제품 개요

AI 가속
인텔® 제온® 스케일러블 프로세서

인텔® 제온® 스케일러블 프로세서의 내장된 가속기로 전체 AI 파이프라인의 성능 부스트

70%의
데이터 센터 AI 추론이
인텔® 제온® 프로세서에서 실행 중[1]

10개 중 9개 기업
애플리케이션이
2025년까지
AI 포함 예정[2]

AI 는 데이터 분석 및 클래식한 머신 러닝에서 언어 처리 및 이미지 인식에 이르기까지 넓은 범위의 워크로드와 사용 사례를 포괄합니다. 인텔® 제온® 스케일러블 프로세서는 전체 AI 파이프라인을 위한 유연한 컴퓨팅 성능뿐만 아니라 데이터 과학, 모델 학습 및 딥 러닝 추론 등의 특정 AI 워크로드를 위한 내장형 가속기를 포함하고 있습니다.

AI 는 딥 러닝보다 더 크며, 점점 더 거대해지고 있습니다

AI 는 아직 초기 단계로 모든 면에서 급성장 중입니다. 기존의 머신 러닝 알고리즘과 딥 러닝 모델은 핵심 엔터프라이즈 애플리케이션에서 자동화된 음성 비서 기능에 이르기까지 비즈니스가 이루어지는 방법에 대한 기본적인 빌딩 블록을 갖춰가고 있습니다. AI 작업은 데이터 과학에서 교육, 유효성 검사 및 최종 배포까지 흐르는 긴 개발 파이프라인에 따라 다양한 규모로 작업할 수 있습니다. 각 단계에는 자체 개발 툴체인, 프레임워크 및 워크로드가 있습니다. 이 모든 단계는 특별한 병목 현상을 일으키고 추가적인 컴퓨팅 리소스를 필요로 하게 됩니다. 인텔 제온 프로세서는 AI 가속을 위한 내장된 가속기를 포함하고 있어 이러한 문제를 극복하고 AI 파이프라인 전반에 걸친 성능을 향상할 수 있습니다.

AI 는 실제로는 단지 수학입니다. 엄청난 양의 계산입니다

어마어마한 양의 수학 계산은 모든 AI 작업 및 운영의 핵심입니다. 모델링 데이터 및 머신 러닝 알고리즘과 같은 많은 데이터 과학 작업은 통계, 대수학 그리고 복잡한 벡터 수학에서 실행됩니다. 딥 러닝 AI 는 방대한 양의 행렬 곱셈이 필요합니다. 이러한 모든 AI 애플리케이션은 CPU, GPU, FPGA 및 워크로드별로 맞춤 제작된 ASIC를 포함한 대규모 데이터 세트와 광범위한 프로세싱 리소스를 포함하는 엄청난 작업입니다.

인텔® 어드밴스드 벡터 익스텐션 512(인텔® AVX-512) - AI 속도를 향상하는 치트키

인텔 제온 코어는 웹사이트를 위한 SSL 암호화를 해시하고, 대규모 데이터베이스를 처리하며, 제약 연구와 칩 디자인 또는 포뮬러원 엔진용 시뮬레이션을 실행할 수 있습니다. 인텔 제온 코어는 만능이지만, 딥 러닝 학습(전체 AI 파이프라인의 일부)에서는 전용 가속기만큼 빠르지 않습니다. 이는 CPU가 한 번에 하나의 계산으로 순차적인 작업을 처리하기 때문입니다. 다른 프로세서 유형은 작업을 병렬로 처리할 수 있습니다. 즉, 동시에 다수의 계산을 처리할 수 있습니다.

인텔® AVX-512는 각 클록 주기에 더 많은 작업을 할당하여 CPU의 구조적 한계를 극복합니다. 이를 통해 CPU는 병렬 프로세서처럼 더 많은 작업을 수행할 수 있습니다.

고객 성공 사례 – 인텔® 제온® 스케일러블 프로세서의 가속

Tencent Cloud는 3세대 인텔 제온 스케일러블 프로세서를 통해 실시간 음성 합성을 제공합니다.

자세한 내용 확인하기 ›

BeeKeeperAI는 데이터 개인정보를 보호하면서 임상 AI 알고리즘을 개발합니다.

