GPU 에 내장된 행렬 연산에 특화된 코어를 사용한 인공 신경망의 가속

박상수, 이원혁, 정기석
한양대학교

po092000@hanyang.ac.kr, louisle111@naver.com, kchung@hanyang.ac.kr

Acceleration of artificial neural network using cores specialized in matrix multiplication in GPU

Sang-Soo Park, Won-Hyuk Lee, Ki-Seok Chung
Hanyang University, Seoul, Korea

요 약

인공 신경망은 컴퓨터 비전, 음성인식, 번역 등 다양한 분야에서 적용되어 우수한 성능으로 학계 및 산업계에서 큰 주목을 받고 있다. 향후 인공지능은 자율주행차, 드론, 스마트 팩토리 등 다양한 분야에 적용될 것으로 예상되며, 이를 위한 전용 하드웨어 개발의 연구가 활발하게 진행되고 있다. 이러한 전용 하드웨어 중 GPU 내부에 행렬 연산에 특화된 하드웨어를 포함하는 연구가 주목을 받고 있으며, 제품이 출시되어 사용되고 있다. GPU 내부의 행렬 연산에 특화된 하드웨어를 사용한 인공 신경망의 가속은, 기존 GPU 내부의 연산 장치를 사용한 가속 방법에 비해 적은 전력 소모로 더 빠르게 가속 가능하다. 본 논문에서는 이러한 하드웨어 중 하나인 텐서 코어를 사용하여 인공 신경망을 가속하고, GPU 내부의 연산 장치를 사용한 가속 방법과의 성능, 효율을 비교하였다.

Ⅰ. 서 론

최근 컴퓨터 비전, 음성인식, 변역 시스템 등 산업과 학계 전반에 걸쳐 인공지능의 관심이 급증하고 있다 [1-2]. 인공지능은 생물학의 신경망에서 영감을 받은 알고리즘으로 컨볼루션 신경망 (Convolutional Neural Network, CNN), 재귀 신경망 (Recurrent Neural Network, RNN), 적대적 신경망 (Generative Adversarial Network, GAN)등 다양한 형태의 신경망이 연구되고 있다.

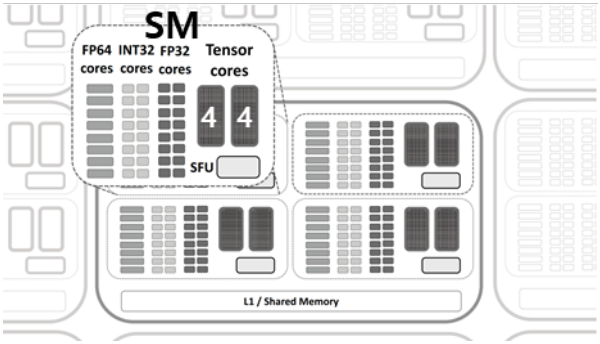
현재 대부분의 연구에서는 Graphic Processing Unit (GPU)를 이용하여 인공 신경망의 학습과 추론을 가속하고 있다 [3]. 인공 신경망의 대표적인 모델 중 하나인 컨볼루션 신경망은 뉴런의 연결이 많은 특성으로 인하여 연산량이 많지만, 다수의 연산 장치를 포함하고 있는 GPU 에서 병렬로 계산되어 빠르게 가속하는 것이 가능하다. 하지만, GPU 를 사용하여 인공지능을 가속하는 과정에서 불필요한 오버헤드가 존재한다 [4]. 이러한 문제를 해결하기 위해 인공지능에서 사용되는 행렬 연산을 적은 오버헤드로 연산 가능한 전용 가속기 연구가 활발하게 진행되고 있다 [4-5].

본 논문에서는 인공 신경망의 연산에 특화된 텐서 코어 (Tensor Core)를 포함하고 있는 NVIDIA 社의 Volta GPU 를 사용하여 실험을 진행하였다. 텐서 코어는 GPU 내부에 내장된 행렬 연산에 특화된 하드웨어로, GPU 에서 사용되는 연산 장치인 쿠다 코어 (CUDA Core)보다 빠르게 인공 신경망을 가속하는 것이 가능하다. 실험에서는 텐서 코어와 쿠다 코어를 사용하여 인공 신경망의 학습에서 소요되는 시간을 측정하고, 전용 하드웨어를 통해 효율적으로 가속할 수 있는 것을 확인하였다.

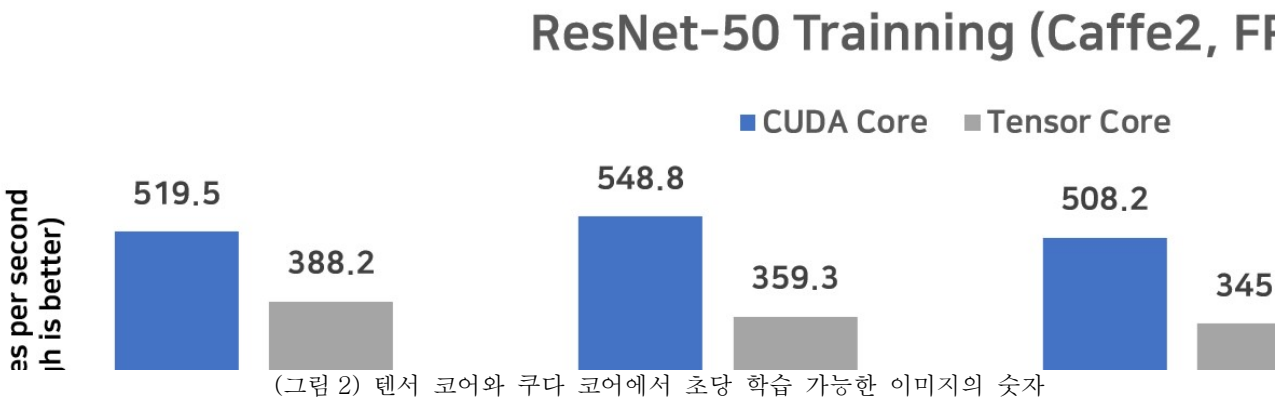
Ⅱ. 본론

2.1 텐서 코어 (Tensor core)

