엣지 환경에서 LLM 학습을 위한 전력 소비 분석 연구

최다은, 이재연, 강동기*

전북대학교

de2572@naver.com, wodus7792@naver.com, dongkikang@jbnu.ac.kr

A Study on Power Consumption Analysis for LLM Training in Edge Environments

Da Eun Choi, Jae Yeon Lee, and Dong Ki Kang*

Jeonbuk National University

요 약

최근 엣지 디바이스 기반의 실시간 생성형 인공지능 수요가 증가함에 따라, 전력 및 자원이 제한된 환경에서의 온디바이스 파인튜닝 기술이 주목받고 있다. 본 연구에서는 NVIDIA Jetson Xavier NX와 Orin Nano에서 GPT-2 Small 및 DistilGPT2 모델의 학습을 수행하고, 소요 시간과 에너지 소비를 측정·분석하였다. 실험 결과는 엣지 환경에서의 에너지 효율적인 LLM 배포 프레임워크 설계에 기여할 수 있을 것으로 기대된다.

Ⅰ. 서 론

최근 거대언어모델(LLM, Large Language Model)을 중심으로 하여 임베디드를 위한 생성형 인공지능 기술이 급속도로 발전됨에 따라, 로봇, 드론 및 IoT 센서 등 엣지 디바이스를 기반으로 하는 실시간 인공지능 서비스 시장이 크게 확대되고 있다 [1][2]. 이러한 엣지 환경에서는 중앙 서버와의 통신 지연을 줄이고 데이터 프라이버시를 보장하는 것이 중요하기 때문에, 디바이스 자체에서 모델의 추론뿐 아니라 학습까지 수행하는 온디바이스 파인튜닝(On-Device Fine-Tuning) 기술을 접목해야 할 필요성이 있다. 그러나 데이터센터의 고성능 컴퓨팅 서버들과 달리 엣지 디바이스에서는 전력 및 연산 자원이 제한적이거나 배터리 용량 한계가 존재하므로, 전력 소비량 및 공급 현황을 고려하여 적절한 모델 학습 할당을 수행할 수 있는 프레임워크 개발이 요구된다 [3].

본 논문에서는 엣지 디바이스인 NVIDIA Jetson Xavier NX 및 Jetson Orin Nano를 기반으로 인공지능 모델에 대한 학습을 수행하였으며, 학습 소요 시간, 소비 전력 및 에너지 소비량을 측정 및 분석하였다. 인공지능 모델은 GPT-2 모델을 기반으로 하여 경량화 기술을 통해 파라미터 수를 줄인 GPT-2 Small [4] 및 DistilGPT2 [5] 모델을 채택하였다. 데이터셋으로는 다양한 분야의 위키피디아 문서를 기반으로 구성한 SQuAD(Stanford Question Answering Dataset)를 활용하였다 [6]. 도출된 실험 결과 및 분석 내용은 향후 LLM 배포를 위한 엣지 프레임워크의 에너지 효율성 개선 연구에 활용될 수 있을 것으로 기대한다.

Ⅱ. 본론

기존의 온디바이스 전력 소비 연구는 주로 딥러닝 모델의 추론 과정에서 발생하는 전력 소모 특성의 정량적 측정에 초점을 맞추어 왔다. 선행 연구 [7]과 [8]에서는 Jetson Xavier NX 및 Jetson Orin Nano 디바이스 위에서 객체 탐지 모델 추론 작업을 실행하였으며, 그에 따른 평균 소비 전력 크기를 측정하였다. 관련 수치는 표 1에 나타나 있다.

그러나 기존 연구들은 공통적으로 추론 중심의 워크로드에 기반하여 수행되었기 때문에, 실제로 연산 부하가 집중되는 학습 혹은 파인튜닝 과정에서의 전력 소비 특성에 대한 정밀한 분석은 부족한 실정이다. 특히 LLM과 같이 높은 연산을 수반하는 학습 환경에서는, 디바이스별 전력 소비 양상 및 연산 안정성이 현저히 달라질 수 있다.

