2

**졸업 프로젝트 지정 과제 공고문**

|  |  |  |  |
| --- | --- | --- | --- |
| **과제명** | 자연어 처리 모델에서 GPU multi stream을 사용한 학습 및 추론 성능 개선 | | |
| **과제 내용** | 최근 좋은 고성능의 자연어 처리 모델들은 Transformer 모델 기반으로, Transformer 레이어가 여러 번 반복되는 모델 구조로 구성되어 있다. 반복되는 Transformer 레이어의 연산들은 계산이 많이 필요한 연산들(computation bound)과, 계산량에 비하여 메모리 Read, Wrtie 비중이 높은 연산(memory bound)들로 다양하게 구성되어 있다.  이러한 구조를 활용하여 학습 및 추론에 필요한 Batch를 Micro batch라는 작은 단위로 분할하고 Multi Stream을 사용해서 데이터 병렬 학습처럼 실행한다. 성능 개선을 위하여 하나의 Stream에서 계산 비중이 높은 연산을 수행하는 동안 다른 Stream에서는 메모리 비중이 높은 연산을 동시에 수행하여 각 Stream들이 서로다른 Stream 에게 미치는 영향을 최소화 한다.  - 필요로 하는 지식  1. 딥러닝 프레임워크 (PyTorch 등) 에 대한 지식 2. GPU 아키텍처의 이해 3. 자연어 처리 모델에 대한 이해 4. 모델 분산 학습 기법에 대한 이해 | | |
| **평가 방법** |  | | |
| **비고** | 1. Pytorch 등 딥러닝 프레임워크에 대해 공부할 수 있는 자료 제공 2. GPU 아키텍처를 이해할 수 있는 자료제공 3. Transformer Model에 대한 간단한 교육(방학기간 동안) 4. 모델 분산학습에 대한 교육 및 NVIDIA의 Visual Profiler에 대한 간단   한 교육제공 | | |
| **프로젝트 유형** | A형:산업체 발주 주제 ( )  B형:산업체 참여 정부과제의 주제 ( )  C형: 산업체 수요 과제 ( ) | **참여 인원** |  |
| **지도교수** | 서지원 | **연락처** | 02-2220-2400 |

**한양대학교 공과대학 컴퓨터소프트웨어학부**