텐서 코어는 행렬의 곱셈과 덧셈 (Matrix-multiply-and-accumulate) 하드웨어로, 텐서 코어 1 개당 한 사이클에 64 번의 부동소수점 기반의 곱셈과 덧셈 (Floating-point multiply and accumulate) 연산을 수행한다. 그림 1 과 같이 GPU 의 Streaming Multipleprocessor (SM) 내부에 8 개의 텐서 코어가 포함되어 있으며, 실험에 사용된 Volta GPU 는 80 개의 SM 이 존재하므로 한 사이클에 4,096 번의 부동소수점 곱셈과 덧셈 연산이 가능하다. 텐서 코어는 인공 신경망에서 자주 사용되는 C=A*B+C 와 같은 형태의



(그림 1) 텐서 코어가 포함된 Volta GPU [4]



(그림 2) 텐서 코어와 쿠다 코어에서 초당 학습 가능한 이미지의 숫자

행렬의 곱셈과 덧셈을 한번에 수행하는 것이 가능하며, 이외에도 mixed precision 의 사용이 가능하다 [5]. mixed precision 은 행렬의 연산에서 입력을 16bit 의 부동소수점 데이터로 하여 16bit 곱셈을 하고, 32bit 의 덧셈기를 사용하여 계산하는 방법으로 feature 나 weight 를 표현하는데 필요한 bit 의 숫자를 줄여 불필요한 메모리 사용량을 줄이는 것이 가능하다.

2.2 실험 결과 및 분석

본 논문에서는 Intel 사의 i9-7900 와 32GB 의 메모리, Titan V GPU 를 사용하여 실험을 진행하였다. 실험에서 사용한 Titan V GPU 는 5,120 개의 쿠다 코어와 640 개의 텐서 코어를 포함하고 있으며, 텐서 코어는 125 Tera floating-point operation per second (TFLOPS), 쿠다 코어는 27.6 TFLOPS 의 연산 성능이 가능하다.

실험에서는 Caffe2 딥러닝 프레임워크 [6]에 16bit half-precision 데이터를 입력으로 하는 mixed precision 을 사용하였으며, ImageNet 을 사용하여 ResNet-50 을 학습하는 과정에서 소요되는 시간을 Batch 단위로 특정하였다. 이때 Batch 의 크기를 조절하여 다양한 크기에서 실험을 진행하였다.

표 1) 쿠다 코어와 텐서 코어의 학습 소요시간

배치 크기	96	64	48	32
텐서 코어 (ms)	1.9	1.82	1.97	2.25
쿠다코어 (ms)	2.58	2.78	2.89	3.31

표 1 은 텐서 코어와 CUDA 코어에서 학습 과정에서 배치 크기에 따른 소요 시간을 나타내며, 이를 사용하여 초당 학습 가능한 이미지의 숫자로 표현한 것은 그림 2 와 같다. 텐서 코어는 쿠다 코어보다 4.53 배 연산 성능이 좋음에도 불구하고, 실제 학습의 성능은 평균 1.45 배 정도 개선됨을 확인하였다. 이는 딥러닝 프레임워크에서 텐서 코어를 사용하기 위해 텐서의 레이아웃을 변환하는 과정에서 오버헤드가 존재하기 때문이다.

표 2) 쿠다 코어와 텐서 코어의 학습 효율성

배치 크기	96	64	48	32
텐서 코어	1.73	1.83	1.69	1.48
쿠다 코어	1.29	1.20	1.15	1.01

표 2 는 초당 학습가능한 이미지의 숫자를 전력소모로 나눈 값을 나타내며, 학습의 효율성을 의미한다. 텐서 코어를 사용하는 것이 쿠다 코어를 사용하는 것보다 학습 과정에서 최대 1.53 배 효율성이 높은 것을 확인하였다.

Ⅲ. 결론

인공 신경망을 가속하기 위한 전용 가속기 개발 연구가 활발하게 진행되고 있다. 본 논문에서는 행렬 연산에 특화된 가속 하드웨어를 포함하고 있는 Volta GPU 를 사용하여 텐서 코어와 쿠다 코어의 성능을 비교하였다. 실험을 통해 텐서 코어는 쿠다 코어 보다 평균 1.45 배 학습 성능, 1.53 배 학습 효율성이 높은 것을 확인하였다.

ACKNOWLEDGMENT

이 연구는 산업통산자원부 및 산업기술평가원 (KEIT) 연구비 지원에 의한 연구임 (10076583).

참 고 문 헌

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Huang, Po-Sen, et al. "Deep learning for monaural speech separation." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014.

[3] Bergstra, James, et al. "Theano: Deep learning on gpus with python." NIPS 2011, BigLearning Workshop, Granada, Spain. Vol. 3. Microtome Publishing., 2011.

[4] Markidis, Stefano, et al. "Nvidia tensor core programmability, performance & precision." 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE, 2018.

[5] NVIDIA TESLA V100 GPU ARCHITECTURE, 2017.

[6] Caffe2, https://caffe2.ai/