	Jetson Xavier NX	Jetson Orin Nano
전체 평균 전력(W)	13.92W	10.4W
연산 전력 (W)	10.54W	5.91W
연산 전력 비율 (%)	약 75.8%	약 56.8%

표 1. Xavier NX와 Orin Nano의 모델 추론 작업 소비 전력 현황

본 연구에서는 이러한 한계를 보완하고자, Jetson Xavier NX와 Orin Nano 디바이스를 대상으로 LLM 학습 과정을 수행하며, 전력 효율성과 연산 안정성을 비교·분석하였다. 이를 통해 추론 단계에 국한되지 않은 학습 기반의 전력 소비 특성을 제시하고, 엣지 컴퓨팅 환경에서의 하드웨어 선택에 실질적인 기준을 제공하는 것을 목표로 한다.

Ⅲ. 실험결과

표 2은 본 논문에서 사용한 두 기기의 실험 환경을 요약한 것이다.

Jetson Xavier NX	GPU 384 NVIDIA CUDA Cores, SD Card 128GB, Jetpack 5, SSD FireCuda 530 1TB
Jetson Orin Nano	GPU 1024 NVIDIA CUDA Cores, SD Card 64GB, Jetpack 6, SSD FireCuda 530 1TB
LLM 및 데이터셋	LLM: GPT-2 Small, DistilGPT2 학습 epochs: 2 데이터셋: SQuAD2.0 (학습 데이터 10,000개)

표 2. 서버 환경 설정

Jetson Xavier NX의 경우 하드웨어 제약으로 인해 Jetpack 버전 5를 설치하였으며, Jetson Orin Nano는 별도 제약이 존재하지 않아 최신 버전인 Jetpack 버전 6을 설치하였다.

표 3와 4는 각각 표 2의 환경을 기반으로 실험을 수행 후 측정한 학습 소요 시간, 평균 소비 전력 및 전체 소비 전력량을 보인다.

GPT-2 Small 및 DistilGPT2 모델 학습 결과를 종합하면, Orin Nano는 Xavier NX에 비해 학습 시간은 평균 약 30% 단축되었고 평균 전력 소비는 약 20% 감소하였으며, 전체 소비 전력량은 약 44% 절감되어 전반적으로 더 높은 에너지 효율과 연산 성능을 보였다.

	Jetson Xavier NX	Jetson Orin Nano
학습 소요 시간	242분 41초	195분 26초
전체 소비 전력량	57.82 [Wh]	36.33 [Wh]
평균 소비 전력	14.29 [W]	11.15 [W]

표 3. GPT2-Small 모델 학습 측정 결과

	Jetson Xavier NX	Jetson Orin Nano
학습 소요 시간	264분 56초	156분 54초
전체 소비 전력량	59.05 [Wh]	28.91 [Wh]
평균 소비 전력	13.37 [W]	11.05 [W]

표 4. DistilGPT2 모델 학습 측정 결과

그림 1은 GPT2-Small 모델 학습 중 시간에 따른 소비 전력 변화를 보여준다. Xavier NX는 전체 학습 시간 동안 14.5 [W] 내외로 높은 소비 전력을 보이지만, Orin Nano는 약 11 [W] 정도의 낮은 수준을 유지하였다.

그림 2는 DistilGPT2 모델 학습 과정에서 측정된 시간 경과에 따른 소비 전력 변화를 시각화한 그래프이다. Xavier NX는 전체 학습 시간 동안 평균 약 13~14.5W 수준의 전력 소비를 유지하였으나, 일부 구간에서는 전력 하강 현상이 반복적으로 발생하였다. 이는 Xavier NX가 전력 소모량이 일정 임계치를 초과하거나, 시스템 온도가 일정 수준 이상으로 상승했을 때 자동으로 GPU 클럭을 낮추거나 전력 공급을 제한하는 동적 전력 관리 메커니즘이 작동하는 결과로 해석할 수 있다 [9]. 반면 Orin Nano는 전 구간에서 11 [W] 내외의 안정적인 소비 전력 패턴을 유지하였으며, 전력 변동 폭도 작게 나타났다.

