rebellions_

리벨리온 소프트웨어 스택

Aug 05, 2024

서론

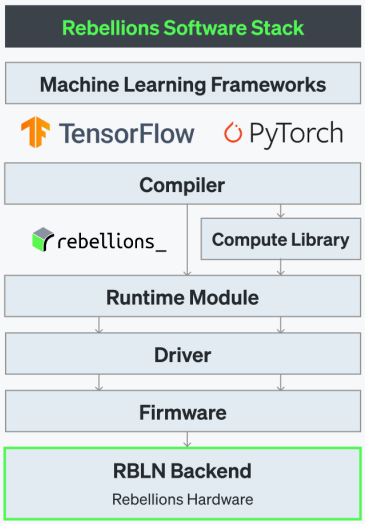
인공지능(AI)의 급속한 확산과 다양한 활용 분야의 등장으로 인해, 짧은 처리 시간과 높은 에너지 효율성을 최적화한 하드웨어 수요가 증가하고 있습니다. AI 가속기(AI 칩)는 AI 알고리즘의 실행 속도를 높이고 전력 소모를 줄이며 기존의 CPU나 GPU보다 복잡한 연산을 더욱 효율적으로 처리하도록 설계된 하드웨어입니다.

리벨리온의 ATOM™은 딥러닝 모델 처리를 최적화하도록 설계된 SoC(System-on-Chip)입니다. 8개의 강력한 뉴럴 엔진이 높은 성능과 낮은 처리 시간을 위해 설계된 메모리 아키텍처 내에서 연산을 수행합니다. 이러한 하드웨어의 잠재력을 최대한 발휘하기 위해서는 최적화된 소프트웨어가 필수적입니다.

소프트웨어는 하드웨어가 최대 성능을 발휘할 수 있도록 작동을 보장하는 핵심 동력으로. 리소스 관리, 데이터 흐름 최적화, 알고리즘의 효율적 실행을 담당합니다. 또한 GPU에서 리벨리온 칩으로 전환하는 사용자를 위해 호환성을 극대화하여 간편한 통합 경험을 제공합니다. 리벨리온의 소프트웨어 스택은 사용 편의성과 신뢰성을 특히 강조하며, 개발자와 엔지니어가 쉽게 접근할 수 있도록 포괄적인 사용자 문서와 SDK를 제공합니다.

본 문서에서는 리벨리온의 소프트웨어 스택의 주요 구성 요소와 핵심 기능을 소개하며, 이를 통해 ATOM™이 달성한 압도적인 연산 성능과 대폭 절감된 전력 소모에 대해 설명합니다.또한 YOLOv6 객체 탐지(Object Detection) 모델을 대상으로 한 NVIDIA RTX A5000과의 성능 비교 결과를 제시하여, 고성능이면서도 에너지 효율적인 풀스택(Full-Stack) AI 솔루션을 제공하는 리벨리온의 지속적인 혁신을 보여줍니다.

리벨리온 소프트웨어 스택



[Figure 1. 리벨리온의 소프트웨어 스택]

ATOM™이 복잡한 AI 워크로드를 처리하기 위한 강력한 연산 성능을 제공한다면, 리벨리온의 소프트웨어 스택은 이러한 하드웨어 아키텍처의 강점을 극대화하도록 모델 실행을 최적화합니다. 이를 통해 ATOM™의 잠재력을 최대한 효율적으로 발휘할 수 있도록 지원합니다.

RBLN SDK는 리벨리온의 독자적 컴파일러, 연산 라이브러리(Compute Library), 런타임, 드라이버, 펌웨어로 구성되어 있으며, TensorFlow, PyTorch, Hugging Face 등 다양한 프레임워크에서 사전 학습된(pre-trained) 모델을 리벨리온의 칩 기반 서빙 환경에 매끄럽게 통합할 수 있도록 설계되었습니다. 이 소프트웨어 스택의 모든 구성 요소는 최저 지연(latency)을 실현하기 위해 긴밀하게 연동되며, 고성능 AI 추론 환경을 위한 완성형 통합 플랫폼을 제공합니다.

프레임워크 지원

TensorFlow, PyTorch 등 주요 프레임워크와 Hugging Face 모델을 포함한 200개 이상의 레퍼런스 모델 및 연산을 지원하는 RBLN SDK는, 개발자가 리벨리온의 칩 환경으로 원활히 이전할 수 있도록 보장합니다. 이를 통해 대형 언어 모델(LLM)과 디퓨전 모델은 물론, 비전 및 음성 기반의 다양한 인기 모델들도 매끄럽게 구동됩니다.

