AI Accelerator

영남대학교
차세대 컴퓨터 시스템 연구실
석사과정생 이원호

FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks

19th USENIX Conference on
File and Storage Technologies(FAST),2021

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

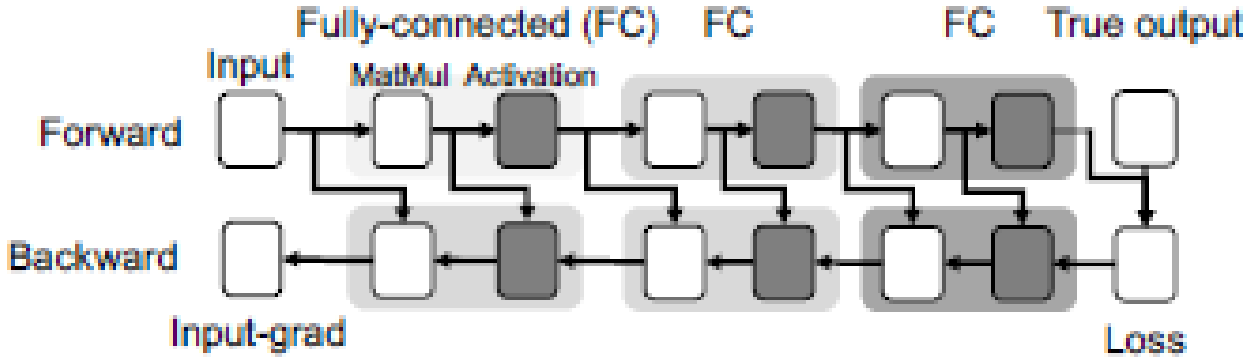
- DNN 구조가 많은 영역에 적용되며, 정확성 요구
- GPU같은 학습 플랫폼에서 DRAM 메모리 부족
- 배치 사이즈와 DNN 크기에 제한이 생김
  - 기존 연구의 해결 방안 intermediate data(feature maps)를 host memory로 offloading
  - Memory bandwidth 와 memory capacity 문제로 성능이 일관적이지 않음

DNN training

- 심층 신경망 모델 학습 과정에서 빈번히 사용되는 역전파 방식은 다음과 같은 데이터 재사용 형태를 보여준다.

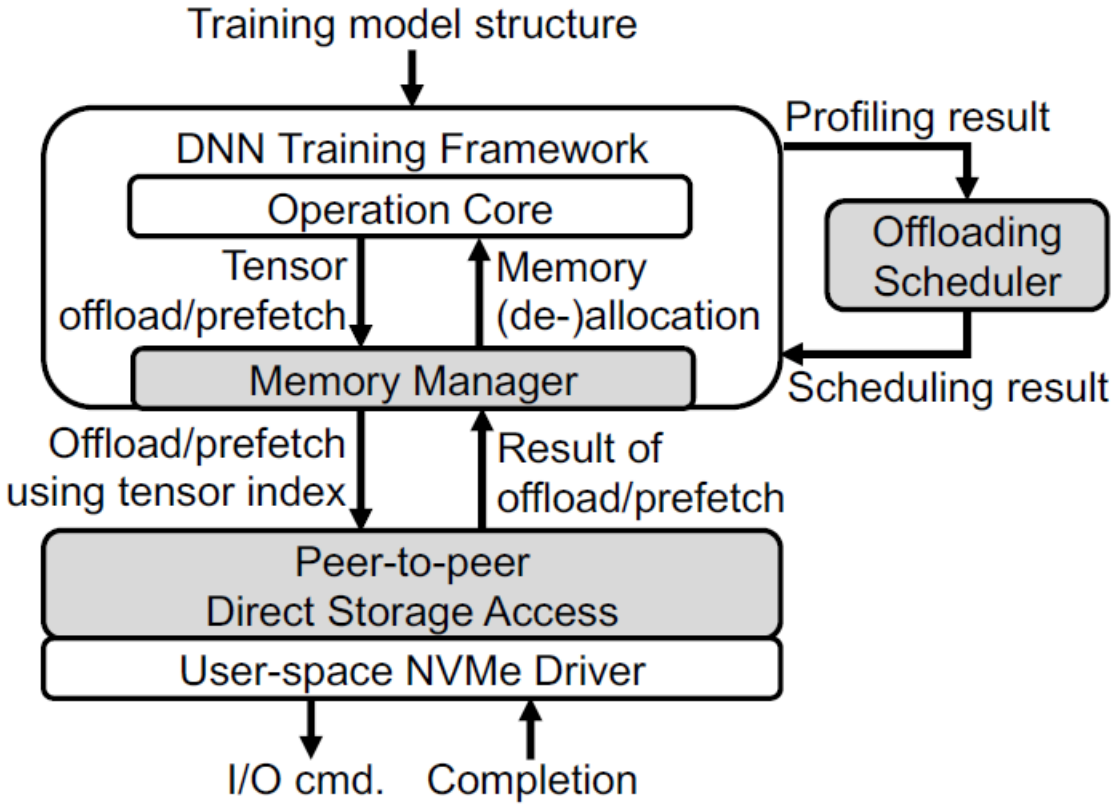
- Forward 과정에서 사용된 신경망의 파라미터를 업데이트하기 위해 backward 과정에서 재사용이 일어남.

