ScienceDMZ 기반 고성능 네트워킹 환경에서 분산 AI 추론 성능 비교 연구

김동학[1], 문정훈[2]
한국과학기술정보연구원

{dhkim,jhmoon}@kisti.re.kr

A Comparative Study of Distributed AI Inference Performance in ScienceDMZ-based High Performance Networking Environments

Dong Hak Kim, Jeong Hoon Moon
Korea Institute of Science and Technology Information

요 약

본 논문은 과학기술분야에서 관측, 실험장비의 비약적인 발전으로 과학빅데이터가 폭발적으로 증가하고 있으며 IT 를 비롯한 인공지능 기술의 발전으로 과학빅데이터의 분석 기법이 고도화 되고 있는 가운데 과학빅데이터에 대해 인공지능 기반의 GPU 와 FPGA 의 분석 및 추론 성능에 대한 비교 연구이다. 특히 과학빅데이터의 고속전송, 인공지능 기반의 분석, 공유 기술은 4 차산업 시대의 핵심 기술이며, 이를 위한 다양한 IT 인프라와 기술체계가 구현되고 있는데, 본 논문에서는 이러한 여러 기술 체계 중 빅데이터 고속도로 체계인 ScienceDMZ 기반의 고속전송 전용 노드인 DTN(Data Transfer Node) 상에서의 인공지능 학습과 추론 성능 비교연구로서 GPU 기반 DTN 과 FPGA 기반 DTN 시스템의 성능을 비교 분석 하였다. 또한, 분석한 결과를 토대로 FPGA 가속기 카드가 적절히 활용될 수 있는 분야를 제시함과 동시에 ScienceDMZ 기반의 고성능 자원화 환경을 구축하는 방향성을 제시한다. 향후 ScienceDMZ 기반의 연구플랫폼 간 대규모 분산 환경에서의 인공지능 학습 및 추론을 위한 컴퓨팅 환경을 제안하였다.

Ⅰ. 서 론

최근들어 과학기술분야에서 실험/관측 장비의 비약적인 발전과 연구개발 기술의 고도화를 통하여 과학빅데이터의 생산이 폭발적으로 증가하고 있다. 특히 IT 기술의 발전으로 이러한 과학빅데이터의 고속전송을 위한 빅데이터 전달 체계, 분석을 위한 인공지능 기반의 컴퓨팅 체계, 적용을 위한 공유 체계 등이 빠르게 발전하고 있다. 이러한 연구 패러다임의 변화는 빅데이터 기반의 4 차 산업으로 확대되어 특히 인공지능 기반의 빅데이터에 대한 활용 수준이 향후 국가 경쟁력으로 자리잡고 있는 실정이다.

본 논문에서는 과학기술분야 빅데이터의 고속전송을 위한 빅데이터 고속도로 체계와 분석을 위한 인공지능 기반의 컴퓨팅 체계를 연계한 ScienceDMZ 기반의 연구플랫폼 환경에서 빅데이터 기반의 인공지능 추론 성능을 비교 분석한다.

ScienceDMZ 는 빅데이터 고속도로 체계로서 빅데이터를 위한 고속의 전용망을 보안적으로 상호 신뢰 구간의 협의가 된 기관들과 대학들 사이에 구성하고, 데이터 전송 전용 노드인 DTN(Data Transfer Node)을 통하여 빅데이터를 고속으로 전송하는 기술 체계이다. 빅데이터의 고속전송을 위해 상호 신뢰 구간에 대한 전용망의 구성은 일반 트래픽과 연구 트래픽을 분리하여 패킷로스의 발생을 최소화한다. 트래픽 분리를 통한 전용망의 구성은 빅데이터의 전송에 있어서 주어진 대역폭 대비 90% 이상의 전송(Throughput) 성능을 나타낸다[1]

분산되어 있는 각각의 ScienceDMZ 환경에 대하여 DTN 은 데이터 전송 가속기로서의 역할 뿐만 아니라 함께 구축된 CPU, GPU. FPGA 등을 컴퓨팅 자원으로 활용이 가능하며, 특히 빅데이터 고속도로 체계인 ScienceDMZ 환경에서의 컴퓨팅 자원의 역할은 빠르게 전송 받은 빅데이터에 대한 빠른 분석과 결과의 빠른 공유가 가능하므로 향후 빅데이터 시대에 매우 유용한 IT 인프라 체계로서 활용이 가능하다.

