딥러닝 기반 추천 모델의 GPU-가속 학습 성능 측정

전민기[1], 김순재[2], 이원준[2]
[1]고려대학교 사이버국방학과
[2]고려대학교 정보보호대학원
jmg08020@korea.ac.kr, sjkim94@korea.ac.kr, wlee@korea.ac.kr

Performance Measurement of GPU-accelerated Training of a Deep Learning-based Recommendation Model

Mingi Jeon[1], Sunjae Kim[2], Wonjun Lee[2]
[1]Dept. of Cyber Defense, Korea University
[2]School of Cybersecurity, Korea University

요 약

본 연구는 딥러닝 기반 추천 모델 중 하나인 DLRM(Deep Learning Recommendation Model)의 학습 시간을 CPU 단일 환경과 GPU-가속 환경에서 비교한다. GPU를 사용하는 경우 일반적으로 더 빠른 학습이 기대되나, 배치 크기가 작아 GPU의 병렬 연산을 효율적으로 활용하지 못하는 경우 CPU만 사용하는 학습이 오히려 빠를 수도 있음을 실험을 통해 확인하였다. 학습 시간 외에 배치 크기는 자원 활용률에도 영향을 미치며, 이는 딥러닝 기반 모델의 학습과 추론에 도입 환경과 워크로드를 고려하여 실행 하드웨어를 선택할 필요가 있음을 시사한다.

1. 서론

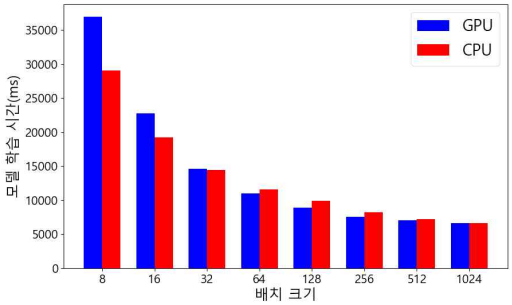
추천 시스템(recommender system)은 사용자에게 개인화된 콘텐츠를 제공하며, 검색, 이커머스, 소셜 네트워크, 비디오 스트리밍을 비롯한 대규모 웹 서비스에서 널리 활용되고 있다. 딥러닝 기반의 추천 모델은 전통적인 규칙 기반 알고리즘보다 사용자의 행동을 정확하게 예측할 수 있어 최근 각광받고 있다[1]. 이러한 추천 시스템은 사용자와 추천 아이템의 업데이트를 반영할 수 있어야 하는데, 실행에 많은 자원이 필요한 딥러닝 기반 추천 모델의 경우 빠르고 비용 효율적인 학습이 요구된다.

딥러닝 모델의 빠른 학습을 위해 병렬 연산에 특화된 GPU를 활용하는 경우가 잦다. 배치 크기는 한 번에 처리하는 데이터의 양을 뜻하며, CPU만 사용하는지, GPU도 함께 사용하는지 여부에 따라 적절한 배치 크기가 달라질 수 있다. GPU를 사용하는 경우 배치 크기가 클수록 유리하고, 데이터를 GPU까지 이동시키는 비용이 존재하므로 배치 크기를 작게 설정하는 경우 오히려 학습이 느려질 수 있다.

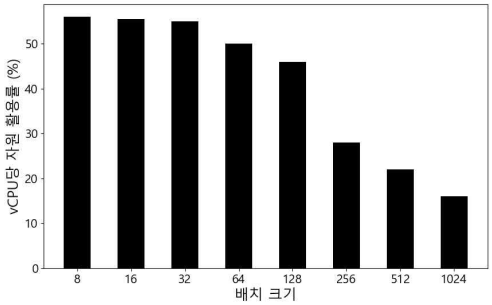
본 연구는 이를 실험을 통해 확인하기 위해 Facebook Research에서 제안, 공개한 DLRM(Deep Learning Recommendation Model)[2]을 실제 GPU 테스트베드에서 학습하고 그 성능을 측정하였다. DLRM은 연속형 데이터를 위한 MLP 단계, 범주형 데이터를 위한 임베딩 조회(lookup), 두 특성(feature)의 상호작용(interaction)과 신경망(neural networks)을 포함하는 모델로, 추천 시스템에 요구되는 사용자와 아이템 간의 관계를 학습하는 딥러닝 기반 추천 모델이다. 다른 딥러닝 기반 추천 모델(DLRM-RMC2, DLRM-RMC3 등) 대비 임베딩 테이블 연산-집약적인 것으로 알려져 있다[2].