- 재사용을 위한 파라미터들을 메모리에서 유지하려고 하지만, 용량 문제가 발생



제안 사항

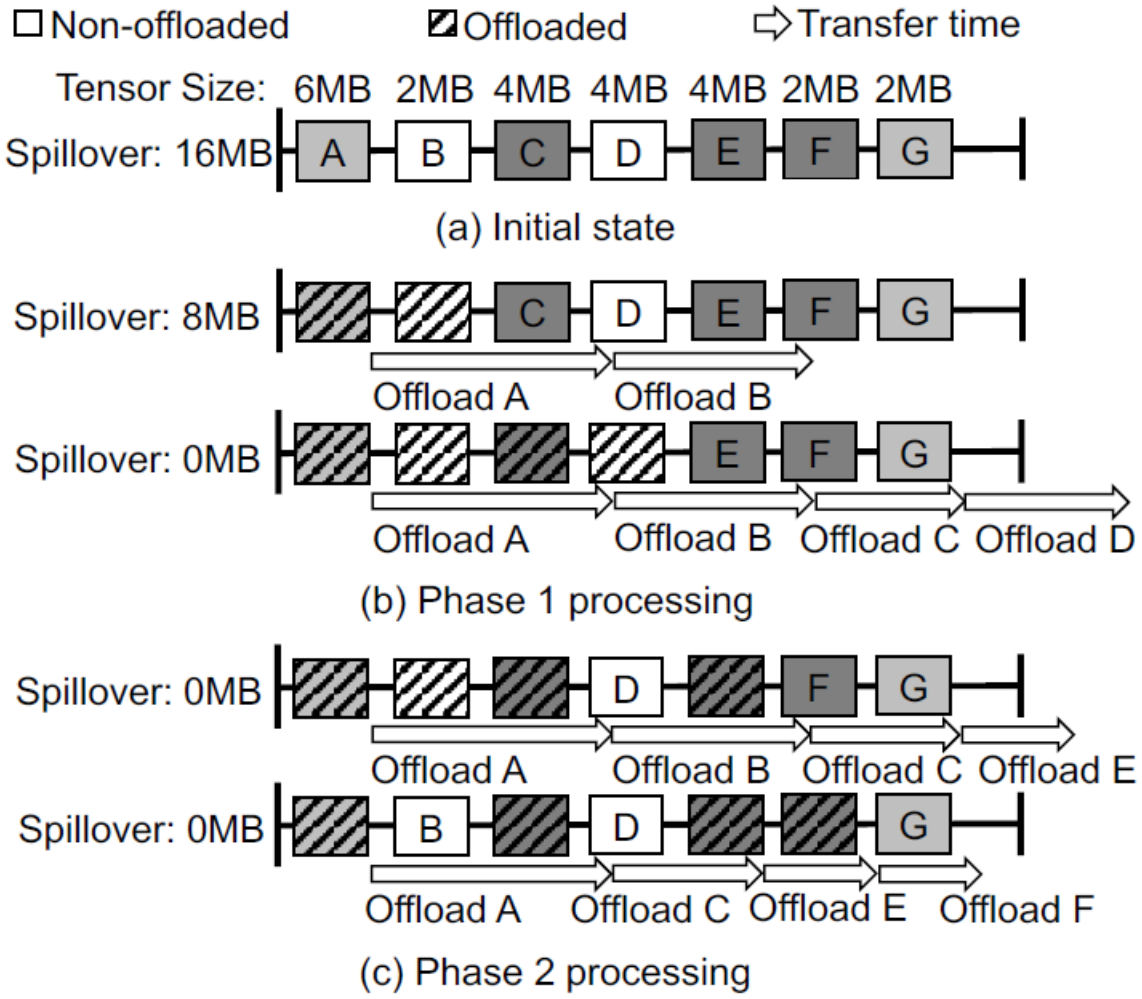
- 입력에 대한 buffering 대신 NVMe SSD를 backing store로 활용하는 FlashNeuron 구조 제안
- 해당 구조에 적절한 offloading scheduler 제안
- CPU에 대한 접근을 최소화시키고자 GPU와 SSD 간의 데이터 전송을 활용 (lightweight user-level I/O stack)



Offloading Scheduling

- Phase1 – Linear Tensor Selection

- Phase2 – Compression-aware Tensor Selection
  - Phase1에서 선택된 Tensor들 중 비압축 텐서는 후보에서 제외
  - 압축률이 높은 텐서에게 우선순위
  - Data transfer time을 다시 계산해 전체 transfer time이 기존 forward pass의 전체 실행시간보다 짧으면 종료
  - Compression-friendly tensor가 없어도 종료



Cop-Flash: Utilizing hybrid storage to construct a large, efficient, and durable computational storage for DNN training

IEEE 15th International Conference on Cloud Computing (CLOUD),2022

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

- 클라우드에서 연속적으로 데이터를 처리할 때 storage에서 컴퓨팅을 구분할 때 발생하는 문제가 많음

- 기존 연구는 동일 저장장치를 통해 최적화를 진행했지만, CSD(Computational Storage Device)에서 DNN 학습의 요구 사항을 고려하지 않음

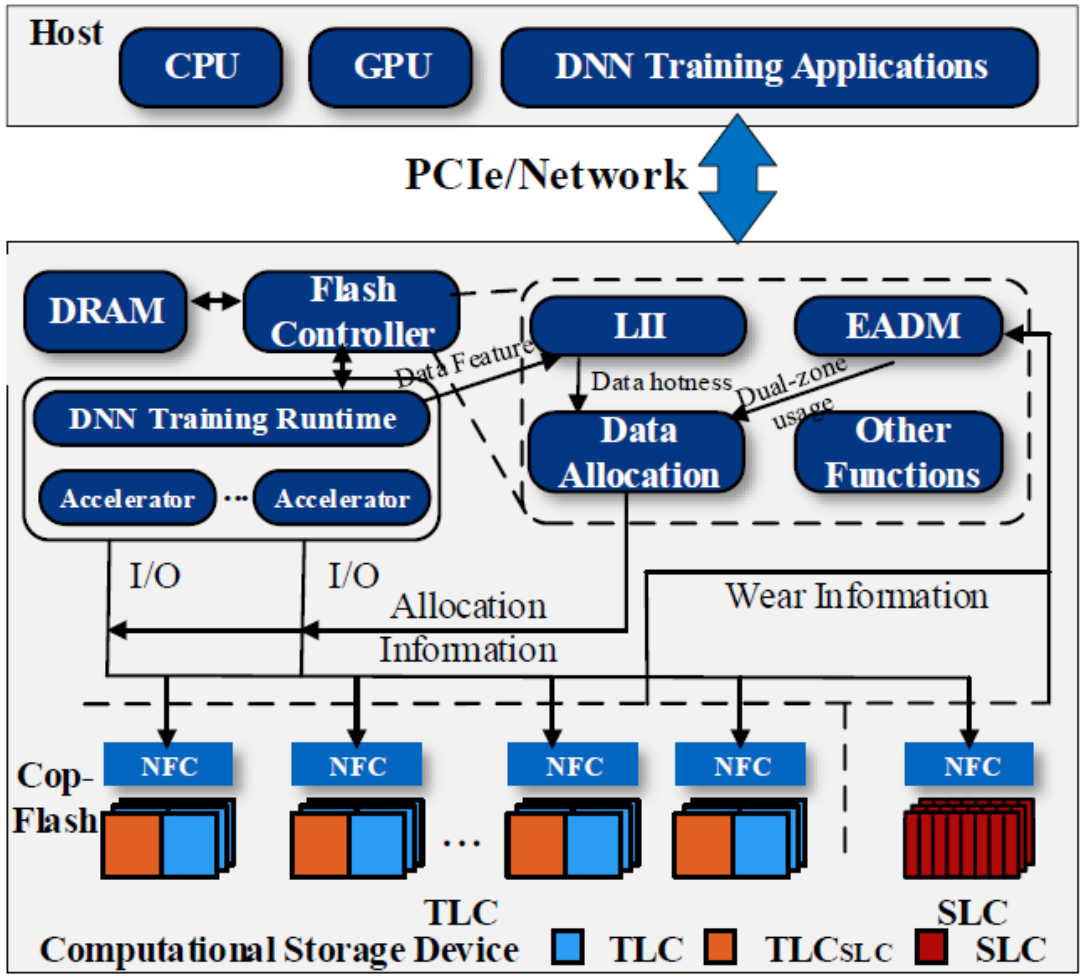
  - 많은 원시 데이터, 임시 데이터, 중간 가중치 저장을 위한 대용량 저장소
  - Data I/O latency를 줄이기 위한 고대역폭 저장소
  - 빠른 마모를 견딜 수 있는 reliability 제공
  - Random write와 non-page-aligned write로 인한 mismatch 문제 관리

