거대 언어 모델을 위한 학습 시간 예측 연구 분석

정진우, 고영훈, 양경식, 유혁

고려대학교

{jwjeong, yhgo}@os.korea.ac.kr, g_yang@korea.ac.kr, chuckyoo@os.korea.ac.kr

Analysis of Training Time Prediction Techniques for Large Language Models

Jinwoo Jeong, Younghun Go, Gyeongsik Yang, Chuck Yoo

Korea Univ.

요 약

거대 언어 모델(LLM)은 자연어 처리와 다양한 산업 분야에서 활발히 사용되고 있다. LLM 학습은 높은 GPU 자원 요구와 긴 학습 시간을 지니고 있으며, 사용자는 경험적으로 GPU를 선택하게 되어 그 과정에서 GPU 낭비를 초래한다. 이를 개선하기 위해, 딥러닝 학습 시 소요되는 시간을 예측하기 위한 연구가 수행되어져 왔다. 기존 연구들은 주로 CNN, RNN 등 전통적인 딥러닝 모델의 학습 시간을 예측했으나, 트랜스포머 기반 LLM에 대한 학습 시간 예측은 고려하지 않는다. 본 연구는 기존 학습 시간 예측 기법들을 비교 분석하고, LLM에 특화된 학습 시간 예측을 위한 핵심 요구사항을 도출한다. 이를 통해 향후 LLM의 효율적인 자원 관리와 학습 최적화 기술을 개발하고자 한다.

Ⅰ. 서 론

현재 거대 언어 모델(Large Language Model, LLM)은 자연어 처리 및 생성, 번역, 요약 등 다양한 분야에서 널리 사용되고 있다. 특히 LLM을 확장하여 의료, 자동차 등 기술 집약적 산업에[1][2] 적용하려는 시도는 산업계에서 주목하고 있다. 그러나 LLM 연구는 대규모 GPU 자원 및 많은 훈련 시간 등 상당한 경제적 비용을 요구함으로 연구와 산업 현장에서 큰 경제적 부담으로 작용하고 있다. 이에 LLM 모델을 학습함에 있어 자원 낭비를 방지하고 효율적인 관리 방안이 요구된다.

이에 현재까지 다양한 연구들에서 딥러닝 모델의 학습 시간 예측을 시도하는 방법론이 제시되었다[3-6]. 예를 들어, 모델의 구조, 데이터셋의 크기, 하이퍼파라미터, 학습을 수행하는 환경(GPU) 등을 기반으로 학습 시간을 예측하는 기법들이 있다. 그러나 이러한 기법들은 주로 CNN, RNN, LSTM 등 전통적인 딥러닝 모델을 대상으로 개발되었으며, 트랜스포머 기반의 LLM에 직접 적용하기에는 한계가 있다. 가령, LLM은 트랜스포머 아키텍처의 인코더-디코더 구조를 사용하며, 레이어의 수, 파라미터의 규모, 어텐션 메커니즘 등에서 전통적 모델과 큰 차이를 보인다. 또한, 학습 속도를 개선하기 위한 PEFT 등 LLM 훈련에 특화된 기술들이 적용되기 때문에, 기존의 예측 모델이 이러한 복잡성을 충분히 반영하지 못한다.

이러한 배경에서 우리는 기존 연구들을 면밀히 비교 분석하고, LLM에 특화된 학습 시간 예측을 달성하기 위해 해결해야 할 핵심 요구사항을 도출하고자 한다. 이를 통해 향후 LLM의 빠르고 자원 효율적인 학습을 지원하는 초기 예측 기술을 연구하고, 거대 언어 모델의 파인튜닝(fine-tuning) 등 활용을 촉진하는 기술을 개발하고자 한다.

Ⅱ. LLM 및 기존 딥러닝 모델의 비교

LLM은 트랜스포머 아키텍처를 기반으로 임베딩, 어텐션, 인코더-디코더, 피드포워드 신경망 등을 주요 요소로 한다[8]. 먼저 LLM은 입력 값인 텍스트 데이터를 보다 상위 차원의 벡터로 변환하여 처리하며, 이 과정에 임베딩 레이어를 사용한다. 임베딩 레이어는 입력 텍스트를 고차원 벡터로 변환하여 모델이 단어의 의미와 문맥을 이해할 수 있도록 한다[8].

또한 self-attention 메커니즘을 통해 입력값을 구성하는 단어 사이의 상관성과 단어의 중요성을 평가한다. 구체적으로 self-attention 메커니즘은 문맥적 의미를 파악하고, 각 단어가 다른 단어와의 관계에서 얼마나 중요한지를 판단하여 중요한 단어에 더 높은 가중치를 부여한다. 그 결과는 일반적으로 피드포워드 신경망으로 전달되는데, 이 신경망은 비선형 활성 함수를 기반으로 입력값을 새로운 값으로 생성한다.

