병렬프로그래밍 5주차

팀명: TEAM\_C\_TENCORE

팀원: 소프트웨어융합대학 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

대학생 팀 발표 내용:

이전 발표에서 논문의 그래프 해석이 잘못된 것을 바로 잡았습니다. 이전에는 16384의 크기 이상부터 cublass 라이브러리보다 cutlass with Tensor Core가 더 빠르게 행렬처리를 한다고 해석 하였는데 다시 논문을 확인한 결과 행렬의 크기가 8192이며 cublass 라이브러리를 이용한 계산 방법이 가장 빠르게 행렬 연산을 수행 한 것으로 확인하였습니다.

TensorCore를 이용한 행렬 연산의 방법에서 gpu의 총 쓰레드 보다 행렬의 크기가 더 크다면 행렬을 Tile로 나누어 곱샘 연산을 진행하고 그 결과들을 축적하게 됩니다. 이에 따라서 순차적으로 알고리즘을 구성하면 총 5번의 루프를 발생시킵니다. 이것을 병렬화 시켰을 때에도 3개의 루프가 남아있습니다.

Tensor Core 행렬 연산 속도에 메모리가 많은 영향을 끼치는 것을 확인하였습니다. 행렬을 타일로 나누고 타일을 gpu에 어떤 메모리에 저장되는지에 따라서 속도 차이가 약 4배정도 차이가 날수 있다는 것을 확인하였습니다.(matrix 크기가 8192일 때 wmma(naive)와 cublas 결과 비교)

마지막으로 Wmma와 cublas를 이용한 TensorCore 코드를 샘플을 살펴보고 TensorCore를 테스트할 서버 환경을 확인 하였습니다.

미팅 내용:

Cuda를 이용하여 strassen알고리즘을 구현할 수 있다는 내용을 대학원생팀으로부터 전달받았습니다.

최종 보고서에 cpu에서 기본 행렬 곱샘 연산과 Strassen 알고리즘, gpu에서 cuda 기본 행렬 연산,cuda core를 이용한 cublas 연산,wmma, tensor core 를 이용한 cublas연산을 각각 수행하여 TFLOPS를 측정하기로 하였습니다. 추가적으로 시간이 된다면 cutlass를 이용한 연산도 추가 하기로 하였습니다.

미팅 사진:

