



인텔 의료, 헬스 업계 인공지능 실전 매뉴얼

인텔® 제온® 확장성 플랫폼
고효율 실행 AI



AI 실행을 가속화하고 싶다면, www.intel.com/ai 를 방문해 주십시오





목차 CONTENTS

업계 동향

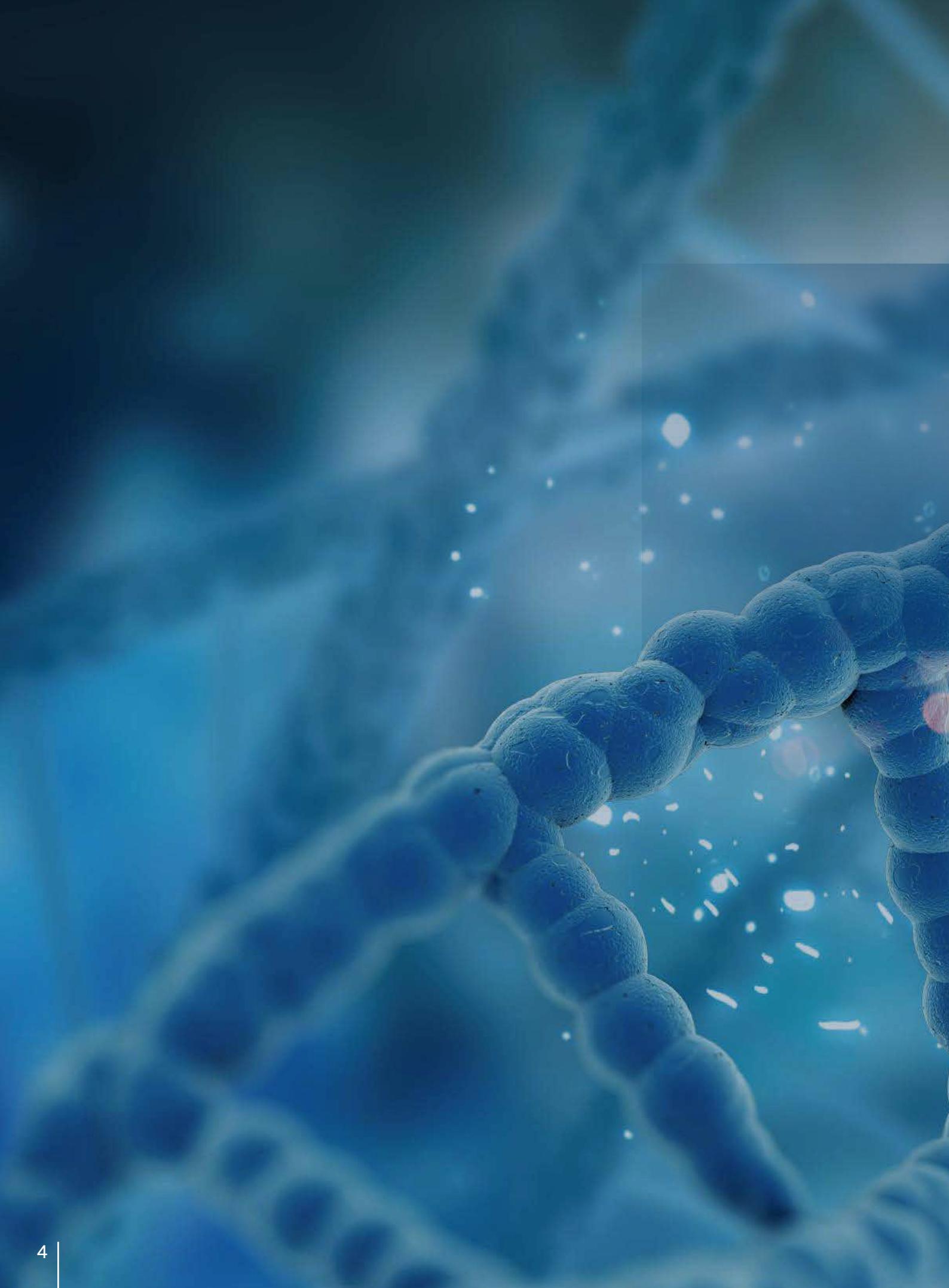
- 의료, 헬스 산업에서 인공지능의 발전과 응용
 - 의료, 헬스 영역에서 인공지능의 발전
 - 의료, 헬스 영역에서 인공지능의 응용 시나리오

실전편

- 의료업계의 이미지 분할 AI 추론 응용 가속화
 - 의료 영상 처리 중 이미지 분할
 - U-net 분할 네트워크 최적화 방법
 - 2세대 인텔® 제온® 확장성 CPU를 기반으로 구축한 Dense U-net 이미지 분할 방식
 - 응용 사례
- 더 효율적인 의료 영상 분석을 가능하게 하는 AI + Cloud
 - 의료 영역에서의 의료 영상 분석
 - 최적화 AI 모델 효율
 - AI 기술과 클라우드 서비스를 이용해 의료 치료 보조 기능을 향상시킨 AccuRad
- 조직 분석을 가속화한 AI 기술
 - 의료 영역의 조직 분석
 - 딥러닝 기반 조직 분석 방법의 최적화
 - AI 기술을 이용해 자궁경부암 검사 효율을 향상시킨 장평생물정보기술
- 약물 연구 개발을 돋는 AI 기술
 - 약물 선별을 가속화한 딥러닝 방식
 - 인텔® 제온® 확장성 CPU 기반 최적화
 - 딥러닝을 이용해 약물 연구개발 효율을 향상시킨 노바티스
- 의료 업계에 응용된 AI 기반 화상 인식 기술
 - 스마트 의료와 화상 인식 기술
 - 외래 진찰 약품 지급에 딥러닝 기술을 이용한 해방군총병원

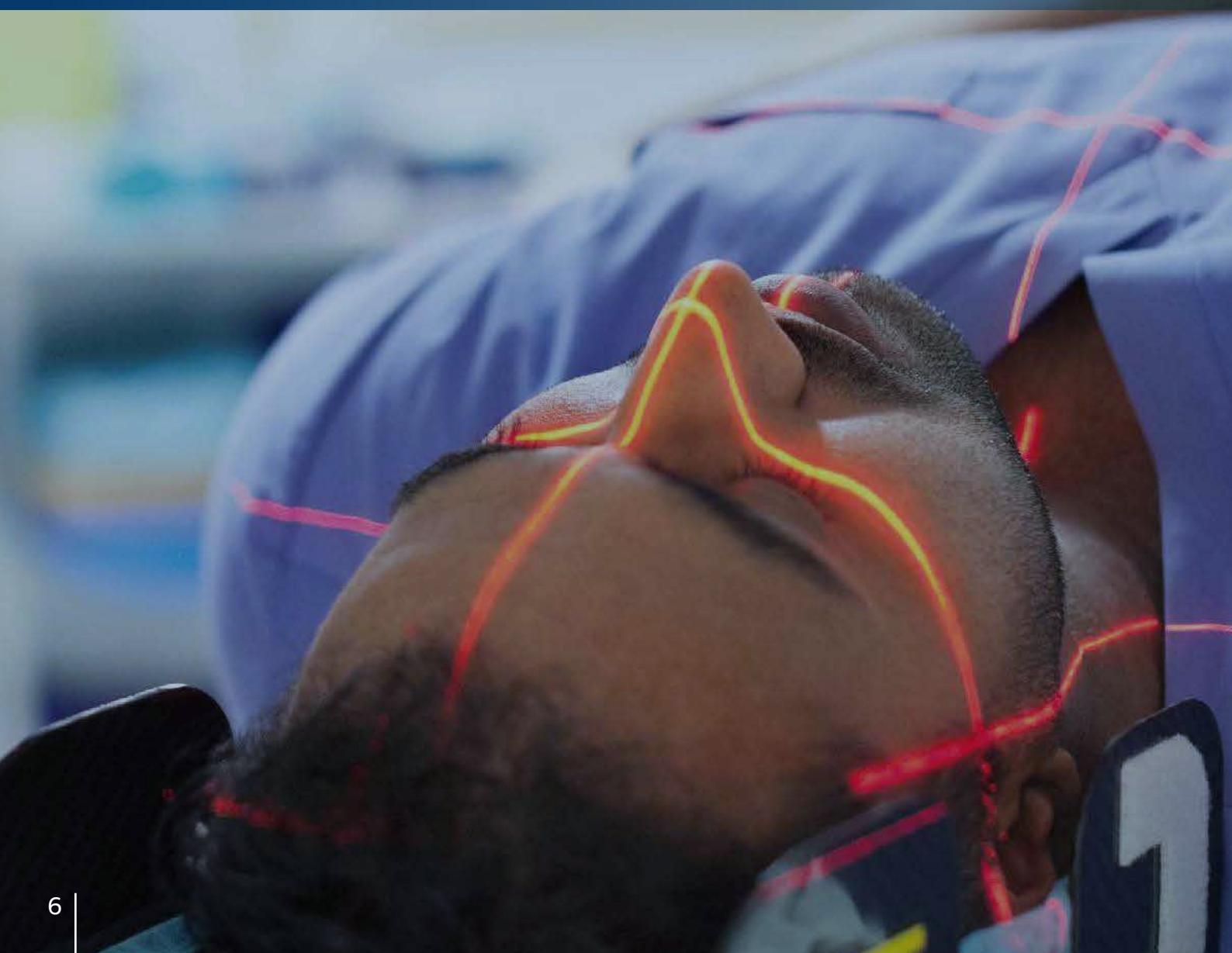
기술편

- 하드웨어 제품
 - 2 세대 인텔® 제온® 확장성 CPU
 - 인텔® 옵테인™ 데이터 센터급 영구 메모리
 - 인텔® 옵테인™ SSD와 인텔® QLC 3D NAND 기술 기반 인텔® SSD
- 소프트웨어와 아키텍처
 - DNN 지향 인텔® MKL
 - 인텔® 아키텍처 최적화 지향 Caffe
 - 인텔® 아키텍처 최적화 지향 TensorFlow
 - 인텔® 아키텍처 최적화 지향 Python 배포 패키지
 - OpenVINO™ 툴킷



업계 동향

의료, 헬스 산업에서 인공지능의 발전과 응용



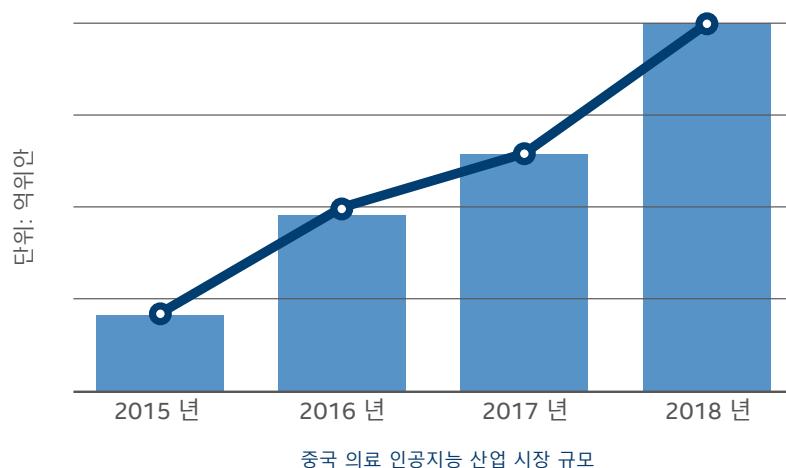
의료, 헬스 영역에서 인공지능의 발전

의료 인공지능 시장 동향

알고리즘이 한층 더 성숙해지고 향상되며 데이터가 누적됨에 따라 인공지능 (Artificial Intelligence, AI) 역시 빠르게 발전했습니다. 특히 딥러닝 분야에서 급속한 발전을 보이고 있으며, 그 응용 분야는 갈수록 특정 산업에 더 초점을 맞추고 있습니다. 2018년 9월 중국 정보통신연구원이 발표한 《2018 세계 인공지능 산업 보고서》에 따르면, 인공지능은 의료 헬스, 금융, 비즈니스, 교육과 안보 등과 같은 중국의 유형별 수직 업계 중 다양한 분야에 깊이 침투해 있음을 알 수 있습니다. 그 중, 전체 AI 기업 중 의료 헬스 분야 관련 AI 기업이 차지하는 비율은 22%¹로 가장 높습니다.

의료 헬스 업계는 주요 응용 영역 중 하나로서 인공지능 기술에 대한 투자 역시 빠른 속도로 성장하고 있습니다. 전망산업연구원이 발표한 《2018-2023년 중국 인공지능 업계 시장 전망과 투자 전략 분석 보고서》에 따르면, 2016년 중국 의료 분야의 인공지능 시장 규모는 96.61 억 위안으로 37.9% 성장했습니다. 2017년은 130억 위안을 초과해 40.7% 성장했으며, 2018년은 200억 위안을 달성할 거라 기대하고 있습니다. 이렇게 고속 성장할 수 있었던 것은 한편으로는 중국 의료 시장의 절박한 수요 덕분이기도 하지만, 또 한편으로는 최근 의료 분야의 인공지능 기술 발전과 관련 정책에 대한 지원에서 비롯된 것이기도 합니다.

중국의 의료 인공지능 응용 영역은 글로벌 기준과는 다소 다르게 세분화되어 있습니다. Global Market Insight의 통계 자료에 따르면, 약물 연구개발이 전세계 의료 인공지능 시장에서 차지하는 비율은 최대 35%까지 다다릅니다. 의학 영상 인공지능 (점유율 25%)이 그 뒤를 바짝 따르고 있으며, 40% 이상의 속도로 발전하고 있어 2024년에는 25억 달러 규모에 다다를 것으로 예상하고 있습니다.



¹ 《2018 세계 인공지능산업 보고서》: <http://www.semi.org.cn/siip/pdf/20180920p2.pdf>

² 전망산업연구원, 《2018-2023년 중국 인공지능업계 시장 전망과 투자 전략 분석 보고서》, 2018년 <https://bg.qianzhan.com/report/detail/300/190314-389cc4a4.html>

³ Global Market Insights report. 2018년 4월 www.elecfans.com/rengongzhineng/592041.html

게놈학 분석은 인공지능 응용에서 또 다른 중요한 영역을 차지합니다. 2022년이 세분화된 시장의 규모는 중국에서만 300억 위안⁴ 가까이 달할 것으로 예상하고 있습니다. 한 단계 더 나아가 염기서열과 인공지능을 결합하면 그 발전이 분명히 가속화되고, 시원싱 시간을 축소하게 될 뿐만 아니라, 시원싱 비용을 크게 절감하게 되면서 의료업계의 성장을 가능케 합니다.

중국 정부의 정책적 격려는 의료 인공지능 애플리케이션 정착을 가속화하는 중요한 요소 중 하나입니다. 2015년 이래로 중국 정부의 관련 부서는 인재 양성, 기술 혁신, 표준 관리감독, 업계 융합, 제품 정착 등 측면에서 인공지능 발전을 촉진하기 위한 20 항목에 가까운 정책을 내놓고 있습니다. 2017년 3월, "인공지능"은 처음으로 정부 업무 보고서에 기록되었습니다. 같은 해 7월에는 국무원이 《차세대 인공지능 발전 계획》을 정식 발행하면서 차세대 인공지능 발전 전략과 목표를 세단계로 명확히 짚었습니다. 또 같은 해 10월에는 "인공지능"이 전국대표회의 보고서에 "인터넷, 빅데이터, 인공지능과 실물 경제를 깊이 있게 융합한다"라고 기록하면서 중국 디지털 경제 발전 방향을 명확히 했습니다. 같은 해 12월에는 중화인민공화국 공업정보화부가 《차세대 인공지능 산업 발전 촉진 3년 행동 계획(2018-2020년)》을 발표하며 향후 3년간 인공지능의 중점 발전 방향과 목표를 상세히 계획했습니다. 2018년 1월에는 국가표준화위원회의 지도 하에 《인공지능 표준화 백서(2018 버전)》이 발표되었습니다. 그리고 4월에는 국무원이 《인터넷+의료, 헬스》 발전 촉진에 대한 의견》을 발행하면서 "인터넷+인공지능 애플리케이션 서비스를 "헬스 중국" 실행 전략의 중요한 조치로 삼고 의료와 헬스 관련 인공지능 기술, 의료용 로봇, 대형 의료 설비 등 연구개발을 중점적으로 지원할 것임을 밝혔습니다.

의료 인공지능 응용 동향

의료, 헬스 분야에서 인공지능의 응용 범위는 매우 넓습니다. 의료 영상부터 보조 진단, 질병 예측, 건강 관리, 약물 연구개발 등 여러 사이클에서 중요한 작용을 합니다. 그 중 인공지능은 의료 영상, 보조 진단,

질병 예측에 사용되며, 의원 또는 기타 의료 기관에 주로 서비스됩니다. 질병 검사 측면에 주로 집중되며, 진단 정확도를 어떻게 향상시키느냐에 초점이 맞추어져 있습니다. 하지만 위음성이 존재하는 상황에서 누락되지 않도록 의사가 모든 필름을 자세히 검토해야 하다 보니 의사의 업무량이 눈에 띄게 줄어들지는 않았습니다.

앞으로는 인공지능이 여러 등급의 의료 기관에서 다양하게 응용되고, 기초 병원 또는 제3자 검진센터에서는 주로 보조 검사와 보조 진단에 응용될 것이며, 3등급 의원에서는 의사 업무 효율 향상을 위해 주로 응용될 것입니다. 건강 관리 측면에서는 단체와 개인이 지불하는 건강 검진을 지원하는데 주로 응용됩니다. 약물 연구개발 영역에서는 인공지능 응용이 다른 특성을 지니며, 인공지능 기술 회사와 대형 제약회사, 의약 연구기관 모두가 힘을 합쳐 추진해야 합니다.

인공지능이 의료, 헬스 영역에서 빠르게 응용되고 있지만, 데이터, 모델 등의 원인 때문에 아직도 여러 측면에서 문제에 직면해 있습니다.

- 데이터 양의 문제. 모델이 복잡할수록 파라미터가 많아지고 필요한 트레이닝 양도 커집니다. 하지만 여러 복잡한 임상 시나리오에서 필요 한 대량의 신뢰할 수 있는 데이터는 얻기 어려운 실정입니다.
- 데이터 품질 향상. 신뢰할 수 있고, 품질이 높은 데이터를 입력하는 것은 매우 중요합니다. 오류 정정, 데이터 유실 알림 등 툴을 이용해 데이터 수집의 품질을 향상시킬 수 있습니다.
- 임상 업무 프로세스 융합. 딥러닝을 기존 전자 진료 기록 시스템 관리에 융합해 임상 전문의의 업무 효율을 향상시킵니다.
- 법제화 및 규범화. 여러 해마다 데이터를 왜곡해 딥러닝 모델의 결과 등 데이터 보안에 영향을 미치는 문제와 관련해 상응하는 법규를 제정해 분석 모델을 보호해야 합니다.
- 모델의 해석 가능성 문제. 딥러닝 모델은 블랙 박스이다 보니 어떻게 결론을 내리는지에 대해 명확한 해석이 없고, 의사결정 모델의 권위성 역시 아직 검증이 필요합니다.
- 모델의 보편성 문제. 우선 모델마다 편차가 있습니다. 예를 들어, 백인 환자의 데이터를 채택해 트레이닝한

모델을 다른 인종의 환자에게 적용할 경우 효과가 이상적이지 않을 것입니다. 두 번째는 모델의 상호운용성이 떨어지다 보니 두 개의 다른 전자 기록 정보 시스템의 딥러닝 모델에 적용하기 어렵습니다.

- 모델 안전 문제. 설령 평소에 트레이닝을 많이 한 이미지 처리 모델이라도 입력한 이미지 교란으로 좋지 않은 영향을 받을 수 있습니다. 이 교란은 사람이 발견할 수 없습니다. 그 밖에 데이터의 "아주 작은 차이"가 "아주 큰 차이"의 예측 결과를 가져올 수 있습니다. 예를 들어, 환자의 전자 기록 데이터 중 테스트 값은 약간만 변경해도 입원 사망률에 대한 예측 결과에 지대한 영향을 미칠 수 있습니다.

이러한 문제를 겨냥해 의료와 인공지능 분야의 전문가는 여러 대응 조치를 내놓음으로써 응용 환경을 개선했습니다.

- 대규모, 다양화된 헬스 데이터 수집. 인종, 민족, 언어와 사회 경제적 지위가 다양한 환자의 데이터를 광범위하게 수집한 뒤 집성해 표준화했습니다.
- 데이터 품질 향상. 신뢰할 수 있고, 품질이 높은 데이터를 입력하는 것은 매우 중요합니다. 오류 정정, 데이터 유실 알림 등 툴을 이용해 데이터 수집의 품질을 향상시킬 수 있습니다.
- 임상 업무 프로세스 융합. 딥러닝을 기존 전자 진료 기록 시스템 관리에 융합해 임상 전문의의 업무 효율을 향상시킵니다.
- 법제화 및 규범화. 여러 해마다 데이터를 왜곡해 딥러닝 모델의 결과 등 데이터 보안에 영향을 미치는 문제와 관련해 상응하는 법규를 제정해 분석 모델을 보호해야 합니다.

⁴ 전망산업연구원, 2018-2023년 중국 염기서열 업계 시장 전망과 투자 전략 계획 보고서, 2018년
<https://bg.qianzhan.com/trends/detail/506/180411-e7daa2c4.html>

의료, 헬스 영역에서 인공지능의 응용 시나리오

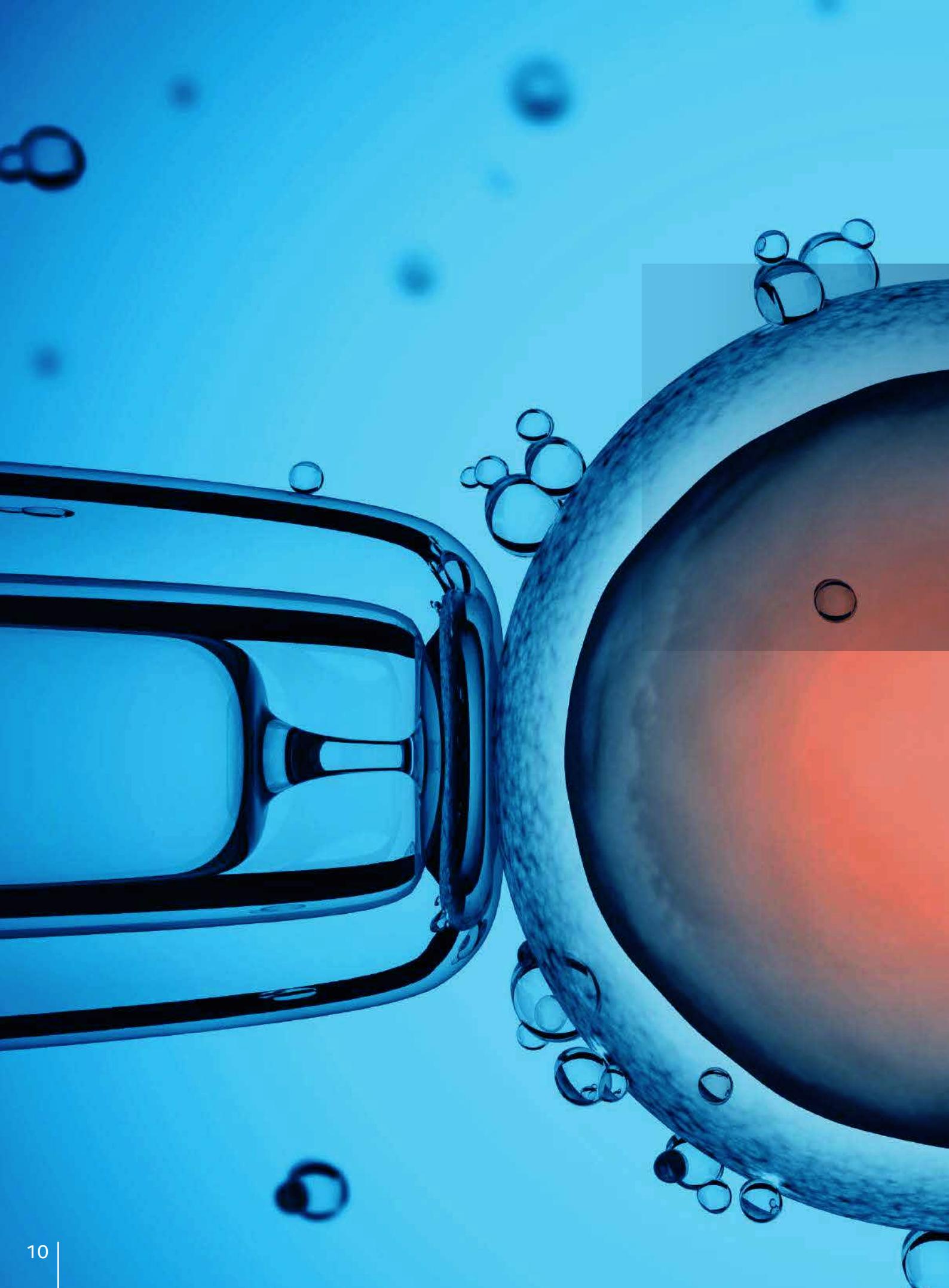
의료, 헬스가 인공지능 응용이 정착할 수 있는 잠재력을 지닌 영역 중 하나라는 점은 업계 모두 인정하는 사실입니다. 응용이 끊임없이 심화됨에 따라, 인공지능은 아래 여러 의료, 헬스 응용 시나리오에서 실력을 과시했습니다.

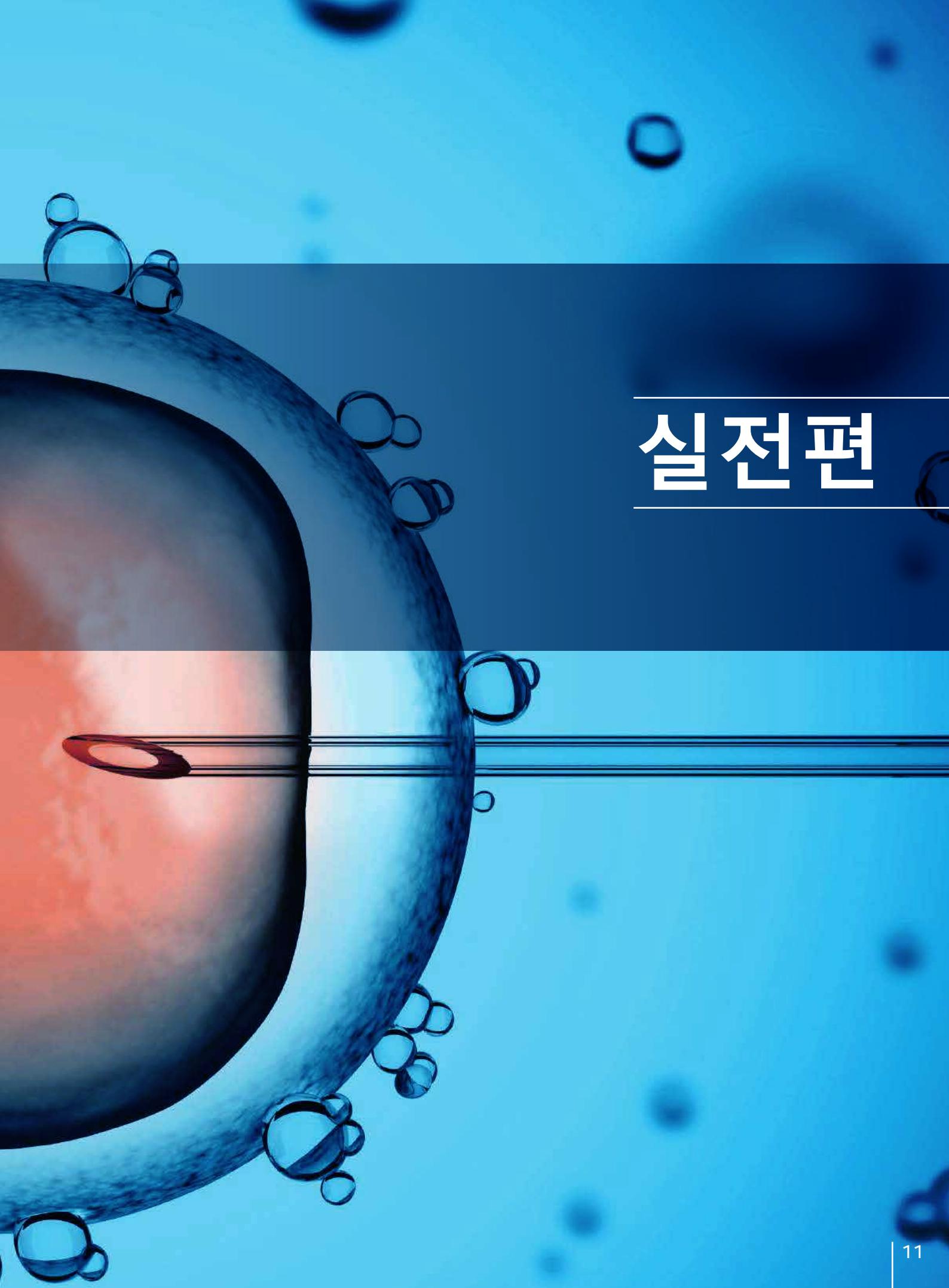
- **만성질환 관리와 질병 감지:** 환자의 징후에 따라(잠재해 있는) 만성질환의 리스크를 예측하고 조기 개입을 통해 환자의 진료 비용을 대체적으로 절감합니다.
- **임상 예측 분석:** 예를 들어, 전자 진료 기록 데이터에 기반한 원내 감염질환(예: 패혈증) 리스크 평가는 운용 모델에 따라 환자의 재입원률을 예측하고, 재무

모델에 따라 끼워 팔만한 서비스 방안 등을 제정합니다.

- **병력 검색과 품질 컨트롤:** 의료 파일 중 중요한 정보를 정확하게 추출해 의료 실제 인식을 진행하고 더 나아가 탄력적으로 전량 전자 진료 기록 검색을 실현합니다.
- **가상 현실 서포트:** 가상 현실 대화를 통해 검진 활동에 참여해 환자와 더 효과적으로 소통함으로써 환자 가 병인을 명확히 이해할 수 있게 합니다.
- **스마트 진료:** 음성과 스크린 등 여러 쌍방향 소통 방식을 통해 원내 네비게이션, 진료 서비스를 제공하고, 환자 분류, 건강 상담, 건강 교육 등 서비스의 수준을 향상시킵니다.
- **화상 인식:** 예를 들어, 스캔 기술과 OCR 기술 또는 이미지 처리 소프트웨어를 이용해 병력 또는 약품 외 포장의 정보를 판독해 관련 자료를 빠르게 취합니다.
- **영상 보조 진단:** 방사선과 전문의가 정상적인 영상을 빠르게 선별할 수 있게 돋기 때문에 의사의 업무 처리량이 증가합니다. 분석 영상의 정확도가 향상되고, 진단 결과 보고 시간이 단축되며, 의료 시스템의 진단 능력이 향상됩니다.
- **병리 분석:** 예를 들어, 암 세포를 고효율적으로 정확하게 감지하고, 암 방사선 요법 포인트 등을 명확히 설명합니다.
- **계놈학 분석:** 염기서열 비용을 대폭 절감하고, 방대한 규모의 게놈 데이터 분석을 빠르고 정확하게 실현하는데 사용합니다. 암 등 질병의 진단과 치료 등에는 훌륭한 도구가 됩니다.
- **약물 발견:** 약물 연구 개발 효율을 가속화해 비용을 절감합니다.

다음 "실전편"에서는 인텔이 Neusoft와 지멘스, 해방군총병원, AccuRad, 장평생물정보기술 등 파트너사 및 고객과 의료 인공지능 영역에서 응용한 사례의 배경과 실시 과정, 얻게 된 경험과 성과를 상세히 소개하고, 여러 응용 시나리오와 결합해 그에 맞는 소프트웨어와 하드웨어 구성을 추천할 것입니다.





실전편

의료업계의 이미지 분할 AI 추론 응용 가속화



의료 영상 처리 중 이미지 분할

전통적인 의료 영상 이미지 분할 방법

컴퓨터 비전 중 이미지 분할⁵이란 물체의 테두리나 선 같은 이미지의 자연적인 경계를 가리킵니다. 이미지를 여러 구역으로 나누는 목적은 이미지 표현 형식을 간소화하거나 변경해 쉽게 해독하고 분석하려 하는 것입니다. 컴퓨터 방식 중, 이 과정은 일반적으로 이미지 중 모든 화소에 태그를 추가해 동일한 태그를 지닌 화소는 공통적인 시각적 특성을 지니도록 디콘스트럭션되어 있습니다. 예를 들어, 색상과 밝기, 무늬 등이 그렇습니다. 그렇게 계산해 얻은 일정 구역의 화소 특성은 모두 유사하지만, 인접 구역은 큰 차이를 갖게 됩니다.

컴퓨터 비전 기술의 중요한 부분으로서, 이미지 분할은 이미 의료 영상 처리, 안면 인식, 공업용 로봇, 지능형 교통, 지문 인식, GPS 등 여러 업계와 영역에서 광범위하게 응용되었습니다. 의료 영상 처리 영역에서도 이미지 분할은 종양과 기타 병리 매핑, 조직 부피 측량, 해부학 연구, 컴퓨터 보조 수술, 치료 방안 제정 및 임상 보조 진단 등 여러 시나리오에서 그 가치를 증명했습니다.

전통적인 이미지 분할 방법은 주로 다음 몇 가지 방법이 있습니다.

- 클러스터 기반 방식: 클러스터링 기법은 K-평균값 기반 알고리즘으로 이미지를 K개의 클러스터로 반복 분할합니다. 이 알고리즘 중, 분할 이미지 화소와 클러스터 중심 사이에는 거리 편차가 있으며, 거리 편차는 일반적으로 색상과 밝기, 무늬, 위치 등 지표를 채택합니다. 이 알고리즘은 수렴성이 좋은 편입니다.
- 임계값 기반 방식: 이 방식은 이미지의 1개 또는 여러 개의 그레이스케일 임계값을 계산한 뒤, 모든 화소의 그레이스케일 값과 임계값을 비교해 마지막에 분류하는 방식입니다.
- 경계 기반 방식: 이 방식은 이미지 중 자연 경계의 그레이스케일, 색상, 무늬 등 특성의 돌연변이 유발성으로 이미지를 분할합니다. 일반적으로 경계 기반 분할 방식은 그레이스케일 값 경계에 따라 감지하기 때문에 경계 그레이스케일 값이 도약형으로 변화할 경우, 이미지 경계로 판단합니다.
- 구역 기반 방식: 이 방식은 이미지의 유사성에 따라 이미지를 분할합니다. 판단 원칙은 근접한 화소 지점의 그레이스케일, 색상, 무늬 등 특성에 유사성이 존재하는지 여부입니다. 유사성이 존재한다면, 화소 지점의 집합을 확대합니다.

딥러닝 기반 이미지 분할 방식

최근 몇 년간 AI 기술, 특히 이미지 영역이 빠르게 발전함에 따라 AI 기술을 기반으로 한 화상 인식, 이미지 처리 애플리케이션이 이미 여러 시나리오에서 사용되고 있으며, 유형별 의료 영상 분석에 인류를 초월하는 식별 능력을 제공했습니다. 콘볼루션 뉴럴 네트워크(Convolutional Neural Network, CNN)와 유사한 모델로서 현재 AI 기반 이미지 분할 기술 중 자주 볼 수 있는 네트워크 모델입니다. 그 중 전량 콘볼루션 네트워크(Fully Convolutional Network, FCN), U-net과 V-net은 자주 볼 수 있는 딥러닝 기반 이미지 분할 방식입니다.

⁵ 영상 분할 설명 참고 문헌: Linda G. Shapiro and George C. Stockman (2001): "Computer Vision", pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13-030796-3

■ FCN

CNN의 전형적인 용도는 임무를 분류하는 것입니다. 이미지 처리에 있어 CNN은 단일 유형 태그를 출력합니다. 생물 의학 이미지 분할 처리에서 기대하는 출력 형태는 포지션을 포함해야 합니다. 즉, 유형별 태그가 모든 화소에 분배되어 있어야 합니다. 콘볼루션 뉴럴 네트워크의 업그레이드 확장 버전으로서 그림 1-1-1과 같이 FCN⁶은 인코딩, 해독한 네트워크 구조 모드를 따르며, 콘볼루션 계층과 풀링 계층을 직렬 연결했습니다. 콘볼루션 계층과 최대 풀링 계층은 최초 이미지의 공간 차원을 효과적으로 낮추었습니다. 뿐만 아니라, FCN은 AlexNet을 네트워크의 인코더로 사용하고, 다중 전위 콘볼루션 중복 확장 방식을 채택해 인코더의 마지막 콘볼루션 계층이 출력한 특징 그림에서 그림이 입력 이미지의 해상도를 회복할 때까지 업샘플링하면 화소 등급별 이미지 분할을 실현할 수 있습니다.

■ U-net

FCN 네트워크의 개선 버전으로, U-net은 선명 한 U 형 구조를 갖 주고 있으 며, 위상 도는 1 - 1 - 2 와 같 습 니 다. 모든 Encoder마다 4회씩 업샘플링을 진행하며, 그렇기 때문에 분할 그림 회복 경계 등 정보 가 더 정확 합 니 다. 또 한 , 동 일 한 stage에서 U-net은 항상 도약 연결 (skip connection)을 채택하지, 고급 의미론 특징에서 감독과 loss 역전을 진행하지 않습니다. 그러면 마지막에 얻은 특성도가 더 많은 저레벨 (low-level) 특성을 융합 할 수 있을 뿐만 아니라 다른 척도의 특성이 융합되어 멀티스케일 예측 (Multi-Scale Prediction) 과 멀티스케일 예측 (Deep Supervision) 를 진행할 수 있습니다. 그 밖에 도 , U - net은 네트워크 뒷 부분에 앞부분과 유사한 네트워크를 보충해 U형 구조를 형성했습니다. 그 중 풀링 연산자는 업샘플링 연산자로 대체하기 때문에 출력 해상도가 증가했습니다. 게다가 포지션을 위해 모델은 경로를 축소한 고해상도 특성과 업샘플링 출력을 결합합니다. 연속 콘볼루션 계층은 relu 활성화 함수를 선택해 초기 이미지에 다운샘플링 작업을 진행해 더 정확한 출력을 획득 할 수 있습니다.

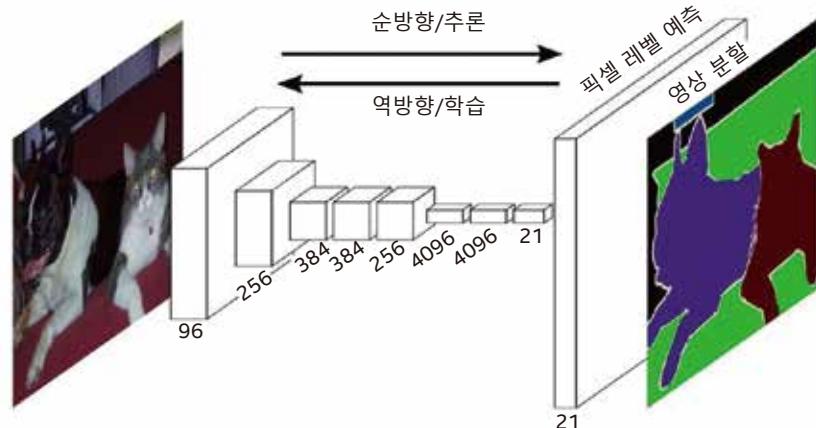


그림 1-1-1 FCN 방식 구성도

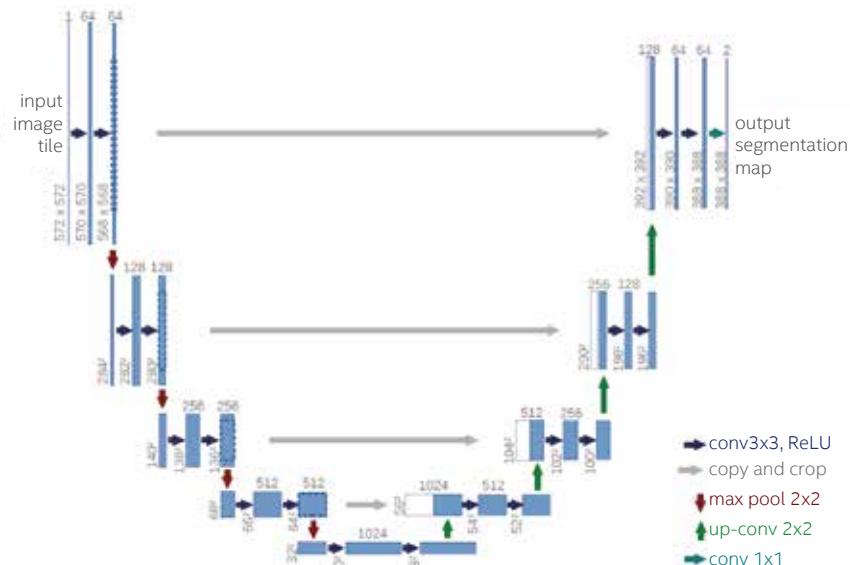


그림 1-1-2 U-Net 토플로지

의료 영상은 실제 응용에서도 독특한 특성을 지닙니다. 일반적으로 흉부단순촬영은 흉부 CT, 안저 검사는 안저 OCT, 모두 전신이 아닌 하나의 지정 기관만을 촬영하는 검사임을 알 수 있습니다. 기관 자체 구조가 고정적이기 때문에 의미 정보는 특별하지 않습니다. 그렇기 때문에 고급 의미 정보와 낮은

등급의 특성이 매우 중요한 것처럼 보이며, U-Net의 U형 구조와 도약 연결은 이러한 시나리오에서 더 큰 역할을 발휘할 수 있습니다. 최근에는 U-Net이 의료 영상 분할 영역에서 뛰어난 효과를 보였으며, 이미 여러 상황에서 충분히 증명했습니다.

⁶ FCN 관련 기술 설명 출처-UC Berkeley jonlong, shelhamer 와 trevor 의 논문 《Fully Convolutional Networks for Semantic Segmentation》: https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf

■ V-net

V-net은 그림1-1-3과 같은 3D 버전의 U-net으로, U-net과 유사한 위상학적 형태를 갖고 있으며 3D 구조의 의료 영상 분할에 적용됩니다. V-net은 3D 이미지 기반 end-to-end 이미지 의미 분할을 실현할 수 있으며, 레지듀얼 러닝과 유사한 trick을 통해 네트워크를 개선할 수 있습니다.

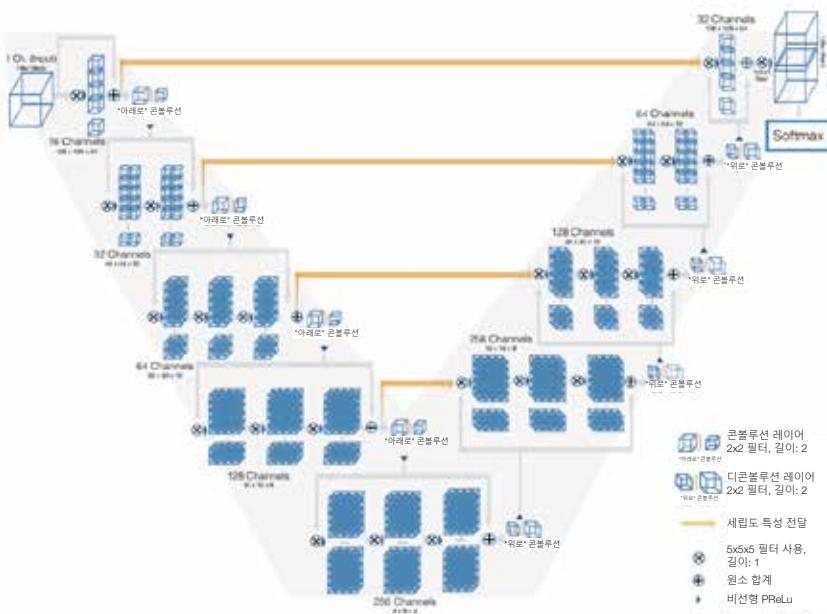


그림 1-1-3 V-Net 위상학적 사고

소프트웨어/하드웨어 구성 제안

의료 업계에서 딥러닝 기반 이미지 분할 방식을 구축하려면 다음의 인텔® 아키텍처 플랫폼 기반 소프트웨어/하드웨어 구성을 참조해 완성할 수 있습니다.

명칭	사양
CPU	인텔® 제온® 골드 6240 CPU 또는 그 이상
하이퍼스레딩	ON
터보 가속	ON
메모리	16GB DDR4 2666MHz* 12 및 이상
저장	인텔® SSD D5 P4320 시리즈 및 이상
운영 체제	CentOS Linux 7.6 또는 최신 버전
Linux 코어	3.10.0 또는 최신 버전
컴파일러	GCC 4.8.5 또는 최신 버전
Python 버전	Python 3.6 또는 최신 버전
Tensorflow 버전	R1.13.1 또는 최신 버전
OpenVINO™ 툴킷	2019 R1 또는 최신 버전
Keras 버전	2.1.3 또는 최신 버전

U-net 분할 네트워크 최적화 방법

인텔® 아키텍처 기반 최적화 방법

전통적인 CNN 이미지 분할 방법을 의료 이미지에 사용하면 종종 다음과 같은 문제가 생깁니다.

- CNN은 일반적으로 분류에 응용되며, 생물 의료 영상은 분할 및 포지션에 더 집중합니다.
- CNN은 다량의 트레이닝 데이터를 획득 해야 하는데 의료 이미지는 대규모의 데이터를 획득하기가 매우 어렵습니다.

과거에는 위와 같은 문제와 맞닥뜨릴 경우, 일반적으로 슬라이딩 윈도우 방식을 채택해 분류해야 할 화소를 주변의 일부 영역에 모두 입력했습니다. 이 방법의 장점은 두 가지입니다. 첫째, 이 방법은 슬라이딩 윈도우에서 포지셔닝 작업을 동시에 완료할 수 있습니다. 둘째, 모든 동작마다 화소 주변의 영역을 취하기 때문에 트레이닝 데이터 양을 대대적으로 증가시킬 수 있습니다. 하지만 장점도 있습니다. 첫째, 슬라이딩 윈도우를 통해 블록끼리 중첩이 되다 보니 트레이닝과 주론 속도가 다소 느려집니다. 둘째, 네트워크를 부분 정확도와 획득한 컨테스트 사이에서 취사선택해야 하다 보니 만약 슬라이딩 윈도우의 블록이 너무 크면 더 많은 팔링 계층이 필요하고 포지션 정확도가 하락합니다. 반대로 블록이 또 너무 작으면 네트워크는 아주 작은 일부 컨테스트밖에 볼 수 없게 됩니다.

인텔® 아키텍처 플랫폼을 기반으로 전개한 일련의 최적화는 또 다른 측면에서 사용자가 이상의 문제를 해결할 수 있게 도와 줍니다. 최적화 방법은 다음과 같습니다. CPU 코어 수량 조정, 동일하지 않은 메모리 접근 아키텍처 (Non-Uniform Memory Access Architecture, NUMA) 기술 및 DNN 지향 인텔® MKL (Intel® Math Kernel Library for Deep Neural Networks, 인텔® MKL-DNN)을 도입해 U-net 이미지 분할 방식에 복합적인 최적화를 제공합니다. 최적화 절차는 아래와 같습니다.

■ 환경 변수 설정

우선, 아래와 같이 환경 변수를 설정해야 하며, 명령은 다음을 포함합니다. 시스템의 캐시를 정리해 CPU를 최고 성능 모드로 설정하고 최고 주파수에서 실행해 CPU의 터보 부스트를 엽니다.

```
1. echo 1 > /proc/sys/vm/compact_memory
2. echo 3 > /proc/sys/vm/drop_caches
3. echo 100 > /sys/devices/system/cpu/intel_pstate/min_perf_pct
4. echo 0 > /sys/devices/system/cpu/intel_pstate/no_turbo
5. echo 0 > /proc/sys/kernel numa_balancing
6. cpupower frequency --set -g performance
7.
8. export KMP_BLOCKTIME=1
9. export KMP_AFFINITY=granularity=fine,verbose,compact,1,0
10. export KMP_NUM_THREADS=20
```

- KMP_BLOCKTIME을 1로 설정합니다. 이는 스레드 현재 임무를 실행 완료하고 휴면 상태에 들어가기 전까지 대기하는 시간 설정으로 일반적으로 1 밀리 초로 설정합니다.
- KMP_AFFINITY를 Compact로 설정합니다. 이는 해당 모드에서 스레드 바인딩이 컴퓨팅 코어의 컴퓨팅에 따라 우선 순위를 정해 동일한 코어를 우선 바인딩하고, 그 다음으로 동일 CPU의 다음 코어를 바인딩하는 것을 의미합니다. 이러한 바인딩 방식은 스레드 사이에 데이터 교류가 이루어지거나 공용 데이터를 갖춘 상황에 적용되며, 다양한 계층의 캐시를 충분히 이용할 수 있다는 점이 장점입니다.
- OMP_NUM_THREADS를 20으로 설정하면 스레드의 수량을 필요한 물리 코어 수로 병행 설정하게 됩니다.

■ 테스트 코드에 스레드 컨트롤 추가

```
1. config = tf.ConfigProto()
2. config.allow_soft_placement = True
3. config.intra_op_parallelism_threads = FLAGS.num_intra_threads
4. config.inter_op_parallelism_threads = FLAGS.num_inter_threads
```

명령 설정에서 보는 바와 같이 `tf.ConfigProto()`를 초기 설정 할 때, `intra_op_parallelism_threads` 파라미터와 `inter_op_parallelism_threads` 파라미터를 설정해 연산자 op로 병렬 컴퓨팅한 모든 스레드 개수를 컨트롤할 수도 있습니다. 두 방법의 차이는 다음과 같습니다.

