서울대학교 SEOUL NATIONAL UNIVERSITY **보 도 자 료** http://www.snu.ac.kr

2021. 3. 8.(월)

연구책임자 이재진 교수(02-880-1863) / 교신저자
연구진 정우근 연구원 / 제1저자

AI 분야 딥 러닝 핵심 소프트웨어 개발

- Google 과 NVIDIA를 능가하는 딥 러닝 핵심 소프트웨어 기술 -

o 최근 **인공지능(AI)과 빅 데이터 분야는 딥 러닝 기술이 필수**인데, 국내 연구진이 이 **딥 러닝 기술에 핵심인 딥 러닝 컴파일러 프레임워크 기술을 개발**했다. 딥 러닝 컴파일러 프레임워크는 주어진 딥 러닝 모델의 추론과 학습 성능을 높이는데 핵심인 소프트웨어이다.

o 서울대학교 데이터사이언스대학원 데이터사이언스학과 및 공과대학 컴퓨터공학부 이재진 교수 연구팀은 AI 분야의 핵심 소프트웨어인 딥 러닝 컴파일러 프레임워크를 개발, 기존 최고의 성능을 제공하는 Google 과 NVIDIA의 상용 딥 러닝 컴파일러 프레임워크보다 높은 성능을 달성하였음을 발표했다.

o 딥 러닝 기술은 현실적인 시간 안에 추론과 학습을 하기 위해서 NVIDIA GPU와 같은 고성능 AI 반도체의 지원이 필요하다. 하지만 AI 반도체 제조사가 제공하는 상용 딥 러닝 컴파일러 프레임워크에 크게 의존하여야 한다는 한계점이 있다.

o 이재진 교수 연구팀이 개발한 딥 러닝 컴파일러 프레임워크는 공개된 GPU 하드웨어 정보만을 사용, 주어진 딥 러닝 모델과 모델을 실행할 GPU에 최적화된 코드를 직접 생성하는 방식으로 고성능을 달성했다.

o 연구팀이 개발한 기술을 널리 사용되고 있는 여러 개의 딥 러닝 벤치마크 모델(ResNet, BERT 등)에 적용하여 테스트한 결과, Google의 TensorFlow XLA, NVIDIA의 TensorRT, Apache의 TVM과 같은 최첨단 딥 러닝 컴파일러 프레임워크와 성능이 비슷하거나 더 높은 성능을 달성할 수 있음을 보였다. 연구팀은 개발한 기술을 공개 소프트웨어화 할 예정이다.

o 이 연구에 주된 기여를 한 컴퓨터공학부 정우근 연구원은 “지금까지 기술들은 국외 하드웨어 제조사가 소스 코드를 공개하지 않은 상용 딥 러닝 컴파일러 프레임워크에 크게 의존하여 국내 기술 발전에 한계가 있었다”며 “이번 연구를 통해 이미 국외에서 상용화되거나 앞선 기술이라도 다른 독창적인 방법을 사용하여 충분히 더 앞선 기술을 개발할 수 있음을 보였다”고 설명했다.

o 서울대 이재진 교수는 “본 연구성과는 딥 러닝 분야의 최첨단 핵심 소프트웨어 기술을 국내에서 확보한 고무적인 사례”라며 “현재 국내를 비롯하여 전 세계적으로 AI 반도체 개발 열풍이 불고 있는데, 본 연구성과는 개발된 AI 반도체의 활용과 상용화에 필수적인 기술이다”고 설명했다.

o 연구 결과는 올해 6월 개최 예정인 프로그래밍 언어 분야 국제 학술대회인 PLDI(Programming Language Design and Implementation)에서 발표될 예정이다.

[붙임] 1. 연구결과 2. 용어설명 3. 그림설명

연 구 결 과

A Deep Learning Optimization Framework for Versatile GPU Workloads

Wookeun Jung, Thanh Tuan Dao, and Jaejin Lee
(PLDI' 21, conditionally accepted with shepherding)

PyTorch, TensorFlow등의 널리 사용되는 딥 러닝 프레임워크들은 NVIDIA가 제공하는 cuDNN 라이브러리에 크게 의존하고 있다. 그러나 cuDNN과 같이 미리 구현된 라이브러리들은 딥 러닝 연산자의 종류가 다양해지고 하드웨어가 다양해질수록 높은 성능을 달성하기 어렵다는 한계가 있으며, fusion 등의 최적화를 적용할 수 없다는 한계가 있다.