스토리 읽기 ›

복잡한 CPU 명령어, 단순한 전략: 싸이클당 더 많은 작업 수행으로 더 스마트한 작업이 가능합니다

인텔 AVX-512의 익스텐션은 CPU가 무엇을 어떻게 할지를 지시하는 명령어 집합입니다. CPU의 작업 방식은 매우 복잡하지만, AVX-512의 기본 로직은 매우 단순합니다. 먼저, 여러 단계를 가능한 적은 단계의 작업으로 압축합니다. 두 번째는 CPU가 클록 주기별로 더 많은 작업을 수행할 수 있도록 도와줍니다.

단계가 줄어들면 처리 속도는 더 빨라집니다

연산은 매우 스마트하면서도 우아해질 수 있습니다. 인텔 AVX-512는 많은 수의 스마트하고 아름다운 계산을 사용하여 일반적인 컴퓨팅 작업을 더 적은 단계로 압축, 결합 및 융합합니다. 아주 단순한 예를 들자면, 다섯 번의 연산이 필요한 3x3x3x3x3을 계산하도록 CPU에게 명령할 수도 있지만, 이렇게 하는 대신 CPU가 한번의 연산으로 수행할 수 있도록 3^5에 대한 명령을 할 수도 있습니다. AVX-512는 이러한 로직을 갖추고 가장 어려운 AI 작업 일부를 포함하여 수백 개의 워크로드별 작업에 이 로직을 적용합니다.

여덟개씩 세는 게 하나씩 세는 것보다 훨씬 빠릅니다

AVX-512에서 “512”는 클록 주기별로 CPU가 처리할 수 있는 비트 수를 증가시키는 두 번째 방법을 의미합니다. 40년 전에는 16비트 PC가 상당히 인상적이었습니다. 하지만 곧 32비트 머신이 그 자리를 차지하였습니다. 지금은 스마트폰이 64비트로 실행되고 있습니다. 비트 수는 클록 주기당 CPU가 처리할 수 있는 레지스터의 수, (레지스터는 데이터를 가지고 있는 메모리 슬롯)를 의미합니다. AVX-512는 512비트로 레지스터 수를 확장합니다. 상상이 되시나요? 애플리케이션이 인텔 AVX-512를 사용할 경우 레지스터 수를 확장하여 CPU의 기본 64비트 속도보다 최대 8배 빠르게 실행됩니다. 96까지 숫자를 셀 때, 1, 2, 3… 으로 세는 것과 8, 16, 24… 으로 세는 것을 비교하는 것과 같습니다.

인텔® 딥 러닝 부스트(인텔® DL 부스트) - 신경망을 위해 더욱 스마트해진 연산

딥 러닝 AI 는 엄청난 양의 행렬 곱셈을 이용하여 신경망 모델을 훈련하고 추론이라는 방법을 사용하여 이 모델을 실제 작업에 적용합니다. 추론하는 동안 컴퓨터는 수신하는 데이터(예: 음성을 포함한 오디오 신호)를 모델(이 경우 음성 인식 모델)과 비교하고 해당 데이터가 무엇을 의미하는지를 추론합니다. 추론은 객체 인식, 이미지 분할, 텍스트 인식 및 거의 모든 기타 딥 러닝 AI 작업에 사용됩니다.

딥 러닝 모델을 훈련하는 데에는 몇 시간 또는 몇 일이 걸릴 수 있습니다. 딥 러닝 추론은 1초 미만부터 수 분까지 걸릴 수 있으며, 모델의 복잡성과 얼만큼 정확한 결과를 원하는지에 따라 달라집니다. 훈련이나 추론을 데이터 센터 수준의 컴퓨팅 정도로 확장하는 경우 시간, 에너지 및 성능 예산은 엄청나게 증가하게 됩니다.

인텔 DL 부스트는 여러 가지 인텔 AVX-512 명령어를 사용하여 딥 러닝 워크로드를 가속화 합니다. 세 개의 작업을 하나의 벡터 신경망 명령어(VNNI) 집합으로 결합함으로써 클록당 연산의 수를 줄입니다. 인텔 DL 부스트는 또한 INT8 정밀도를 이용하여 딥 러닝 워크로드를 가속화 합니다.

앞으로의 발전은 AI 성능을 더욱 가속화시킬 것입니다

4세대 인텔® 제온® 스케일러블 프로세서에는 딥 러닝 워크로드의 핵심인 행렬 곱셈 전용 가속기가 내장되어 있습니다. 인텔® AMX (인텔® 어드밴스드 매트릭스 익스텐션)에는 큰 행렬을 단일 연산으로 바꾸는 새로운 명령어 세트와 각 코어에 큰 데이터 청크를 저장하는 2차원 레지스터 파일이 결합되어 있습니다.

