

Alibaba 클라우드 주소 정확화의 AI 추론 성능을 향상시키는 인텔® AMX(인텔® Advanced Matrix Extensions)

인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서는 전체적인 추론 성능이 이전 세대보다 2.48배 더 우수합니다.¹



딥 러닝(DL)은 중요한 인공지능(AI) 기법으로서 컴퓨터 비전(CV), 자연어 처리(NLP) 및 추천 시스템과 같은 여러 분야에서 널리 구현되어 왔습니다. 하지만 데이터가 폭발적으로 증가하고 DL 모델이 점점 복잡해짐에 따라 추론을 프로덕션 환경에서 사용하려면 어려울 수 있습니다. 사용자는 하드웨어, 소프트웨어 및 알고리즘 최적화를 통해 성능 개선과 전체적인 비용 절감을 기대합니다. DL 추론을 최적화하면 사용자가 더 복잡한 DL 모델을 채택하여 정확성을 개선하면서도 대기 시간을 똑같이 유지할 수 있습니다.

주소 정확 서비스의 성능을 개선하기 위해, Alibaba Cloud의 머신 러닝 플랫폼(PAI) 및 DAMO 아카데미(Alibaba Academy for Discovery, Adventure, Momentum and Outlook) NLP 팀은 인텔과 협업했습니다. 인텔® AMX를 탑재한 4세대 인텔® 제온® 스케일러블 프로세서를 최적화 도구와 함께 사용하자 이전 세대 플랫폼을 사용할 때보다 전체적인 추론 성능이 최대 2.48배까지 개선되었습니다.¹

Alibaba 클라우드 주소 정확

주소 정확화는 자동으로 우편 주소를 표준화하고 수정하고 확인하는 프로세스입니다. 이 프로세스는 물류, 전자상거래, 소매업 및 금융을 포함한 여러 산업에서 사용됩니다. Alibaba 클라우드 주소 정확화는 Alibaba DAMO 아카데미의 NLP 팀이 Alibaba Cloud의 거대한 주소 컬렉션을 기반으로 개발한 효율적인 표준 AaaS(서비스형 주소 알고리즘)입니다.² 더 빨라진 전체적인 성능은 Alibaba Cloud 고객의 비즈니스 실적 개선으로 이어집니다. 이 AaaS는 주소 데이터 프로세싱을 위한 원스톱 폐쇄 루프 서비스 플랫폼입니다. AaaS는 NLP 알고리즘을 사용하여 비즈니스 시스템에 등록된 주소 데이터를 수정하고 완성하고 정규화하고 구조화하고 라벨링합니다. 그리고 20가지가 넘는 주소 서비스를 제공하며,³ 퍼블릭, 프라이빗 또는 하이브리드 클라우드에 배포할 수 있습니다. Alibaba Cloud의 목표는 다음과 같습니다.

- 데이터 정확 및 모델 추론 같은 여러 워크로드를 전체적으로 고려하여 플랫폼의 원스톱 성능 가속
- 기존 하드웨어 리소스를 효율적으로 사용하고 퍼블릭, 프라이빗 및 하이브리드 클라우드에서 고객의 서버 리소스를 안전하게 사용하여 하드웨어 비용 절감

인텔® 기술로 Alibaba 서비스 최적화

BERT(Bidirectional Encoder Representations from Transformers) 모델은 인공지능(AI) 프로그램이 텍스트에서 불명료한 단어의 문맥을 이해하는 데 도움이 되는 자연어 처리(NLP) 딥 러닝 기법입니다. Alibaba는 BERT를 자사 주소 정확 서비스의 검색 모듈로 사용합니다.⁴ BERT는 멀티태스크 벡터 리콜과 미세 정렬(fine sorting)에 사용됩니다. 인텔은 솔루션 성능을 크게 가속하는 데 도움이 될 수 있는 다양한 솔루션을 제공합니다.

인텔® AMX

4세대 인텔® 제온® 스케일러블 프로세서에는 Alibaba 클라우드 주소 정화 솔루션이 뛰어난 성능과 비용 대비 효과와 확장성을 달성하는 데 도움이 되는 인텔® AMX라는 내장형 가속기가 있습니다. 인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서는 추천 시스템, 자연어 처리 및 소매업 전자상거래 소프트웨어 솔루션을 포함하는 광범위한 DL 이용 사례에 배포할 수 있습니다.