RBLN 컴파일러

RBLN 컴파일러는 모델을 ATOM™에서 실행 가능한 명령어로 변환합니다. 컴파일러는 프론트엔드(Frontend Compiler)와 백엔드(Backend Compiler) 두 주요 구성 요소로 이루어져 있습니다. 프론트엔드 컴파일러는 딥러닝 모델을 중간 표현(Intermediate Representation, IR)으로 변환하고 최적화한 후 이를 백엔드 컴파일러로 전달합니다. 백엔드 컴파일러는 이러한 IR을 추가로 최적화하여, 하드웨어에서 실행할 수 있는 커맨드 스트림(Command Stream), 프로그램 바이너리, 그리고 직렬화된 가중치(Serialized Weights)를 생성합니다.

1. 프론트엔드 컴파일러

프론트엔드 컴파일러는 모델을 통합된 그래프 형태의 IR로 변환합니다. 변환된 IR의 노드들은 백엔드 컴파일러로 전달되기 전에 최적화 과정을 거칩니다. 이 과정에는 불필요한 함수 및 노드 제거, 노드와 가중치의 바인딩, 데이터 및 커널 로드/스토어 오버헤드를 줄이기 위한 노드 병합, 전송 가능한 노드에 대한 주석 추가, 그리고 효율적인 병렬화를 위한 그래프 분할이 포함됩니다.

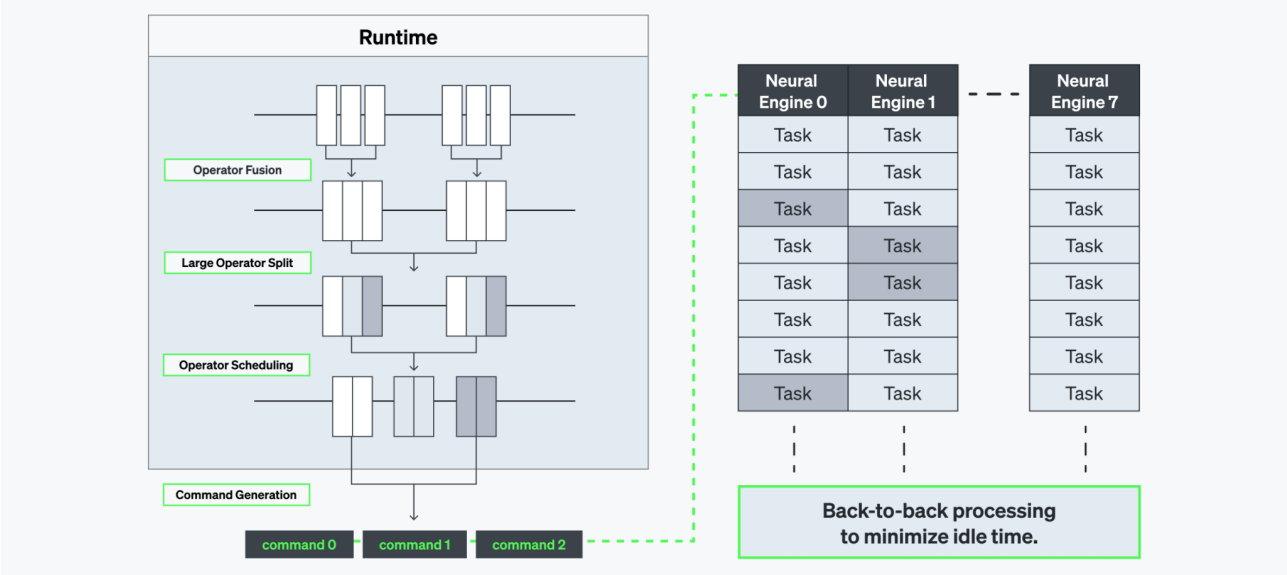
2. 백엔드 컴파일러

백엔드 컴파일러는 상위 수준의 코드가 하드웨어에서 효율적으로 실행되도록 보장합니다. IR 연산을 받아 연산과 메모리 간의 리소스 사용을 최적화하며, 최대 활용률을 달성하기 위한 다양한 기술을 적용합니다. 주요 기술은 다음과 같습니다.