제안 사항

- Lifetime-based I/O Identifier (LII)
  - 저장소 이종성의 이점을 최대화하고, garbage collectio의 영향을 줄이기 위해 data lifetime에 따라 data hotness를 규정함

- Erase-aware Adaptive Dual-zone Management (EADM)
  - SLC와 TLC간의 마모도 차이를 계산해 동적 할당 진행
  - 대역폭 성능 향상
  - 시스템 신뢰성 보장

Overall Architecture



A^3: Accelerating Attention Mechanisms in Neural Networks with Approximation

IEEE International Symposium on
High Performance Computer Architecture(HPCA),2020

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

- 기존 가속기들은 DNN, CNN, RNN과 같은 형태의 신경망에 대해서만 최적화를 진행시키려 함
  - 최근 NLP 분야에서 활용되는 트랜스포머는 attention 기반의 작업이므로 이에 대한 고려가 필요함

- Attention 메커니즘은 기존 신경망과 다른 형태의 연산이며 높은 복잡도로 인해, 연산에 대한 특별한 고려가 필요함
  - 기존 하드웨어는 Dense multiplication과 softmax를 통해 구현됨

제안 사항

- Attention 메커니즘에 대해 algorithmic approximation + hardware specialization을 통해 가속

- Algorithmic approximation
  - 모든 타겟이 항상 균일한 연관성을 가지지 않음

  → approximate candidate selection algorithm
  - 검색 타겟 추정을 통해 연산량 감소

- Specialized hardware pipeline
  - 병렬화를 통해 추정된 attention mechanism을 가속
  - Attention 메커니즘에 대한 최적화

DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation

55th IEEE/ACM International Symposium on Microarchitecture(MICRO),2022

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

- 기존 시스템은 다중 입력, 다중 출력을 지원하기 위해 병렬화에 최적화 되어있음

- 자연어 생성의 경우 다중 입력이 들어오지만, 출력은 하나의 토큰에 대해서만 진행함
  - 기존 GPU를 사용하는 시스템이 순차처리에 적합하지 않음

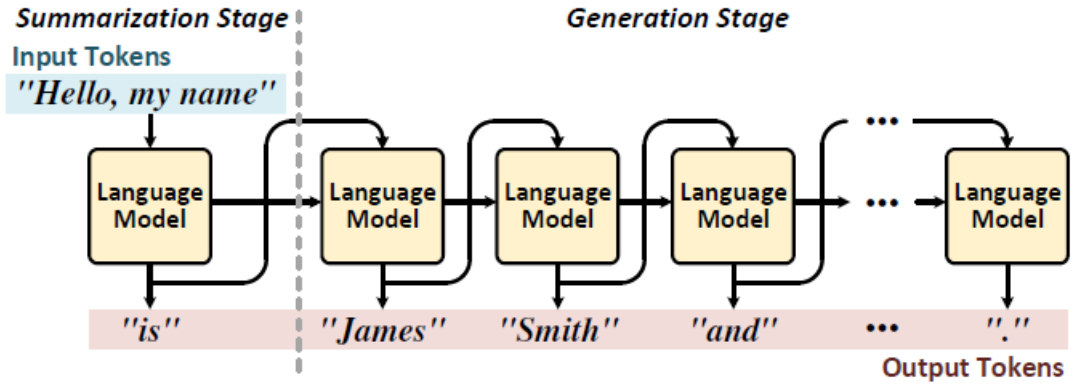
- 기존 연구는 트랜스포머 모델을 가속하기 위해 attention mechanism (matrix multiplication + softmax)을 최적화하는 방식을 택했지만, 언어 모델은 트랜스포머 전체 구조에 대한 고려가 필요함