- `intra_op_parallelism_threads`는 연산자 op의 내부 병렬을 컨트롤합니다. 연산자 op가 단일 연산자이고 행렬 곱셈, `reduce_sum`과 같이 내부에서 병렬 실행이 가능한 경우 `intra_op_parallelism_threads` 파라미터를 설정해 병렬을 실행할 수 있으며, `intra`는 내부를 의미합니다.
- `inter_op_parallelism_threads`는 여러 연산자 op 사이의 병렬 컴퓨팅을 컨트롤합니다. 연산자 op가 여럿이고 서로 독립적이며 연산자와 연산자 사이에 직접적인 경로 없이 Path할 경우, Tensorflow는 시범적으로 컴퓨팅하고 `inter_op_parallelism_threads` 파라미터로 수량을 컨트롤한 스레드 풀 하나를 사용합니다.

통상적으로 `intra_op_parallelism_threads`는 단일 CPU의 물리적 코어 수량으로 설정하지만 `inter_op_parallelism_threads`는 1 또는 2로 설정합니다.

■ NUMA 특성을 이용해 CPU 컴퓨팅 리소스 사용 컨트롤

데이터 센터가 사용하는 서버는 일반적으로 2대 또는 그 이상의 CPU를 구성하며, 대다수가 NUMA 기술을 사용해 여러 서버를 단일 시스템처럼 운영합니다. CPU가 자신의 로컬 기억 장치에 접근하는 속도는 타지의 기억 장치에 접근하는 속도보다 빠릅니다. 이러한 시스템에서 더 뛰어난 컴퓨팅 성능을 획득하기 위해서는 일부 특정 명령을 통해 컨트롤해야 합니다. Numactl이 바로 과정을 컨트롤하고 기억을 공유하는 일종의 기술 매커니즘으로, Linux 시스템에서 광범위하게 사용되는 컴퓨팅 리소스 컨트롤 방식입니다. 구체적인 사용 방법은 아래와 같습니다.



그림1-2-1 NUMA 특성을 이용해 CPU 컴퓨팅 리소스 사용 컨트롤

1. `numactl -C 0-19,40-59 -m 0 python3 test.py`

위의 명령 표현은 `test.py`를 실행할 때 CPU#CPU0 중 0-19와 40-59 코어, CPU#CPU0와 상응하는 근단 메모리를 사용한 것입니다.

■ 인텔® MKL-DNN 최적화 지향 TensorFlow 채택

사용자가 범용 CPU 플랫폼에서 고효율 AI 컴퓨팅을 진행할 수 있도록 인텔은 여러 주류의 딥러닝 소스 아키텍처에 대해 다양한 최적화를 진행했습니다. 이것은 현재 공업계와 학계에서 아주 광범위하게 사용되는 TensorFlow입니다.

인텔® MKL-DNN 최적화를 사용한 프리미티브(Primitive)를 통해 인텔은 TensorFlow를 최적화했습니다. 인텔® MKL-DNN은 TensorFlow 1.2에서 추가된 것입니다. CNN 기반 모델을 트레이닝할 때 성능이 눈에 띄게 향상하는 것 외에도, 인텔® MKL-DNN은 사용해 컴파일링을 진행하면 인텔® 고급 벡터 확장 명령어 조합(Intel® Advanced Vector Extensions, 인텔® AVX), 인텔® AVX 2와 인텔® AVX-512가 최적화를 진행한 2진법 파일을 생성할 수 있으며, 그를 통해 최적화를 거쳤고 대다수의 현대(2011년 이후) CPU와 호환할 수 있는 파일을 얻을 수 있습니다.

참고문헌:

- https://www.tensorflow.org/guide/performance/overview?hl=zh_cn
- <https://software.intel.com/zh-cn/articles/tensorflowoptimizations-on-modern-intel-architecture>

* 인텔® MKL-DNN과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

U-net을 인텔® 아키텍처를 기반으로 최적화한 테스트와 결과

이상 네 가지 측면의 최적화를 통해, U-net은 인텔®아키텍처 기반 CPU 플랫폼에서 성능이 현저히 향상했으며, 테스트 결과는 다음 그림⁷과 같습니다.



그림 1-2-2 인텔®아키텍처 기반 최적화 전후 성능 비교

OpenVINO™ 툴킷 기반 인텔® 배포 버전으로 U-net을 한 단계 더 최적화

고객의 실제 응용 시나리오 수요를 만족시키기 위해 상기 결과를 기초로 인텔은 OpenVINO™ 툴킷 기반 인텔® 배포 버전(이하 "OpenVINO™ 툴킷")으로 U-net 이미지 분할 방식을 한 단계 더 최적화했습니다. 구체적인 최적화 절차는 다음과 같습니다.

■ 모델 전환

기존의 모델은 Keras를 기반으로 트레이닝을 진행하기 때문에 생성된 모델은 hdf5 형식이 됩니다. 이 형식의 모델은 바로 OpenVINO™ 툴킷의 입력이 될 수 없고, 우선 형식을 전환해야 합니다. 작업 명령은 다음과 같습니다.

```
1. git clone https://github.com/amir-abdi/keras_to_tensorflow.git
2. cd keras_to_tensorflow
3. python3 keras_to_tensorflow.py --input_model=./unet/unet_membrane.hdf5 --
   output_model=./unet/full_unet.pb #<--模型转换
```

■ 모델을 OpenVINO™ 툴킷의 mo.py를 통해 xml 파일과 bin 파일로 전환

명령은 다음과 같습니다.

```
1. python3 /opt/intel/openvino/deployment_tools/model_optimizer/mo.py --framework:tf --
   input_model:full_unet.pb --data_type:FP32 --output_dir:/ --input_shape:[28,512,512,1]
```

■ Inference Engine을 통해 모델 추론 진행

명령은 다음과 같습니다.

```
1. python3 segmentation_demo.py -m /home/worker/unet/full_unet.xml -
   -i /home/worker/0.png -l /home/worker/openvino/intel64/Release/libcpu_extension.so
```

그 중, 추론한 코드는 다음 로직 모듈을 포함합니다.

```
1. #Load input data ##数据预处理及导入，包括数据格式统一（医学图像dcm格式转化为jpg）：执行图像缩放、多通道扩增、归一化处理等操作
2. #Loading model to the plugin net = IENetwork.from_ir(model=model_xml, weights=model_bin)
   ##其中 xml 文件为网络结构，bin 文件为权重参数
3. input_blob = next(iter(net.inputs))##确定模型的输入
4. out_blob = next(iter(net.outputs))##确定模型的输出
5. exec_net = plugin.load(network=net)##模型导入
6. # Start sync inference
7. res = exec_net.infer(inputs=[input_blob:images])##数据推理过程
8. # Processing output blob
9. res = res[out_blob]##提取推理结果
10. #Visualization result
```

⁷ 테스트 구성: CPU: 투웨이 인텔® 제온® 골드 6148 CPU, 2.40GHz 코어/스레드: 20/40, 메모리: 16GB DDR4 2666MHz * 12, 하드 디스크: 인텔® SSD SC2BB480G7, BIOS: SE5C620.86B.02.01.00008.031920191559, 운영 체제: CentOS Linux 7.6, Linux 커널: 3.10.0-957.21.3.el7.x86_64, gcc 버전: 7.2, Python 버전: Python 3.6, Tensorflow 버전: R1.13.1

OpenVINO™ 툴킷 기반 최적화 결과

최적화 결과는 그림 1-2-3과 같습니다. 가장 왼쪽 열은 대뇌 CT 원도, 중간 열은 최적화되지 않은 이미지 분할 결과, 가장 오른쪽 열은 OpenVINO™ 툴킷을 통해 최적화해 생성한 이미지 분할 결과입니다. 우리는 OpenVINO™ 툴킷을 통해 최적화해 생성한 이미지 분할 결과가 정확도 면에서 최적화하지 않은 경우와 기본적으로 일치하지만 추론 속도 측면에서는 최적화하지 않았을 경우⁸보다 아주 높음을 알 수 있습니다.

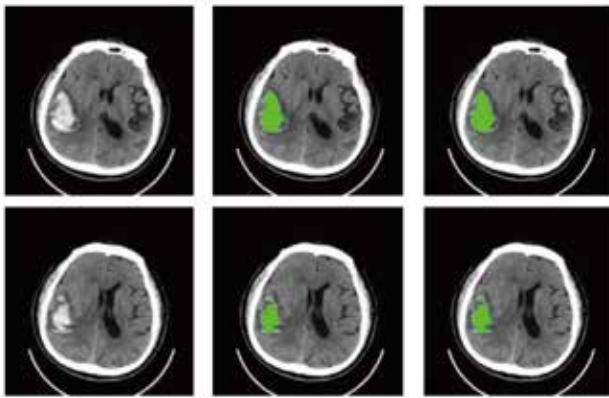


그림 1-2-3 OpenVINO™ 툴킷을 기반으로 U-net에 최적화한 결과

* OpenVINO™ 툴킷 기술과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

2세대 인텔® 제온® 확장성 CPU를 기반으로 구축한 Dense U-net 이미지 분할 방식

인텔® 딥러닝 가속(Intel® Deep Learning Boost, 인텔® DL Boost) 기술

2세대 인텔® 제온® 확장성 CPU는 최적화된 마이크로 아키텍처와 더 많은 커널 및 더 빠른 메모리 채널의 컴퓨팅 성능을 향상시켰으며, AI 지향 애플리케이션에는 전면적인 가속 기능을 제공했습니다. 특히 집성화된 인텔® 딥러닝 가속 기술(VNNI 명령어 조합)에 INT8에 대한 지원을 추가해 사용자에게 고효율의 INT8 딥러닝 추론 가속 기능을 제공했으며, 이 기능은 U-net 이미지 분할 방식의 실행 효율을 효과적으로 향상시켰습니다.

인텔® 딥러닝 가속 기술은 VNNI 명령어 조합을 통해 8자리 또는 16자리 저정밀도 수치 곱셈을 지원하며, 이는 다량의 행렬 곱셈을 실행해야 하는 딥러닝 컴퓨팅에게 있어 매우 중요한 부분입니다. 이 기술을 도입하면, 사용자가 INT8 추론을 진행할 때, 시스템 메모리

수요를 최대 75%⁹ 감소시킬 수 있습니다. 또한 메모리와 필요한 대역폭의 감소는 저수치 정밀도 연산 속도를 가속화해 시스템 전체 성능을 대폭 향상시켰습니다.

과거의 FP32 모델과 비교하면, INT8 모델은 수치 정밀도와 동적 범위가 더 작기 때문에 이미지 분할 등 딥러닝에서 INT8 추론 방식을 선택하려면 컴퓨팅 실행 시 정보 유실 문제를 집중적으로 해결해야 합니다. 일반적으로 INT8 추론 기능은 양자화 교정 방식을 통해 추론 대기 INT8 모델을 형성함으로써 FP32가 정보 유실을 최소화하는 전제 하에서 INT8로 전환하는 목표를 실현합니다.

이미지 분석 응용을 예로 들면, 고정밀도 수치에서 저정밀도 데이터로 전환은 실제적으로 하나의 컴퓨팅하며 축소되는 과정입니다. 바꿔 말하면, 축소 범위를 어떻게 확정하느냐는 정보 유실 최소화를 실현하는 핵심입니다. FP32에서 INT8로 매핑하는 과정 중, 데이터세트에 따라 교정하는 방식을 선택해 매핑 축소 파라미터를 확정합니다. 파라미터를 확정한 뒤, 플랫폼은 지원한 INT8 작업 리스트에 따라 다시 이미지를 분석하고 양자화/역양자화 작업을 실행합니다. 양자화 작업은 FP32가 S8(부호가 있는 INT8) 또는 U8(부호가 없는 INT8)의 양자화에 사용되며, 역양자화 작업은 역방향으로 작업을 실행합니다.

OpenVINO™ 툴킷을 기반으로 FP32 모델을 INT8 모델로 전환

일반적으로 뉴럴 네트워크를 통해 트레이닝된 모델은 SPFP 정밀도입니다. 즉 FP32는 사용자가 이 모델을 실제 응용 시나리오에 직접 배치하고, 양자화 기술을 통해 저정밀도 모델을 획득할 수 있습니다. 예를 들어, INT8 모델은 모델 정밀도 보장을 기본으로 해 더 효율이 높은 모델 추론 애플리케이션을 제공할 수 있으며, 일반 상황에서 모델 정밀도의 손실은 1% 미만입니다.

OpenVINO™ 툴킷은 2018 R4 버전부터 시작해 FP32 모델부터 INT8 모델까지 전환 기능을 제공했고, 2019 R1 버전부터는 2세대 인텔® 제온® 확장성 CPU로 집성화된 인텔® 딥러닝 가속 기술을 지원합니다.

우선 OpenVINO™ 툴킷은 트레이닝이 잘 되고, 개방형 뉴럴 네트워크 교환(Open Neural Network Exchange, ONNX)으로 트레이닝된 모델을 전환하고 최적화해 FP32 형식의 xml 파일과 bin 파일을 생성합니다. 최적화는 노드 융합, 다량 정규화 제거와 상수 폴딩 등 방식을 포함합니다. 그런 다음 OpenVINO™ 툴킷 중 전환 툴(Calibration Tool)로 FP32 형식의 파일을 INT8 형식의 xml 파일과 bin 파일로 전환합니다. 전환 과정 중에는 소량의 검증 데이터세트가 필요하며, 전환 양자화 과정 중인 통계 데이터를 이후 추론 시 정밀도가 영향을 받지 않도록 저장합니다. 위의 전환 프로세스는 오프라인으로 진행되며, 1회만 전환하면 됩니다. 자세한 방법은 그림 1-3-1과 같습니다.

⁸ 관련 검증 테스트 구성: CPU: 투웨이 인텔® 제온® 골드 6148 CPU, 2.40GHz 코어/스레드: 20/40, 메모리: 16GB DDR4 2666MHz* 12, 하드 디스크: 인텔® SSD SC2BB480G7, BIOS: SE5C620.86B.02.01.0008.031920191559, 운영 체제: CentOS Linux 7.6, Linux 커널: 3.10.0-957.21.3.el7.x86_64, gcc 버전: 4.8.5, Python 버전: Python 3.6, OpenVINO™ 툴킷: 2019 R1, Keras: 2.1.3

⁹ 데이터 소스 출처: <https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training>

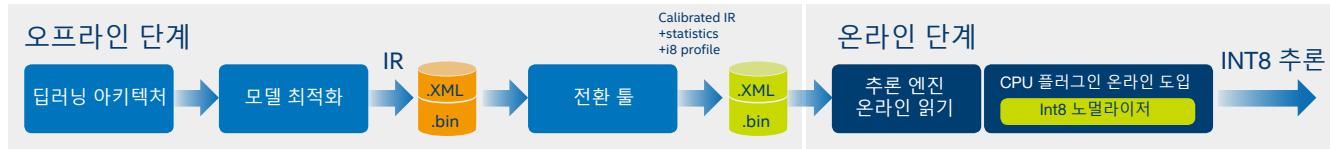


그림 1-3-1 OpenVINO™ 툴킷 기반 FP32 모델을 INT8 모델로 전환

위에 따라 모델을 전환한 이후의 초기 모델 성능은 다음 그림과 같습니다.

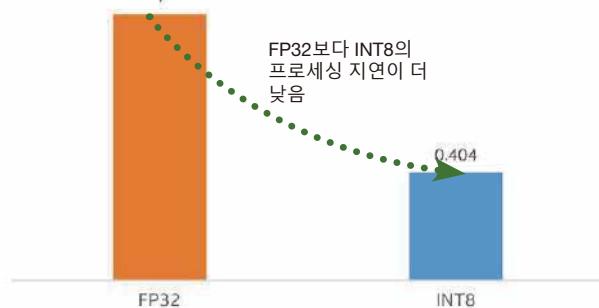


그림 1-3-2 FP32와 INT8의 지연 성능 비교

이 두 가지 모델을 통해 성능을 분석하면, FP32 모델은 리오더 작업(Reorder Ops)이 다량의 실행 시간을 차지했고, INT8 모델은 리샘플 작업(Resample Ops)이 FP32의 작업을 지원한다는 것을 알 수 있습니다. 그리고 연속 작업(concat Ops) 실행 시간이 과도하게 길면, 본래 점유율이 가장 높았던 콘볼루션 작업(Convolution Ops)이 전체 모델 운용 중 차지하는 시간 비율이 오히려 줄어듭니다. 그렇기 때문에 한 단계 더 최적화해야 합니다.

그림 1-3-3과 같이 최적화를 거친 모델은 지연 시간이 대폭 하락합니다.



그림 1-3-3 최적화 이후 INT8 모델의 지연 시간 성능 비교

이때 INT8 모델을 다시 심층 분석하면, 이전과 비교했을 때 현저히 향상되었음을 알 수 있습니다. 단, 최적화 이후 모델 중, Concat Ops가 차지하는 실행 시간은 긴 편이며, 모델의 스루풋을 더 향상시키기 위해 Concat Ops를 특정 최적화해야 합니다. 이제는 인텔® MKL-DNN의 프리미티브를 사용하지 않고 맞춤화를 진행해야 합니다. 상세 코드는 다음과 같습니다.

```

1. for (size_t i = 0; i < num_src; i++) {
2.     const MKLDNNMemory& src_mem = getParentEdgeAt(i)->getMemory();
3.     channels.push_back(src_num.GetDims()[1]);
4.     src_ptrs.push_back(reinterpret_cast<const uint8_t*>(src_mem.GetData()));
5.     dst_ptrs.push_back(dst_ptr + channels_size);
6.     channels_size += src_mem.GetDims()[1];
7. }
8.
9. parallel_for(iter_count, [&](int i){
10.     for (int j = 0; j < src_ptrs.size(); j++) {
11.         memcpy(dst_ptrs[i] + (i * channels_size), src_ptrs[j] + i * channels[j], channels[j]);
12.     }
13. });
  
```



그림 1-3-4 한층 더 최적화한 INT8 모델의 지연 시간 성능 비교

성능 분석에서 알 수 있듯이, 이때 모델 운용 비율이 가장 높은 프리미티브는 콘볼루션 작업이 되어 이 사례 중 Dense U-net 모델이 본래 보유한 효과와 완전히 부합하게 됩니다.

응용 사례

Neusoft eStroke 혈전용해술 색전제거술 영상 플랫폼

■ 배경

뇌졸중은 오랜 기간 사람들의 건강을 위협해온 적입니다. 추산에 따르면, 매년 200만 명의 뇌졸중 환자가 전국적으로 발생하고 있으며, 그 중 65세 이하가 50%를 차지한다고 합니다. 이 데이터는 중국의 뇌졸중 환자의 저령화 현상이 심각해지고 있음을 의미합니다. 이러한 현상은 매년 13%의 속도로 상승하고 있으며, 재발율 역시 17.7%¹⁰에 도달해 환자와 사회에 심각한 부담을 주고 있습니다. 뇌졸중을 효과적으로 치료할 수 있는 중요한 수단은 혈전용해술과 색전제거술이지만, 이 방법은 대뇌 의료 영상을 빠르고 정확하게 판독해야 실현 가능한 부분입니다. 뇌졸중을 치료할 수 있는 골든 타임이 30분 뿐이다 보니 기본적으로 환자를 상급 병원으로 옮길 시간이 부족합니다. 그런데 응급 구조 상황은 대부분 기초 병원에서 발생합니다. 하지만 기초 병원은 기술이 부족해 혈전용해술, 색전제거술을 진행할 수 있는 확률이 낮은 편입니다. 그 뿐만 아니라 의사의 판독 수준이 천차만별이고, 전문 영상의학전문의도 부족하다 보니 중심 병원의 영상의학전문가는 몸이 열개라도 부족한 상황입니다. 이러한 이유 때문에 뇌졸중 혈전용해술, 색전제거술을 효과적으로 지도하기 어렵고, 구조 가능한 조직을 식별하기 어려워져 환자는 응급 처치를 받을 기회를 잃게 됩니다.

이러한 문제를 해결하기 위해, 의료 업계에는 기초 병원 전문의의 수준이 부족한 상황에서도 의료 영상을 빠르게 분석할 수 있는 툴이 필요합니다. 이제 딥러닝 기반 의료 영상 판독 방식이 의료 기관에 점차 도입되면서 위의 문제를 해결할 수 있게 도와주고 있습니다. Neusoft 스마트 의료 연구원과 선양 Neusoft 의료 시스템 유한회사(이하 "Neusoft")가 대중 파트너와 협력해 만든 고품질의 eStroke 혈전용해술

색전제거술 영상 플랫폼은 급성 뇌졸중 정맥 혈전용해술과 동맥 색전제거술 치료를 위해 더 정확한 가이드를 제공할 수 있습니다.

■ 방법과 성과

eStroke 혈전용해술 색전제거술 영상 플랫폼은 하혈성 뇌졸중 반암부, 미세뇌출혈, 뇌의 측부 순환에 대한 양적 평가를 기반으로 한 클라우드 서비스 플랫폼으로서 혈전용해술, 색전제거술 멀티 모달리티 영상을 정확하게 평가할 수 있을 뿐만 아니라 다음과 같은 장점을 지니고 있습니다.

- 멀티 모달리티 영상 학장비를 지원하며, 전산화 단층촬영(Computed Tomography, CT), 자기공명영상(Magnetic Resonance Imaging, MRI) 이미지 등 16채널 이상의 MDCT와 1.5T 이상의 MRI가 포함됩니다.
- 프로세스 전체 자동화로, 병원 디바이스 스캐닝부터 시작해 영상 후처리 분석, 출력 영상 진단 보고까지 모든 작업에 인력이 필요하지 않습니다.
- 인터넷 의료 치료 기술 응용 연구 플랫폼 등 외부 치료 시스템을 도입해 심혈관계 질환 원격 응급 처치, 모바일 응급 처치, 고위험군 스마트 조기 경보 및 개입, 심혈관계 질환 연합 처리, 가상 수술 등 기술 연구개발과 엔지니어링을 지원합니다.

Neusoft와 인텔은 eStroke 혈전용해술 색전제거술 영상 플랫폼을 저장 장치로 하고, U-net 모델을 기반으로 플랫폼 중 뇌졸중 의료 영상을 이미지 분할 처리했습니다. eStroke 플랫폼에 따라 이미지를 주입한 CBF, CBV, MTT와 TMAX(대뇌 혈류량, 혈액뇌용량, 평균 통과 시간과 잔류 함수의 티맥스)로 계산하고, 이상의 파라미터가 좌우 뇌순환을 거친 이후의 대칭성을 결합해 그림 1-4-1과 같이 의료 진단에 사용하는 허혈성 음영과 경색 코어의 위치를 더 정확히 추론해낼 수 있습니다.

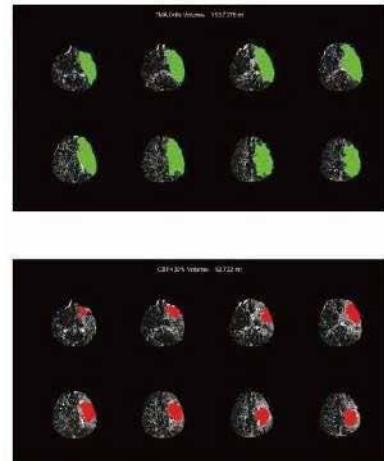


그림 1-4-1 TMAX&CBF 비정상 구역을 통한 허혈성 음영과 경색 코어 구역 컴퓨팅

이 방안은 인텔®아키텍처 최적화 지향 TensorFlow(인텔® MKL-DNN 기반 최적화)와 OpenVINO™ 툴킷을 채택해 최적화를 진행함으로써 U-net 모델 기반의 딥러닝 추론의 정확성을 보장하고, 추론 시간을 대폭 감소했습니다. 이러한 장점은 촌각을 다투는 뇌졸중 치료에 두말할 것 없이 중요한 의미를 부여합니다. 그림 1-4-2처럼 추론 정확도가 기본적으로 일치하는 상황에서 툴을 최적화한 방안과 최적화하지 않은 방안을 비교합니다. 그 결과 추론 지연이 각각 72.6%, 85.4%¹¹ 감소했습니다.



