병렬프로그래밍 4주차

팀명: TEAM\_C\_TENCORE

팀원:소프트융합대학 20143148 이호중

소프트웨어융합대학 20135357 한승탁

컴퓨터 공학과 20135114 김용화

콘텐츠IT 20125345 조준형

목표: TENSOR CORE 기반의 행렬곱 가속화 및 딥러닝 적용 방식 분석

기간:2018.09.01 ~ 2018.12.20

미팅내용:

11월 12일 대학원생들이 대학생팀에게 준 과제인 400\*400의 FP 데이터셋 3개를 만들었고 과 TENSOR CORE에 대하여 조사하고 다음과 같이 발표하였습니다.

TENSORCORE는 4\*4 행렬 곱하고 더하는 연산을 GPU click cycle마다 실행하게 됩니다. FP16 데이터를 입력으로 받고 곱샘 연산을 하고 FP32로 축적시켜 결과를 만듭니다.

FP32 연산과 비교하여 FP16은 메모리와 주파수 등이 반정도 필요하게 되어서 결과적으로 데이터 전달이 더 빠르게 이루어집니다. TENSOR CORE는 CLOCK 마다 64번에 곱샘 연산을 시행하고 TENSOR CORE 8개로 이루어진 하나의 SM에서는 1024(64\*8\*2(곱샘+덧샘)) 번에 연산을 시행하게됩니다.

TENSOR CORE를 사용하기 위해서 NVIDIA는 3가지 라이브러리 WMMA,cuBLAS,cuTLASS를 제공합니다.CuBLASS는 4가지 조건을 만족하여야 하며 그렇지 않을 경우에는 쿠다코어에서 동작하게 됩니다. WMMA는 CUDA 9 에서 가장 낮은 레벨에 TENSOR CORE인터페이스 입니다. WMMA는 오직 16\*16 연산만을 제공합니다. 따라서 WMMA를 사용한다면 데이터 셋을 16\*16에 맞게 변경할 필요가 있습니다. CUTALSS는 최적의 WMMA를 구현을 지원하며 GPU메모리 대기시간을 숨기기 위하여 소프트웨어 파이프 라이닝을 제공합니다.

CUDA 버전에 따라 또 어떤 라이브러리를 사용하는 것인지에 따라서 또 행렬의 크기에 따라서

초당 Tflops/s가 다르다는 것을 논문을 통하여 확인하였습니다.

발표 이후 대학원생들과 논의를 통하여 최종 보고서에 CPU 최적 (슈트라센 알고리즘)과 CUDA에서 cuBlass,WMMA,CUTLASS 모두를 사용하여 정해진 데이터 셋을 연산하는 시간을 비교하는 것을 목표로 하였습니다. 다음 주 까지 슈트라센 알고리즘과 WMMA와 cuBlass, CUTLASS를 사용하는 코드를 작성할 것으로 하였고 다음 텐서코어를 설치하는 pc환경에 CUDA 버전을 맞추는 것으로 하였습니다.

추가적으로 CUDA 행렬 연산에서 쓰레드의 인덱스만을 사용하여 행에 따라 병렬화 했던 방식에서 블록 인덱스도 같이 사용하여 행과 열 모두에서 병렬화가 될 수 있도록 수정하였습니다.

미팅사진:

