|  |
| --- |
| **SW프로젝트 제안서** |

|  |  |  |
| --- | --- | --- |
| **프로젝트명** | 자연어 처리 언어모델(BERT, T5)의 양자화를 통한 추론성능 개선 | |
| **프로젝트**  **유형** | 지정 주제 ( O )  자유 주제 ( ) | 졸업 작품 ( O )  졸업 논문 ( ) |
| **프로젝트**  **요약** | 현재 자연어 처리 모델의 크기는 계속 증가하는 추세이다. 모델의 크기가 커질수록 메모리 사이즈, 추론에 걸리는 시간이 증가하는 문제로 실제 서비스에 사용되기 위해서 모델 최적화 기술들의 필요성이 대두되고 있으며 대표적인 최적화 기술로 quantization이 있다. 프로젝트의 목표는 다양한 방식의 quantization을 적용하고 모델의 성능에 미치는 영향을 분석하는 것이다. | |
| **R&D 산출물** | SW ( O ), HW ( ), 특허 ( ), 논문 ( ), 프로그램등록 ( ) | |
| **지도교수** | 서지원 | |
| **예상기간** | 2022.9.9. – 2023.5.30. | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **전공** | **학번** | **학년** | **이름** | **연락처** |
| 컴퓨터소프트웨어학부 | 2017030337 | 4 | 조한빛 | vsj@naver.com  010-8820-6971 |
| 컴퓨터소프트웨어학부 | 2017030546 | 4 | 황태경 | plaire@hanyang.ac.kr |
| 컴퓨터소프트웨어학부 | 2018008904 | 4 | 이성진 | lsjg9909@hanyang.ac.kr |

⦁연락처는 이메일과 전화번호를 모두 쓰되 반드시 수신 가능한 것으로 기입   
(연락을 받지 못해 불이익을 당할 수 있음)

**목 차**

**1. 프로젝트 배경 및 목표**

**2. 프로젝트 주요 내용**

**3. 추진 계획**

**4. 결론**

**5. 참고 문헌**

**1. 프로젝트 배경 및 목표**

본 프로젝트는 자연어 처리 모델에 다양한 방식의 quantization기법을 적용하여 모델 성능에 미치는 영향을 분석하는 것이다.

Transformer, BERT와 같은 모델이 등장한 이후, 자연어 처리 모델의 크기가 점점 증가하는 추세이다. 모델의 크기가 증가함에 따라 다양한 문제들이 발생할 수 있게 되었다. 대표적으로 메모리 사이즈가 증가하는 문제가 있다. 모델 학습에 사용되는 배치크기가 점점 커지면서 학습에 필요한 메모리 사이즈가 증가하게 되었다. 또 다른 문제로는 gradient 계산 량이 많아지면서 추론에 걸리는 시간이 증가하는 것이다. 자원이 한정된 모바일, 임베디드 환경에서 거대한 모델을 사용하기에 제한이 된다는 점도 문제가 된다. 이러한 상황에서 모델의 성능을 유지하면서 크기를 줄이는 모델압축 기법의 필요성이 대두되었다.

프로젝트에서는 모델압축기법 중 하나인 quantization기술을 다양한 방식으로 모델에 적용하여 성능을 분석해보고 모델크기, 추론성능 등을 비교하여 효과적인 quantization 방식을 찾아내는 것을 목표로 한다.

**2. 프로젝트 내용**

프로젝트에서 모델에 quantization기법을 3가지 방식으로 적용한다. Layer별로 다른 자료형을 적용, Layer내에서 channel 별로 다른 자료형을 적용, singed/unsigned 자료형을 적용해본다. 이를 위해서 quantization기법에 대한 이해가 필요하고 대규모 자연어 처리 모델(BERT, T5)의 구조를 이해하여 어느 부분에 quantization을 적용할 수 있는지 알고 있어야한다. 또한 적용할 quantization기법의 오픈소스 코드를 필요로 한다. quantization기법은 실수형 변수를 bit수가 적은 정수형 변수로 변환하는 모델 압축 방법으로서 일반적으로 딥러닝 모델에서 FP32 자료형으로 사용되는 weight, activation 값들을 INT8과 같은 정수형 자료형으로 변환하여 메모리 사용량을 줄이고 처리 속도를 향상시키는 효과를 얻을 수 있다. 그러나 같은 quantization을 적용하더라도 하드웨어의 특성에 따라 추론 성능 향상 정도가 다르다. 또한 실수형 변수를 정수형 변수로 변환할 때 round연산을 진행하면서 실제 값이 완전히 반영되지 않고 오차가 발생한다. weight값의 미세한 변화가 추론 성능에 큰 영향을 미치는 task의 경우 이러한 오차로 인해 모델의 정확도가 크게 떨어질 수 있다. 따라서 프로젝트에서는 다양한 양자화 방식을 적용함으로써 추론 성능을 크게 저하하지 않으면서 자연어 처리 모델에 맞는 양자화 방식을 찾도록 한다.

**3. 프로젝트 추진 계획**

프로젝트는 크게 3단계로 나누어 진행할 계획이다.

1. pytorch 오픈소스 BERT구조 및 quantization코드를 이해한다.

2. 양자화를 적용한 모델 코드를 작성한다.

3. 서버를 이용해서 모델 학습 및 추론 성능을 비교한다.

**4. 결론**

Quantization이 적용된 새로운 자연어 처리 모델은 기존 모델에 비해 메모리 사용량 감소, 처리 속도 향상 등의 효과를 얻을 것으로 기대한다. 향상된 모델은 메모리, CPU, GPU등의 자원이 한정된 모바일/임베디드 환경의 배포에도 이용할 수 있을 것이다. 또한 quantization이 적용된 모델에 pruning, knowledge distillation같은 다른 모델 압축기법을 추가로 적용하여 더 성능이 뛰어난 모델을 생성하는 데 이용할 수 있다.

**5. 참고 문헌**

[1] <https://gaussian37.github.io/dl-concept-quantization/> 2022.9.5

[2] [딥러닝 모델 압축 방법론과 BERT 압축 – AI PLUS Tech Blog (est.ai)](https://blog.est.ai/2020/03/%EB%94%A5%EB%9F%AC%EB%8B%9D-%EB%AA%A8%EB%8D%B8-%EC%95%95%EC%B6%95-%EB%B0%A9%EB%B2%95%EB%A1%A0%EA%B3%BC-bert-%EC%95%95%EC%B6%95/) 2022.9.5

[3] Sheng Shen, et al. “Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT” arXiv preprint arXiv:1909.05840 (2019)

[4] Ofir Zafrir, et al. “Q8BERT: Quantized 8Bit BERT” arXiv preprint arXiv:1910.06188 (2019)

[5] Amir Gholami, et al. “A Survey of Quantization Methods for Efficient Neural Network Inference” arXiv preprint arXiv:2103.13630 (2021)