그림 1-4-2 Neusoft U-net 이미지 분할 방안별 성능 비교

¹⁰ 데이터 출처: 《안후이성 뇌졸중 등급별 진료 가이드(2015 버전)》

¹¹ 이 데이터에 사용한 테스트 구성: CPU: 투웨이 인텔® 제온® 골드 6148 CPU, 2.40GHz 코어/스레드: 20/40, 메모리: 16GB DDR4 2666MHz * 12, 하드 디스크: 인텔®SSD SC2BB480G7, BIOS: SE5C620.86B.02.01.0008.031920191559, 운영 체제: CentOS Linux 7.6, Linux 커널: 3.10.0-957.21.3.el7.x86_64, gcc 버전: 7.2(Tensorflow) & 4.8.5(OpenVINO), Python 버전: Python 3.6, Tensorflow 버전: R1.13.1, OpenVINO™ 툴킷: 2019 R1, Keras: 2.1.3

인텔® 딥러닝 가속 기술을 이용한 지멘스, 심혈관 질환 치료 중 AI 응용 가속화

■ 배경

심혈관질환은 오랫동안 인류의 건강을 위협해온 적입니다. 통계에 따르면, 매년 심혈관 질환으로 인한 사망자수가 1,800만명¹²에 달한다고 합니다. 심장 자기공명영상 검사(MRI)를 이용한 심장 자기공명영상(Cardiac Magnetic Resonance, CMR) 이미지 양적 측정은 줄곧 심장 기능과 심실 용량, 심근 조직 평가에 있어 가장 중요한 기준이었습니다. 전통적으로 심혈관 전문의는 경험에 근거해 MRI 영상을 판독하기 때문에 시간과 에너지 소모가 많았고, 오류율도 높은 편이었을 뿐만 아니라 이미지를 해석할 때 주관이 많이 개입되다 오진하는 경우가 많았습니다.

이제 지멘스 의료는 혁신적인 의료 AI 응용 연구를 진행하고, 그 성과를 심장병학과 방사능 활영 영상 분석 실제 응용에 도입합니다. 하지만 이 AI 기능을 의료 분야에 실제로 어떻게 응용하느냐는 또 다른 문제입니다.

임상 치료에는 지연이 허용되지 않기 때문에 AI 응용을 하려면 우선 여러 유형의 검사 기기의 데이터와 동기화를 유지해야 하며, 높은 AI 추론 스플릿과 저지연을 보장해야만 AI 기반 의료 시스템을 더 많은 환자에게 서비스할 수 있습니다. 또한 시간을 절약할 뿐만 아니라 측정과 진단 사이에 일치성과 정확도를 향상시키기 위해 AI 응용을 임상 치료 프로세스와 최대한 융합해야 합니다.

그리기 위해 지멘스 의료는 인텔과 함께 범용 CPU 플랫폼을 기반으로 MRI 영상 판단과 측정을 진행해 고효율의 AI 추론 작업을 실행했습니다. 두 기업은 딥러닝 방식으로 MRI 심혈관 의료 영상을 AI 판단 연구했을 뿐만 아니라 최신 2세대 인텔® 제온® 확장성 CPU 플랫폼과 OpenVINO™ 툴킷 등을 기반으로 최적화 작업을 진행해 추론 속도를 대폭 향상시킴으로써 임상 의료 치료 분야에 강력한 지지대를 제공했습니다.

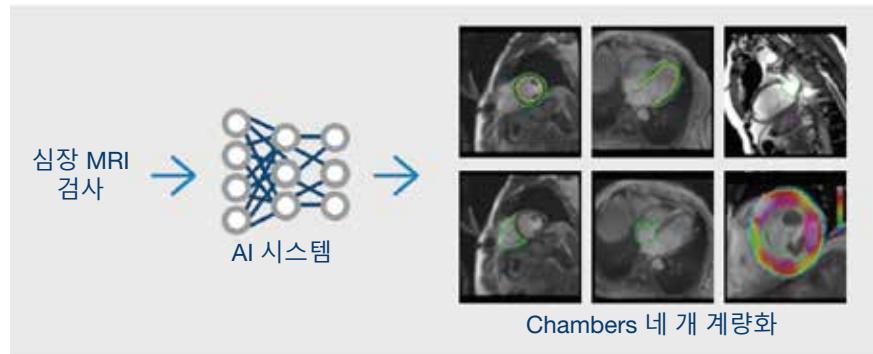


그림 1-4-3 지멘스 의료가 인텔과 함께 구축한 심장 MRI의 AI 분석 능력

■ 방법과 성과

이번 사례에서는 지멘스 의료와 인텔이 합작해 최신 2세대 인텔® 제온® 확장성 CPU를 기반으로 구성한 심실 검사와 양자화 모델을 최적화했습니다. 이 AI 모델은 Dense U-net을 기반으로 하고 있어, 심장의 좌우 심실에 시蔓 тек 세그먼테이션을 진행하고 네 개의 chambers까지 확장할 수도 있습니다. AI 모델의 입력은 흥苁 뛰는 심장의 MRI 이미지를 축적한 것이며, 출력은 심장의 구역과 구조를 식별한 것입니다. 그 종 구조는 모두 색상으로 부호화됩니다. 이렇게 하면 본래 사람이 직접 식별해 태그해야 했던 과정이 AI로 실행되기 때문에 영상 판독 속도가 빨라집니다. 전체 업무 프로세스는 그림 1-4-3과 같습니다.

2세대 인텔® 제온® 확장성 CPU은 이 AI 모델의 추론 작업에 고효율에 유연하고 확장성이 높은 플랫폼을 제공했습니다. OpenVINO™ 툴킷과 긴밀하게 결합한 덕분에 비전 애플리케이션의 딥러닝 추론 작업을 효과적으로 가속화하고, 치료 과정 중 매우 중요한 부분을 차지하는 진단과 의사 결정의 속도와 정확도를 향상시켰습니다. 그 뿐만 아니라, CPU 집약 인텔® 딥러닝 가속 기술은 최신 VNNI를 보유하고 있어 딥러닝의 여러 컴퓨팅 밀집형 작업을 한 단계 더 가속화할 수 있기 때문에 이미지 분류, 이미지 분할, 객체 탐지 등 인텔® CPU 플랫폼에서 AI 응용의 추론 효율을 향상시킬 수 있습니다. 인텔® 딥러닝 가속 기술은 INT8을 지원하기 때문에 FP32 트레이닝 모델을 INT8로

전환할 수 있고 정확도를 유지할 뿐만 아니라 추론 속도도 대폭 향상시킬 수 있습니다.

이번 사례에서 DNN(예: Dense U-net)은 트레이닝을 거친 뒤 심장 구역 식별에 사용되었습니다. 그리고 뉴럴 네트워크의 가중치는 일반적으로 부동 소수점 수(FP32)로 표시되기 때문에 모델은 보통 FP32 정밀도를 통해 트레이닝과 추론을 진행합니다. INT8 역시 마찬가지로 정확도를 아주 작게 잃는(일반적으로 <0.5%, 이번 사례에서는 <0.001%) 상황에서 추론 속도¹³를 향상시킬 수 있습니다.

인텔® 딥러닝 가속 기술과 OpenVINO™ 툴킷이 제공하는 FP32부터 INT8까지의 변환 툴(Calibration Tool)을 통해 인텔은 지멘스가 정확도를 유지하며 더 빠른 속도로 추론 연산을 할 수 있도록 도와줍니다. 그림 1-4-4는 AI를 이용한 심장 이미지 분할을 나타내고 있습니다. 왼쪽 그림은 AI 모델로 심장의 여러 구조를 분할한 것이고, 오른쪽 그림의 위쪽은 INT8 모델을 사용하지 않은 전통적인 ONNX 출력 이미지, 아래쪽은 INT8 모델을 사용한 출력 이미지입니다. 두 그림의 정밀도가 기본적으로 일치하고 있음을 직관적으로 알 수 있습니다.

^{12, 13} 데이터 출처: Journal of the American College of Cardiology, 2017.

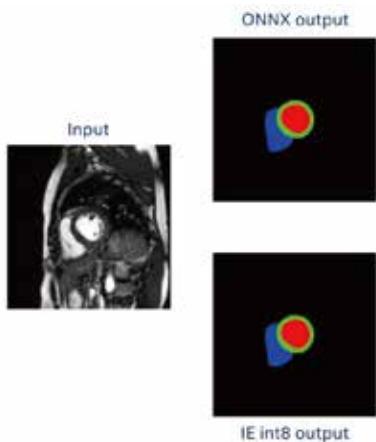


그림 1-4-4 INT8 모델 사용 전후 출력 정밀도 비교

추론 속도 측면에서 보았을 때, 2세대 인텔® 제온® 확장성 CPU, 인텔® 딥러닝 가속 기술 및 OpenVINO™ 툴킷을 기반으로 최적화를 진행하면, 심장 MRI의 AI 분석 기능이 대폭 증가합니다. 심장 MRI 영상의 처리 속도가 향상되어 200 FPS(초당 촬영률)에 도달한다는 것은 1초도 되지 않는 시간 안에 완전한 심장 MRI 검사를 분석까지 완료할 수 있음을 의미합니다. 그 덕분에 심장 MRI를 임상 분야에서 근실시간으로 응용할 수 있는 가능성이 열렸습니다. 또한 최적화를 거친 솔루션은 모델을 양자화 및 실행하면, 정밀도가 거의 감소하지 않는 상황에서 성능이 최적화하지 않은 방안보다 5.5배¹⁴ 향상합니다.

인텔의 기술과 제품을 이용해 딥러닝 모델을 최적화하고 CT 이미지 추론 성능을 향상시킨 제너럴일렉트릭

■ 배경

전산화 단층촬영(Computed Tomography, CT) 검사는 현대 의료 분야에서 가장 자주 사용되는 검사 수단 중 하나입니다. CT는 X선을 통해 신체 표면층을 스캔해 해당 부위의 단층면이나 입체적인 이미지를 얻어 신체의 병변 상황을 발견하는 방식입니다. CT 검사는 임상적으로 매우 중요한 의미를 지니고 있지만 전통적으로 경험이 많은 전문의의 판독에 의존한 방식이다 보니 효율이 낮고, 전문의의 주관적 판단으로 오진을 하거나 진단을 누락하는 경우가 자주 발생합니다.

이제 제너럴일렉트릭(이하 "GE 의료")은 딥러닝 방식을 이용해 CT 절단 이미지를 분류하고 태그하기 때문에 앞으로는 전문의가 미세한 병소를 찾아내 연구나 임상 비교에 사용하기 쉬워졌습니다. 2018년 국제광공학회(SPIE)에서 GE 의료는 AI 기반 구조 분류기 관련 논문을 발표했습니다. 논문에서 GE 의료의 CT 이미징 전문가는 Python 언어와 TensorFlow 아키텍처, Keras 라이브러리를 사용해 새로운 AI 모델을 만들고 트레이닝했습니다. GE 의료는 인텔과 함께 인텔® 제온® CPU, 인텔® 딥러닝 배치 툴(Intel® Deep Learning Deployment Toolkit, 인텔® DLDT) 등 제품과 기술 협작을 통해 CT 추론 지향 솔루션을 최적화했습니다.

■ 방법과 성과

우선 인텔® DLDT로 딥러닝 모델을 최적화하고, 인텔® 제온® CPU 플랫폼으로 더 뛰어난 추론 성능을 선보였습니다.

인텔® DLDT는 OpenVINO™ 툴킷 중 딥러닝 모델에 전문적으로 사용되는 추론 가속 부품입니다. 이 툴을 이용해 트레이닝 컨버전스한 모델은 다양한 인텔® CPU 플랫폼에서 데이터 처리 능력을 향상시키고, 데이터 처리 지연 시간을 줄일 수 있습니다. 그 덕분에 다양한 주류 딥러닝 소스 아키텍처 트레이닝이 잘 된 모델을 전환하고 최적화해 딥러닝 아키텍처에서 독립적인 bin 파일과 xml 파일을 생성할 수 있습니다. bin 파일은 딥러닝 모델의 가중치를 2진법 형식으로 저장하는데 사용되고, xml 파일은 딥러닝 모델의 네트워크 구조를 설명하는데 사용되며 이 두 가지를 결합해 모델을 해석합니다. 그렇기 때문에 모델의 대표 파일은 딥러닝 아키텍처에 의존할 필요 없이 손쉽게 배치를 진행할 수 있습니다. 뿐만 아니라 이 두 개의 파일을 생성하는 과정 중에는 모델에 상수 폴딩, Batch 층 융합, 수평방향층 융합, 빙 노드 삭제 등 모델 최적화 작업을 진행할 수도 있습니다.

그림 1-4-5에서 보는 바와 같이, 인텔® DLDT는 GE 의료가 TensorFlow 등 아키텍처 트레이닝으로 획득한 모델을 손쉽게 도입할 수 있습니다.

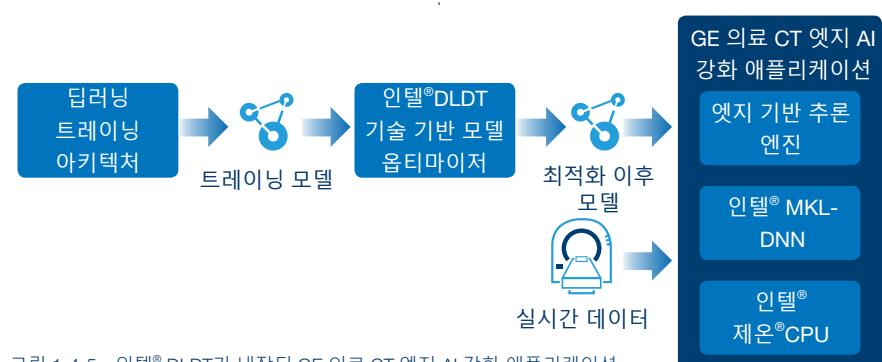


그림 1-4-5 인텔® DLDT가 내장된 GE 의료 CT 엣지 AI 강화 애플리케이션

¹⁴ 이 데이터에 사용한 테스트 구성: CPU: 투웨이 인텔® 제온® 플래티넘 8280 CPU, 2.70GHz 코어/스레드: 28/56, HT: ON, Turbo: ON, 메모리: 192GB DDR4 2933, 하드 디스크: 인텔® SSD SC2KG48, BIOS: SE5C620.86B.02.01.0008.031920191559, 운영 체제: CentOS Linux 7.6.1810, Linux 커널: 4.19.5-1.el7.elrepo.x86_64, gcc 버전: 4.8.5, OpenVINO™ 툴킷: 2019 R1, 워크로드: Dense U-Net

인텔® DLDT를 이용해 모델을 전환 및 최적화한 뒤, 최적화한 모델을 GE 의료 CT 엣지 AI 강화 애플리케이션에 도입합니다. 이 애플리케이션은 인텔® 제온®CPU 플랫폼과 인텔® MKL-DNN을 기초로 하고 엣지를 기반으로 구축된 강력한 추론 엔진입니다.

이 최적화 방안이 실제 효력이 있는지 검증하기 위해 양측은 일련의 성능 테스트를 진행했습니다. 이 데이터세트는 8,834개의 CT 스캔 이미지를 보유하고 있습니다. GE 의료는 모델 최적화 이후, 4개 미만의 CPU 코어를 사용하고도 추론 엔진의 초당 이미지 처리 수량이 100장¹⁵에 다다를 것이라 기대하고 있습니다.



그림 1-4-6 多核心帶來了推論效能的穩步擴展

그림 1-4-6 멀티 코어 덕분에 단계별로 확장된 추론 성능

테스트 결과에서 볼 수 있듯이, 단일 코어만 가능 한 인텔® 제온®CPU E5-2650 v4에서 최적화를 거친 모델은 추론 스크립트를 최적화 이전의 14배까지 향상시킬 수 있습니다. 또한 인텔® 제온®CPU의 멀티 코어 성능은 GE 의료 추론 엔진의 효율도 대폭 향상시킵니다. 그림 1-4-6과 같이, 4개의 CPU 코어를 사용하면, 추론 엔진의 초당 처리 가능한 이미지 수가 초기 기대값의 6배에 달하는 596장까지 향상됩니다.

요약

의료 이미지 분할과 객체 탐지는 AI 의료 영역 응용 분야에 중요한 갈래 중 하나입니다. 훌륭한 이미지 분할 모델은 의료 기관의 의료 영상 판독 효율을 효과적으로 향상시켜줄 뿐만 아니라, 임상 치료 능력을 강화하고 질병 치료 확률을 향상시키며 환자의 대기 시간을 단축하고 의료 기관 영상과 자료 부족으로 인한 여러 문제를 보충해 줍니다.

같은 AI 응용이라 하더라도 의료 영역의 이미지 분할이 다른 이미지 처리 영역과 다른 점은 실효성에 대한 요구 사항이 높다는 점입니다. 환자에게 주어진 골든 타임은 보통 몇십분 밖에 되지 않습니다. 그렇기 때문에 이미지 분할 AI 응용의 추론 효율이 충분히 높지 않으면 귀중한 응급 처치 시간을 지연할 수 있습니다. 다양한 업계와 다양한 시나리오의 사례에서 증명하듯이 인텔® 제온®확장성 CPU와 2세대 인텔® 제온®확장성 CPU, 인텔® 딥러닝 가속 기술 및 OpenVINO™ 툴킷 등의 제품 기술은 딥러닝 모델의 추론 효율을 효과적으로 향상시킬 수 있습니다.

Neusoft의 응용 사례에서는 OpenVINO™ 툴킷이 U-net 모델의 높은 정확도를 보장함과 동시에 추론 시간을 대폭 단축해 뇌졸중 응급 처치를 위한 시간을 벌 수 있었습니다. 지멘스의 응용 사례에서는 인텔® 딥러닝 가속 기술이 INT8을 지원한 덕분에 추론 성능이 한 단계 더 강화되면서 즉시성에 대한 이용자의 요구를 만족시켰습니다. GE 의료의 응용 사례에서는 OpenVINO™ 툴킷의 인텔® DLDT 툴이 다양한 주류 딥러닝 아키텍처의 최적화 기능을 제공하면서 이용자가 더 다양한 최적화 방안을 선택할 수 있게 되었습니다.

무엇보다 중요한 것은 현재 사례 방안 중 대부분이 기존의 인텔® 제온®CPU와 인텔® 제온®확장성 CPU 등 하드웨어 제품으로 진행되었다는 점입니다. 앞으로 이용자는 2세대 인텔® 제온®확장성 CPU, 인텔® 옵테인™데이터 센터급 영구 메모리처럼 더 강하고, AI 영역에서 더 두드러지는 차세대 인텔 제품과 기술을 선택해 솔루션을 구성할 수 있을 것입니다. 인텔은 앞으로도 AI 응용이 앞서 나가는 제품과 기술을 기반으로 의료 분야에서 혁신을 거쳐 정착함으로써 사람들의 건강한 생활을 위해 기여할 수 있는 방법을 탐색할 것입니다.

¹⁵ 이 데이터에 사용한 테스트 구성: CPU: 인텔® 제온® CPU E5-2650 v4, 2.20GHz 코어/스레드: 12/24, HT: ON, Turbo: ON, 메모리: 264GB, 하드 디스크 480GB, 운영 체제: CentOS Linux 7.4.1708, Linux 커널: 3.10.0-693.el7.x86_64, gcc 버전: 4.8.5, 워크로드: 8,834개의 CT 이미지 스캔 데이터세트가 포함되어 있습니다

더 효율적인 의료 영상 분석을 가능하게 하는 AI + Cloud





의료 영역에서의 의료 영상 분석

의료 영상 분석이 직면한 도전

익히 알고 있듯이 높은 수준의 치료를 실행할 수 있는 전제 조건은 병세를 정확히 파악하고 정밀하게 분석하는 것입니다. 오래전에는 의술이 뛰어난 의사가 보고, 듣고, 묻고, 자르는 것으로 병세를 추정했습니다. 하지만 오늘날은 다양한 의료 설비와 정보 시스템, 특히 의료 영상 설비를 이용하면서 의사가 치료 과정을 더 능숙하게 제어하고 환자에게 우수한 의료 서비스를 제공할 수 있게 되었습니다. 지금은 중대형 의료 기관에 X-ray와 CT, 핵자기공명 등 기기가 보급되어 있으며, 기초 의료 기관에서도 환자들은 다양한 의료 영상 검사를 진행할 수 있게 되었습니다.

의료 영상 기기와 시스템은 이처럼 빠르게 자리를 잡았지만, 소위 "소프트웨어"는 이를 따라잡기 어려워 보입니다. 예를 들어, 의료 영상 분석은 영상학과 전문의가 비교적 높은 수준의 전문성을 지니고 있어야 합니다. 임상 의료, 의료 영상학 등 방면의 전문 지식을 섭렵하고 있어야 할 뿐만 아니라 방사선학과 CT, 핵자기공명, 초음파 공학 등 관련 기술을 숙지하고, 다양한 영상 분석 기술을 운용해 질병을 판단해야 합니다.

상황이 이렇다 보니 아무리 의료 기관에 의료 영상 기기가 보급되어 있다 하더라도 일부 외진 지역이나 기초 의료 기관의 경우, 기기는 충분한데 영상을 판독할 수 있는 전문의가 부족한 난감한 상황이 종종 벌어지기도 합니다. 일부 성을 예로 들어보면, 여러 대의 의료 영상 기기가 이미 현, 마을 단위의 의료 기관에 배치되어 있지만 환자의 검사 결과를 현지 전문의가 명확하게 판단하고 분석할 수 없어 영상 파일을 사진 촬영이나 스캔 등 방식을 통해 상위 의료 기관에 전송해 판독하는 설정입니다. 영상 파일의 품질과 왜곡 여부를 보장할 수 없는 경우도 있어서 환자의 병세 판단을 지연하거나 오진하는 일이 종종 발생합니다.

뿐만 아니라 의료 기관별로 정보화 시스템이 독립적으로 형성되어 있어 데이터 기준이 천차만별입니다. 예를 들어, 의료영상전송시스템(Picture Archiving and Communication Systems, PACS)에 저장된 의료 영상 데이터는 서로 연결되지 않아 개별 정보의 "고립"을 형성하는 꼴이 되었습니다. 이로 인해 외진 지역에 거주하는 환자는 기초 의료 기관에서 병세를 효과적으로 분석받을 수 없게 되고, 대형 병원까지 장거리를 이동해야 되는 것도 모자라 중복 검사를 받아야 하는 곤란한 상황이 일어나는 등 모순적인 위험이 존재하고 있습니다.

의료 영상 분석에서 "클라우드 기술+빅데이터"의 응용

클라우드 컴퓨팅 기술을 빠른 발전은 정보 고립 문제를 점차 해결했습니다. 그림 2-1-1과 같이 점점 더 많은 의료 기관이 의료 기기와 의료 서비스 과정을 모두 클라우드 방식을 통해 연결하기 시작했고, 또 이를 기반으로 전체 의술 협력 플랫폼, 영상 협력 플랫폼 등을 구축했습니다. 서비스로서의 플랫폼(Platform as a Service, PaaS) 또는 서비스로서의 소프트웨어(Software as a service, SaaS) 방식으로 등급별 의료 기관의 다양한 요구를 만족시키고 있는 것입니다.



그림 2-1-1 의료기기를 연합한 클라우드 서비스

예를 들어, 전체 의술 협력 서비스 플랫폼은 클라우드 서비스를 도입함으로써 등급별 의료 기관이 단자와 플랫폼을 넘어서는 의술 기능 응용을 실현할 수 있게 되었습니다. 영상 협력 플랫폼은 중대형 의료 기관의 의료 영상 전문가가 각 지역에서 전달해온 영상 데이터를 언제 어디서든 처리할 수 있게 하고, 난치병은 합동 진단해 의료 자원을 효율적으로 공유할 수 있게 했습니다.

의료 영상 데이터는 클라우드 컴퓨팅과 빅데이터 기술 상호 연결을 기반으로 하기 때문에 모든 의료 기관이 과도한 검사와 중복 치료 등 문제를 방지할 수 있을 뿐만 아니라 데이터 고립 현상에서 벗어나게 해주어 의료 기관 전체를 빈틈 없이 연결함으로써 의료 서비스 품질을 향상시켰습니다. 또한 영상 데이터 축적과 분석을 통해 AI를 기반으로 한 의료 영상 분석 응용을 나날이 성숙시키고 있습니다. 이제 클라우드 기술+AI 기반의 의료 영상 분석 기능은 대부분의 의료 기관에 배치되었고, 긍정적인 피드백을 받고 있습니다.

AI 기반 의료 영상 분석

클라우드 서비스와 빅데이터 시스템을 통해 수집한 다양한 데이터는 표적 검출 뉴럴 네트워크 등의 AI 모델이 다양한 트레이닝 견본을 얻을 수 있게 해주고, AI를 기반으로 한 스마트화 보조 진단 시스템은 의료 기관의 치료 능력을 효과적으로 향상시켜 줍니다.

폐암 조기 발견을 예로 들어 보면, 폐암은 목숨을 위협하는 무서운 악성 종양이지만 초기에는 증상이 없어 간과하기 쉬운 사르코이드증입니다. 사르코이드증을 초기에 발견(양성 또는 악성)하면 폐암으로 인한 사망률을 효과적으로 낮출 수 있습니다. 사르코이드증은 아주 미세해 육안으로 적시에 정확하게 발견하기엔 어려움이 많습니다. 그렇다 보니 폐암은 보통 증기나 말기에 발견되고 환자는 최적의 치료 적정기를 놓치게 됩니다.

