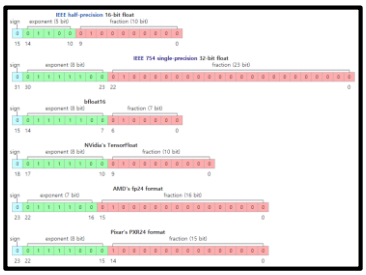
학부생 연구기회 프로그램 (UROP) 공고

◆ 담당교수 : 이재진	◆ 연구실명 : 천둥 연구실
◆ UROP 연구 과제명 : 딥 러닝에서의 Low-precision / Quantization 기술 분석 및 개발	
◆ 모집대상 : C++, Python, 딥 러닝에 익숙한 사람	
◆ 모집기간 : ~ 2022년 6월 말	

연구 배경



수를 표현하는 다양한 방법

- 32비트 부동소수점 표현에서 벗어나 딥 러닝 추론 및 학습을 시도하는 연구가 활발하게 진행되고 있음
  - 16비트 부동소수점, BF16 등의 low-precision 표현 사용
  - 8-bit integer 등으로 양자화(quantization)
  - 여러 표현을 각 phase에서 섞어서 사용
- 모델의 정확도를 비슷한 수준으로 유지하거나 약간만 희생
- 추론 및 학습 성능은 크게 향상됨
  - 실행시간, 모델 크기, 런타임때 메모리 사용량 등

연구 내용

- 기존 low-precision / quantization 딥 러닝 기술 학습
  - 관련 논문 읽고 분석
- PyTorch 등의 딥 러닝 프레임워크에서 재현
  - 지원되지 않는 경우 Python, C++, CUDA를 활용하여 직접 구현
- 모델의 품질 및 성능을 분석하고 개선방안 모색

사전 지식

- Linux 사용 가능해야 함, C++, Python 코드를 읽거나 작성할 수 있어야 함
- 선택 : PyTorch, TensorFlow 등 딥 러닝 프레임워크 사용 경험
- 조건에 충족 되지 않더라도 배워 가면서 진행할 수 있음

서울대학교 컴퓨터공학부
Seoul National University
Dept. of Computer Science and Engineering

THUNDER Research Group
Seoul National University
서울대학교 천둥 연구실

문의 박대영 (daeyoung@aces.snu.ac.kr)