본론 1 절에서는 논문에서는 이와같은 과학 빅데이터를 초고속으로 전송하는 연구플랫폼을 구축하기 위해 진행했던 실험을 상세히 설명한다. 2 절에서는 진행했던 실험의 구성과 함께 분석한 실험 결과를 상세히 설명한다. 실험에서 사용된 하드웨어 군과 함께 실험 목표, 실험 방법 등을 상세히 설명하며 이를 통해 향후 고성능 학습 및 추론을 위한 컴퓨팅 환경을 제안하기 위해 ScienceDMZ 기반 최적의 연구플랫폼을 구축하는 방향성을 제시하고자 한다.

마지막으로 3 절에서는 본 연구 실험을 통해 분석한 결과와 함께 FPGA 가속기 카드가 적절히 활용될 수 있는 분야 등을 논의하며 향후 연구방향을 제시한다.

Ⅱ. 본론

2-1. 실험 구성

본 논문에서는 'CIFAR-10' 데이터 셋과 Kaggle 에 공개되어 있는 'CIFAR Images Classification using CNN' 라는 소스코드를

[1] 주저자
[2] 교신저자

활용하여 실험을 진행하였다. 기계학습과 컴퓨터 비전 알고리즘에 이미지 분류 실험[2]으로 자주 활용되는 CIFAR-10 데이터 셋은 32 * 32 크기의 모델 학습용 이미지 5 만장과 모델 평가용 이미지 1 만장, 도합 6 만 장의 이미지로 구성되어 있으며 10 개의 클래스로 각각 분류되어 있다. 또한 본 실험에 사용된 Kaggle 코드는 학습되지 않은 컨볼루젼 신경망 모델을 사용하여 학습용 이미지 데이터 셋을 통해 학습을 진행하고 학습 과정이 끝나면 평가용 이미지 데이터 셋을 통해 모델 평가를 진행하는 매커니즘의 코드이다.

본 논문에서 진행한 실험은 DTN (Data Transfer Node) 장비 기반에서 FPGA 가속기 카드[3]와 GPU 간 머신러닝 학습 연산 성능을 비교하고 분산되어 있는 GPU 기반의 DTN 과 FPGA[4] 기반의 DTN 에 같은 어플리케이션을 동작시켰을 때 머신러닝 연산 성능을 비교하고자 한다.

< 표 1 > 실험 결과 도표

	Epoch 당 평균 소요시간 (sec)	샘플 당 평균 소요시간 (us/sample)	모델 학습 최대 정확도 (acc)	모델 학습 최소 실패율 (loss)	평가 샘플 평균 정확도 (val_acc)	평가 샘플 평균 실패율 (val_loss)
CPU Intel Xeon Gold 6226R	42	840	0.9771	0.0751	0.7255	1.8957
GPU NVIDIA RTX 2080 Ti	7	140	0.9750	0.0781	0.7208	1.8430
FPGA Xilinx Alveo U200	4	82	0.9590	0.1136	0.7185	0.8820

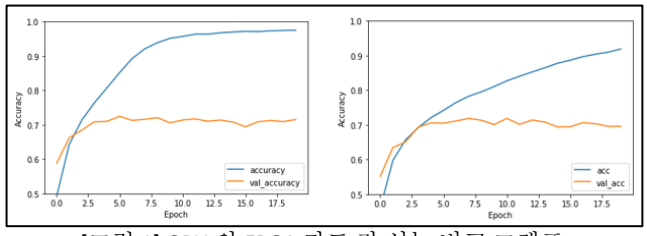
이와 같은 연구 목표를 달성하기 위해 고성능 데이터 처리 관점에서 CPU, GPU, FPGA 하드웨어 장비 스펙트럼 간 데이터 처리 성능을 중심으로 실험을 구성하였고, 향후 ScienceDMZ 기반의 고성능 네트워킹 연구 플랫폼을 구현하기 위해 인공지능 머신러닝 연산 성능을 바탕으로 실험을 진행하였다. 따라서 본 실험에서는 CIFAR-10 데이터 셋을 기반으로 이미지 분류 실험을 진행하고 그에 따른 각 Epoch 당 모델의 분류 소요시간, 이미지 샘플 당 분류 속도, 모델 이미지 분류 정확성 등을 중심으로 결과치를 산출하여 분석하였다.

2-2. 실험 및 분석 결과

본 논문에서 진행한 실험은 Cent OS 7 운영체제 기반의 Dell PowerEdge 740 모델인 DTN (Data Transfer Node) 장비 2 대를 사용하였으며 IPMI 를 사용하여 원격으로 장비를 제어하며 실험을 수행하였다. 또한 DTN 장비에 기계학습 연산을 수행하기 위해 Tensorflow, Caffe Framework 환경을 구축하고 활용한 Kaggle 코드를 FPGA 가속기 카드 연산에 활용할 수 있도록 일부 수정하여 실험을 진행하였다. AI 이미지 분류 연산에 사용된 신경망 모델은 Microsoft 사에서 발표한 resnet50[5], Google 의 googlenet 모델을 각각 사용하였다.