- **분할(Partitioning):** 여러 카드 구성을 사용하는 경우, 모델을 더 작은 구성 요소로 나누어 여러 장치에 분산하여 효율적인 병렬 처리를 구현합니다.
- **융합(Fusion):** 연산자(operator)를 병합하여 불필요한 중간 활성화 데이터 전송을 줄이고, Neural Engine에서 최적화된 계산을 가능하게 합니다.
- **분리(Splitting):** 장치 내에서 연산을 나누어 파이프라이닝과 스케줄링을 최적화하면서 성능을 극대화합니다.
- **타일링(Tiling):** 분리된 작업을 Neural Engine 전체에 분산하여 가장 효율적인 성능을 달성합니다.

백엔드 컴파일러가 생성하는 결과물은 다음과 같습니다.

1. **커맨드 스트림: 칩의 여러 레이어에서 워크로드 실행을 제어하는 커맨드 프로세서(Command Processor)용 명령 세트**
2. **프로그램 바이너리(Program Binary): 연산 라이브러리의 프로그램 스트림을 기반으로 한 뉴럴 엔진용 고도로 최적화된 명령어**
3. **커널 직렬화(Kernal Serialization): FP32 가중치를 FP16으로 변환하여 뉴럴 엔진에 최적화된 형식으로 직렬화**



[Figure 2. 모델 컴파일 과정]

연산 라이브러리

연산 라이브러리는 모델 추론에 필요한 고도로 최적화된 저수준 연산(low-level operations) 집합으로 구성되어 있습니다. 이 연산들은 뉴럴 엔진 내 산술 논리 장치(ALU)의 프로그래머블 구성 요소를 형성하며, 컴파일러의 명령에 따라 프로그램 바이너리를 준비합니다.

RBLN SDK는 전통적인 합성곱 신경망(CNN)부터 최신 생성형 AI(GenAI) 모델까지 모두 지원합니다. 여기에는 수백 개의 GEMM(General Matrix Multiply), 정규화(normalization), 비선형 활성화 함수가 포함됩니다. 뉴럴 엔진의 높은 유연성 덕분에 지원되는 저수준 연산의 목록은 계속 확장되고 있으며, 다양한 AI 애플리케이션 가속을 가능하게 합니다.

런타임 모듈

런타임 모듈은 컴파일된 모델과 하드웨어 사이의 중간 계층으로, 실제 프로그램 실행을 관리합니다. 컴파일러가 생성한 실행 가능한 명령어를 준비하고, 메모리와 뉴럴 엔진 간 데이터 전송을 제어하며, 실행 과정을 모니터링하여 성능을 최적화합니다.

Driver

드라이버는 커널 모드 드라이버(KMD)와 사용자 모드 드라이버(UMD)로 구성되어 있으며, 하드웨어 접근을 안전하고 효율적으로 제공합니다. KMD는 운영체제가 하드웨어를 인식하도록 하고, UMD가 사용할 수 있는 API를 노출합니다. 또한 컴파일러 스택에서 생성된 명령 스트림을 디바이스로 전달합니다. UMD는 사용자 공간에서 실행되며, 애플리케이션과 하드웨어 간의 상호작용을 관리하는 중간 역할을 수행합니다.

펌웨어

펌웨어는 ATOM™의 가장 하위 레벨 소프트웨어 구성 요소로, 소프트웨어와 하드웨어를 직접 연결하는 최종 인터페이스 역할을 합니다. SoC 상에 위치한 커맨드 프로세서의 작업을 제어하며, 메모리 계층 전반에 걸쳐 실제 AI 워크로드(커맨드 스트림)를 조율합니다. 또한 하드웨어의 상태를 실시간으로 모니터링합니다.

소프트웨어로 극대화된 유연한 아키텍처

ATOM™은 다양한 워크로드와 애플리케이션에 적응할 수 있도록 설계된 유연한 아키텍처를 갖추고 있습니다.

Future-Proofing

새로운 모델과 알고리즘이 등장함에 따라, 새로운 저수준 연산이 필요할 수 있습니다. 리벨리온의 지속적인 소프트웨어 지원을 통해, ATOM™은 별도의

성능 및 효율성 최적화

모델은 단계별로 다른 연산(예: 합성곱, 행렬 곱셈, 풀링, 활성화 함수)을 요구합니다. RBLN 컴파일러는 각 단계의 특성에 맞게 자원을 동적으로 할당하

전용 하드웨어 없이도 다양한 연산을 유연하게 지원하며 효율적으로 실행할 수 있습니다.