Transformer based text generation

- 2가지 스테이지를 가짐
  - 요약 스테이지
  - 생성 스테이지

- 요약을 통해 입력 토큰을 병렬처리해 중간 임베딩 벡터 생성

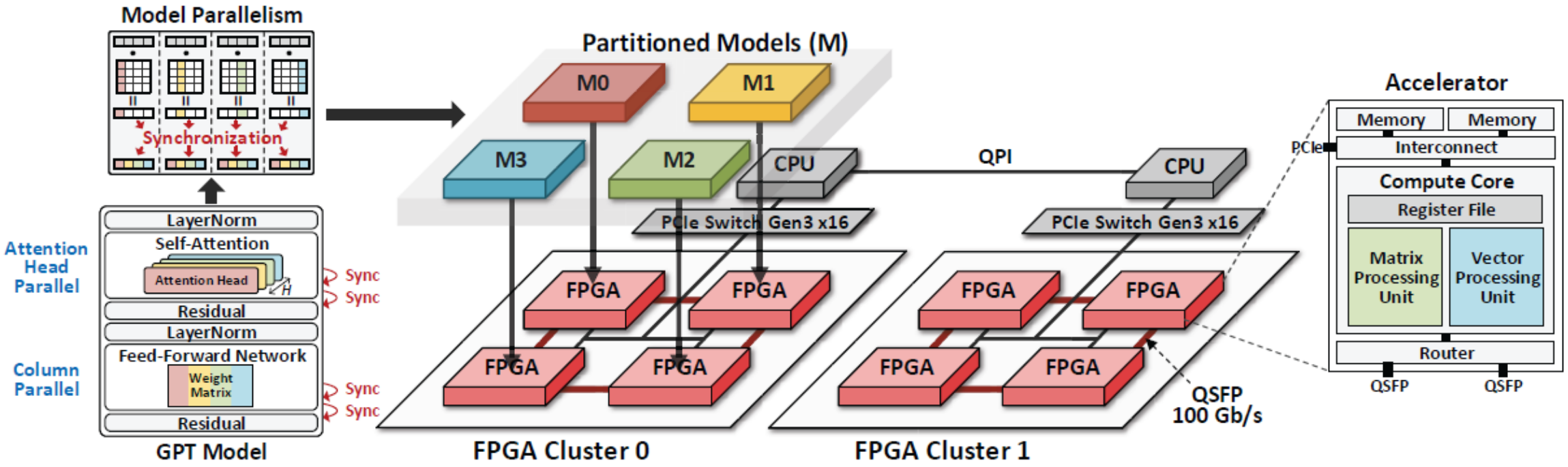
- 생성 단계에서 임베딩 벡터를 통해 출력 토큰을 " 단일"로 생성



제안 사항

- 단일 토큰 처리에 최적화된 코어
- 효율적인 tiling scheme와 dataflow 이용
- 여러 시스템 디바이스에서 model parallelism을 사용
- FPGA로 구성해 지속적으로 변하는 transformer 모델에 대한 적용 가능

Overall Architecture



Summary (Hardware)

- DRAM의 용량 부족을 스토리지로 해결하려는 접근법
- CSD (Computational Storage Device)와 같은 접근 방식
- 인공지능 모델의 구조에 따른 고려사항
  - 인공지능에서 빈번히 사용되는 matrix multiplication, softmax 등의 최적화 고려
  - 최신 인공지능 모델(transformer based)에 따른 최적화

DUET: A Compiler-Runtime Subgraph Scheduling Approach for Tensor Programs on a Coupled CPU-GPU Architecture