데이터 센터 아키텍처의 새로운 표준

확장성을 위한 멀티타일 Soc

물리적으로 타일식, 논리적으로 모듈리식

범용 및 전용 가속 엔진

클라우드, 마이크로서비스 및 AI 워크로드를 위한 설계

성능 코어 아키텍처

워크로드에 특수화된 가속

고급 메모리 및 I/O 전환을 사용하는 선구적인 기술

DDR5와 HBM

PCIe 5.0

향상된 가상화 기능

인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서

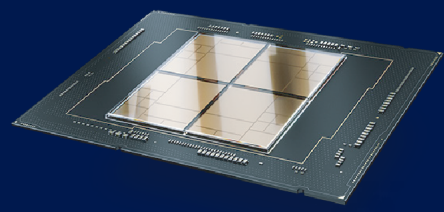


그림 1. 인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서

인텔® AMX

새로운 내장형 가속 엔진

2세대 인텔® 제온® 스케일러블 프로세서	3세대 인텔® 제온® 스케일러블 프로세서	4세대 인텔® 제온® 스케일러블 프로세서
인텔® 딥 러닝 부스트(첫 도입) 인텔® AVX-512(인텔® Advanced Vector Extensions 512) (VNNI/INT8)	인텔® DL Boost 인텔® AVX-512: VNNI/INT8(CPX/ICX) 및 BFloat16(CPX)	인텔® AMX INT8 및 BFloat16 지원 인텔® AVX-512(VNNI/INT8)
주요 이점 <ul style="list-style-type: none"> ▪ 광범위한 하드웨어(전용 실리콘/TILE와 행렬 곱셈 명령어 집합/TMUL) 및 소프트웨어(시장 관련 프레임워크, 툴킷 및 라이브러리 등) 최적화로 인텔® 제온® 스케일러블 프로세서에 내장된 AI 가속 개선 ▪ 인텔® AMX, INT8(추론) 및 BFloat16(학습/추론) 데이터형식 지원 		
대상 워크로드/용도 <ul style="list-style-type: none"> <li style="width: 33%;">▪ 이미지 인식 <li style="width: 33%;">▪ 기계/언어 번역 <li style="width: 33%;">▪ 자연어 처리(NLP) <li style="width: 33%;">▪ 미디어 분석 <li style="width: 33%;">▪ 추천 시스템 <li style="width: 33%;">▪ 강화 학습 <li style="width: 33%;">▪ 미디어 프로세싱 및 전송 		
정의 <ul style="list-style-type: none"> ▪ 이전 세대 인텔® 제온® 스케일러블 프로세스 대비 AI/딥 러닝 추론 및 학습 워크로드 성능 크게 향상 		

그림 2. 인텔® AMX 개요

Blade: 추론 최적화를 위한 일반적인 도구

Alibaba 클라우드 주소 정화 솔루션은 Alibaba 클라우드 머신 러닝 PAI 팀이 내놓은 일반적인 추론 최적화 도구인 Blade를 사용하여 주소 정화의 추론 성능을 최적화합니다. Blade는 계산 그래프 최적화, 인텔® oneDNN(인텔® oneAPI Deep Neural Network Library) 같은 최적화 라이브러리, BladeDISC 컴파일러, Blade 고성능 연산자 라이브러리, 인텔® 커스텀 백엔드 및 Blade 믹스드 프리시전 등 여러 최적화 방법을 통합합니다.

Blade에 인텔® 커스텀 백엔드 통합

인텔® 커스텀 백엔드⁵는 Blade의 소프트웨어 백엔드로서 모델의 양자화 및 희소화 추론 성능을 높입니다. 인텔® 커스텀 백엔드는 크게 세 개의 최적화 단계를 포함합니다. 첫째, 기본 캐시 전략을 사용하여 메모리를 최적화하고, 둘째, 그래프 융합을 최적화하고, 마지막으로 연산자 수준에서 커스텀 및 스파스 커널을 포함한 효율적인 오퍼레이터 라이브러리를 만듭니다.