더 빠른 AI 는 인텔 제온 프로세서로 자동화됩니다

인텔 제온 스케일러블 프로세서의 AI 가속기능은 CPU의 ISA (명령어 집합 아키텍처)에 내장되어 있습니다. 이것은 AI 가속기능이 모든 소프트웨어에 준비되어 있고 사용가능하다는 것을 의미합니다. 우리는 데이터 과학자들과 AI 개발자들이 사용 중인 툴을 재코딩하고 인텔 AVX-512용으로 다시 컴파일 하는것을 바라지 않습니다. 인텔이 이런 일을 가능하게 합니다.

인텔의 소프트웨어 엔지니어는 지속적으로 오픈 소스 AI 툴체인을 최적화하고 최적화한 내용을 커뮤니티에 배포합니다. 예를 들어, TensorFlow 2.9는 인텔® oneDNN (인텔® oneAPI Deep Neural Network Library) 최적화를 기본으로 제공합니다. TensorFlow 최신 버전을 다운받으면, 인텔 최적화의 잇점을 바로 사용할 수 있습니다.

AI 파이프라인의 기타 애플리케이션의 경우, 데이터 과학자와 개발자는 인텔의 3세대 인텔 제온 스케일러블 프로세서용 ISA의 모든 내장형 가속기에 대한 이점이 있는 무료 오픈소스 인텔 배포판과 라이브러리 및 개발 환경을 다운로드할 수 있습니다.

기본적으로 인텔 하드웨어에서 더욱 빠른 AI 구현은 사용자가 이미 사용하고 작업하고 있는 툴에 대해 인텔 버전을 다운로드하는 것 만으로도 가능해 집니다.

자세히 알아보기

인텔 제온 스케일러블 프로세서의 AI 및 딥 러닝 ›
인텔 AVX-512 ›
인텔 입 러닝 부스트 ›
인텔 AI 분석 툴킷 ›

AI 파이프라인 애플리케이션의 소프트웨어 최적화 이점

38~200배
더 빠른 사이킷런
사이킷런 용
인텔 익스텐션[3]
이용 시

~90배
더 빠른 판다스
인텔® 배포판
모딘[3] 이용 시

최대
3배 더 빠른
TensorFlow
인텔 oneDNN[3]
이용 시

3세대 인텔® 제온® 스케일러블 프로세서의 AI 가속화

딥 러닝 AI 워크로드를 위한 속도 향상

최대
1.74배
더 높은 INT8 추론 처리량
이전 세대[4] 대비 3세대 인텔® 제온® 스케일러블 프로세서의 인텔 DL 부스트를 이용하는 BERT-Large SQuAD

최대
1.59배
더 높은 INT8 실시간 추론 처리량
이전 세대[5] 대비 **3세대 인텔® 제온®** 스케일러블 프로세서의 인텔® DL 부스트 이용 시

최대
4.5배
INT8[6]에서 더 많아진 초당 이미지 및 BF16[7] 객체 감지에서 최대 6배 더 많아진 초당 이미지
(SSD-ResNet-34)
4세대 인텔® 제온 스케일러블 프로세서의 인텔® AMX 사용시

지금 AI 및 머신 러닝용 인텔 최적화로 클라우드 또는 자체 인프라에서 AI 워크로드 가속화를 시작해 보십시오.

자세히 보기 ›

intel XEON

1. 2021년 12월 기준으로 AI 추론 워크로드를 실행하는 데이터 센터 서버의 전 세계적으로 설치한 기본 인텔 마켓 모델링에 따라 달라집니다.
2. "IDC FutureScape: 전 세계 IT 산업의 2020년 예측(2019년 10월 기준)." Doc #US45599219.
3. 부스트 판다스, 사이킷런 및 TensorFlow 성능으로 변화하는 One-Line 코드(2021년 7월)." intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html
4. intel.com/3gen-xeon-config에서 [123]을 참조하십시오. 결과는 상이할 수 있습니다.
5. intel.com/3gen-xeon-config에서 [122]를 참조하십시오. 결과는 상이할 수 있습니다.
6. edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/에서 [41]을 참조하십시오. 결과는 상이할 수 있습니다.
7. edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/에서 [42]를 참조하십시오. 결과는 상이할 수 있습니다.

0922/MP/CMD/PDF