**졸업 프로젝트 지정 과제 공고문**

|  |  |  |  |
| --- | --- | --- | --- |
| **과제명** | 자연어 처리 언어모델(BERT, T5)의 양자화를 통한 추론성능 개선 | | |
| **과제 내용** | 현재 자연어 처리 모델의 크기는 계속 증가하는 추세이다. 모델의 크기가 증가함에 따라 실제 서비스에서 다양한 최적화 기술들의 필요성이 대두되고있다. 대표적인 모델 최적화 기술들 중 하나로 quantization이 있으며, 이는 실제 서비스에서의 임베디드 및 모바일 배포를 위해 자주 사용된다.  Quantization에 동일한 자료형을 사용할 지라도 그 활용 방법에 따라  symmetric과 asymetric으로 나뉘며, 자료형의 선택 기준에 따라 다시  single-precision과 mixed-precision으로 나뉜다.  프로젝트 목표는 다음 사항들이 모델의 성능에 어떤 영향을 미치는지  분석하는 것이다.   * Layer 별로 다른 자료형을 적용   예시) 첫 번째와 마지막 layer는 INT8을 사용하고, 나머지 layer에서는 INT4를 사용하여 양자화   * Layer 내에서, channel 별로 다른 자료형을 적용   특정 convolutional layer 내에서 output channel 별로 다른 자료형을 사용하여 양자화   * Signed/Unsigned 자료형의 응용   양자화 된 모델에서 0 값을 어떻게 활용할 것인가를 결정  (해당 내용은 교육을 통해 의미 파악 가능)  [심화] 위 내용을 완료한 이후 심화 연구를 원할 시, Apache TVM을 이용하여 양자화 된 모델을 최적화하고 성능을 분석하는 연구를 할 수 있다. | | |
| **평가 방법** |  | | |
| **비고** | 아래 교육을 여름 방학 동안 진행한다.   * PyTorch framework에 대한 기본 교육 자료 제공 * 자연어 처리 모델에 대한 기본 교육자료 제공 * INT 양자화에 대한 교육  및 코드 리뷰  제공 * PyTorch로 개발된 quantization 기법(HAWQ-V3)에 대한 코드 제공 및 멘토링  (<https://github.com/Zhen-Dong/HAWQ>) 참고 | | |
| **프로젝트 유형** | A형:산업체 발주 주제 ( )  B형:산업체 참여 정부과제의 주제 ( )  C형: 산업체 제안 과제 ( ) | **참여 인원** |  |
| **지도교수** | 서지원 | **연락처** | 02-2220-2400 |

**한양대학교 공과대학 컴퓨터소프트웨어학부**