IEEE International Parallel and Distributed Processing Symposium (IPDPS),2021

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

- 기존 DL 프레임워크와 컴파일러는 CPU와 GPU간 격리를 통해 DL 추론 속도를 최적화시키는 데 집중
- 이는 CPU와 GPU의 결합을 통한 성능 향상의 기회를 놓침
- 이런 DNN의 특징이 이종 장치들 사이에서 높은 성능향상을 이끌어 낼 수 있음
- DNN 연산
  - 복잡한 연산 패턴
  - 다른 형태의 요소들

제안 사항

- Coarse-grained partitioning strategy
  - DNN 연산 그래프를 작은 그래프로 분할 → 높은 연산량을 적은 통신 볼륨을 통해 활용가능
- Compiler-aware profiling method
  - Loop 최적화를 통해 스케쥴링 선택 성능 향상
- Greedy-correction subgraph scheduling
  - 개발자의 추가 입력없이 CPU와 GPU에 자동 연산 매핑

DnD: A Cross-Architecture Deep Neural Network Decompiler

31st USENIX Security Symposium(Security), 2022

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

- 최근 DNN 사용이 꾸준히 증가했고, 특히 엣지 디바이스에서 구체적인 DNN 컴파일러의 활용빈도가 높다.
- 디컴파일러의 부재로 인해 아래와 같은 문제점들이 발생
- 기존 보안 기술들 (model extraction, white-box adversarial sample generation, model patching, hardening)은 컴파일된 DNN에 적용할 수 없음
- DNN의 이진 코드에서 시작하는 DNN의 high-level 표현을 복구할 수 없음

제안 사항

- 전용 루프 분석과 함께 심볼릭 실행을 통해 분석된 이진 코드를 새로운 IR에 적용
- 새로운 IR은 컴파일러에서 ISA-agnostic한 방식으로 고수준의 수학적 DNN 연산을 표현
- 추출된 해당 연산을 기존 DNN 수학 연산 구조(하드웨어가 지원하는 or 기존 연산 라이브러리)와 매칭시킴
- 규정된 모든 DNN 연산의 hyper-parameter와 parameter를 복구할 수 있음

One-Shot tuner for deep learning compilers

Proceedings of the 31st ACM SIGPLAN International Conference on Compiler,2022

차세대 컴퓨터 시스템 연구실

이원호

문제점과 동기

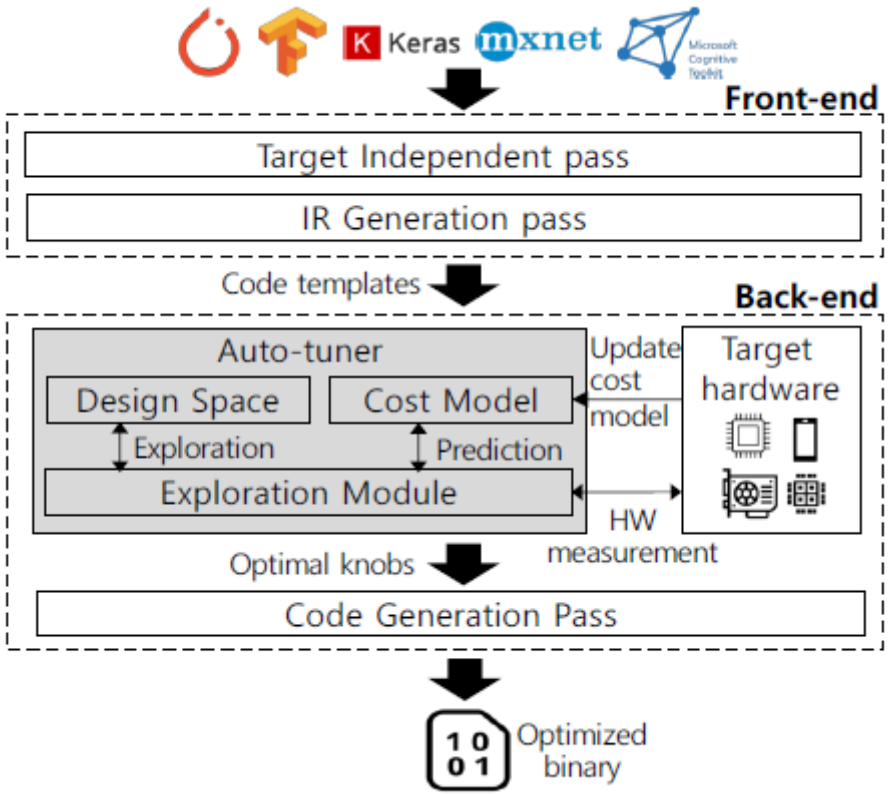
- DL 프레임워크의 최적화 영역에서 auto-tuning DL compiler는 많은 영향을 끼침
- 이는 기존 hand-tuned 라이브러리보다 좋은 성능을 나타냄

