

동기식 분산 딥러닝 환경에서 배치 사이즈 변화에 따른 모델 학습 성능 분석

김예랑, 김형준, 유현창
고려대학교 정보대학 컴퓨터학과
{hs01151116, ledzep0830, yuhc}@korea.ac.kr

A Performance Analysis of Model Training Due to Different Batch Sizes in Synchronous Distributed Deep Learning Environments

Yerang Kim, HyungJun Kim, Heonchang Yu
Dept. of Computer Science and Engineering, Korea University

요약

동기식 분산 딥러닝 기법은 그래디언트 계산 작업을 다수의 워커가 나누어 병렬 처리함으로써 모델 학습 과정을 효율적으로 단축시킨다. 배치 사이즈는 이터레이션 단위로 처리하는 데이터 개수를 의미하며, 학습 속도 및 학습 모델의 품질에 영향을 미치는 중요한 요소이다. 멀티 GPU 환경에서 작동하는 분산 학습의 경우, 가용 GPU 메모리 용량이 커짐에 따라 선택 가능한 배치 사이즈의 상한이 증가한다. 하지만 배치 사이즈가 학습 속도 및 학습 모델 품질에 미치는 영향은 GPU 활용률, 총 에포크 수, 모델 파라미터 개수 등 다양한 변수에 영향을 받으므로 최적값을 찾기 쉽지 않다. 본 연구는 동기식 분산 딥러닝 환경에서 실험을 통해 최적의 배치 사이즈 선택에 영향을 미치는 주요 요인을 분석한다.

1. 서론

AI 분야에서는 거대 모델 및 대용량 데이터셋을 통한 성능 향상 시도가 이어짐에 따라 학습 시간 단축이 중요한 과제로 떠올랐다. 이에 따라 멀티 GPU 환경에서의 동기식 분산 딥러닝이 표준적 방법으로 사용되고 있다. 멀티 GPU 환경에서 분산 학습을 적용하는 경우 배치 사이즈 확장이 용이하며, 배치 사이즈 선택에 따라 학습 속도 및 학습 모델 품질이 달라진다. 따라서, 본 연구는 배치 사이즈 변화가 동기식 분산 딥러닝을 통한 모델 학습 시간 및 정확도에 미치는 영향을 분석하고 실험을 통해 검증한다.

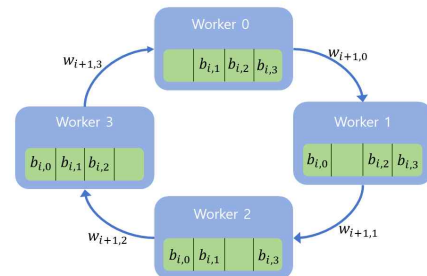
2. 동기식 분산 딥러닝 환경에서의 학습

분산 학습 방식으로 여러 노드에서 학습을 진행하면 워커 노드의 수가 증가함에 따라 단일 GPU의 메모리 용량에 국한되지 않고 글로벌 배치 사이즈를 증가시킬 수 있다. 하지만 배치의 크기를 증가시킬 때의 정확도와 손실, 그리고 속도 간의 trade-off가 존재하기에 분산 학습 환경에서의 최적의 배치 사이즈 탐색은 중요한 과제이다.

2.1 Horovod

본 연구의 실험에서는 동기식 분산 딥러닝 프레임워크인 Horovod를 사용하였다. Horovod는 연산을 수행하는 모든 노드의 그래디언트 연산이 끝난 후 평균이 종합되는 동기식 처리 방식이며, 노드간 상호 그래디언트 취합 및 반영이 이루어지는 Ring-AllReduce 방식이다.

Ring-AllReduce 방식은 Horovod에서 사용하는 동기식 분산 학습 기법으로, 효율적인 대역폭 활용을 장점으로 갖는 링 토폴로지 상에서의 AllReduce 집합 통신 기법이다. 자세한 방식은 그림 1과 같다.



(그림 1) Ring-AllReduce 방식 구조도

3. 실험

3.1 실험 배경 및 실험 환경

다중 워커로 분산 학습을 수행하는 환경에서 미니 배치 크기의 최적성에 영향을 미치는 요인을 워커 개수, 실행 시간, 모델 정확도로 나누어 이 요인들 간의 상호 관계를 분석하는 것을 실험의 목적으로 하였다. 또한 모든 실험은 CIFAR-10 데이터셋을 통해 ResNet-18 모델을 학습하는 방식으로 수행하였다.