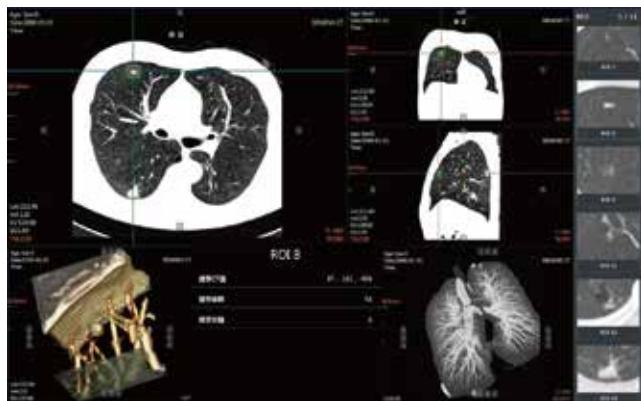


그림 2-1-2 저선량 CT를 이용한 폐 소결절 스마트화 보조 진단

그림 2-1-2와 같이 일부 의료 기관에서는 의료 영상 AI 분석 응용의 일환으로 저선량 CT를 이용해 폐의 소결절에 스마트화 보조 진단을 진행하고 있습니다. 실제 적용 결과 데이터에서 볼 수 있듯이, 정량의 모니터링 민감도(검출 감도)가 이미 95%에 다다랐고, 검사 시간 역시 사람이 진행했을 때 소요되는 시간인 10분이 초단위¹⁶까지 축소되었습니다. AI 모델을 통해 사르코이드증을 식별한 뒤 전문의에게 넘겨 심층 진단을 진행하며, 효율과 정밀도가 대폭 향상됩니다.

현재 표적 검출 뉴럴 네트워크는 광범위하게 운용되는 의료 영상 AI 분석 응용 분야 중 하나로, X-ray, CT 이미징 등 의료 영상을 고효율로 정확하게 검사할 수 있습니다.

표적 검출 뉴럴 네트워크

전형적인 표적 검출 뉴럴 네트워크로는 R-CNN, Fast R-CNN, SPP-NET, R-FCN¹⁷ 등이 있습니다. R-FCN은 최근 의료 영상 분야에서 자주 볼 수 있는 표적 검출 뉴럴 네트워크 모델입니다.

¹⁶ 데이터 출처: AccuRad 내부 테스트 데이터: <https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yinggu-case-study-medical.html>

¹⁷ R-FCN 관련 기술 설명 출처: Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, <https://arxiv.org/pdf/1605.06409v2.pdf>

전형적인 R-FCN 구조는 그림 2-1-3과 같으며, 우선 처리해야 할 영상 사진을 전처리한 뒤, ResNet-101 네트워크와 같은 사전에 트레이닝 완료된 컨볼루션 뉴럴 네트워크(CNN)에 발송합니다. 이 네트워크의 가장 마지막 부분인 컨볼루션 층이 획득한 특징 맵(feature map)에서 3개의 갈래를 끄집어냅니다. 첫 번째 갈래는 특징 맵을 RPN(Region Proposal Network)에 도입하고, 그에 상응하는 관심영역(Region Of Interest, ROI)을 획득합니다. 두 번째 갈래는 이 특징 맵에서 분류에 사용되는 다차원 포지션 센시티브 스코어 맵(position-sensitive score map)을 획득합니다. 세 번째 갈래는 이 특징 맵에서 회귀에 사용되는 다차원 포지션 센시티브 스코어 맵을 획득합니다. 마지막으로 두 개의 다차원 포지션 센시티브 스코어 맵에서 각각 포지션 센시티브 ROI 폴링(Position-Sensitive ROI Pooling)을 실행하고, 그를 통해 상응하는 유형과 포지션 정보를 획득합니다.

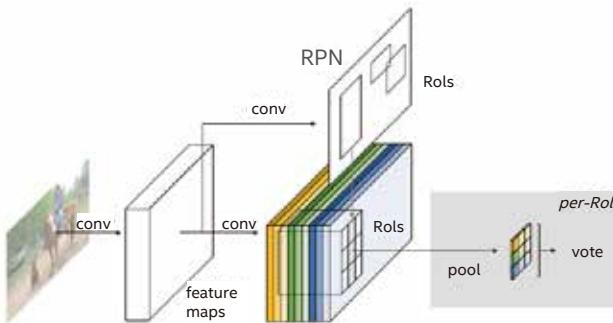


그림 2-1-3 전형적인 R-FCN 구조

Faster R-CNN 같은 다른 표적 검출 뉴럴 네트워크와 비교했을 때, R-FCN은 검출 속도가 더 빠르고 검출 정밀도 역시 더 높은 특징¹⁸을 지니고 있습니다.

소프트웨어/하드웨어 구성 제안

AI 기반 의료 영상 분석 방안 구축은 아래의 인텔® 아키텍처 플랫폼 기반 소프트웨어/하드웨어 구성을 참조해 완료할 수 있습니다.

파라미터	설명
CPU	인텔® 제온® 골드 6240 CPU 또는 그 이상
하이퍼스레딩	ON
터보 가속	ON
메모리	16GB DDR4 2666MHz* 12 및 이상
저장	인텔® SSD D5 P4320 시리즈 및 이상
운영 체제	CentOS Linux 7.6 또는 최신 버전
Linux 코어	3.10.0 또는 최신 버전
컴파일러	GCC 4.8.5 또는 최신 버전
Caffe 버전	인텔® 아키텍처 최적화 지향 Caffe 1.1.6 또는 최신 버전

최적화 AI 모델 효율

인텔® 아키텍처 기반 CPU 플랫폼의 최적화

인텔® 제온® 확장성 CPU, 2세대 인텔® 제온® 확장성 CPU 등이 포함된 인텔® 아키텍처 CPU 플랫폼은 AI + Cloud 기반 스마트 의료 영상 분석 시스템에 강력한 범용 컴퓨팅 기능을 부여할 수 있을 뿐만 아니라 더 시급한 병렬 컴퓨팅 기능도 제공할 수 있습니다. 딥러닝 모델 추론 과정 중에는 높은 수준의 병렬 컴퓨팅 능력을 요구하는 경우가 종종 있습니다. 인텔® 제온® 확장성 CPU는 인텔® AVX -512를 도입함으로써 더 효과적인 SIMD(Single Instruction Multiple Data)를 제공해 시스템의 병렬 컴퓨팅 가속 능력을 강화했습니다.

게다가 인텔® 매스 커널 라이브러리(Intel® Math Kernel Library, 인텔® MKL), 인텔® MKL-DNN이 추가되면 AI 모델의 업무 효율을 한층 더 향상시킬 수 있습니다. 주로 아래의 세 가지 측면에서 인공지능 모델 성능이 향상됩니다.

- Cache Blocking 기술 최적화 데이터 캐시를 사용해 데이터 적중률을 향상시킵니다.
- 뉴럴 네트워크에서 자주 사용하는 연산자에 대해 병렬화 및 벡터화 최적화를 진행합니다.
- Winograd 알고리즘 최적화를 사용합니다.

최신형 2세대 인텔® 제온® 확장성 CPU에 추가된 인텔® 딥러닝 가속 기술은 딥러닝 추론이 INT8을 사용해 더 뛰어나게 성능을 표현할 수 있게 해줍니다.

인텔® 제온® 확장성 CPU 플랫폼에서 단일 폭 흉부 Dicom 데이터로 R-FCN 모델을 실행하는 경우를 예로 들면, 애플리케이션의 데이터를 통해 볼 수 있듯이 인텔® 제온® 골드 6148 CPU를 최적화하면 성능을 5배¹⁹ 가까이 향상시킬 수 있습니다.

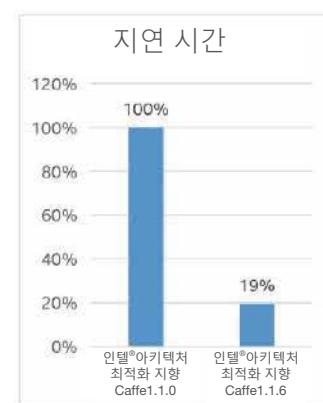


그림 2-2-1 단일 폭 흉부 Dicom 데이터로 R-FCN 모델 처리 지연 실행 비교

¹⁸ R-FCN 성능 데이터는 다음을 참조하십시오. Jifeng Dai, Yi Li, Kaiming He, Jian Sun, R-FCN: Object Detection via Region-based Fully Convolutional Networks, <https://arxiv.org/pdf/1605.06409v2.pdf>

¹⁹ 성능 테스트 결과는 [2019년 4월 10]에 진행한 테스트이며, 테스트 구성은 다음과 같습니다. 투웨이 인텔® 제온® 골드 6148 CPU, 20 코어/40 스레드, HT/Turbo 가동, 192GB 메모리(12 slots / 16GB / 2666MHz) 탑재, CentOS 7.6, BIOS: SE5C620.86B.02.01.0008.031920191559(uncode: 0x200005e), Kernel 버전: 3.10.0-957.21.3.el7.x86_64, 컴파일러 GCC 4.8.5입니다. 테스트 그룹은 인텔® MKL-DNN 0.12 버전을, 대조 그룹은 인텔® MKL-DNN 0.18 버전 사용. 아키텍처 인텔® 아키텍처 최적화 지향 Caffe 1.1.0, 대조 그룹은 인텔® 아키텍처 최적화 지향 Caffe 1.1.6. Minibatch=1으로 구성해 완료했습니다.

인텔® 아키텍처 최적화 지향 Caffe

버클리 비전 앤 러닝 센터(Berkeley Vision and Learning Center, BLVC) 버전의 Caffe²⁰와 비교하면 인텔® 아키텍처 최적화 지향 Caffe²¹는 인텔® 아키텍처 지향 제품을 거냥해 다양 최적화를 진행했고, 인텔® MKL, 인텔® MKL-DNN과 인텔® AVX-512 지원까지 추가되어 딥러닝 모델별로 뛰어난 성능을 보이고 있으며, 추론 효율 또한 높은 편입니다.

인텔® 아키텍처 CPU의 컴퓨팅 리소스가 충분히 이용될 수 있도록 일반적으로 추론을 실행하기 전 다음과 같이 일부 환경 변수를 설정할 수 있습니다.

1. export OMP_NUM_THREADS=36

여기에서 OMP_NUM_THREADS는 사용할 스레드 수를 가리킵니다.

BLVC Caffe에 실시한 성능 분석을 통해 인텔® 아키텍처 최적화 지향 Caffe는 다음 몇 가지 측면에서 최적화를 진행했습니다.

■ 코드 벡터화 최적화

최적화 내용은 다음과 같습니다.

- 블 라 스 (BLAS) 라이브러리를 아틀라스(ATLAS)에서 인텔® MKL-DNN으로 전환해 범용 행렬 곱셈(GEMM) 등을 최적화하면 벡터화와 다중 스레드화된 워크로드에 더 적합해지고 캐시량도 향상됩니다.
- Xbyak just-in-time(JIT) 어셈블러를 사용해 컴파일링 과정을 실행합니다. x86/x64 JIT 어셈블러로서 Xbyak은 MMX™ 기술, 인텔® 스트리밍 SIMD 확장(Intel® Streaming SIMD Extensions, 인텔® SSE), 인텔® AVX 시리즈 기술 등 인텔® 아키텍처의 명령어 조합을 지원합니다. 또한 코드 실행 과정 중 인텔® 아키텍처 최적화 지향 Caffe가 벡터화율을 향상시킬 수 있도록 도와줍니다.
- GNU Compiler Collection(GCC)과 Open Multi-Processing(OpenMP)에 코드 벡터화를 진행합니다. 벡터화율이 향상되면 SIMD 명령어 더 많은 데이터를 동시에 처리하기 유리하기 때문에 데이터

병렬 이용률이 향상됩니다. 게다가 코드를 벡터화 처리하면 딥러닝 모델 중 풀링 계층의 성능도 향상시킬 수 있습니다.

■ 일반 코드 최적화

최적화 내용은 다음과 같습니다.

- 프로그래밍 복잡성 감소
- 컴퓨팅 수량 감소
- 사이클 전개.

예를 들어 코드 최적화 과정 중 스칼라 최적화 스킬을 채택하면, 코드는 다음과 같습니다.

```
1. for (int h_col = 0; h_col < height_col; ++h_col) {
2.     for (int w_col = 0; w_col < width_col; ++w_col) {
3.         int h_im = h_col * stride_h - pad_h + h_offset;
4.         int w_im = w_col * stride_w - pad_w + w_offset;}
```

코드 세그먼트의 세 번째 열은 h_im 컴퓨팅 관련 열로 아래와 같이 이 열을 최내층에서 시프트아웃할 수 있습니다.

```
1. for (int h_col = 0; h_col < height_col; ++h_col) {
2.     int h_im = h_col * stride_h - pad_h + h_offset;
3.     for (int w_col = 0; w_col < width_col; ++w_col) {
4.         int w_im = w_col * stride_w - pad_w + w_offset;}}
```

■ 기타 인텔® 아키텍처 CPU 기반 최적화 조치

최적화 내용은 다음과 같습니다.

- im2col_cpu/col2im_cpu 실행 효율을 개선합니다. im2col_cpu 함수는 딥러닝 컴퓨팅 상용 함수로 최적화한 BLAS 라이브러리를 사용하면 GEMM 방식으로 직접 컨볼루션을 실행할 수 있습니다. im2col_cpu에 BLVC Caffe 코드 중 아래와 같이 최적화를 실행할 수 있습니다.

```
1. for (int c_col = 0; c_col < channels_col; ++c_col)
2.     for (int h_col = 0; h_col < height_col; ++h_col)
3.         for (int w_col = 0; w_col < width_col; ++w_col)
4.             data_col[c_col*height_col+h_col*width_col+w_col] = // ...
```

그 중 네 차례의 연산(두 번은 덧셈, 두 번은 곱셈)은 단일 인덱스 증분 연산으로 대체함으로써 연산 효율을 향상시킬 수 있습니다.

- 배치 정규화의 복잡성 감소
- 특정 CPU/시스템 최적화 방법
- 모든 컴퓨팅 스레드를 하나의 코어로

잠금 처리해 스레드 이동을 방지하면 다음 환경 변수로 설정해 실현할 수 있습니다.

```
1. export KMP_AFFINITY=granularity=fine,compact,1,0
```

인접한 스레드를 치밀하게 설정하면 GEMM 작업 성능을 향상시킬 수 있습니다. 스레드는 동일한 LLC를 공유할 수 있기 때문에 사전에 취득한 캐시 라인을 데이터에 다시 사용해 효율을 높일 수 있습니다.

■ OpenMP를 통해 코드 병렬화 실현

OpenMP 다중 스레드 병렬 처리 방법을 채택하면 뉴럴 네트워크의 추론 효율을 향상시킬 수 있습니다. 예를 들어, 풀링 계층 중 단일 풀링 계층은 단일 특성도 처리에 적용됩니다. 하지만 풀링 계층을 OpenMP 다중 스레드와 병렬 실행하면 독립적인 이미지 이미지 때문에 여러 스레드로 여러 개의 이미지를 동시에 병렬 처리할 수 있어 효율이 향상됩니다. 코드는 아래와 같습니다.

```
1. #ifdef _OPENMP
2. #pragma omp parallel for collapse(2)
3. #endif
4. for (int image = 0; image < num_batches; ++image)
5.     for (int channel = 0; channel < num_channels; ++channel)
6.         generator_fn(bottom_data, top_data, top_count, image, image+1,
7.                         mask, channel, channel+1, this, use_top_mask);
8. }
```

collapse(2) clause, OpenMP #pragma omp parallel을 이용하면 두 개의 for-loop 내포된 어구까지 확장할 수 있고, 다시 배치 반복 이미지와 이미지 채널 두 개의 사이클을 하나의 사이클로 통합한 뒤 이 사이클을 병렬화 처리할 수 있다는 걸 알 수 있습니다.

인텔® 아키텍처 최적화 지향 Caffe는 일련의 최적화 방법과 기술을 통해 성능면에서 BLVC Caffe보다 크게 성장했습니다. 우리는 테스트를 통해 인텔® 아키텍처 최적화 지향 Caffe의 워크로드 실행 시간이 기존 시간보다 10% 단축되고, 전체 실행 성능이 기존보다 10배 이상 향상되었음을 알 수 있었습니다.

* 인텔® 아키텍처 최적화 지향 Caffe 기술과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

²⁰ 이 버전의 소스 코드는 <https://github.com/BVLC/caffe>를 참조하십시오

²¹ 이 버전의 소스 코드는 <https://github.com/intel/caffe>를 참조하십시오

²² 테스트 관련 데이터 및 인텔® 아키텍처 최적화 지향 Caffe의 최적화 방법에 대한 더 많은 정보는 다음을 참조하십시오.《Caffe* Optimized for Intel® Architecture: Applying Modern Code Techniques》: <https://software.intel.com/en-us/articles/caffeo-optimized-for-intel-architecture-applying-modern-code-techniques>

AI 기술과 클라우드 서비스를 이용해 의료 치료 보조 기능을 향상시킨 AccuRad

배경

의료 자원 구성의 불균형으로 다양한 의료 기관의 의료 영상 후처리, 후분석 능력 또한 천차만별입니다. 데이터가 서로 연결되어 있지도 않아 의료 자원의 이용 효율 역시 효과적으로 향상되기 어려운 면이 있습니다. 20년 가까이 의료 영상 코어 기술을 연구해온 AccuRad 네트워크 테크놀로지 유한회사(이하 "AccuRad")는 전문 의료 영상 코어 기술과 제품에 클라우드 컴퓨팅과 빅데이터, AI 기술을 결합시켜 고효율의 스마트한 의료 스마트화 보조 진단 기능을 형성함으로써 여러 의료 기관이 치료 효율과 품질을 높일 수 있도록 도왔습니다.

AccuRad는 의료 영상 분석 처리 능력 불균형 문제를 해결하기 위해 클라우드 컴퓨팅 등의 방식으로 의료 영상 데이터를 효과적으로 취합하고, AI 기반 데이터 분석 능력을 형성할 뿐만 아니라 더 나아가 자원 공유와 AI 두 가지 능력을 이용해 의료 기관별 의료 영상 분석 기능 차이를 반드시 없애야 한다고 판단했습니다.

이를 위해 AccuRad는 메디컬 클라우드를 배치하고, 혁신적인 의료기기 사물인터넷 기술 AMOL을 이용해 여러 기기에서 유래된 다양한 의료 영상 데이터를 연결하기 시작했습니다. AccuRad는 딥러닝을 의료 영상 처리에 머신러닝을 도입해 표적 검출 뉴럴 네트워크 모델을 기반으로 최신형 Cloud IDT 서비스를 구축해 검출률을 높이고, 의사 결정은 낫혔으며, 업무 효율은 높이는 등 여러 방면에서 효과를 거두었습니다.

AccuRad가 이 시스템을 더 확실하게 배치하고 정착화할 수 있도록 인텔은 인텔® 제온® 확장형 CPU 등 최신 플랫폼 제품과 기술을 제공함으로써 Cloud IDT 서비스의 인텔® 아키텍처 플랫폼으로 이전을 도왔고, 또한 Caffe, TensorFlow 등 딥러닝 아키텍처를 배치 및 최적화했습니다. 양측의 협업과 노력 덕분에 의료 스마트화 보조 진단 시스템은 검사 시간과 정확도 등 여러 측면에서 사용자에게 호평을 받고 있습니다.

방법과 성과

새로운 방안에서, AccuRad는 표적 검출 뉴럴 네트워크 모델을 기반으로 일련의 의료 영상 분석 처리 애플리케이션을 구축하고, 인텔® 아키텍처 CPU를 채택해 고효율의 모델 추론을 실행했습니다. 또 한편으로는 AccuRad Cloud IDT 스마트 애플리케이션을 의료 영상 처리 및 분석 클라우드 컴퓨팅 @ iMAGES 코어 엔진 등과 결합함으로써 강력한 영상 빅데이터 온라인 스마트 처리 능력을 향상시켰습니다.

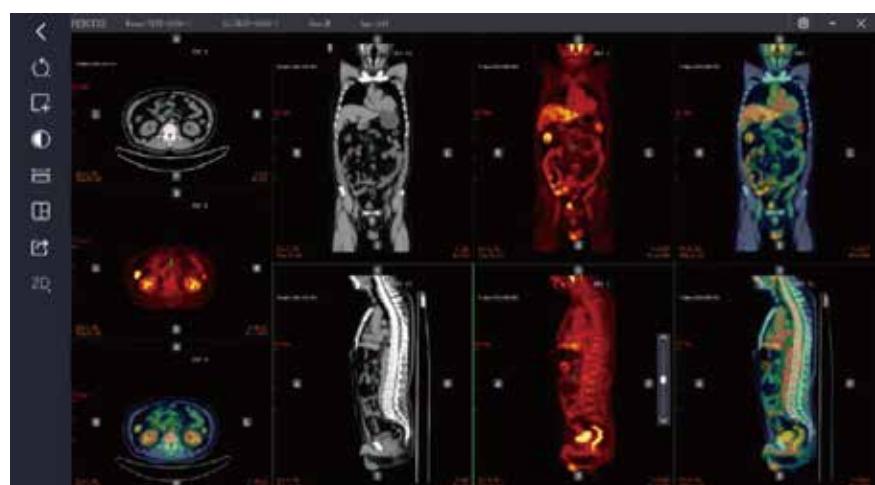


그림 2-3-1: 클라우드 PET-CT 융합

그림 2-3-1과 같이 인텔® 아키텍처 기반 CPU가 제공하는 강력한 해시 레이트와 @iMAGES 코어 엔진이 제공하는 클라우드 기반 펫시티(Positron Emission Tomography CT, PET-CT) 융합 능력을 결합하면 형태학과 기능 기반의 "열지도"를 제공할 수 있을 뿐만 아니라, 영상에 준양자화 SUV(Standard Uptake Value) 분석도 할 수 있습니다. 이런 영상은 Cloud IDT 스마트 시스템 중 R-FCN 표적 검출 뉴럴 네트워크를 통해 종양 등 질병 감별과 정량 분석도 실행할 수 있습니다.

인텔은 뛰어난 하드웨어 성능을 기초로, Caffe, TensorFlow 등 인공지능 아키텍처를 최적화해 AccuRad Cloud IDT 스마트 시스템의 실행 효율을 향상시켰습니다. R-FCN 모델 최적화를 통해 모델 자르기 융합은 성능을 30% 가까이 향상시켰고, OpenMP 다중 스레드 실현 방안을 최적화한 뒤에는 성능이 다시 40-50%²³ 향상되었습니다.

²³ 데이터 출처: AccuRad 내부 테스트 데이터: <https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yinggu-case-study-medical.html>, 사용한 테스트 구성: CPU: 투웨이 인텔® 제온® 골드 6148 CPU, 2.40GHz 코어/스레드: 20/40, HT: ON, Turbo: ON, 메모리: 192GB DDR4 2666, 하드 디스크: 인텔® SSD SC2KB48 네트워크 어댑터: 인텔® 이더넷 융합 네트워크 어댑터 XC710 BIOS: SE5C620.86B.02.01.0008.031920191559, 운영 체제: CentOS Linux 7.6, Linux 커널: 3.10.0-957.21.3.el7.x86_64, gcc 버전: 4.8.5, Caffe 버전: 인텔® 아키텍처 최적화 지향 Caffe 1.1.6 워크로드: R-FCN

그 밖에도 인텔® 제온® 확장성 CPU는 범용 컴퓨팅 기능과 병렬 컴퓨팅 기능 면에서 해시 레이트를 지원하며, 스마트 시스템이 여러 플랫폼이 분산된 기존의 임무를 처리할 수 있도록 도와주기도 합니다. 예를 들어 데이터 통계를 모델 추론과 함께 통합함으로써 사용자가 개인

클라우드에 더 많은 가상 컴퓨터를 배치해 총소유비용(Total Cost of Ownership, TCO)을 절감할 수 있게 도와줍니다.

AccuRad는 이미 AI + Cloud 모드에 기초해 사르코이드증 진단, 갈비뼈 골절 진단, 폐결핵 진단 등 일련의 스마트 보조 진단 시스템을 구축했으며 아래 표와 같습니다.