- 넓은 검색 공간으로 인해 반복적인 하드웨어 측정이 일어나고 auto-tuning time을 길어지게 함

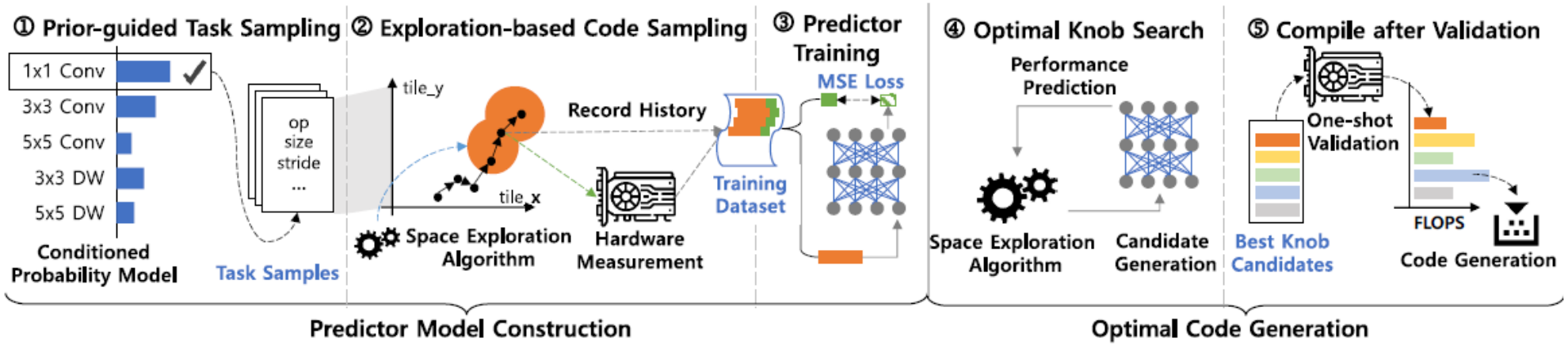
제안 사항

- Auto-tuning overhead를 줄이기 위해 Neural-Predictor 제안
- 반복적인 검색과 하드웨어 측정없이 최적화된 텐서 연산 코드 생성
- 효율적인 학습을 위해 입력 표현의 확장
  - Task-specific 정보를 포함
  - Data sampling method를 안내

Auto-tuning DL compiler architecture



Overall Architecture



Summary (Compiler)

- 기존 이종 디바이스에 대한 고려를 격리가 아닌 결합을 통해 최적화
- 디컴파일러의 제안으로 기존 adversarial attack과 같은 보안 공격으로 부터 신뢰성을 확보하고 안정적인 모델 사용
- 연산을 위해 기존 hand-tuned library를 사용하는 컴파일러 대신 auto-tuning을 통해 컴파일러의 성능 향상과 그에 따르는 비용의 감소

Reference

- FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks, 19th USENIX Conference on File and Storage Technologies(FAST),2021
- Cop-Flash: Utilizing hybrid storage to construct a large, efficient, and durable computational storage for DNN training, IEEE 15th International Conference on Cloud Computing (CLOUD),2022
- A^3: Accelerating Attention Mechanisms in Neural Networks with Approximation, IEEE International Symposium on High Performance Computer Architecture(HPCA),2020
- DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation, 55th IEEE/ACM International Symposium on Microarchitecture(MICRO),2022
- DUET: A Compiler-Runtime Subgraph Scheduling Approach for Tensor Programs on a Coupled CPU-GPU Architecture, IEEE International Parallel and Distributed Processing Symposium (IPDPS),2021
- DnD: A Cross-Architecture Deep Neural Network Decompiler, 31st USENIX Security Symposium(Security), 2022
- One-Shot tuner for deep learning compilers, Proceedings of the 31st ACM SIGPLAN International Conference on Compiler,2022