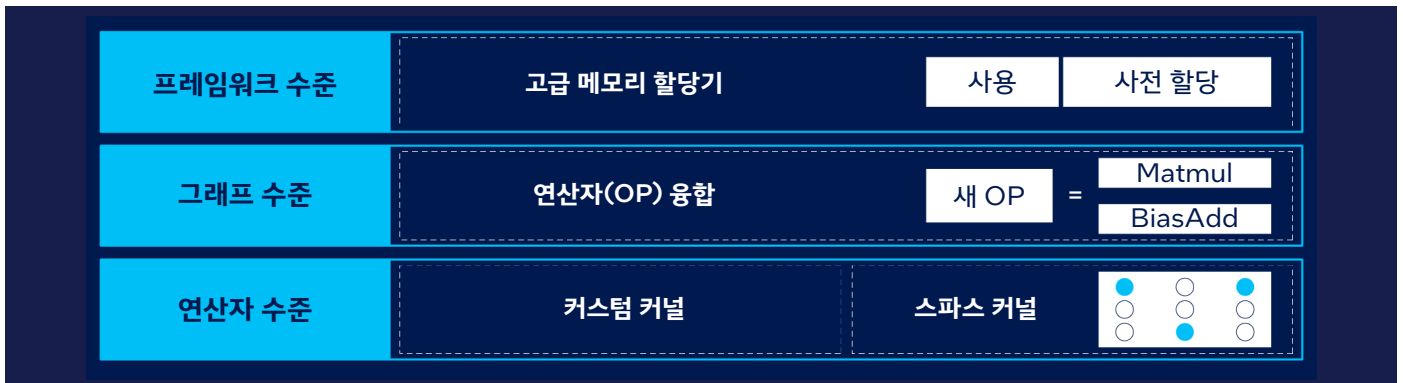


그림 3. 인텔® 커스텀 백엔드의 구조

인텔® 제온® 스케일러블 프로세서는 2세대부터 VNNI를 INT8 데이터 형식에 따라 AI 성능을 최적화하고 모델 양자화 솔루션에서 널리 사용되는 INT8 양자화를 위해 특별히 제공하기 시작했습니다.

인텔® AMX는 INT8의 기능을 크게 개선하며, 인텔® oneDNN을 사용하여 지원됩니다. 인텔® AMX 기반 INT8 양자화는 모델 성능을 VNNI에 비해 크게 개선할 수 있습니다. 그림 4에는 인텔® AMX의 작동 방식이 설명되어 있습니다.