AccuRad 가 AI+Cloud 에 기초해 구축한 스마트 외래 진찰 시스템의 능력²⁴

사르코이드증 진단	다수의 전문의가 작성한 CT 데이터를 기초로 딥러닝 기술과 3D 입체 사진 처리 기술을 이용해 특정 DNN과 이미지 알고리즘을 설계하면 흉부 CT 데이터에서 3mm 이상 사르코이드증의 위치를 측정하고, 결정의 크기와 악성 지표를 컴퓨팅합니다. 검사 정확도는 95%입니다.
늑골 골절 진단	흉부 투시 X선 사진 데이터에 기초한 전자동 스마트 탐지 시스템의 주요 목적은 늑골 골절 탐지입니다. 딥러닝 기술과 이미지 처리 기술을 이용해 골절 위치를 측정하고 이미지에 자동 표시합니다. 탐지 정확도는 90% 이상으로 전문의가 골절을 빠르게 발견하고 진단할 수 있게 도와줍니다.
폐결핵 진단	흉부 투시 X선 사진 폐결핵 자동 탐지 시스템은 첨단 이미지 처리 기술과 AI 머신러닝 알고리즘을 기반으로 흉부 투시 X선 사진 고해상도 디지털 이미지를 스캔해 병소로 의심되는 부분을 탐지하고 평가하며, 탐지 결과를 간단하고 빠르게 구현합니다. 민감도가 86%에 달하기 때문에 의료진에게 중요한 정보가 될 것입니다.
폐렴 AI 탐지	폐렴 AI 탐지는 흉부 X-ray로 폐렴 의심 병소 위치를 탐지해 낼 수 있는 방식으로, 주로 전문의가 자주 진단하는 폐부 염증 선별에 응용되어 전문의의 진단 효율을 높여 줍니다. 폐렴 탐지 지표의 민감도는 82%입니다.
흉부 필름 검사	흉부 필름 검사는 흉부 X-ray로 정상적인 흉부 투시 X선 사진을 전체 흉부 투시 X선 사진에서 골라내 검사 작업량을 줄여 주기 때문에 전문의는 비정상 데이터를 명확히 진단하는 데만 집중할 수 있습니다. 이 서비스는 주로 검사 시나리오(예: 건강검진)에 응용되고, 흉부 필름 검사 지표는 그 민감성이 99%, 특이성이 22%까지 다다를 수 있습니다.
진폐증 스마트검사	흉부 투시 X선 사진 진폐 자동 탐지 시스템은 첨단 이미지 처리 기술과 AI 머신러닝 알고리즘을 기반으로 합니다. 특정 DNN과 이미지 알고리즘을 이용해 흉부 투시 X선 사진 고해상도 디지털 이미지를 스캔하고 병소로 의심되는 부분을 탐지한 뒤 탐지 결과를 간단하고 빠르게 구현합니다. 민감도는 92%에 달합니다.
유방암 스마트 조기 선별	딥러닝을 기초로 다수의 전문의가 작성한 유방촬영 필름 데이터에 특정 DNN을 설계해 유방촬영 필름에서 비정상 종양과 석회화 등 병변을 자동 식별해냅니다. 석회화 반점과 종양의 민감도는 95%까지 다다르며, 석회화 식별 민감도는 92%입니다.
사카라이드 과형성 병변 검사	광각안저촬영은 안저 데이터에 기반한 전자동 스마트 탐지 시스템으로, 주요 목적은 사카라이드 탐지 및 레벨별 예측에 있습니다. 데이터 대조 검사와 의학 로직 모듈을 구축해 안저 이미지에서 여러 가지 안저 병변과 질병을 구현해 전문의가 조기에 환자를 발견할 수 있도록 돋고 오진 확률도 낮춰 줍니다.
소장폐쇄 스마트 식별	일반 장폐색증 데이터의 전자동 스마트 탐지의 주요 목적은 장 공동 내적 액면 탐지입니다. 특정 AI 스마트 알고리즘과 이미지 알고리즘을 설계해 복부 입위 이미지에서 여러 가지 폐색 병변과 질병을 구현해 냈으며, 다양한 형식의 검사를 제공함으로써 전문의가 급성 복통 환자를 효과적으로 선별할 수 있게 도왔습니다. 덕분에 오진 확률도 낮아졌습니다.
CTA 관상 동맥 스마트 진단	CTA 관상 동맥 전자동 진단은 CTA 박층 이미지를 스마트하게 처리 분석해 프로세스 전체가 자동화된 관상 동맥 분할과 세그먼트 탐지, 패치 분류 탐지, 협착 분석, 석회화 지수를 구현했습니다. 또한 VR과 단면의 시각 교구를 결합해 최종적으로 자동화, 구조화 보고 출력력을 구현함으로써 기존의 CTA 관상 동맥 진단보다 효율을 6배 이상 향상시킵니다.

²⁴ AccuRad의 AI 응용 소개 및 관련 데이터 출처: AccuRad Medical Real Cloud AI 공식 홈페이지. <http://ai.yizhen.cn/#Page03>

요약

데이터를 통한 의료 정보화는 인텔과 AccuRad를 포함한 협력 파트너 모두가 꿈꾸는 미래입니다. 클라우드 컴퓨팅과 IoT, 빅데이터 및 AI 등에 기초한 기술의 의료 정보화와 스마트화 응용은 이미 광범위하게 구현되었고, 의료 영상 데이터 실시간 컴퓨팅 구현과 의료 비주얼 데이터 인공지능 연구 등 여러 측면에서도 새로운 진전을 이루었습니다. 실제 배치해 실시한 여러 의료 기관에서도 긍정적인 피드백을 얻었습니다.

인텔® 아키텍처 기반 플랫폼에서 현재 주류 AI 아키텍처의 잠재력을 끊임없이 발굴하기 위해, 인텔은 이 아키텍처에서 다방면의 최적화 작업을 펼쳐 왔습니다. 인텔® 아키텍처 최적화 지향 Caffe 아키텍처는 코드 벡터화와 OpenMP 병렬화 등 최적화 수단을 통해 모델의 전체 성능이 BLVC Caffe와 비교했을 때 대폭 향상되었고, AccuRad Cloud IDT 스마트 애플리케이션과 의료 영상 처리 및 분석 클라우드 컴퓨팅@ iMAGES 코어 엔진 등 애플리케이션을 결합한 뒤에는 사르코이드증 진단 등 중요한 시나리오에서 "AI+Cloud" 기반 스마트 보조 진단 시스템 능력을 구축했습니다.

2세대 인텔® 제온® 확장성 CPU와 인텔® 옵테인™ 데이터 센터급 영구 메모리 등 새로운 세대의 인텔 기술과 제품이 대량으로 출시되면서, 인텔® 아키텍처 플랫폼에 기초해 구축한 의료 영상 분석 솔루션은 더 강력한 성능과 우수한 AI 능력을 선보일 수 있을 것입니다. 인텔은 앞으로도 더 많은 파트너와 지속적으로 협력하면서 더 다양하고 앞서가는 제품과 기술을 의료 정보화에 결합해 정밀 의학, 스마트 의학 발전을 촉진함으로써 정보화, 디지털화, 스마트화가 의료 서비스 수준을 효과적으로 향상시켜 환자에게 편안하고 친근한 의료 헬스 서비스를 제공할 계획입니다.

조직 분석을 가속화한 AI 기술



의료 영역의 조직 분석

전통적인 조직 분석 방법이 직면한 문제점

조직 검사는 일부 병변 조직 또는 장기를 일련의 처리를 거쳐 미크론 조각으로 만든 뒤 유리에 붙여 염색 처리해 병리학과에 전달하면, 병리학과 전문의가 현미경을 이용해 검사를 진행하고 병리 변화를 관찰해 진단을 내리고 예후 평가를 과정을 의미합니다. 조직 검사는 아주 복잡하고 도전적인 특성이 두드러지는 작업이기 때문에 병리학 전문의가 되려면 수만장을 판독한 다년간의 경험과 풍부한 전문 지식을 보유해야 합니다. 통계에 따르면, 현재 전국 병리학과 전문의는 만 명²⁵이 채 되지 않는다고 합니다.

또한, 인력을 통한 검사는 주관적 판단이 배제될 수 없어 동일한 환자라도 하더라도 전문의마다 판단이 다르고, 오진을하거나 진단을 누락하는 경우도 자주 발생하게 됩니다. 게다가 실제 조직 검사에서 환자의 조직을 40배의 확대 배율로 디지털화하면 단일 조직의 화소는 백만 화소를 초과할 수도 있습니다. 따라서 병리학과 전문의는 수백만 화소에 달하는 사진 여러장을 연속으로 관찰하고, 미세한 부분에서 이상한 점을 짚어내야 하니 시간과 에너지 소모가 많을 뿐만 아니라 오진과 진단 누락하는 경우가 잦았습니다. 판독 시간이 길어 환자 대기 시간까지 지연되어 때로는 환자의 병세를 악화시키기도 했습니다.

AI 기반 조직 분석 방법

AI에 기초한 이미지 처리와 분석 기술이 크게 발전하면서 여러 의료 기관은 딥러닝 또는 머신러닝 기반 조직 분석 방법 발전에 힘을 기울였고, 그 결과 좋은 성과를 얻었습니다. 예를 들어, ResNet50 네트워크로 진행한 딥러닝 모델 트레이닝은 종양 병리 조직 판별 작업에 사용할 수 있습니다. 작업을 통해 얻은 종양 예측 서모그램에 여전히 소음 문제가 해결되지 않았지만, 이미 병리학과 전문의처럼 다양한 확대 배율로 조직 이미지를 검사할 수 있는 수준에 도달했습니다. 실험을 통해 알 수 있듯이, 의료 기관은 Deep 계열 네트워크 모델 하나를 트레이닝해 전문적인 검사 기술과 상상을 초월한 검사 속도, 무기한 작업 시간을 보유할 수 있게 되었습니다.

뉴욕 대학교의 최신 연구 결과에 따르면, 다양한 디지털화 조직 이미지 트레이닝 Inception v3 딥러닝 모델을 이용하면 암조직과 정상 조직 식별 정확도가 99%에 다다르며, 선암과 편평세포암종을 구분하는 정확도 역시 97%²⁶에 이른다고 합니다.

이제 CNN을 기반으로 한 정렬 알고리즘과 표적 검출 알고리즘은 모두 장족의 발전을 이루었습니다. 딥러닝의 대표 방법 중 하나로서, LeNet과 ZFNet, VGGNet, ResNet 등 CNN의 대표는 이미 이미지 정렬과 얼굴 인식, 표적 위치추정, 이미지 분석 등 분야에서 광범위하게 사용되고 있습니다.

²⁵ 이 데이터의 출처는 한 대중 매체의 보도입니다. <https://www.cn-healthcare.com/article/20141118/content-463705.html>

²⁶ 데이터 소스 출처: Coudray N, Moreira A L, Sakellaropoulos T, et al. Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning[J].bioRxiv, 2017.

■ 정렬 CNN

의료 이미지 검사 결과를 보면 정렬 상황이 뚜렷하게 나타나는 경우가 종종 있습니다. 예를 들어, 음성이면 정상이고 양성이면 정상이 아닌 것입니다. 그런데 검사가 기대하는 결과가 0이나 1같은 불연속형 숫자이어서 전형적인 정렬 문제를 일으킨다는 것을 알 수 있습니다. 따라서 이진 분류와 유사한 정렬 알고리즘을 이용하면 CNN은 의료 기관이 문제가 있는 부분이나 조직을 초보적이지만 질적으로 선별한 뒤 정량 분석과 판독을 진행할 수 있도록 도와줍니다.

로지스틱 회귀같은 전형적인 이진 분류 알고리즘은 일종의 일반적인 선형 회귀 분석 모델입니다. 조직 사진에 따라 환자의 암 발병 여부를 검사하는 경우, 환자의 연령이 높아짐에 따라 발견된 어느 세포의 수가 x 개를 초과하면 암으로 판단할 수 있습니다. 이는 임계값이 x 인 선형 함수로 표현된 것이며, $y = \text{연령}(n)*a + \text{초기값}(b)$ 이라고 할 때, $y >= x$ 면 암으로 판단합니다.

하지만 실제 상황에서 이 함수는 훨씬 더 복잡해집니다. 연령 외에도 이상 세포의 크기, 상태 등 요소가 판단의 근거로

작용하며, 이때 선형 함수는 하나의 다변량 선형 함수로 바뀝니다.

$$y = n*a + m*c + o*d + \dots + b$$

위에서 서술한 바와 같이, 정렬 문제는 일련의 불연속형 결과를 입력해야 하므로 선형 함수에 활성화 함수를 추가해 출력 결과가 불연속성을 띄도록 해야 합니다. 뉴럴 네트워크의 경우에는 활성화 함수가 뉴럴 네트워크에 비선형 요소를 추가해 뉴럴 네트워크가 비교적 복잡한 문제를 잘 해결할 수 있도록 도와줍니다. 자주 볼 수 있는 활성화 함수는 Sigmoid 함수, tanh 함수, ReLU 함수 등이 있으며, 로지스틱 회귀는 경사하강법 반복 해결 방법을 채택해 최소화된 손실함수를 획득합니다.

일반적으로 이진 분류 알고리즘 기반 CNN 이미지 정렬은 그림 3-1-1과 같은 주요 모듈을 보유하고 있으며, 이미지 읽기와 전처리, 이미지 트레이닝, 반복 최적화와 이미지 예측을 포함합니다. 그 중, CNN 기반 모델 트레이닝은 콘볼루션층과 풀링 계층, FC 등으로 구성되어 있어 크로스 엔트로피 손실함수나 MBGD 경사하강법 알고리즘 또는 BGD 경사하강법 알고리즘을 채택할 수 있습니다.

실제 응용에서는 레지듀얼 네트워크(Residual Net, ResNet) 역시 자주 볼 수 있는 정렬 CNN 중 하나로 2D 이미지 정렬과 검사, 위치추정에 매우 뛰어납니다. 다른 CNN과 비교했을 때, ResNet은 네트워크에 직접 연결 채널을 추가했기 때문에 그림 3-1-2처럼 입력 정보가 후면 층에 직접 전달되는 것을 허용합니다.

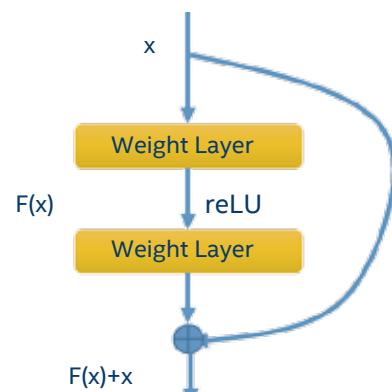


그림 3-1-2 ResNet의 잔차 구조

이 구조(잔차 구조)는 전형적인 CNN 네트워크 구조가 정보를 전달할 때 존재하는 정보 유실과 훼손, 소실 등의 문제를 어느 정도 해결했습니다. 이 문제는 Deep 모델의 층 수를 너무 많이 구성할 수 없는 원인 중 하나이기도 합니다. 하지만 ResNet을 채택하면, 트레이닝 모델의 층 수를 대폭 증가할 수 있고, 덕분에 정렬 정확도도 향상됩니다.

■ 표적 검출 뉴럴 네트워크

표적 검출 뉴럴 네트워크는 지정한 사진에서 물체의 위치를 정확하게 찾아내고, 물체의 유형까지 태그합니다. 자주 볼 수 있는 표적 검출 뉴럴 네트워크로는 R-CNN, Fast R-CNN, SPP-NET, R-FCN 등이 있습니다.

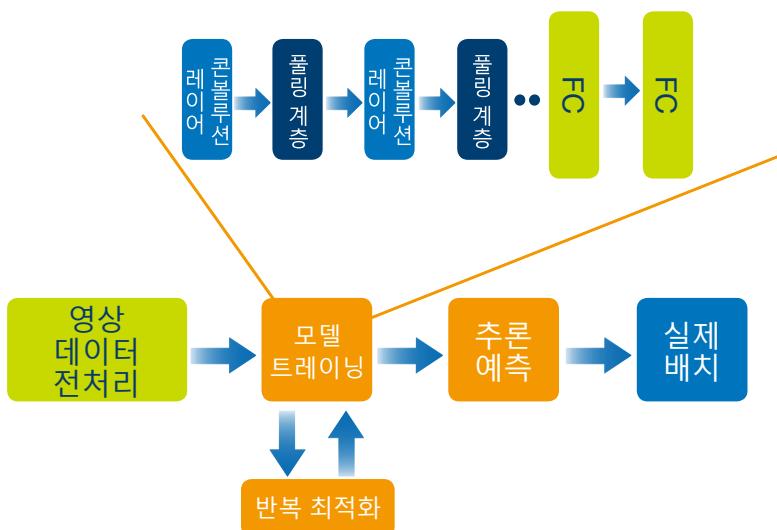


그림 3-1-1 이진 분류 알고리즘에 기초한 CNN 이미지 분류 구성 모듈

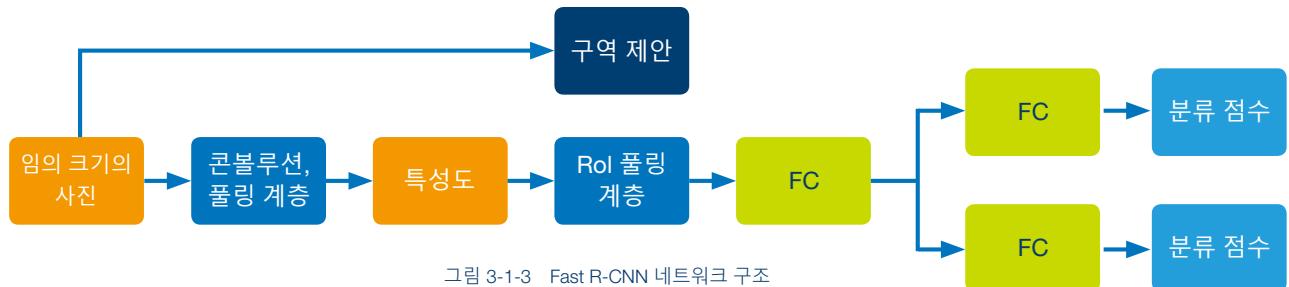


그림 3-1-3 Fast R-CNN 네트워크 구조

R-CNN은 전형적인 딥러닝 객체 탐지 알고리즘으로 다음 작업 프로세스를 기초로 합니다. 우선, R-CNN은 selective search 방법을 기초로 해 초기 그림에 크기가 일정한 후보 지역 수천 개를 생성하고, CNN 네트워크를 입력합니다. 이 네트워크 모델로 얻은 특성 벡터를 여러 유형의 서포트 벡터 머신(Support Vector Machines, SVM) 분류기를 통해 타깃마다 1개의 SVM 분류기를 트레이닝하고 특성 벡터에서 이 타깃에 속하는 확률을 추단합니다. R-CNN은 Arc의 회귀 모델을 설정해 포지션의 정확도를 올리고, Arc 회귀 모델로 Arc의 정확한 위치를 최적화합니다.

R-CNN을 실제 응용할 때 트레이닝, 추론, 테스트 속도가 느리고, 트레이닝에 필요한 공간이 큰 문제점을 해결하기 위해 Fast R-CNN은 다음 방법을 채택했고 그 덕분에 R-CNN보다 높은 효과를 얻을 수 있었습니다. 방법은 다음과 같습니다.

- 우선 전체 이미지를 정규화한 뒤 CNN 네트워크로 보냅니다.
- 콘볼루션 층에서는 후보 지역 특성 추출을 진행하지 않고, 마지막 풀링 계층에 후보 지역 좌표 정보를 추가해 특성 추출 컴퓨팅을 진행합니다.
- CNN 네트워크에서는 타깃과 후보를 함께 바운딩 박스 회귀합니다.

또한 Faster R-CNN은 특성 추출(feature extraction)과 proposal 추출, bounding box regression(rect refine), classification을 모두 하나의 네트워크에 통합하기 때문에 포괄성이 비교적 크게 향상되고 검사 속도 측면에서 특히 두드러집니다.

소프트웨어/하드웨어 구성 제안

AI 기반 조직 분석 방안 구축은 아래의 인텔®아키텍처 플랫폼 기반 소프트웨어/하드웨어 구성을 참조해 완료할 수 있습니다.

파라미터	설명
CPU	인텔® 제온®골드 6240 CPU 또는 그 이상
하이퍼스레딩	ON
터보 가속	ON
메모리	16GB DDR4 2666MHz* 12 및 이상
저장	인텔®SSD D5 P4320 시리즈 및 이상
운영 체제	CentOS Linux 7.6 또는 최신 버전
Linux 코어	3.10.0 또는 최신 버전
컴파일러	GCC 4.8.5 또는 최신 버전
Caffe 버전	인텔®아키텍처 최적화 지향 Caffe 1.1.6 또는 최신 버전

딥러닝 기반 조직 분석 방법의 최적화

인텔® 아키텍처 CPU 기반 최적화 방법

인텔®CPU 플랫폼에서 딥러닝 기반 조직 분석 방법을 만들고 최적화한 사용자는 다음 몇 가지 측면에서 이익을 취할 수 있습니다.

- 조직 이미지 파일의 용량은 흔히 수십 수백 MB에 달합니다. 전통적으로 저장 공간의 한계 때문에 트레이닝 도중 설정한 Batch Size는 모두 작은 편이고, 그로 인해 트레이닝 시간도 증가합니다. 하지만 인텔®아키텍처 CPU 기반 플랫폼을 사용하면 서버가 대형 메모리(보편적으로 수 TB부터 수십 TB까지 보유)를 보유하게 되어 Batch_Size를 100 이상 설정 가능해 트레이닝 속도를 빠르게 할 수 있습니다.
- 3D XPoint™ 스토리지를 기반으로 만든 인텔®옵테인™ 데이터 센터급 영구 메모리를 도입하면, 제온 확장성 플랫폼의 장점이 더 강화됩니다. 가격이 비싼 동적 램(Dynamic Random-Access Memory, DRAM)과 비교했을 때, 인텔®옵테인™ 데이터 센터급 영구 메모리는 대용량이고 비휘발성이며, 적은 비용으로 용량을 확장할 수 있습니다. 이러한 장점 덕분에 모델 트레이닝과 추론한 서버의 메모리 밀도 및 컴퓨팅 효율을 효과적으로 올릴 수 있을 뿐만 아니라 TCO를 대폭 절감할 수 있습니다.

- 인텔® 제온® 확장성 CPU의 혁신적인 마이크로아키텍처는 더 많은 수의 코어와 동시성이 더 높은 스레드, 더 왕성한 고속 캐시가 포함되어 있고, 그것이 집약한 다량의 하드웨어 강화 기술이 어울려져 있습니다. 특히 인텔® AVX- 512 등은 모두 AI 응용을 위해 더 강력한 해시 레이트를 제공합니다.

* 인텔® 옵테인™ 데이터 센터급 영구 메모리 기술과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

인텔® 아키텍처 최적화 지향 Caffe

Caffe는 일종의 상용 딥러닝 아키텍처로 동영상과 이미지 처리 등 영역의 AI 트레이닝과 추론 작업에서 폭넓게 사용되고 있습니다. Caffe에 기초한 딥러닝 모델의 업무 효율을 높이고 또 최적화하기 위해, 인텔은 인텔® 아키텍처 특성에 기반해 Caffe를 다양으로 최적화했습니다.

최적화 작업은 다음을 포함합니다.

■ 전형적인 ResNet 네트워크를 겨냥해 진행한 최적화

인텔® 아키텍처 최적화 지향 Caffe는 ResNet 시리즈 모델의 특성을 이용해 컴퓨팅과 메모리 접근으로 인해 발생하는 비용을 절감했습니다. 그림 3-2-1은 전형적인 ResNet의 잔차 구조입니다. 그림의 왼쪽 절반 부분에서 볼 수 있듯이 아랫 부분의 1*1 Stride-2 콘볼루션 층 2개는 활성화 작업을 절반만 소모했습니다. 최적화 방안은 바인딩 층 설정을 그림의 오른쪽 절반 부분처럼 1*1의 풀링 계층 1개를 직접 연결 채널에 추가해 컴퓨팅 양을 절반으로 절감했습니다.

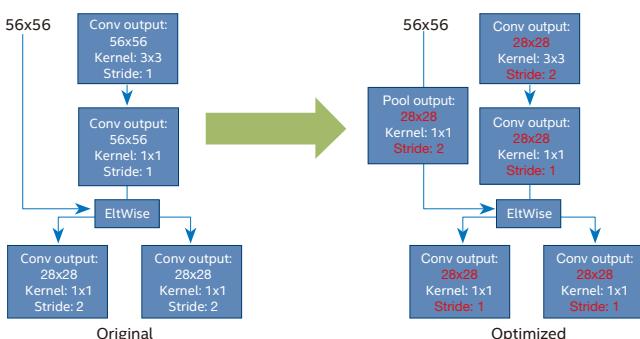


그림 3-2-1 인텔® 아키텍처 최적화 지향 Caffe가 ResNet 네트워크에 진행한 최적화 방안

■ 레이어 융합 기술

인텔® 아키텍처 최적화 지향 Caffe는 명령어 조합 벡터화, 스레드급 병렬 최적화 외에도, Caffe 아키텍처에 BN+Scale과 Conv+Sum, Conv+Relu, BN inplace, sparse fusion 같은 더 효과적인 레이어 융합(Layer Fusion) 최적화 수단을 도입했습니다. 이런 수단은 ResNet50 같은 뉴럴 네트워크의 성능을 극대화해줍니다. 그림 3-2-2에서 보는 바와 같이, 이것은 잔차 구조의 Conv 층을 Eltwise 층과 융합한 것입니다. 그림의 왼쪽 절반 부분의 콘볼루션 층(res2a_branch2c와 Eltwise 층 res2a_relu가 하나의 새로운 콘볼루션 층(res2a_branch2c(그림의 오른쪽 절반 부분))로 융합되면서 ResNet 유형 네트워크 모델의 성능을 효과적으로 향상시킵니다.