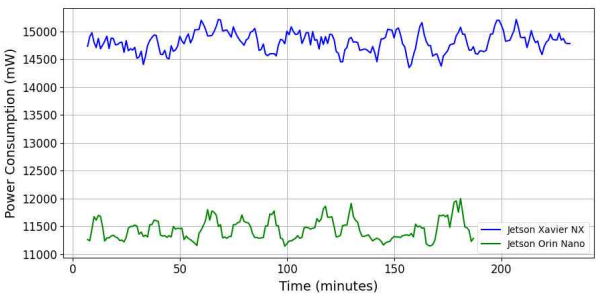

그림 1 GPT2-Small 모델의 시간별 소비 전력 변화 추이

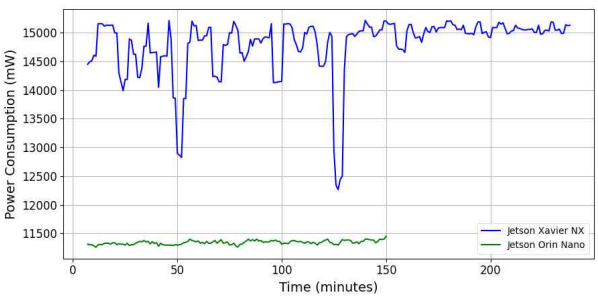

그림 2 DistilGPT2 모델의 시간별 소비 전력 변화 추이

그림1과 2의 결과를 통해 전반적으로 Orin Nano가 더 높은 전력 효율성과 안정성을 제공함을 확인할 수 있었다. 본 연구 결과를 통해, 엣지 디바이스 환경에서의 온디바이스 파인튜닝시 모델 구조 최적화뿐 아니라 적절한 기기 선택이 에너지 효율성에 결정적인 영향을 끼친다는 것을 알 수 있다.

Ⅳ. 결론

본 연구에서는 Jetson Xavier NX 및 Jetson Orin Nano 엣지 디바이스위에서 GPT-2 Small 및 DistilGPT2 모델을 파인튜닝할 때, 도출되는 학습 소요 시간, 평균 소비 전력 및 전체 소비 전력량을 측정하고 분석하였다. Orin Nano는 Xavier NX 대비 평균 소비 전력이 약 20% 낮고 학습 시간은 약 30% 단축되었으며 전체 소비 전력량은 약 44% 절감되는 등 에너지 효율성에서 우위를 보였다. 본 연구 결과는 향후 LLM 배포를 위한 엣지 프레임워크의 에너지 효율성 개선 연구에 활용될 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부와 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2022-0-01067)

참 고 문 헌

[1] https://venturebeat.com/business/openinfer-raises-8m-for-ai-inference-at-the-edge/

[2] https://venturebeat.com/ai/liquid-ai-is-revolutionizing-llms-to-work-on-edge-devices-like-smartphones-with-new-hyena-edge-model/

[3] M. Shafique, "A Cross-Layer Approach to Energy-Efficient and Secure EdgeAI: Architectures, Systems and Applications," 2024 5th CPSSI International Symposium on Cyber-Physical Systems (Applications and Theory) (CPSAT), Tehran, Iran, Islamic Republic of, 2024, pp. 1-1, doi: 10.1109/CPSAT64082.2024.10745418

[4] https://huggingface.co/openai-community/gpt2

[5] https://huggingface.co/distilbert/distilgpt2

[6] https://rajpurkar.github.io/SQuAD-explorer/

[7] J. Lee, P. Wang, R. Xu, S. Jain, V. Dasari, N. Weston, Y. Li, S. Bagchi, and S. Chaterji, "Virtuoso: Energy- and Latency-aware Streamlining of Streaming Videos on Systems-on-Chips," ACM Transactions on Design Automation of Electronic Systems, vol. 28, no. 3, pp. 1 - 32, Apr. 2023. [Online]. Available: https://doi.org/10.1145/3564289

[8] M. Rouchou, "Development of an Adaptive Pipeline for Object Detection Training and Benchmarking," M.S. thesis, Inst. of Technology, Univ. of Tartu, 2024.

[9] A. Dutt, "Evaluating the energy impact of device and workload parameters for DNN inference on edge," M.S. thesis, Dept. of Computer Science, Stony Brook Univ., NY, USA, 2023. [Online]. Available: https://commons.library.stonybrook.edu/electronic-disserations-theses/46/