이기종 Jetson 플랫폼 기반 경량 LLM 성능 분석 연구

이재연, 최다은, 강동기*

전북대학교

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

Performance Analysis of Lightweight LLMs on Heterogeneous Jetson Platforms

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

Jeonbuk Univ.

요 약

최근 엣지 디바이스(Edge Device) 환경에서 경량화된 거대 언어 모델(Lightweight Large Language Model)을 배포하고 효율적으로 추론 및 학습을 수행할 수 있게 하는 다양한 연구들이 활발히 제시되고 있다. 본 논문에서는 이기종(Heterogeneous) NVIDIA Jetson 플랫폼 및 다양한 경량 LLM을 기반으로 하여 모델 추론 및 학습 성능을 측정하였고 정량적인 분석을 제시하였다. 본 연구의 결과는 IoT 환경에서 생성형 인공지능 서비스를 제공하는 엣지 지능형 프레임워크 설계를 위한 기초 자료로써 활용될 수 있다.

Ⅰ. 서론

최근 GPT(Generative Pre-trained Transformers)와 같은 거대 언어 모델(LLM: Large Language Model)의 추론 성능이 비약적으로 향상됨에 따라, 다양한 도메인에서 LLM 기반의 생성형 인공지능 응용 서비스가 폭넓게 개발되고 배포되고 있다 [1]. 특히 기존의 클라우드 중심 인프라에서 벗어나 자원이 제한된 엣지 환경에서 LLM을 효율적으로 운용하려는 연구가 활발히 진행되고 있으며, 이는 단순한 모델 추론뿐 아니라 대상 디바이스 상에서 직접 수행되는 온디바이스 파인튜닝(On-Device Fine-Tuning)과 같은 모델 학습 기법에 대한 수요 증가로 이어지고 있다 [2].

그러나 지속적인 프로세서 및 하드웨어 가속기의 발전에도 불구하고, 엣지 디바이스는 여전히 CPU, GPU, 메모리 등에서 제한적인 자원을 보유하고 있어, LLM의 추론 및 학습과 같은 고연산 작업을 수행하는 데 구조적인 제약이 있다 [3]. 이러한 한계를 극복하기 위해서는 추론 성능을 가능한 한 유지하면서도 모델 파라미터의 크기를 효과적으로 축소할 수 있는 모델 경량화(Model Compression) 기법이 필요하며, 동시에 이기종(Heterogeneous) 엣지 디바이스들이 지닌 다양한 하드웨어 사양을 고려한 자원 최적화 및 할당 전략(Resource-Aware Scheduling and Deployment)이 요구된다.

본 논문에서는 이러한 연구의 출발점으로서, 이기종 NVIDIA Jetson 플랫폼에 경량화된 LLM을 배포하고 각각의 환경에서 모델의 추론과 학습을 수행하며 성능을 비교·분석하는 실험을 설계하였다. 실험을 통해 도출된 결과는 향후 엣지 디바이스 기반의 IoT 환경에서 생성형 인공지능 서비스를 보다 효과적으로 관리 및 운용할 수 있는 프레임워크의 설계와 구현에 기여할 수 있을 것으로 기대한다.

Ⅱ. 본론

본 연구에서는 NVIDIA Jetson 플랫폼 위에서 Transformer 기반 경량 LLM을 학습시킬 때 도출되는 성능을 측정 및 분석하고자 한다. Jetson 플랫폼으로서 Jetson Xavier NX 디바이스 [4]와 Jetson Orin Nano 디바이스 [5]를 채택하여 실험을 진행하였다. 표 1에서는 각 디바이스의 하드웨어 사양을 요약하고 있다. 모델 설치를 위하여 SD 카드 64GB 및 SSD (FireCuda 530, 1TB) 디바이스를 장착하였다. 표 2에서는 학습 성능을 분석하기 위해 설치한 GPT-2 기반의 TinyGPT2 [6] , DistilGPT2 [7] 및 GPT-2 Small [8] 모델과 추론 성능을 분석하기 위한 TinyLlama-1.1b [9] , Phi-2 [10] , Qwen1.5-1.8b [11] 의 특성을 요약하고 있다. TinyGPT2는 경량화 기법을 적용하여 GPT2 모델의 파라미터 수를 크게 줄인 초소형 모델이며, DistilGPT2는 지식 증류 [12] 기법을 적용하여 GPT-2 모델의 크기를 절반으로 줄이면서도 학습 성능은 유사한 수준으로 유지할 수 있게 한다. GPT-2 Small은 별도의 경량화 기법 적용 없이 파라미터 수만을 줄인 모델로서 엣지 디바이스 및 모바일 환경에서 학습 수행이 가능하다. 추론 작업은 학습 작업보다 적은 자원을 요구하므로, 추론 실험에는 더 큰 규모의 모델을 사용하였다. 데이터 세트로는 주어진 질문에 대해 정답 구간을 찾는 질의응답(QnA) 형식으로 구성된