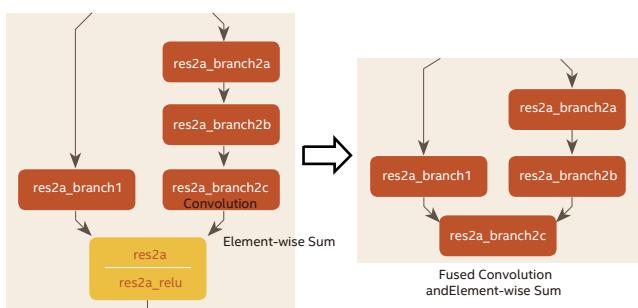


그림 3-2-2 Conv 층과 Eltwise의 레이어 융합

인텔® 아키텍처 최적화 지향 Caffe는 INT8에 대한 지원도 훌륭하고, calibration 툴도 제공하기 때문에 사용자가 뉴럴 네트워크를 INT8로 자연스럽게 전환하도록 도와주기 때문에 성능 역시 대폭 향상됩니다.

일부 테스트를 통해 알 수 있듯이 BVLC Caffe를 사용한 경우와 비교했을 때, 인텔® 아키텍처 최적화 지향 Caffe는 인텔® 제온® 확장성 CPU에서 운용할 때, 레이어 융합 기술을 추가하고 ResNet50 콘볼루션 뉴럴 네트워크를 사용해 동등한 평가 환경에서 AI 추론을 실행하면 그림 3-2-3과 같이 단위 시간의 추론 성능을 전자의 51배 이상 향상시킬 수 있으며, 추론 시간은 전자의 4.7%²⁷까지 단축됩니다.

²⁷ 이 데이터의 출처는 《Highly Efficient 8-bit Low Precision Inference of Convolutional Neural Networks with IntelCaffe》에서 추출한 것입니다. <https://arxiv.org/pdf/1805.08691.pdf>, 테스트 구성은 다음과 같습니다. 콘볼루션 모델: ResNet50, 하드웨어: AWS single-socket c5.18xlarge

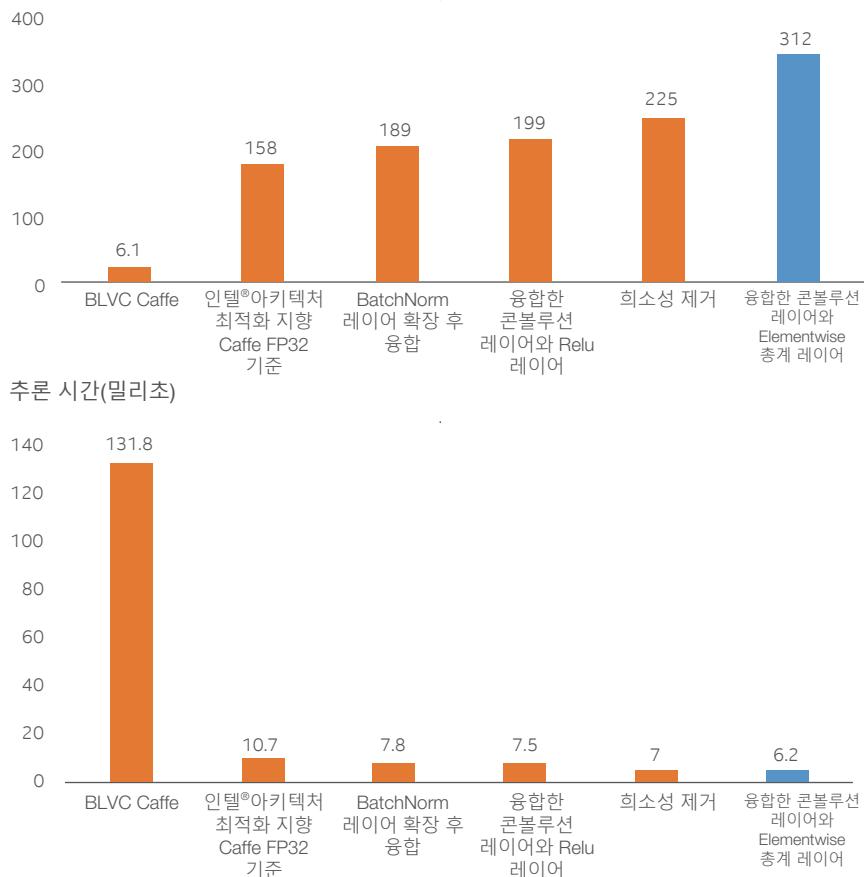


그림 3-2-3 인텔® 아키텍처 최적화 지향 Caffe를 인텔® 제온® 확장성 CPU에서 최적화 방안을 추가한 이후 추론 속도와 추론 시간 및 성능 BLVC Caffe와 비교

인텔® 딥러닝 가속 기술

최신 2세대 인텔® 제온® 확장성 CPU에 INT8를 최적화 지원한 인텔® 딥러닝 가속 기술을 추가하면, 예측 정확도에 영향을 미치지 않으면서 여러 딥러닝 모델이 INT8를 사용할 때의 추론 속도를 가속화할 수 있습니다. 게다가 사용자의 딥러닝 응용 업무 효율도 향상됩니다.

이미지 정렬, 객체 탐지 등의 딥러닝 시나리오에서 INT8 등 정밀도가 떨어지는 수치를 FP32로 대체하는 것은 성능을 최적화하는 좋은 방법 중 하나입니다. 낮은 정밀도 수치는 고속 캐시를 더 유용하게 사용해 메모리 데이터 전송 효율을 높이고, 대역폭 병목 현상을 줄이며, 컴퓨팅과 스토리지 리소스를 충분히 이용함과

동시에 시스템 동력도 효과적으로 낮출 수 있습니다. 또한 동일한 리소스가 지원되어도 INT8은 딥러닝 추론에 더 많은 초당 작업 수(Operations Per Second, OPS)를 제공할 수 있습니다.

인텔® 딥러닝 가속 기술은 VNNI 명령어 조합을 통해 8자리 또는 16자리 저정밀도 수치 곱셈에 사용되는 여러 개의 최신 FMA 커널 명령을 제공하며, 이는 다양한 행렬 곱셈을 실행해야 하는 딥러닝 컴퓨팅에게 있어 매우 중요한 부분입니다. 이 기술을 도입하면, 사용자가 INT8 추론을 진행할 때, 시스템 메모리 수요를 최대 75%²⁸ 감소시킬 수 있습니다. 또한 메모리와 필요한 대역폭의 감소는 저수치 정밀도 연산 속도를 가속화해 시스템 전체 성능을 대폭 향상시켰습니다.

* 인텔® 제온® 확장성 CPU와 인텔® 딥러닝 가속 기술과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

툴을 이용한 모델 정확도 최적화 방법

■ 유사도 측정 툴

딥러닝에서는 유사도 측정(Similarity) 툴을 사용해 두 개의 특성 값 간의 유사도를 판단할 수 있습니다. 툴이 다르면 서로 다른 차원에서 유사도 측정을 진행할 수 있으며, 비교적 흔한 유형은 다음과 같습니다.

- 유클리디안 거리(Euclidean Distance): 가장 흔한 거리 측정 유형으로 좌표계의 2점 사이의 절대 거리를 컴퓨팅합니다. 거리가 멀수록 유사도는 낮아집니다.
- 벡터 공간 코사인 유사도(Cosine Similarity): 벡터 공간 중 벡터 끼인 각 두 개의 코사인 값을 사용해 객체 간의 차이를 가늠합니다. 거리 측정과 비교하면 코사인 유사도는 두 가지 벡터의 방향성의 차이에 더 중점을 두었으며, 끼인 각이 작을수록 유사도가 높아집니다.
- 표준화 유클리디안 거리(Standardized Euclidean distance): 유클리디안 거리의 업그레이드형으로 특성 간의 거리를 컴퓨팅하기 전 우선 각 무게를 표준화 컴퓨팅해야 합니다.
- 마할라노비스 거리(Mahalanobis distance): 점과 하나의 분포 지점 사이의 거리를 표시하는데 사용됩니다. 간단히 말해, 단일 샘플과 어느 샘플 사이의 거리가 가장 가까우면 이 샘플 세트에 속합니다.

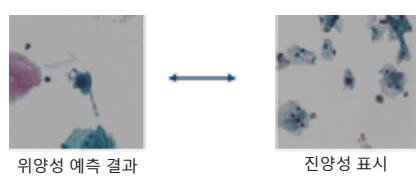


그림 3-2-4 유사도 측정 툴을 이용한 분석 예측 실태의 원인

²⁸ 데이터 소스 출처: <https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training>

유사도 측정 툴을 이용하면 모델 트레이닝 정확도를 높이는 일련의 방법을 원활하게 설계하고 조합할 수 있습니다. 예를 들어 두 개의 특성 간의 유클리디안 거리를 컴퓨팅해 예측 실패 원인을 분석할 수 있는 것입니다. 그림 3-2-4와 같이, 위양성 샘플이 특성 추출 층과 어느 양성 태그 사이에서 가장 가까운지 측정하면 판독 오류의 주요 원인을 도출해낼 수 있습니다.

■ LRP 툴

전통적으로 딥러닝 모델의 층 간 정보 전송과 로직은 줄곧 블랙 박스처럼 백트래킹하기 어려웠습니다. 하지만 LRP(Layer-wise Relevance Propagation) 툴을 이용하면 사용자가 이 곤혹스러운 상황을 해결할 수 있도록 일정 정도에서 도움을 줍니다. LRP 툴은 컴퓨팅 상관성을 이용하며, 상관성을 한층 뒤로 전파하기 때문에 백트래킹성이 좋은 편입니다. 게다가 이 메커니즘을 이용하면 어느 요소가 예측 결과에 미치는 작용이 큰지도 시스템이 도출해낼 수 있기 때문에 모델의 정확도가 향상됩니다.

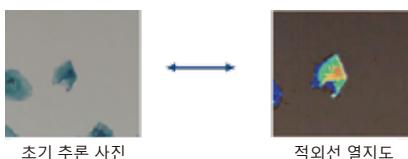


그림 3-2-5 LRP를 이용해 다른 화소가 추론 효과에 미치는 영향 검사

그림 3-2-5에서 보는 바와 같이, 의료 이미지 분석 예측에 AI를 응용할 때 LRP 툴을 이용할 경우, 화소에 따라 추론 결과에 미치는 효과를 볼 수 있을 뿐만 아니라 열지도를 형성하기 때문에 마지막 예측 결과에 어느 화소가 더 큰 영향을 미치는지 도출하도록 돋습니다.

AI 기술을 이용해 자궁경부암 검사 효율을 향상시킨 장평생물정보기술

배경

자궁경부암은 현재 여성의 건강을 심각하게 위협하는 악성 종양 중 하나입니다. 통계에 따르면, 2018년도에만 570,000명에 달하는 여성 암환자 중 자궁경부암이 6.6%로 이미 여성암 중 네 번째로 치명적인 질환²⁹이 되었습니다. 그런데 자궁경부암은 발병 원인을 확인할 수 있어 조기에 발견해 예방할 수 있는 유일한 암 질환입니다. LBP(Liquid-Based Cytologic Preparation) 검사는 간단하고 조작이 쉬우며 정확도가 높기 때문에 조기에 암병변을 발견해 조기 진단과 적시치료를 도와주고 암세포 확산을 막아줍니다.

현재 중국은 매년 수천만 개의 자궁 LBP 도보 표본이 새롭게 생성되고, 이는 의료 기관의 병리 분석 능력에는 큰 도전이 되었습니다. 장평생물정보기술은 인텔과 함께 첨단 AI 기술을 이용해 자궁 LBP 조직 기반 자궁경부암 검사 AI 솔루션을 구축하고 최적화하면서 자궁경부암 예방과 치료가 효과적으로 이루어지도록 최선을 다하고 있습니다.

하지만 현재 몇 가지 요소가 방안의 검사 효율과 정확도를 제약해 한 단계 더 발전하기에 어려움이 있습니다. 우선은 데이터 태그 문제가 있습니다. 다른 의료 데이터와 비교했을 때, 조직 분석 데이터는 독특한 부분이 있습니다. 그림 3-3-1처럼 조직 사진은 줌 스케일이 1부터 40배까지 되며, 줌 스케일이 작으면 사진은 기본적으로 태그를 진행할 수 없습니다. 하지만 사진을 20배 심지어 40배까지 확대했을 때, 전체 사진 중 아주 작은 부분을 사람이 태그하기만 해도 이 작은 부분의 모든 문제 세포를 커버할 수 없게 됩니다.

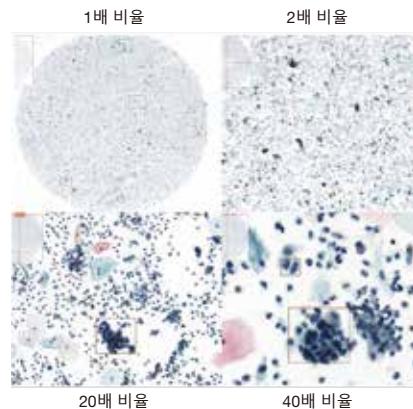


그림 3-3-1 크기별 조직

태그 과정에 태그 불완전 현상이 존재한다는 것도 문제입니다. 태그 작업자는 경우에 따라서 눈에 띠는 가장 심각한 문제 세포만 태그합니다. 그림 3-3-2 왼쪽과 같이, 오른쪽 하단 파란색 틀 안의 악성 종양이 태그되어 나왔지만, 왼쪽 상단의 빨간색 틀의 약한 양성 세포는 태그하지 않았습니다. 그림 3-3-2의 오른쪽에는 태그 위치가 정확하지 않은 상황이 나타났습니다.

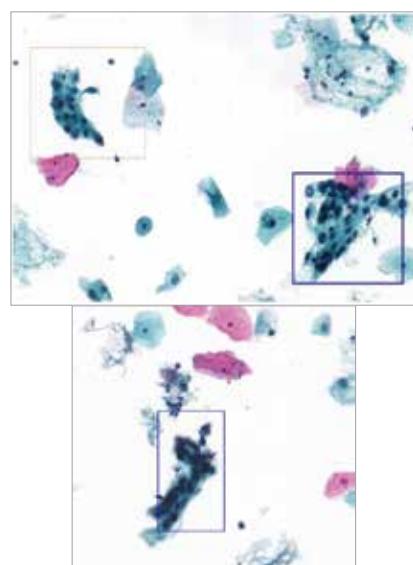


그림 3-3-2 태그가 완전하지 않은 조직 사진

²⁹ 출처: WTO 공식 홈페이지 <http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/>

현재 태그 방안은 통상적으로 양성 세포에만 초점을 맞추고, 음성 세포는 중시하지 않는 점도 문제입니다. 설령 음성 세포에 태그를 진행해도 슬라이스 레벨만 커버할 뿐, 총량의 대다수를 차지하는 음성 세포에는 효과적인 이용 방안이 존재하지 않습니다. 또한 기존의 태그 견본 불균형이 너무 심각하고, 비전형적인 비늘 모양 상피 세포(ASC-US)가 절대 다수를 차지하며, 비늘 모양 세포암(SCC)과 자궁 내막, 트리코모나드 등 견본이 적은 편이어서 학습 효율을 높이기엔 불리합니다.

또 한 가지 관심을 두어야 할 문제는 뉴럴 네트워크 선택입니다. 실제 적용 효과에서 볼 수 있듯이, 현재 일상적으로 사용하는 세포 병변 표적 검출 네트워크는 병변 세포 소재 위치 직사각형 좌표 및 병변 세포의 구체적인 묘사성(The Bethesda system, TBS) 분급을 출력할 수 있습니다. 하지만 단독 표적 검출 네트워크는 태그 완전성 문제를 잘 해결할 수 없습니다. 이런 문제를 해결하기 위해, 장 평생물 정보 기술은 인텔과 함께 다음 몇 가지 차원에서 최적화를 전개해 검사 딥러닝 모델의 업무 효율을 향상시켰습니다.

- 데이터 정리와 전처리 프로세스 최적화
- 2단계 end-to-end 뉴럴 네트워크 구축
- 모델 정확도 최적화 툴 도입.

방법과 성과

장평생물정보기술이 인텔과 연합해 만든 자궁 LBP 조직 기반 자궁경부암 검사 AI 솔루션의 주요 업무 프로세스는 그림 3-3-3과 같습니다. 시스템은 사진 입력 후, 데이터 전처리, 정렬 CNN과 후처리 단계를 거쳐 양성 예측과 음성 예측을 분리 획득합니다. 양성 예측에 대해 방안은 두 번째 단계의 표적 검출 네트워크(Resnet50 기반) 모델 트레이닝을 진행한 뒤, 양성 식별 추론 과정을 진행해 전문의에게 보내 최종 심사합니다.

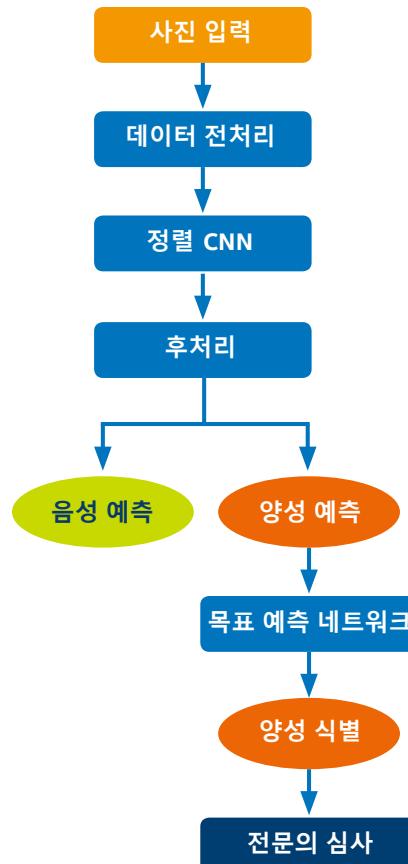


그림 3-3-3 최적화 후의 방안 프로세스

데이터 정리와 전처리 프로세스 최적화 과정 중, 조직 이미지의 줌 스케일이 다른 문제를 해결하기 위해 줌 스케일이 큰 편이고 양성이 세포/세포 블록 레벨로 태그한 조직 이미지를 큰 조직 이미지에서 작은 그림을 재단하는 방식으로 트레이닝 데이터를 획득합니다. 조직 중 견본이 불균형한 문제에 대해 트레이닝 세트는 양성: 음성=1: 5의 비율을 채택했습니다. 양성 태그 견본이 적은 편이면, 방안 역시 견본을 회전해 견본의 다양성을 확대합니다.

또한 음성 세포 견본 이용 효율을 향상시키기 위해 방안의 음성 조직 중 모든 세포가 음성 세포일 경우, 음성 조직의 트레이닝 세트는 모든 음성 조직을 비율대로 랜덤 재단(목적은 조직 엣지 간섭을 제거하기 위함)합니다. 양성 조직의

트레이닝 세트에 대해서는 양성 조직에 태그한 좌표 중심점에 따라 512*512로 랜덤 오프셋 재단한 서브그래프를 추가합니다.

식별 정확도와 효율을 높이기 위해, 방안을 혁신적으로 2단계 end-to-end 뉴럴 네트워크로 구축했습니다. 그 중, 단계 1은 정렬 CNN이고, 단계 2는 표적 검출 뉴럴 네트워크입니다. 그림 3-3-4처럼 정렬 CNN의 주요 역할은 조직마다 생성된 슬라이딩 윈도우에서 이진 분류 추론을 진행하고, 이 조직의 모든 슬라이딩 윈도우 결과를 융합 처리한 뒤 조직급 추론 결과를 얻는 것입니다. 하지만 표적 검출 네트워크는 이전 단계에서 양성으로 확정한 조직에 대해 양성 지역 검출을 진행합니다.

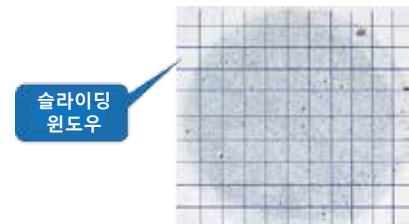


그림 3-3-4 슬라이딩 윈도우 기반 정렬 CNN

모델 트레이닝 과정 중 방안은 다음의 최적화 방안을 채택해 트레이닝 효과를 향상시켰습니다.

- 모델은 Imagenet 데이터 세트에서 우수한 성능을 지닌 ResNet50으로 트레이닝을 진행했습니다.
- 트레이닝 세트를 잘 준비하면 회전한 뒤 중심점에 따라 224*224로 재단해 평균값(Normalize)과 정규화(Scale) 처리를 진행하고, 이어서 모델 트레이닝을 시작합니다.
- 트레이닝 세트 중 플러스 마이너스 견본 수량의 차가 크다는 점을 감안해, 방안은 트레이닝을 완료한 일부 음성 조직과 일부 양성 조직의 서브그래프를 모아 점차 늘려가며 트레이닝 세트에 추가해 가며 반복 트레이닝을 형성합니다. 트레이닝 세트 양성: 음성 비율은 1: 5이며, 이를 통해 모델의 정확도를 한층 향상합니다.

- 방안에는 유사도 측정(Similarity) 툴과 LRP 툴도 추가해 모델의 정확도를 향상시킵니다.

장평생물정보기술은 인텔과 함께 최적화된 자궁 LBP 조직 기반 자궁경부암 검사 AI 솔루션을 평가했고, 5,961장의 정밀 태그 기반 견본을 트레이닝했으며, 246장의 테스트세트에서 모델별로 평가했습니다.

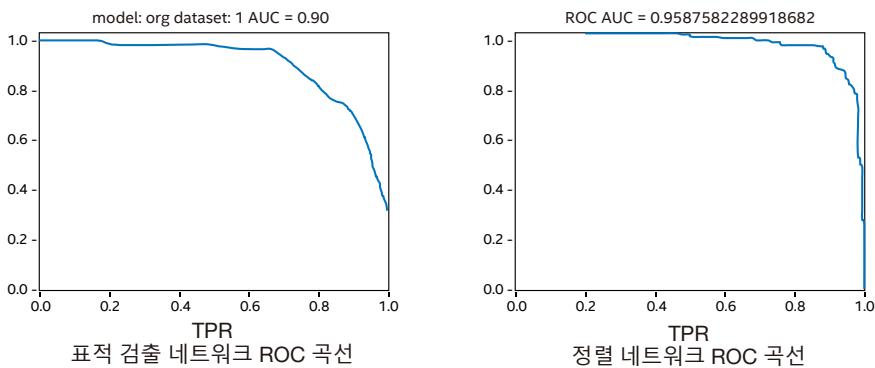


그림 3-3-5 최적화 방안과 전통적인 방안의 정확도 비교

평가 결과에서 볼 수 있듯이, 정렬 네트워크를 추가한 최적화 방안의 정확도는 단독 표적 검출 네트워크 방안보다 대폭 향상되었습니다. 그림 3-3-5에서 볼 수 있듯이, 정렬 네트워크를 추가한 뒤, 민감도(TPR)가 96%일 때 특이도(TNR)가 70% 가까이 됩니다. 단독 표적 검출 네트워크 방안에서는 특이도가 40% 정도³⁰밖에 되지 않는데, 이는 정확도가 대폭 향상³¹ 향상되었음을 의미합니다.

³⁰ 이 데이터의 출처는 장평생물정보기술과 인텔이 배포한 『딥러닝 기반 병리 이미지 분석』입니다

³¹ 데이터에 사용한 테스트 구성: 투웨이 인텔® 제온® 플래티넘 8280 CPU, 2.70GHz 코어/스레드: 28/56, HT: ON, Turbo: ON, 메모리: 192GB DDR4 2933, 하드 디스크: 인텔® SSD SC2KG48, 네트워크 어댑터: 인텔® 이더넷 네트워크 어댑터 X722 for 10GBASE-T, BIOS: SE5C620.86B.02.01.0003.020220190234, 운영 체제: CentOS Linux 7.6, Linux 커널: 3.10.0-957.el7.x86_64, 컴파일러 버전: ICC 18.0.1 20171018, Caffe 버전: 인텔® 아키텍처 최적화 지향 Caffe 1.1.0 워크로드: Resnet50 with 2 classes, 130장의 이미지/초.

요약

딥러닝의 방법을 이용해 조직 이미지 등을 빠르게 검사하면 의료 기관 조직 검사 생산력을 대대적으로 향상할 수 있을 뿐만 아니라, 병리학과 전문의가 부족해서 발생하는 문제를 해결하고 환자에게 더 정밀하고 시기적절하게 치료 방안을 제시할 수 있습니다. 이제 이미지 정렬과 객체 탐지에 기초한 조직 검사 AI 응용은 여러 의료 기관에서 정착되었고, 또 긍정적인 피드백을 얻었습니다.

인텔®아키텍처 CPU 플랫폼, 인텔®아키텍처 최적화 지향 Caffe, 인텔®딥러닝 가속 기술 등이 포함된 인텔의 최첨단 제품과 기술은 이미 여러 응용 시나리오에서 딥러닝 기반 조직 검사 응용 효율이 대폭 향상하도록 도와주고 있습니다. 예를 들어 인텔®아키텍처 CPU 플랫폼은 라지 메모리를 지원해 모델 트레이닝 중 더 큰 Batch_Size를 설정해 트레이닝 효율을 대폭 향상할 수 있게 해줍니다. 인텔®아키텍처 최적화 지향 Caffe와 인텔®딥러닝 가속 기술은 INT8을 지원하기 때문에 추론 효율을 효과적으로 향상시키고, 조직 분석의 실시간 분석 기술을 향상시킬 수 있습니다.