또한 하드웨어 비교군인 CPU 는 Intel Xeon Gold 6226R, GPU 는 NVIDIA Geforce RTX 2080 Ti, FPGA 는 Xilinx 사의 Alveo U200 가속기 카드를 활용하였다. <표 1>은 본 연구의 실험 결과 내용을 도표로 정리하여 나타낸 자료이다.

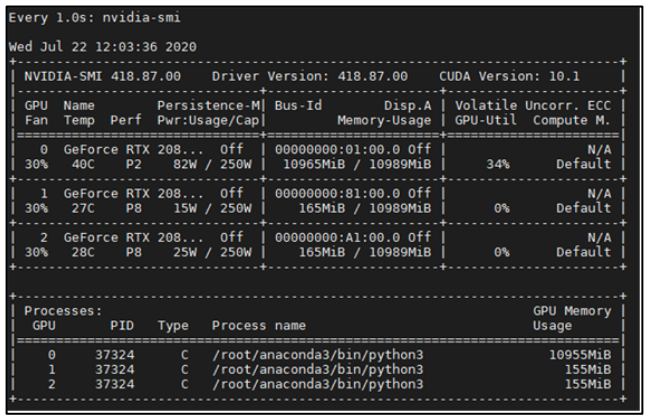
<표 1>과 같이 AI 연산 예제를 사용하여 CPU 만으로 연산을 진행했을 때 학습용 데이터 셋인 5 만 장을 학습하는데 걸린 시간은 평균 42 초가 소요되었으며, 평가용 데이터 셋 1 만 장을 기반으로 모델 평가를 진행했을 땐 평균 0.72 정확도 수준으로 이미지를 클래스별로 분류할 수 있었다.



[그림 1] GPU 와 FPGA 카드 간 성능 비교 그래프

이와 동일한 방법으로 GPU 만을 활용하여 실험을 진행했을 땐 CPU 에 비해 각 Epoch 당 평균 35 초 감소된 모델 학습 소요시간을 확인할 수 있었으며 각 이미지 샘플 당 분류 소요시간 또한 600us/sample 의 속도로 크게 줄어든 것을 확인할 수 있었다. 정확도와 관련된 다른 지표 부분에선 큰 차이를 보이진 못했다.

FPGA 가속기 카드를 장착한 DTN 에서 실험을 진행[6]했을 때엔 각 Epoch 당 평균 4 초로 모델 학습 소요시간이 감소되는 것을 확인할 수 있었으며 각 이미지 샘플 당 분류되는 소요시간 또한 82us/sample, GPU 에 비해 속도가 약 7 배 정도 감소되는 것을 확인할 수 있었다. 정확도 관련한 지표는 CPU 와 GPU 실험 케이스들에 비해 Epoch 초반엔



[그림 2] AI 연산 병렬 GPU 활용 화면

다소 성능이 떨어지는 것을 확인할 수 있었으나 이는 10 번 째 Epoch 부터 정확도가 높아지며 이미지를 분류하는 것을 확인할 수 있다. [그림 1]은 본 실험에서 AI 머신러닝 학습 연산에 GPU 를 사용했을 때의 분류 정확도와 FPGA 가속기 카드를 사용했을 때의 분류 정확도를 그래프로 나타낸 그림이다.

이와 더불어, [그림 2]와 같이 DTN 장비에 3 개의 GPU 를 장착하고 AI 머신러닝 연산을 수행했을 때 재차 실험[7]을 진행하였다.

< 표 2 > 실험 결과 도표

	Epoch 당 평균 소요시간 (sec)	샘플 당 평균 소요시간 (us/sample)	모델 학습 최대 정확도 (acc)	모델 학습 최소 실패율 (loss)	평가 샘플 평균 정확도 (val_acc)	평가 샘플 평균 실패율 (val_loss)
GPU 1 EA NVIDIA RTX 2080 Ti	7	140	0.9853	0.0731	0.7107	1.6423
GPU 3 EA NVIDIA RTX 2080 Ti	7	130	0.9352	0.0856	0.7037	1.7052
FPGA Xilinx Alveo U200	4	86	0.9048	0.2430	0.7063	0.8786

실험 구성은 앞서 <표 1> 의 구성과 동일하며 AI 머신러닝 연산에 GPU 1 기, 3 기, FPGA 카드를 각각 활용하며 결과를 분석하였으며 <표 2>는 분석한 자료를 도표로 정리한 그림이다.