	Jetson Xavier NX	Jetson Orin Nano
출시연도	2020	2023
GPU	384-core NVIDIA Volta	1024-core NVIDIA Ampere
CPU	6-core ARM NVIDIA Carmel (64-bit)	6-core ARM Cortex-A78AE (64-bit)
INT8 연산 성능	21 TOPS	67 TOPS
DL Accelerator	NVDLA 2개	-
메모리	8GB LPDDR4x (bandwidth: 59.7 GB/s)	8GB LPDDR5 (bandwidth: 102 GB/s)
소비전력	최대 20W	최대 25W

표 1. Jetson 디바이스 하드웨어 사양

학습 모델			
모델명	TinyGPT2	DistilGPT2	GPT-2 Small
모델 파라미터 수	2,500만 개	8,200만 개	1억 1,700만 개
주요 특징	경량화	지식 증류 경량화	GPT-2 기본구조
추론 모델			
모델명	TinyLlama -1.1b	Phi-2	Qwen1.5-1.8b
모델 파라미터 수	11억 개	27억 개	18억 개
주요 특징	엣지 최적화	논리 추론 강화	다국어·챗봇

표 2. LLM 별 주요 특성

SQuAD(Stanford Question Answering Dataset)를 사용하였으며, 이는 일반 상식, 역사, 과학, 문학 등 다양한 분야의 위키피디아 문서를 기반으로 구성되어 있다 [13].

Ⅲ. 실험 결과

Jetson 디바이스 및 경량 LLM 별 처리 성능을 비교하기 위하여 모델 학습 소요 시간과 GPU 이용률을 측정하였다. 소요 시간은 별도의 파서(Parser) 모듈을 구현하여 측정하였으며 GPU 이용률은 JetPack에 내장된 tegrastats 커맨드를 실행하여 수집하였다. 모델 학습을 위한 epoch는 각 모델에 대하여 2-epochs 씩 설정하였다.

그림 1은 모델 학습 소요 시간 결과를 보인다. 모델 파라미터 수가 상대적으로 적은 TinyGPT2의 경우 Orin Nano는 Xavier NX 대비 약 40%의 소요 시간 절감을 보여준 반면, 파라미터 수가 더 많은 DistilGPT2의 경우 약 60%의 성능 향상을 보였다. 모델 추론 결과를 나타내는 그림2의 TinyLlama-1.1b에서는 Orin Nano가 Xavier NX 대비 약 1.6배의 소요 시간 절감을 보여주며, Phi-2와 Qwen1.5-1.1b 모델의 경우 약 2.1배의 성능 향상이 나타났다.

해당 결과는 LLM 학습 및 추론에 요구되는 연산량이 많아질수록 Orin Nano의 더 많은 Core 유닛, Ampere 아키텍처의 구조적 연산 효율성, 그리고 상대적으로 높은 메모리 대역폭이 복합적으로 작용하여 성능 향상에 기여함을 보여준다.