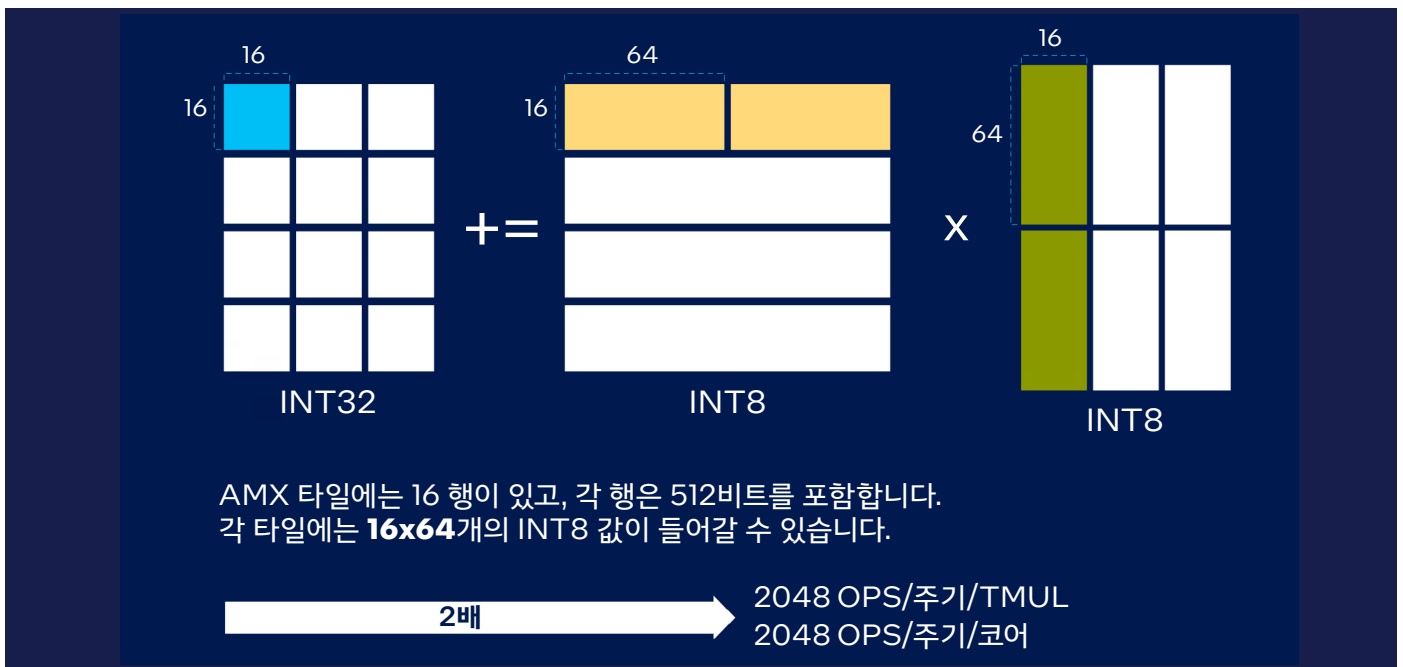


그림 4. 인텔® AMX(인텔® Advanced Matrix Extensions) 기능

성능 및 최적화 게인

Alibaba Cloud와 인텔은 주소 정화 모델을 튜닝하여 모델의 추론 성능을 개선해 인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서를 사용하는 PAI로 성능을 이전 세대 플랫폼 대비 2.48배 높였습니다.¹ 인텔® AMX 기반 인텔 커스텀 백엔드는 도형 크기가 (10 x 53으로) 고정된 4층 BERT 모델을 최적화하여 이 성능 향상을 실현합니다. 그림 5를 참조하십시오.

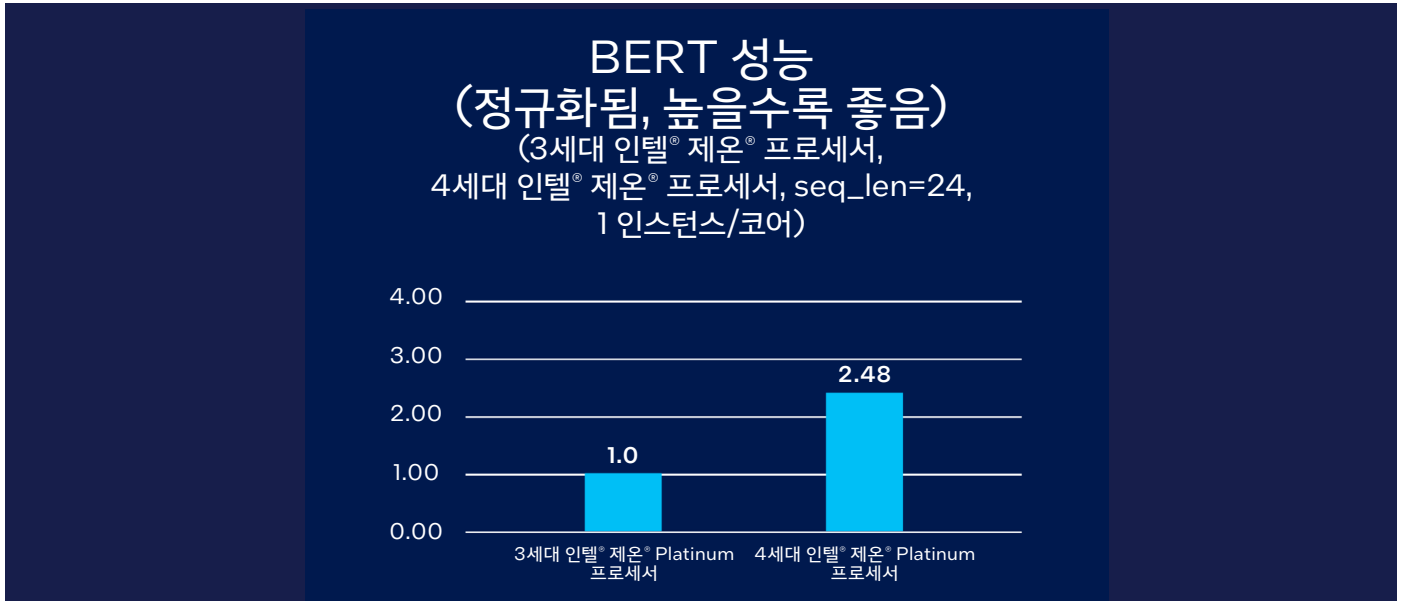


그림 5. BERT 모델의 추론 성능¹

CCKS2021 중국어 NLP 주소 상관 작업을 사용하여 모델을 확인했습니다. 부동 소수점 32(FP32) 기반 최적화 점수는 78.72였고, INT8 기반 최적화 점수는 78.85였습니다.⁶ 점수는 높을수록 좋습니다.