레이어 각각은 정규화 및 residual connection을 사용하여 학습 과정에서의 그라디언트 소실(vanishing gradient) 등에 대응할 수 있다. LLM은 이러한 레이어들을 계층적으로 쌓아 구성되며, 인코더는 입력 데이터를 처리하여 문맥적 정보를 추출하고, 디코더는 응답을 생성한다.

반면 전통적인 딥러닝 모델인 이미지 분류 모델은 CNN과 같이 상대적으로 덜 복잡하고 작은 규모의 모델로서 구성된다. 전통적인 자연어 처리 모델의 경우에도 RNN이나 LSTM 레이어 등을 기반으로 상대적으로 작은 규모로 구성된다. 일반적으로 LLM은 위에서 설명한 트랜스포머 기반 인코더-디코더 구조 등 전통적인 모델과 다른 종류의 레이어로 구성된다. 특히, 모델의 규모는 수십억 개의 파라미터에 이르고, 단일 모델을 하나의 GPU로 로드할 수 없을 정도의 크기를 나타낸다.

Ⅲ. 학습 시간 예측 연구 분석

기존의 학습 시간 예측 연구들은 전통적인 딥러닝 모델(CNN, RNN, MLP)을 대상으로 여러 예측 방법론을 제시했다. 이러한 연구들은 주로 모델의 구조, 학습 환경 등 다양한 요소를 고려하여 학습 시간을 예측하려는 시도를 했다. 표 1의 관련 연구들을 중심으로 각 연구의 주요 특징을 분석해보면 다음과 같다.

먼저, [3]에서는 그래프 신경망(graph neural network)을 활용한 예측 모델을 사용하여 학습 시간을 예측했다. 이 연구는 분산 학습 전략 중 데이터 병렬화에 대해서만 학습 시간을 예측하고, 이미지 분류 및 자연어 처리 모델을 대상으로 한다. 그러나 이기종 GPU 구성에서의 학습 시간 예측이 다뤄지지 않았다는 점에서 다소 제한적이다.

둘째, [4] 연구는 비선형 회귀(non-linear regression) 기법을 사용하여 MLP, CNN, RNN 등의 학습 시간을 예측했다. 이 연구는 이기종 GPU 환경을 고려하여 다양한 하드웨어 성능 차이를 반영함으로써 실제 학습 환경에서의 현실성을 높였다. 그러나 분산 학습은 고려하고 있지 않아, 단일

표1. 학습시간 예측 관련연구 비교

	방법론	이기종 GPU 고려	대상 분산 학습 전략	대상 워크로드
[3]	그래프 신경망	X	데이터 병렬화	이미지 분류 및 자연어 처리 모델
[4]	비선형 회귀	O	X	MLP, CNN, RNN, fully connected
[5]	Profiling + 수식 기반	X	X	ResNet, VGG, MobileNet
[6]	수식 기반	X	3D 병렬화	DNN

GPU를 사용하는 경우에만 예측이 가능하다는 점에서 한계가 있다.

셋째, [5] 연구는 실행하는 모델 학습에 대한 자원 소모량 프로파일링과 그 결과값을 기반으로 수식 기반 모델링을 수행한다. 해당 연구는 ResNet, VGG, MobileNet과 같은 전통적인 CNN 모델들의 학습 시간을 예측하며, CNN 모델들의 학습 패턴을 분석하는 데 중점을 둔다. 그러나 다양한 분산 학습 방식이나 이기종 GPU 환경을 고려하지 않고, 특정한 CNN 모델의 구조에만 적용되는 모델링을 수행하기 때문에, 다른 모델에는 예측의 정확도가 낮다.

마지막으로, [6]은 수식 기반 모델을 사용해 DNN 모델의 학습 시간을 예측하였으며, 특히 데이터 병렬화, 파이프라인 병렬화, 텐서 병렬화를 동시 구동하는 3D 병렬화를 적용하여 분산 학습 환경에서의 예측을 진행한다. 하지만, 이기종 GPU 환경에서의 차이를 반영하지 않았으며, 다양한 하드웨어가 혼재된 환경에서의 성능 차이는 충분히 반영되지 않았다.

Ⅳ. LLM 학습 시간 예측 연구 방향

상기 살펴본 기존 연구의 문제점을 극복하기 위해, LLM을 대상으로 한 학습 시간 연구는 아래 지점에 대한 고려와 극복이 필요하다.