멀티 디바이스 확장성

대규모 GenAI 워크로드에서도 안정적인 성능을 보장하기 위해, 고급 드라이버와 펌웨어가 강력한 디바이스 간 통신을 지원합니다. 또한 디바이스 간 통신 오버헤드를 최소화하는 전략적 파티셔닝을 통해 성능을 최적화합니다.

고 성능을 최적화합니다.

동적 구성(Dynamic Configuration)

응용 프로그램에 따라 속도나 에너지 효율성 등 다양한 우선 순위가 있을 수 있습니다. 리벨리온의 Dynamic Voltage Frequency Scaling (DVFS)는 최소 전력 소비를 목표로 특정 응용 프로그램 요구 사항에 따라 유연한 구성을 가능하게 합니다.

YOLOv6-Large 성능 비교

ATOM™은 에너지 효율적인 성능을 위해 설계되었으며, 다양한 모델을 유연하게 지원합니다. 이를 입증하기 위해 대표적인 AI 모델인 YOLOv6을 사용해 추론을 수행했습니다. YOLOv6은 객체를 동시에 분류하고 탐지하는 최신 합성곱 신경망(CNN) 기반 객체 탐지 모델로, 이전 YOLO 시리즈의 개선 및 최적화를 바탕으로 정확하고 빠른 실시간 탐지를 구현합니다.

테스트는 리벨리온의 AI 가속기 ATOM™을 탑재한 PCIe 카드 RBLN-CA12에서 수행되었으며, NVIDIA RTX A5000 GPU와 성능을 비교했습니다. 비교 지표는 다음과 같습니다.

- Watts: 전력 소비량. 대규모 배포 환경에서는 운영 비용과 효율성에 직접적인 영향을 미칩니다.
- Joules per frame: 전체 에너지 효율을 나타내는 지표로, 프레임당 소비된 에너지량을 나타냅니다.

	ATOM™	A5000
Power Consumption (W)	avg. 40	avg. 110
Energy Consumption (J/Frame)	avg. 0.64	avg. 2.82

[Table 1. 결과]
* Input resolution: 640×640

테이블 1의 결과는 YOLOv6-Large 모델을 실행하는 데 있어 ATOM™의 뛰어난 성능을 명확히 보여줍니다. ATOM™은 RTX A5000에 비해 최대 **2.1배의 성능**과 **4.5배의 에너지 효율성**을 달성하였습니다.

효율성과 성능: 점점 커지는 이점

하드웨어와 소프트웨어 수준에서 자원 활용과 워크로드 균형을 효율적으로 관리함으로써, 칩의 특정 영역에서 과도한 전력 소모나 열 집중 현상을 방지합니다.

RBLN 소프트웨어 스택은 동시성(concurrency)을 효과적으로 제어하여 칩의 뉴럴 엔진이 제공하는 병렬 처리 성능을 극대화합니다. 또한 데이터 플로우와 실행 파이프라인이 최적화되어, 처리 유닛이 항상 데이터를 공급받도록 설계되었습니다. 여기에 저수준 연산의 효율적인 구현이 더해져 연산 부하와 지연(latency)을 줄였습니다.

이러한 효율성과 성능 최적화는 비즈니스적 측면에서도 중요한 의미를 갖습니다. 1,500W급 서버는 RBLN-CA12를 16장 탑재할 수 있지만, GPU는 4장만 장착할 수 있습니다. 따라서 개별 카드 단위에서 확보된 효율성은 다수의 서버를 운영하는 환경에서 더욱 큰 누적 효과를 가져옵니다.

결론

산업 전반에서 AI 활용 사례가 급격히 증가함에 따라, 적절한 하드웨어 설계가 중요해지고 있습니다. 그러나 종종 간과되는 또 하나의 핵심 요소는 **최적화된 소프트웨어**입니다. 리벨리온의 소프트웨어 스택은 하드웨어의 유연성과 성능을 극대화하기 위한 숨은 동력으로 작용합니다.YOLOv6 모델을 실행한 RBLN-CA12와 NVIDIA RTX A5000의 성능 비교 결과는, 최적화된 소프트웨어와 고성능 하드웨어의 결합이 가져오는 뚜렷한 성능 우위를 보여줍니다. 이는 리벨리온이 **최첨단 AI 솔루션을 제공하기 위한 확고한 의지**를 입증합니다.