팀 B

팀 원 : 학부생 (4) 대학원생 (2)

학부생 : 이준범, 양형모, 고동환, 김혜민

대학원생 : 이주상, 조현상

팀 B의 연구 목표

CUDA를 활용한 Convolutional Neural Network (CNN) 성능 향상 기법 분석 및 성능평가

회의 장소 : 공학관 북카페, 학교 근처 카페(TOM N TOMS)

회의 날짜 : 2018년 11월 26일(월), 2018년 12월 01일(토)

회의 시간 : 18:00~19:30, 19:00~20:30

이번 주는 두 번의 미팅을 가졌습니다.

월요일 정기 회의에서 그 동안 분석해오던 CNN c 코드에 대한 정리를 마무리하고 본격적으로 CUDA 코드를 작성하기 위해 CUDA 코드 분석에 들어가기로 결정했습니다. 그런데 지금 같은 패턴대로면 제 시간안에 결과물을 내기 힘들 것 같다는 의견에 해당 주에 한 번 더 모임을 가져 두 번째 모임 전까지 CUDA 코드를 분석해오기로 했습니다. 그리고 두 번째 모임 이후로 본격적으로 코드 작성을 시작하기로 했습니다.

그렇게 토요일 오후 7시에 해당 주의 두 번째 모임을 가져, 분석한 CUDA 코드에 대해 서로 이해했고, 이해하지 못한 부분을 공유했습니다.

다른 부분은 조원들 모두 상당부분을 이해했지만, layer1\_feature\_maps 커널에서 s\_in\_layer 공유메모리에 d\_in\_layer의 타일을 로드 해서 사용하는 부분에서 난관을 겪고 있었습니다. 사용하는 이유로는 uncoalescedglobal memory access를 피해서 속도를 향상시키기 위해서지 않을까 하는 등의 의견이 나왔지만, 동작하는 부분이 잘 이해되지 않아 명확히 이거다 싶은 의견이 나오지 않았습니다.

일단 그 부분은 이해하기만 하면 추가하면 되는 거니 좀 더 살펴보기로 하고, 먼저 2층 3층을 s\_in\_layer를 사용하지 않고 구현해보기로 했습니다.

공유 메모리를 사용하지 않고 글로벌 메모리를 사용해서, CPU에서 1층의 동작과 GPU에서 1층의 동작을 비교해보자 GPU에서의 동작이 기존보다 0.027초 정도 느리게 측정되었습니다. 다만 1층의 계산에서 유독 오버헤드가 많이 발생하는 것일 수도 있고, 2층 3층을 만들면 CPU보다 성능이 나아지지 않을까 하는 의견이 제시되어, 당일 또는 다음 날 동안 2층 3층을 구현해 테스트 해보기로 결정되었습니다.

2018\_12\_01(토) 학교 근처 카페에서 조모임을 가졌을 때의 사진입니다.

월요일 날은 공학관 북카페에서 약 1시간 30분 가량의 모임을 가졌는데, 사진 찍는 것을 깜박해 사진을 남기지 못했습니다.