동일한 컴퓨팅 작업을 각각 GPU 1 기, GPU 3 기, FPGA 카드 1 기로 수행하였을 때 GPU 1 기와 FPGA 가속기 카드 1 기의 성능은 위 <표 1> 도표 결과와 비슷했다. GPU 3 기로 [그림 2]와 같이 구성한 후 연산을 수행한 결과, 각 Epoch 당 평균 학습 소요시간은 7 초로 동일하였으며 샘플 당 평균 소요시간은 미세하게 감소하는 모습을 확인할 수 있었다. 또한 7 번째 Epoch 부터 컨볼루젼 신경망 모델이 약 0.9 이상으로 정확하게 이미지를 분류하는 것을 확인할 수 있었다. 이를 통해, FPGA 가속기 카드는 AI 모델의 학습 연산에 활용되는 면보다 추론 연산에 활용되는 것이 가속기 카드의 장점을 극대화할 수 있을 것으로 판단된다.

Ⅲ. 결론

본 논문에서는 과학 빅데이터를 초고속으로 전송하기 위한 기술 중 하나인 ScienceDMZ 의 전용 노드 DTN 상에서 인공지능 학습과 추론 성능을 비교하기 위한 실험을 진행하였다. CPU, GPU, FPGA 하드웨어 분류에 따른 인공지능 머신러닝 학습 성능을 위주로 실험을 구성하였으며, 병렬 GPU 구성과 FPGA 가속기 카드 간 성능을 확인할 수 있도록 추가적으로 실험을 구성하여 진행하였다. 이를 바탕으로 GPU 기반 DTN 과 FPGA 기반 DTN 시스템 간의 성능을 비교 분석할 수 있었으며 ScienceDMZ 기반의 연구플랫폼 간 대규모 분산 환경에서의 고성능 학습 및 추론을 위한 컴퓨팅 환경을 제안하였다.

본 논문에서 진행한 실험을 통해 과학 빅데이터 고속전송노드 DTN 상에서 인공지능 학습 및 추론 연산 성능에 효율적인 하드웨어 군을 확인할 수 있었다. FPGA 기반의 DTN 에서는 인공지능 모델 학습 연산에서 GPU 기반 DTN 에 비해 우수한 성능을 보였지만 두드러진 차이를 보이지 못한 것을 확인할 수 있었다. 그러므로 본 논문 실험 결과를 바탕으로 판단했을 때는 FPGA 가속기 카드의 활용 분야가 딥러닝 연산보다는 추론 연산을 가속화하는데 적합할 것으로 판단되며, 향후 FPGA 가속기 카드의 장점이 잘 부각되는 컴퓨팅 작업들에 대해 파악할 예정이다. 더불어, 본 연구를 통해 향후 추론에 적합한 Kubeflow 를 활용하여 ScienceDMZ 에 접목시켜 고성능 자원화 환경을 구축할 계획이다.

참 고 문 헌

[1] Dart, Eli et al, “The Science DMZ: A Network Design Pattern for Data-intensive Science”, The International Conference for High Performance Computing, Networking Storage and Analysis (SC13), pp. 173-185, 2014.

[2] Fei Wang, Mengqing Jiang et al, Xiaoou Tang, “Residual Attention Network for Image Classification”, Proceedings for the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 3156-3164, 2017.

[3] Kaiyuan Guo, Shulin et al, “A Survey of FPGA-Based Neural Network Accelerator”, ACM Transactions on Reconfigurable Technology and Systems, Vol. 9, No. 4, Article 11, pp 1-26, 2017.

[4] Y. Ma, Y. Cao, S. Vrudhula and J. Seo, “Performance Modeling for CNN Inference Accelerators on FPGA”, in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 39, No. 4, pp. 843-856, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, arXiv:1512.03385v1, pp 1-12, Dec 2015.

[6] S. Kolala Venkataramanaiah et al., "Automatic Compiler Based FPGA Accelerator for CNN Training," 2019 29th International Conference on Field Programmable Logic and Applications (FPL), pp. 166-172, 2019.

[7] M. Qasaimeh, A. Sagahyroon and T. Shanableh, “FPGA-based Parallel Hardware Architecture for Real-Time Image Classification,” in IEEE Transactions on Computational Imaging, Vol. 1, No. 1, pp. 56-70, March 2015.