트랜스포머의 고유한 특성 반영. LLM의 핵심인 self-attention 메커니즘은 입력 시퀀스 길이가 길어질수록 연산 복잡도가 급격히 증가하며, 자원 소모 역시 크게 늘어난다. 기존 연구는 이러한 특성을 반영하지 못해 학습 시간 예측의 정확도가 떨어진다. 이를 해결하기 위해서는 self-attention 연산의 자원 요구를 세밀히 분석하고, 이를 반영해 자원 소비를 정확히 예측할 수 있는 모델이 필요하다. 이러한 방식은 기존 CNN 등의 전통적인 레이어에 대한 분석에 국한된 모델의 예측 범주를 LLM까지 확장하고, 정확도를 개선할 것으로 기대된다.

분산 학습 시 병렬화 기법에 대한 반영. LLM 학습은 대규모 자원을 필요로 하며, 다수의 GPU 클러스터를 사용하는 분산 학습이 필수적이다. 그러나 3D 병렬화 등 다양한 병렬화 기법을 사용하는 과정에서 통신 오버헤드와 동기화 문제는 학습 시간에 큰 영향을 미친다. 특히, 모델을 학습할 때 활용하는 병렬화 기법의 종류에 따라 1) 자원을 활발히 사용한 시간과 그 양이 상이하며, 2) 수 많은 GPU 간 통신을 학습과정 중 언제 수행하는 지, 얼마나 많은 데이터의 교환을 요구하는 지가 매우 상이하다. 이는 동일한 LLM이라고 하더라도, 분산학습 방식에 따라 자원 소모 및 학습 시간 측면에서 큰 차이를 지님을 의미하며, 예측을 수행하는 것은 더욱 복잡해진다. 그러나 현재 존재하는 연구들은 이러한 차이들을 고려하지 못한다. 따라서, 향후에는 병렬화 기법별 자원 사용 방식에 대한 구체적인 이해를 통해, 정확도의 개선이 필요하다.

이기종 GPU 환경 대응. LLM 학습은 일반적으로 성능이 다른 다양한 제조사, 세대의 GPU가 혼용된 환경에서 이루어진다. 각 GPU에서의 LLM 학습 시간은 매우 상이하며, 더욱이 서로 다른 종류의 GPU가 혼용되어 사용되는 상황은 역시 그 학습 시간이 다르다. 이러한 관점에서, 수 많은 GPU들이 활용될 때 각 GPU의 연산 처리 속도, GPU 내 SM 코어의 수, 메모리의 크기 등을 고려하여 GPU가 혼용된 상황에서 학습 시간을 예측하는 기술이 새롭게 연구될 필요가 있다.

Ⅴ. 결론

본 연구는 현재까지 제안된 딥러닝 모델의 학습 시간을 예측하는 연구들을 다양한 각도에서 비교 분석하고, 최근 다양한 도메인에서 활용되는 LLM을 대상으로 한 학습 시간 예측 기술의 연구 방향을 분석했다. 기존 기술들이 대부분 이미지 분류 및 자연어 처리 등 전통적인 모델에 국한되어 있고, LLM이 요구하는 수준의 대규모 GPU 병렬화 및 이기종 GPU 활용을 전혀 반영하지 못하고 있음을 고려할 때, LLM을 위한 특화된 학습 시간 예측 기술이 필요하다. 향후 본 연구진은 다양한 규모의 LLM을 기반으로 학습 시간 예측 기술을 개발하고, 예측을 통한 자원 활용률 개선 등을 연구할 예정이다.

ACKNOWLEDGMENT

이 논문은 KT의 지원을 받아 수행된 고려대학교-KT 산학공동연구개발 과제의 연구 결과임

참 고 문 헌

[1] Singhal, Karan, et al. "Large language models encode clinical knowledge." Nature 620.7972 (2023): 172-180.

[2] Cui, Can, et al. "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles." Winter Conference on Applications of Computer Vision. 2024.

[3] Yang, Gyeongsik, et al. "Prediction of the resource consumption of distributed deep learning systems." Proceedings of the ACM on Measurement and Analysis of Computing Systems 6.2 (2022): 1-25.

[4] Justus, Daniel, et al. "Predicting the computational cost of deep learning models." 2018 IEEE international conference on big data.

[5] Xie, Zhen, et al. "Centimani: Enabling Fast AI Accelerator Selection for DNN Training with a Novel Performance Predictor." 2024 USENIX ATC. 2024.

[6] Qi, Hang, Evan R. Sparks, and Ameet Talwalkar. "Paleo: A performance model for deep neural networks." International Conference on Learning Representations. 2017.

[7] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[8] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.