본 연구는 주어진 딥 러닝 연산과 GPU 하드웨어 정보를 사용하여 최적의 코드를 생성하는 딥 러닝 최적화 프레임워크인 DeepCuts를 제안한다. DeepCuts는 fusion을 고려한 코드 생성 기법과 성능 모델링을 통해 효율적인 코드를 생성하며, 기존 state-of-the-art 딥 러닝 최적화 프레임워크(Apache TVM, Google TensorFlow XLA, NVIDIA TensorRT)보다 더 높은 성능을 달성한다.

용 어 설 명

1. 딥 러닝 컴파일러 프레임워크

하드웨어 정보와 무관하게 작성된 딥 러닝 응용을 입력으로 받아, 밑단의 하드웨어 가속기(예: GPU)에서 실행하기 위해 최적화된 코드를 생성하는 소프트웨어를 의미함. 대표적으로 Google이 개발한 TensorFlow XLA, NVIDIA가 개발한 TensorRT, Apache의 TVM 등이 있음.

2. GPU

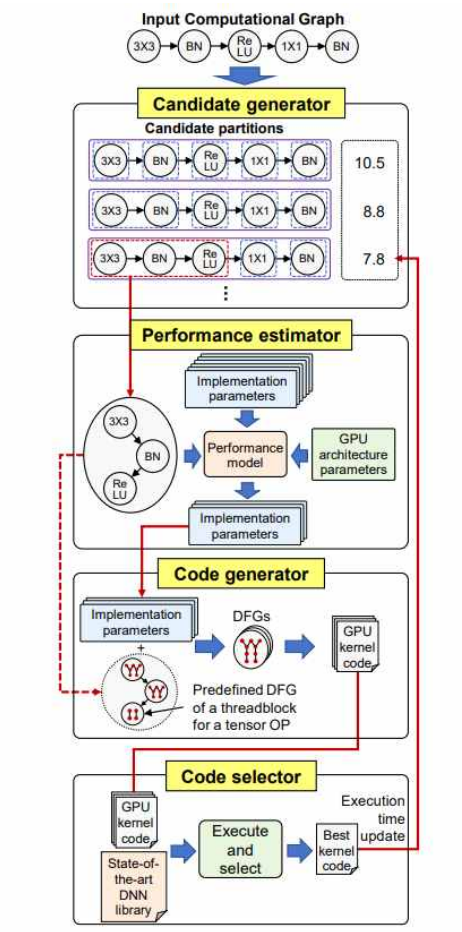
Graphics Processing Unit의 약자로 그래픽 처리에만 사용하였으나 현재는 범용 계산에 이용하여 성능과 전력효율을 높이고 있으며 AI 분야의 딥 러닝의 추론과 학습을 실행하는 사실 상의 표준(de facto standard) 플랫폼

3. PyTorch 와 TensorFlow

AI 분야 딥 러닝 모델의 개발과 실행에 가장 널리 쓰이는 소프트웨어 프레임워크. 이들을 통해 개발된 딥 러닝 모델은 Google의 TensorFlow XLA, NVIDIA의 TensorRT, Apache의 TVM 등을 통해 최적화 된다.

그 림 설 명

DeepCuts의 코드 생성 방식



PyTorch나 TensorFlow에서 딥 러닝 연산을 표현하는 Computation Graph를 입력으로 받아 GPU 프로그램을 생성하는 일련의 과정을 나타낸다. Candidate generator는 최적화 코드 생성을 위한 후보 연산 또는 fusion을 통한 연산의 뭉치를 정한다. Performance estimator는 이들을 받아서 하드웨어(GPU) 정보와 구현 파라미터들을 이용하여 이들 연산이 실제 GPU 코드로 구현되었을 때의 성능을 간단한 예측 모델을 통해 예측한다. 예측 결과에 따라 성능이 아주 나쁘게 나오는 코드 구현을 걸러서 Code generator에 넘긴다. Code generator는 실제 코드를 여러 벌 생성하여 실행해 보고 최종적으로 가장 성능이 좋은 코드를 선택한다.