LLM 추론이 시작되는 경우 열 쓰로틀링이 더 이르게 발생할 수 있으며, 이는 지연 시간 증가와 함께 에너지 효율을 더욱 악화시킨다. 반대로, 열 상태가 상대적으로 안정적인 환경에서는 보수적인 주파수 설정이 불필요한 성능 저하를 초래할 수 있다. 이러한 열 환경의 변화는 정적인 DVFS 정책의 유효 범위를 더욱 제한한다.

이와 같이 LLM 추론은 입력 토큰 수, 문맥 길이, 실행 시점에 따른 단계적 특성, 그리고 열 환경과 같은 다양한 요인에 의해 복합적으로 영향을 받는다. 이러한 요인들은 서로 독립적이지 않으며, 상호작용을 통해 성능 정체와 에너지 소모에 영향을 미친다. 따라서 단일 지표나 단순한 규칙에 기반한 주파수 제어 방식으로는 모든 실행 상황에서 안정적인 에너지 효율을 달성하기 어렵다.

이에 따라 본 논문에서는 이러한 동적이고 복합적인 특성을 갖는 LLM 추론 환경에서, 실행 중 관측되는 상태 정보를 기반으로 주파수 설정을 적응적으로 조절할 수 있는 학습 기반 제어 방식의 필요성을 제기한다. 특히 강화학습(Reinforcement Learning)은 실행 과정에서 관측되는 지연 시간, 성능 정체 비율, 에너지 소모, 그리고 열 상태와 같은 다양한 정보를 상태로 활용하고, 주파수 조절을 행동으로 정의함으로써, 장기적인 에너지 효율을 최적화할 수 있는 가능성을 제공한다.

강화학습 기반 접근은 사전에 모든 입력 조건과 환경 변화를 명시적으로 모델링하지 않더라도, 반복적인 실행을 통해 각 상태에 적합한 주파수 설정을 점진적으로 학습할 수 있다는 장점을 가진다. 이는 입력 특성과 열 환경이 지속적으로 변화하는 모바일 LLM 추론 환경에서, 기존의 휴리스틱 기반 DVFS 정책보다 높은 적응성과 확장성을 제공할 수 있다. 이러한 관점에서 학습 기반 적응형 DVFS 제어는 모바일 환경에서의 전력 및 열 효율적인 LLM 추론을 달성하기 위한 유망한 접근 방식으로 판단된다.

Ⅲ. 결론

본 본 논문에서는 모바일 환경에서 대규모 언어 모델(LLM) 추론이 가지는 전력 및 열 비효율성의 원인을 분석하고, 기존 동적 전압 및 주파수 조절(DVFS) 기법의 한계를 논의하였다. LLM 추론은 입력 토큰 수와 문맥 길이에 따라 실행 특성이 동적으로 변화하며, 이러한 변화는 성능 병목과 에너지 소모 양상에 직접적인 영향을 미친다. 특히 동일한 지연 시간 요구사항을 만족하는 주파수 설정이라 하더라도, 성능 정체의 정도에 따라 에너지 효율이 크게 달라질 수 있음을 관찰하였다.

또한 모바일 기기의 열 환경은 실행 시점과 사용 조건에 따라 지속적으로 변화하며, 이러한 열 상태는 동일한 주파수 설정에서도 상이한 성능 및 전력 소모를 초래한다. 초기 온도가 높은 상태에서 LLM 추론이 시작되는 경우 열 쓰로틀링이 조기에 발생하여 지연 시간 증가와 에너지 효율 저하로 이어질 수 있으며, 반대로 열 상태가 안정적인 환경에서는 보수적인 주파수 설정이 불필요한 성능 저하를 유발할 수 있다. 이러한 특성은 정적인 DVFS 정책의 유효 범위를 제한하는 주요 요인으로 작용한다.

이와 같은 관찰 결과는 LLM 추론이 입력 특성, 실행 단계, 그리고 열 환경과 같은 다양한 요인에 의해 복합적으로 영향을 받는 워크로드임을 시사한다. 따라서 단일 지표나 고정된 규칙에 기반한 기존의 주파수 제어 방식으로는 모든 실행 상황에서 안정적인 에너지 효율을 달성하기 어렵다. 이에 본 논문에서는 실행 중 관측되는 상태 정보를 바탕으로 주파수 설정을 적응적으로 조절할 수 있는 학습 기반 제어 방식의 필요성을 제기하였다.

향후 연구에서는 강화학습 기반 접근을 통해 입력 특성과 열 상태의 변화에 따라 주파수 설정을 동적으로 조정하는 기법을 설계하고, 이를 실제 모바일 LLM 추론 환경에 적용하여 전력 소모, 성능 정체, 그리고 사용자 경험 측면에서의 효과를 정량적으로 평가할 예정이다. 이러한 접근은 모바일 환경에서의 전력 및 열 효율적인 LLM 추론을 달성하기 위한 실질적인 대안이 될 것으로 기대된다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT)(RS-2022-NR070834)

참 고 문 헌

[1] Amirhossein Ahmadi, Hazem A. Abdelhafez, Karthik Pattabiraman, and Matei Ripeanu, “EdgeEngine: A Thermal-Aware Optimization Framework for Edge Inference,” *Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing (SEC '23)*, Wilmington, DE, USA, pp. 67- 79, 2023.

[2] Kyungmin Bin, Seyeon Kim, Sangtae Ha, Song Chong, and Kyunghan Lee, “NeuroBalancer: Balancing System Frequencies with Punctual Laziness for Timely and Energy-Efficient DNN Inferences,” *IEEE Transactions on Mobile Computing*, vol. 24, no. 5, pp. 4339- 4354, May 2025.

[3] Seyeon Kim, Kyungmin Bin, Sangtae Ha, Kyunghan Lee, and Song Chong, “zTT: Learning-Based DVFS with Zero Thermal Throttling for Mobile Devices,” *GetMobile: Mobile Computing and Communications*, vol. 25, no. 4, pp. 30- 34, Mar. 2022.