2. 실험 환경 및 구성

실험 테스트베드 CPU는 Intel Core i7-8700K, 16 GB RAM, GPU는 NVIDIA GeForce RTX 2080 Ti로 구성되어 있으며, Linux v6.5 커널을 사용하였다. DLRM 모델[2]은 PyTorch 라이브러리를 사용하여 구현한 공개 모델을 사용했다. 배치 크기를 8, 16, 32, 64, 128, 256, 512, 1024로 설정하고, 자원 당 배치 크기 별로 10번씩 실험하여 모델 학습 시간의 평균을 기록하였다. CPU만 사용한 학습과 GPU도 활용한 학습을 달리한 것 모두 동일한 환경과 데이터셋을 사용하였다. 학습 시 사용되는 데이터의 크기는 10,000으로 설정하였다. 또한, 추가로 CPU에서 모델의 학습 시간을 측정할 때, 배치 크기 별로 테스트베드의 자원 사용률도 함께 측정하여 배치 크기에 따른 연산 자원의 활용률을 분석하였다.



(그림 1) CPU와 GPU에서 배치 크기 별 학습 시간.



(그림 2) 배치 크기 별 CPU 자원 활용률.

3. 실험 결과 분석 및 논의

그림 1은 CPU만 사용한 경우와 GPU를 사용한 경우 각각의 배치 크기 별 DLRM의 학습 시간을 나타낸다. 배치 크기가 가장 작은 8에서는 CPU와 GPU에서 학습 시간이 각각 29.1, 36.9s였으며, 배치 크기가 가장 큰 1024에서는 CPU와 GPU가 6.6s로 동일했다. 8부터 512까지의 배치 크기에서 CPU와 GPU의 학습 시간 차이는 각각 8000, 3400, 120, 500, 1000, 700, 그리고 150ms 이상으로 기록되었다. 배치 크기 1024에서는 CPU와 GPU가 유사한 학습 시간을 보였다.

딥러닝 기반 추천 모델의 경우 CPU와 GPU의 메모리 접근 패턴과 처리 방식에 따라 학습 속도가 달라진다[3]. 작은 배치 크기인 8, 16, 그리고 32에서 CPU 학습이 더 빠른 것은 GPU의 병렬 연산의 이점보다 GPU로의 데이터 이동 오버헤드가 더 크기 때문이다. 32보다 큰 크기(64, 128, 256, 그리고 512)의 배치 크기에서는 병렬 연산에 장점이 있는 GPU의 학습 시간이 짧다. 배치 크기 1024에서 학습 시간이 유사한 것은 임베딩 테이블-집약적인 DLRM 모델에서 GPU의 병렬 실행 이점 대비 데이터 이동 오버헤드가 다시 높아졌기 때문으로 보인다.

그림 2는 CPU에서 DLRM 학습 시 배치 크기 별 CPU 자원 활용률을 나타낸다. 배치 크기가 작을수록 CPU 자원 활용률이 높은데, 배치 크기 8에서의 CPU 활용률은 vCPU당 56% 이상이었으며, 배치 크기 1024에서는 vCPU당 16%로 감소했다. CPU 자원 활용률 면에서도 배치 크기가 가장 큰 1024에서 가장 적다. 배치 크기와 무관하게 학습 시 사용하는 메모리는 전체 시스템 메모리 중 2.45%였다.

4. 결론 및 향후 연구

본 연구는 여러 배치 크기에 따른 DLRM 모델의 학습 성능을 CPU와 GPU에서 각각 측정하였다. 실험 결과, 배치 크기가 작을 때는 데이터 이동 오버헤드로 인해 GPU 가속을 이용하지 않고 CPU만을 사용한 학습이 더 빨랐으며, 중간 크기의 배치(64~512)에서 GPU 활용의 이점이 확인되었다. 그러나 배치 크기가 작은 경우 CPU 활용률이 높아지므로, 도입 환경과 모델에 따른 적응적으로 하드웨어를 선택할 필요가 있다. 예를 들어, 사용한 인스턴스의 종류와 활용률에 따라 요금이 청구되는 클라우드 서비스의 경우, 배치 크기에 따른 GPU가 탑재된 고성능 인스턴스와 CPU-전용 인스턴스 활용률의 모델을 만들어 비용-효율적으로 클라우드 기반의 모델 학습에 활용할 수 있다.

ACKNOWLEDGMENTS

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. RS-2023-00234719, (SW스타랩) 서비스 연속형 지향 에지 Continuum SW 프레임워크)과 한국연구재단의 지원(No. RS-2024-00338786)을 받아 수행된 연구임.

참고문헌

[1] Udit Gupta et al, “DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference,” in *Proc. of ISCA*, May 2020.

[2] Maxim Naumov et al, “Deep Learning Recommendation Model for Personalization and Recommendation Systems,“ arXiv preprint arXiv:1906.00091, May 2019.

[3] Rishabh Jain et al, “Optimizing CPU Performance for Recommendation Systems At-Scale,” in *Proc. of ISCA*, June 2023.