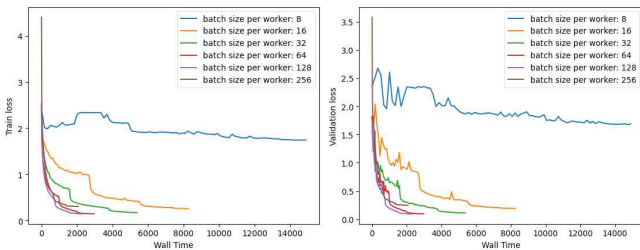
<표 1> 실험 환경 정보

| | |
|---------------------------|---------------------------------|
| VM Instance Type | GN7.XLARGE320 |
| Number of vCPU | 80 |
| Memory | 320GB |
| GPU | 4*NVIDIA T4 |
| Processor | Inter Xeon Cascade Lake(2.5GHz) |
| Private Network Bandwidth | 25Gbps |
| Packets In/Out | 5600k PPS |

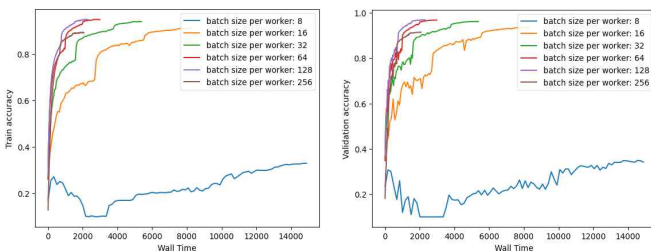
표 1과 같은 환경에서 배치 크기를 8, 16, 32, 64, 128, 256 총 6단계로 나누어 실험을 진행하였다.

3.2 실험 결과 및 분석

(그림 2) 배치 크기 별 loss 추이



(그림 3) 배치 크기 별 accuracy 추이



학습 소요 시간은 8부터 256까지 배치 크기가 커질수록 14,892초에서 8,282초, 5,379초, 2,977초, 2,300초, 2,069초로 감소했다. 배치 크기 16은 8에 비해 학습 시간이 약 55.6% 소요되어, 가장 큰 폭으로 감소했다. 학습 정확도는 배치 크기 8에서 0.3282, 16에서 0.9106으로 약 2.8배 증가하였다. 배치 크기 32, 64, 128, 256에서는 각각 정확도 0.9381, 0.9464, 0.9465, 그리고 0.8907을 기록하였다. 데이터에 대한 모델의 오류를 나타내는 지표인 손실률은 배치 크기 8에서부터 256까지 차례대로

1.7422, 0.2614, 0.1765, 0.1490, 0.1497, 0.3073 0.3073을 기록하였다(그림 2).

배치 크기 16, 32, 64, 128, 256에서 각 단계별로 1.03배, 1.01배, 1.0001배로 정확도의 증가 폭이 점차 줄었으며(그림 3), 배치 크기 128과 256은 증가율이 1 미만(0.94), 즉 감소하였다. 정확도는 이터레이션 횟수와 수렴률의 영향을 받아 결정되는데, 배치 크기 256에서의 정확도 감소는 이터레이션 횟수는 감소하였으나 수렴률 증가가 그 폭에 미치지 못하여 총 학습 정확도가 감소하였기 때문이다. 배치 크기 128에서 256으로 증가 시 학습 손실률이 다시 증가하게 된 것도 같은 원인 때문이다.

4. 결론 및 향후 과제

동기식 분산 딥러닝 환경에서 배치 크기 별 GPU 이용률과 이터레이션 수 변화로 인해 학습 성능을 결정짓는 정확도와 손실률, 그리고 학습 시간이 달라짐을 밝혔다. 본 연구의 실험에서 배치 크기별로 성능 증가 폭과 학습 시간의 감소폭에 차이가 있던 점을 토대로 볼 때, 동기식 분산 딥러닝 환경에서 학습 진행 상황별로 최적의 배치 크기는 고정적이지 않을 것이라는 결론을 내릴 수 있다. 따라서 학습 단계마다 동적인 배치 크기 설정이 가능하도록 하여, 각 단계에서 최적의 GPU 사용과 학습 시간 단축 달성이 본 연구를 비롯한 동기식 분산 딥러닝 분야에서의 향후 과제가 될 것이다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (IITP-2023-RS-2022-00156439)

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2018-0-01405)

참고문헌

- [1] Rakshith, R. M., et al. "Performance analysis of distributed deep learning using horovod for image classification." 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.
- [2] Kunde, Shruti, Amey Pandit, and Rekha Singhal. "Benchmarking performance of RaySGD and Horovod for big data applications." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020.