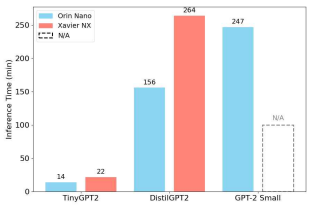

그림 1. LLM 학습 시간 비교

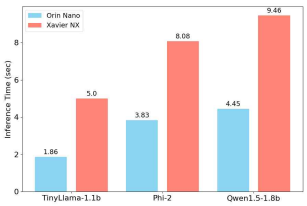

그림 2. LLM 추론 시간 비교

모델 학습 과정에서 가장 큰 모델 파라미터 개수를 가진 GPT-2 Small 모델의 경우 Orin Nano는 정상 학습이 되는 것과 달리, Xavier NX 디바이스에서는 학습 도중 프로세스가 강제 종료되어 실험을 진행할 수 없었다. 이는 Xavier NX GPU의 Volta 아키텍처가 Ampere 아키텍처와 비교하여 GPT-2 Small 모델 학습에 필요한 dense matrix 연산 효율성이 떨어지고 발열로 인한 throttling이 발생하여 GPU 클럭이 쉽게 다운되기 때문으로 분석된다 [14].

평균 GPU			
학습 모델	TinyGPT2	DistilGPT2	GPT-2 Small
Jetson Orin Nano	67.05%	96.73%	96.18%
Jetson Xavier NX	74.01%	91.42%	-
추론 모델	TinyLlama -1.1b	Phi-2	Qwen1.5-1.8b
Jetson Orin Nano	27.17%	33.05%	16.4%
Jetson Xavier NX	9.26%	14.41%	9.26%

표 3. 모델 평균 GPU 이용률 비교

표 3은 Xavier NX와 Orin Nano의 학습 및 추론 시 평균 GPU 이용률을 비교하여 보여준다. 가장 작은 모델인 TinyGPT2 모델은 학습 시 두 디바이스 간 유사한 성능을 나타냈으나, 나머지 모델에서는 Orin Nano가 더 높은 GPU 활용률을 보여주었다. 이는 Orin Nano 디바이스가 높은 메모리 대역폭을 통해 메모리 병목을 줄임으로써 GPU Core의 활용 시간을 늘려줄 뿐만 아니라 여유로운 전력 스케일링(Power Scaling) 설계 구조를 통해 GPU Core 자원을 적극적으로 사용할 수 있기 때문으로 분석된다.

Ⅳ. 결론

본 논문에서는 Jetson Orin Nano 및 Xavier NX 디바이스로 구성된 이기종 NVIDIA Jetson 플랫폼 위에서 경량 LLM 학습 및 추론을 수행하였으며 각 소요 시간과 GPU 이용률을 측정하였다. 실험 결과를 통해 동일한 LLM 및 데이터 세트에 대해서도 Jetson 디바이스의 특성에 따라 학습 및 추론 효율성이 달라질 수 있음을 확인하였다. 본 실험 결과는 향후 대규모의 엣지 인프라 환경에서 최적화된 모델 배포 및 자원 할당을 수행할 수 있는 프레임워크 설계 연구에 이바지할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] https://www.edge-ai-vision.com/2025/03/fine-tuning-llms-for-cost-effective-genai-inference-at-scale/

[2] Gao, Lei et al. “Enabling Efficient On-Device Fine-Tuning of LLMs Using Only Inference Engines” arXiv preprint arXiv:2409.15520 (2024)

[3] Abstreiter, Maximilian et al. “Sometimes Painful but Certainly Promising: Feasibility and Trade-offs of Language Model Inference at the Edge” arXiv preprint arXiv:2503.09114 (2025)

[4] https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-series/

[5] https://www.nvidia.com/ko-kr/autonomous-machines/embedded-systems/jetson-orin/

[6] https://huggingface.co/sshleifer/tiny-gpt2

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

[8] https://openai.com/index/gpt-2-1-5b-release/

[9] https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

[10] https://huggingface.co/microsoft/phi-2

[11] https://huggingface.co/Qwen/Qwen1.5-1.8B

[12] Hinton, Geoffrey et al. “Distilling the Knowledge in a Neural Network” arXiv preprint arXiv:1503.02531 (2015)

[13] https://huggingface.co/datasets/rajpurkar/squad

[14] Foster, Brett et al. “Evaluating Energy Efficiency of GPUs using Machine Learning Benchmarks” 2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) (2023) DOI: 10.1109/IPDPSW59300.2023.00019