요약

Alibaba Cloud는 인텔® AMX가 탑재된 4세대 인텔® 제온® 스케일러블 프로세서를 사용하여 자사 주소 정화 서비스의 AI 추론을 최적화했습니다. 빨라진 전체적인 성능은 Alibaba의 물류, 전자상거래, 에너지, 소매업 및 금융 고객의 비즈니스 실적 개선으로 이어집니다. 인텔® AMX는 Alibaba가 독립형 GPU 같은 전용 가속기를 배포했을 경우에 회사에 발생했을 수 있는 간접비도 줄입니다. 내장형 가속기를 사용하여 Alibaba는 자사 주소 정화 서비스의 총 소유 비용(TCO)을 통제할 수 있습니다.

추가 DL 모델의 전체적인 성능을 높이기 위해, 인텔과 Alibaba는 자사 고객과 함께 소프트웨어 및 하드웨어 통합을 최적화하기 위한 협력을 확장하고 있습니다. 목표는 DL 모델의 성능을 가속하고 인텔 기술의 가치를 최대한 활용하는 것입니다. 인텔은 또한 업계 파트너들과 더 심층적으로 협력하고 AI 기술의 배포와 구현에 기여할 수 있기를 기대합니다.



¹ 구성: 기준: 2022년 10월 19일에 인텔에서 실시한 테스트. 1노드, 2x 3세대 인텔® 제온® Platinum 프로세서, 인텔® 하이퍼스레딩 기술(인텔® HT 기술) 사용, 인텔® 터보부스트 기술 사용, 총 메모리 256GB (16슬롯/16GB/3,200MT/s[작동 속도 3,200MHz]), WLYDCRB1.SYS.0029.P30.2209011945, 0xd00037b, CentOS Linux 8, 4.18.0-305.12.1.el8_4.x86_64, GCC 8.5.0, NLP 톨킷 v0.3, Pytorch 1.11, BERT-mini, INC 1.13, transformer 4.18.0, 1 인스턴스/코어, BS=32, seq_len=24, 데이터 형식: INT8

NEW-1: 2022년 10월 19일에 인텔에서 테스트. 1노드, 2x 4세대 인텔® 제온® Platinum 프로세서, 인텔® HT 기술 사용, 인텔® 터보부스트 기술 사용, 총 메모리 256GB(16 슬롯/16GB/4,800MHz [작동 속도 4,800MHz]), EGSDCRB1.SYS.0090.D03.2210040200, 0x2b0000c0, CentOS Stream 8, 5.15.0-spr.bkc.pc.8.8.5.x86_64, GCC 8.5.0, LP 톨킷 v0.3, Pytorch 1.11, BERT-mini, INC 1.13, transformer 4.18.0, 1 인스턴스/코어, BS=32, seq_len=24, 데이터 형식: INT8

² Alibaba Cloud. "Address Normalization." alibaba.com/product/addresspurification/addrp.

³ Alibaba Cloud. "What is Address Normalization?" https://help.aliyun.com/document_detail/169746.html.

⁴ Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." ACL Anthology. 2019년 6월. <https://aclanthology.org/N19-1423/>.

⁵ GitHub. "Intel Neural Compressor." https://github.com/intel/neural-compressor/commits/inc_with_engine.

⁶ Alibaba Cloud. "'Intel Innovation Master Cup' Deep Learning Challenge Track 3: CCKS2021 Chinese NLP Address Correlation Task." 2021년 11월 <https://tianchi.aliyun.com/competition/entrance/531901/introduction>.

성능은 사용, 구성 및 기타 요인에 따라 다릅니다. www.intel.com/PerformanceIndex에서 자세히 알아보십시오.

성능 결과는 구성에 표시된 날짜의 테스트를 기반으로 하며 공개된 모든 업데이트가 반영되어 있지 않을 수도 있습니다. 구성 백업 상세 정보를 확인하십시오. 어떤 제품 또는 구성 요소도 절대적으로 안전할 수는 없습니다.

비용과 결과는 다를 수 있습니다.

인텔® 기술은 지원되는 하드웨어, 소프트웨어 또는 서비스 활성화가 필요할 수 있습니다.

인텔은 상품성, 특정 목적에의 적합성 또는 비침해에 대한 묵시적인 보증과 성능, 거래 과정 또는 거래 사용으로 발생하는 모든 보증을 포함하여(이에 제한되지 않음) 다른 모든 명시적 또는 묵시적 보증을 부인합니다.

인텔은 타사 데이터를 제어하거나 감사하지 않습니다. 정확성을 평가하려면 기타 소스를 참고해야 합니다.

© 인텔사. 인텔, 인텔 로고 및 기타 인텔 마크는 인텔사 또는 그 자회사의 상표입니다. 기타 명칭 및 브랜드는 해당 소유업체의 자산일 수 있습니다.