이번 사례와 관련된 CPU 플랫폼은 1세대 인텔® 제온® 확장성 CPU지만, 최신 2세대 인텔® 제온® 확장성 CPU와 기타 인텔 신제품과 신기술이 도래함에 따라 사용자는 이 업데이트된 소프트웨어/하드웨어에 기초한 트레이닝과 추론 성능이 더 강력한 AI 애플리케이션을 만들 수 있습니다. 또한 인텔은 더 많은 질환의 치료 시간과 효율을 향상시킬 수 있도록 다양한 딥러닝 모델에 대해 추론 최적화 연구를 계획 중입니다.

약물 연구 개발을 돕는 AI 기술

약물 선별을 가속화한 딥러닝 방식

HCS 기반 표현형 분류

점차 더 많은 신기술이 약물 연구 개발 과정에 운용되고 있습니다. 세포 이미지에 기초한 HCS(High Content Screening) 방식은 현재 시스템 생물학과 약물 연구 개발 영역에서 자주 사용하는 자동화 분석 방법 중 하나이자, AI 기술로 약물을 조기에 발견하는 사이클에서 중요한 역할을 하기도 합니다. 마이크로이미징으로 획득한 이미지 정보를 통해 유전 또는 화학 처리로 유도한 세포 표현형 특성을 분석하고 획득합니다.

이 프로세스에서 세포 이미지에 대한 표현형 검사와 분석, 정렬은 가장 중요한 사이클입니다. 하지만 생물학 분석 과정의 고유한 복잡성과 세포 측정의 고유한 가변성은 세포 이미지의 표현형을 분석하는데 어려움을 주었습니다. 전통적인 세포 표현형 특성 추출의 이미지 분석 방법은 주로 일련의 독립적인 데이터 분석 절차로 구성됩니다. 그림 4-1-1처럼, 초기 이미지를 입력하면 우선 객체 탐지(Object Detection) 방법을 입력하고, 세포 층 레벨 또는 이미지 층 레벨에서 특성을 추출(Feature Extraction)합니다. 그런 다음 이 특성에 전환(선택, 표준화 등)을 진행하고 마지막으로 관련 특성을 귀납해 예측 표현형의 정렬 알고리즘 입력으로 삽습니다.



그림 4-1-1 전통적인 HCS 방법

위의 특성 검사와 분석, 정렬 방법이 이미 다량의 약물 연구 개발 과정에서 성공적으로 응용되고는 있지만, 적지 않은 한계성이 존재합니다. 객체 분할과, 차원 하락, 표현형 분류는 일반적으로 사전 지식이 필요하고, (예: 예상한 몇 가지 표현형 형태, The geometric properties of the expected phenotypes)는 모든 측정 프로세스를 커스텀해야 합니다. 또한 전통적인 HCS 방법 채택과 절차 실행은 방법의 커스텀 및 파라미터 조정과 연관됩니다. 전체 분석 프로세스의 성능 조정 과정 중, 모든 파라미터를 어떻게 연합 최적화해 성능 최적화를 이루느냐는 현재 우리가 직면한 도전입니다. 그러다 보니 전체 효율 향상은 좀 더 시간이 필요합니다. 이런 문제를 해결하기 위해 더 다양한 딥러닝 기반 AI 방법이 점차 세포 이미지에 기초한 HCS 표현형 분류 작업에 도입되고 있습니다.

딥러닝 기반 HCS 방법³²

배경

전통적인 HCS 이미지 분석 방법은 이미지 데이터를 화소 밝기(Pixel Intensity) 같은 추상적 단계로 전환합니다. DNN 등의 딥러닝 방법에는 하나의 아키텍처를 통해 이미지 데이터의 계층에 대해 컴퓨팅과 분석을 진행하지만, 이 방법은 여러 측면에서 수도 정의에 의존하는 특성이 있습니다. 이 방법과 달리 CNN은 자동적으로 이미지에서 특성을 학습하고 추출하기 때문에 세포 이미지의 표현형 예측 효율이 더 좋습니다.

³² 이 섹션 중 CNN와 M-CNN에 기초한 HCS 기술에 대한 세부 정보는 다음을 참조하십시오. Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

CNN 네트워크는 일반적으로 입력 층과 콘볼루션 층, ReLU 층, 풀링 계층, FC 등을 포함하고 있습니다. 그 중 콘볼루션 층은 컴퓨팅 층 입력(예를: 초기 이미지 또는 이전 콘볼루션 층의 출력)과 여러 개의 2차원 콘볼루션 커널 사이의 콘볼루션을 통해 이미지의 2차원 기하학적 정보를 획득합니다. 모든 콘볼루션 커널은 기하학적 무늬(Geometric Pattern) 1개를 코딩하고, 콘볼루션 커널 매핑(또는 휘쳐 대응) 1개를 콘볼루션해 획득할 수 있습니다. 이 매핑은 픽셀 기반 비선형 활성화 함수이며, 후방 콘볼루션 층에 전달되어 더 복잡한 모드를 획득합니다. 결국 콘볼루션 층 출력은 FC에 발송되고, 이전 피드백 방식으로 지정 입력에 예측을 생성합니다.

CNN의 출력 층에 분류 대기 중인 표현형이 N_p 개가 있을 경우, 네트워크는 지정한 입력 이미지 x 에 대해 출력 층에서 j 유닛의 활성화 함수 $a_j(x)$ 를 컴퓨팅하고, 이것을 기초로 벡터 p , p_k 1개를 컴퓨팅하면 확률 품질 함수 1개를 구성해 N_p 개의 분류 대기 표현형을 커버할 수 있습니다.

$$\hat{y} = \operatorname{argmax}_k p(y = k|x)$$

여기에서 k 는 표현형의 번호이며, 이 확률에 따라 표현형의 예측 값을 얻을 수 있습니다.

$$p_k := p(y = k|x) = \frac{\exp(a_k(x))}{\sum_j^{N_p} \exp(a_j(x))}$$

이것만 보아도 층 수와 콘볼루션 층 내 유닛 수량, 콘볼루션 커널과 풀링 인자 크기가 예측 성능에 영향을 미친다는 것을 알 수 있습니다. 세포 표현형 분류에는 문제가 하나 더 있는데, 세포 자체의 크기가 다르다 보니 이미지 데이터에 생기는 공간의 차이가 크다는 것입니다. 그런데 이때 전형적인 CNN 네트워크 구조를 여전히 사용한다면 정확도가 하락할 것입니다.

멀티 스케일 CNN(Multi-scale Convolutional Neural Networks, M-CNN)은 이 문제를 손쉽게 해결할 수 있습니다. 전형적인 CNN 네트워크 구조와 달리 병렬 멀티 스케일 분석 방식을 추가했기 때문에, 스케일

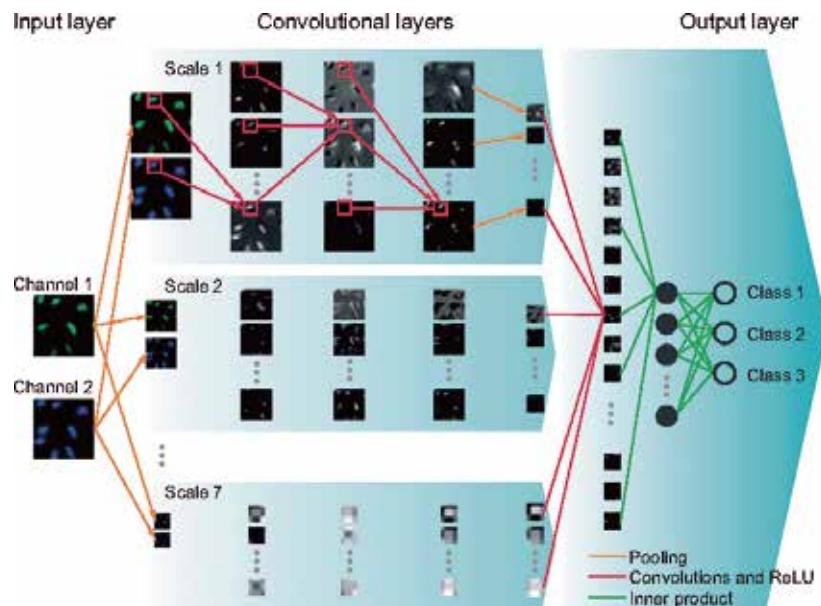


그림 4-1-2 M-CNN 아키텍처 설명도

이미지마다 다른 CNN 네트워크를 사용해 독립적인 방법으로 트레이닝을 진행할 수 있기 때문입니다.

그림 4-1-2는 7개 크기의 M-CNN 네트워크 구조를 표시한 것입니다. 크기는 위에서 아래로 점차 작아집니다. 네트워크는 입력 층에 크기가 다른 7개의 이미지를 입력하고, 콘볼루션 층 3개의 시퀀스를 사용해 모든 스케일의 이미지를 처리합니다. 콘볼루션 경로는 크기별로 분리되어 있지만 크기와 상관없이 마지막 층은 모두 수집 방식을 통해 획득한 콘볼루션 커널 매핑을 가장 굵은 크기에 맞추어 조절하고 링키지해 마지막 콘볼루션 층의 입력으로 삽습니다. 그리고 마지막 출력 층은 모든 표현형이 생성한 확률 값을 출력하게 됩니다.

소프트웨어/하드웨어 구성 제안

AI 기술을 통한 약물 연구개발의 가속화는 인텔®아키텍처 플랫폼 기반 소프트웨어/하드웨어 구성을 참조해 시스템을 배치할 수 있습니다.

파라미터	설명
CPU	인텔® 제온®골드 6240 CPU 또는 그 이상
하이퍼스레딩	ON
터보 가속	ON
메모리	16GB DDR4 2666MHz* 12 및 이상
저장	인텔®SSD D5 P4320 시리즈 및 이상
운영 체제	CentOS Linux 7.6 또는 최신 버전
Linux 코어	3.10.0 또는 최신 버전
컴파일러	GCC 4.8.5 또는 최신 버전
Tensorflow 버전	인텔®아키텍처 최적화된 TensorFlow v1.7.0 또는 최신 버전
Horovod	0.12.1 또는 최신 버전
OpenMPI	3.0.0 또는 최신 버전
ToRSwitch	인텔® Omni-Path 아키텍처

인텔® 제온® 확장성 CPU 기반 최적화

단일 컴퓨팅 노드 트레이닝의 효율 향상

신약 연구 개발은 길게는 수 년이 걸리다 보니 환자들은 언제나 조급한 마음으로 기다리기 마련입니다. M-CNN 네트워크 모델에 기초한 HCS 방식이 약물 개발 작업에 효율적으로 사용되어 연구개발이 가속화되도록 인텔® 제온® 확장성 CPU 최적화 방안을 출시했습니다. 이 제품은 단일 컴퓨팅 노드 스루풋 향상과 멀티플 컴퓨팅 노드 효율 향상 등 여러 방법을 포함하고 있습니다.

우선 단일 컴퓨팅 노드에서 M-CNN 모델을 켜 아래의 트레이닝 코드를 진행합니다.

```

1. python tf_cnn_benchmarks.py
2. --model=mcnn
3. --batch_size=32
4. --data_format=NCHW
5. --data_dir=INPUT_DATA_DIR
6. --data_name=mcnn
7. --num_intra_threads=40
8. --num_inter_threads=2
9. --num_batches=2000
10. --num_warmup_batches=70
11. --display_every=5
12. --momentum=0.9
13. --weight_decay=0.00005
14. --optimizer=momentum
15. --resize_method=bilinear
16. --distortions=False
17. --sync_on_finish=True
18. --device=cpu
19. --mkl=True
20. --kmp_affinity=="granularity=fine,compact,1,0"
21. --variable_update=horovod

```

단일 컴퓨팅 노드에서 M-CNN 모델을 사용하는 경우 메모리 용량이 문제될 수 있는데, 일반적으로 딥러닝 네트워크의 효율은 Batch Size가 증가함에 따라 일정 정도 향상됩니다. HCS에 쓰이는 세포 이미지는 보통 큰 편인데다 멀티스케일까지 결합해 작업할 경우, Batch Size가 일정량 까지 증가하면 필요 한 메모리 용량이 아주 커집니다. 그림 4-2-

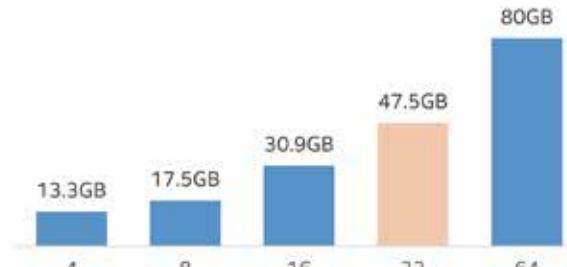


그림 4-2-1 Batch Size별 메모리 소비량

1에서 보는 바와 같이 Batch Size가 32이면, 시스템에 필요한 메모리가 47.5GB까지 커집니다.

인텔® 제온® 확장성 CPU 플랫폼은 대용량 메모리 지원 능력이 탁월하기 때문에 BatchSize 증가로 인한 대용량 메모리 수요를 해결할 수 있습니다. 최적화된 마이크로아키텍처와 더 다양해진 코어 수량, 더 빠르고 커진 용량의 메모리 제어와 조절 능력 덕분에 TensorFlow 아키텍처 기반 M-CNN 방식을 수월하게 진행할 수 있습니다. Broad Bioimage Benchmark Collection 021(BBBC-021) 데이터세트³³를 사용한 테스트에서, 입력한 현미경 이미지 크기가 1024*1280*3이고 Batch Size가 32일 때 단일 TensorFlow WP(Worker)가 처리할 수 있는 속도는 13장/초입니다. 이 속도로 수천만

장에 달하는 이미지의 데이터세트를 처리하기엔 너무 느리기 때문에 효율 향상이 필요합니다.

만약 NUMA(Non Uniform Memory Access Architecture) 기술과 분산형 딥러닝 아키텍처 Horovod에 기초한 가중치 동기화 기술을 도입하면 TensorFlow 아키텍처에서 TensorFlow WP 4개를 동시에 사용할 수 있습니다. 그림 4-2-2처럼 전형적인 컴퓨팅 노드에 배치한 투웨이 인텔® 제온® 확장성 CPU에 컴퓨팅 구역이 4개로 구분되며, 구역마다 TensorFlow WP를 1개씩 실행할 수 있습니다.

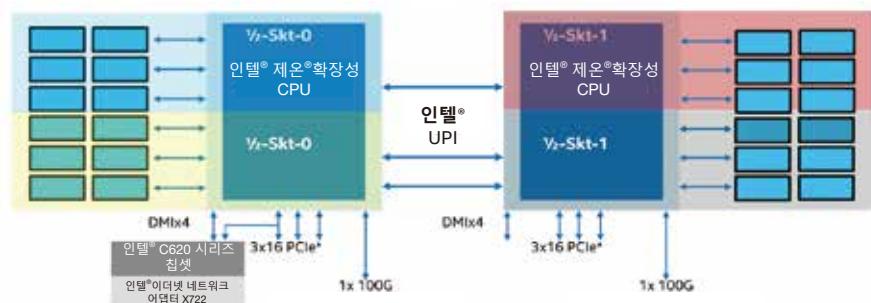


그림 4-2-2 전형적인 컴퓨팅 노드 중 투웨이 인텔® 제온® 확장성 CPU 구분

³³ BBBc-021: Ljosa V, Sokolnicki KL, Carpenter AE, Annotated high-throughput microscopy image sets for validation, Nature Methods, 2012

이러한 NUMA의 기술 특성을 이용하면 CPU의 코어와 메모리를 개별로 바인딩해 트레이닝을 진행할 수 있기 때문에 컴퓨팅 리소스와 스토리지 리소스끼리 경쟁하지 않아도 됩니다. 또한 컴퓨팅 구역마다 인텔®UPI(Intel® Ultra Path Interconnect, 인텔® UPI) 기술을 사용하면 가중치를 동기화할 수 있습니다. 이러한 방식을 통해 트레이닝 모델의 스루풋은 한층 더 향상될 수 있습니다. 그림 4-2-3의 오른쪽을 보면 TensorFlow Wp를 사용했을 때 Batch Size가 32이어도 처리 속도가 16.3장/초로 효율이 25.4% 향상되었음을 알 수 있습니다.

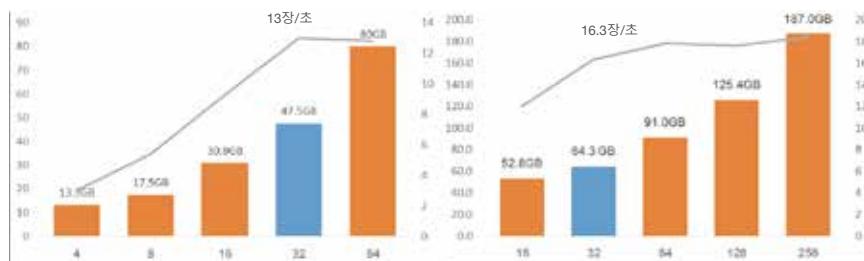


그림 4-2-3 TensorFlow 중 워커 스레드 4개와 단일 워커 스레드의 성능 비교

멀티플 컴퓨팅 노드 트레이닝 효율 향상

분산형 트레이닝 기술 방식을 이용하면 트레이닝 효율을 한층 더 업그레이드할 수 있습니다. 전형적인 TensorFlow 분산형 아키텍처에서는 파라미터 서버 방식으로 기울기를 평균화하다 보니 처리 스레드 모두 워커 스레드나 파라미터 서버가 될 수 있습니다. 전자는 사용자의 데이터 처리 및 트레이닝, 기울기 컴퓨팅하는데 사용되며 이를 파라미터 서버에 전달해 평균화를 진행하기도 합니다.

문제는 이 방식에서 파라미터 서버의 처리 능력이 부족할 경우, 시스템 전체 병목 현상이 발생하기 쉽다는 것입니다. 게다가 성능을 최고로 최적화하기 위해 사용자는 시작할 때부터 적당한 초기 워터 스레드와 파라미터 서버를 지정해야 하는데, 이때 자칫하면 성능이 하락할 수 있습니다. 하지만 새로운 오픈 소스 TensorFlow 분산형 딥러닝 아키텍처 Horovod가 이러한 문제를 효과적으로 해결할 수 있습니다. Horovod가 도입한 Ring-allreduce 알고리즘은 새로운 커뮤니케이션 전략을 구축했는데, 워커 스레드로 기울기를 평균화하기 때문에 파라미터 서버를 추가할 필요가 없게 되었습니다.

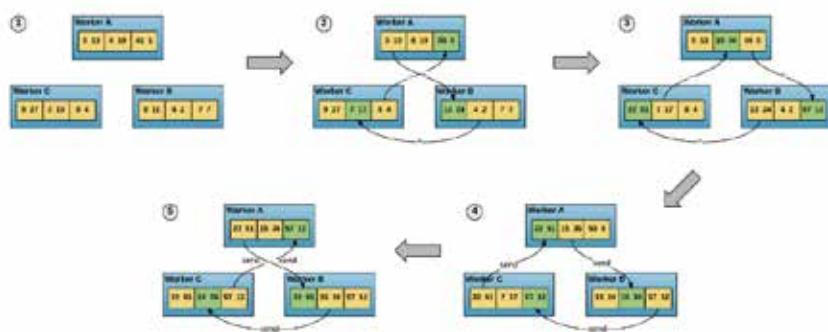


그림 4-2-4 Ring-allreduce 알고리즘 설명도

그림 4-2-4처럼, Ring-allreduce 알고리즘에서는 모든 워커 스레드가 개별 트레이닝 데이터에 따라 우선 기울기 컴퓨팅을 진행해 기울기 정보를 획득합니다. 그런 다음 워커 스레드별로 다른 N-1개의 워커 스레드와 $2*(N-1)$ 회의 통신을 진행하는데, 이 과정에서 워커 스레드별로 데이터 버퍼 존에서 전달한 기울기 정보를 발신 및 수신하며, 수신된 기울기 정보는 WP 버퍼 존에 추가되어 이전 값을 대체하게 됩니다. 모든 워커 스레드는 N-1개의 기울기 메시지를 발신 및 수신한 뒤, 컴퓨팅 업데이트 모델에 필요한 기울기를 수신합니다. 이 방법은 네트워크 능력을 최대화로 이용할 수 있게 해주기 때문에 컴퓨팅 병목 현상³⁴을 방지할 수 있습니다. Horovod는 이 커뮤니케이션 전략을 기초로 개방형 MPI(Open Message Passing Interface, OpenMPI)를 통해 TensorFlow 기반 분산형 시스템을 구축합니다.

이때 Horovod 아키텍처를 채택해도 전달해야 할 기울기 정보는 여전히 볼 수 있습니다. 예를 들어 Broad Bioimage Benchmark Collection* 021(BBBC-021) 데이터세트로 진행한 테스트에서 기울기 정보의 크기는 162.2MB입니다.

인텔® 제온® 확장성 CPU가 지원된 인텔® Omni-Path 아키텍처는 기울기 정보를 더 신속하게 전달하기 때문에 M-CNN 방식의 트레이닝 효율을 전체적으로 높일 수 있습니다. 인텔® Omni-Path 아키텍처는 100Gbps 점대점 방식의 대역폭과 1us급 점대점 방식의 MPI 통신 지연을 겸비했습니다. 게다가 OFA 소프트웨어 인터페이스와 완벽하게 호환 가능하며, RDMA와 PSM 인터페이스를 완벽하게 지원하고, 메시지 패킷 무결성 보장, 동적 링크 확장 등 혁신 기술도 보유로 기울기 정보를 고속 전송할 수 있는 견고한 기초를 다졌습니다. 그림 4-2-5처럼 인텔® 제온® 확장성 CPU를 8개 배치한 노드에 Horovod 아키텍처를 사용하면 동기점 전송이 10Gb 이상 됩니다.

³⁴ 기술 관련 세부 정보는 다음을 참조하십시오. Alex Sergeev, Mike Del Balso, Meet Horovod: Uber's Open Source Distributed Deep Learning Framework

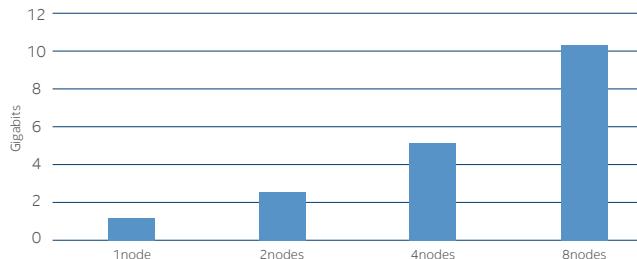


그림 4-2-5 Horovod와 인텔® Omni-Path 아키텍처를 사용한 동기점 전송 10Gb 초과

멀티플 컴퓨팅 노드 트레이닝 효율을 최적화할 수 있는 또 다른 방법은 학습률(Learning Rate, LR)을 수렴하고 조정하는 것입니다. 트레이닝 단계별 LR 크기는 딥러닝에서 아주 중요한 설정 항목인데, LR이 과도하게 크면 진동이 발생하고, 너무 작으면 수렴 속도가 느려져 과도한 맞춤이 발생하기 때문입니다. TensorFlow 아키텍처에 기초한 M-CNN 모델 트레이닝 과정은 아래의 LR 조정 방식으로 성능을 최적화할 수 있습니다.

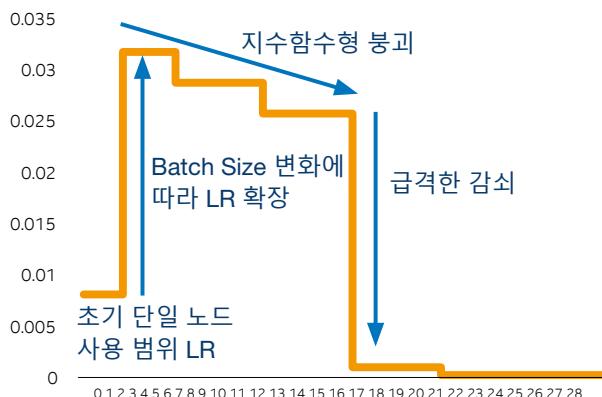


그림 4-2-6 M-CNN 네트워크 트레이닝 과정 중 LR 조정

그림 4-2-6처럼 트레이닝 초기에는 먼저 사용한 단일 노드의 LR을 반복한 뒤, 이것을 전체 Batch Size 파라미터까지 확장합니다. 반복 과정 중, LR은 지수 방식으로 감소하게 되고 14회를 반복할 때부터는 LR이 급격하게 감소³⁵하기 시작합니다.

그렇기 때문에 멀티플 컴퓨팅 노드에서 M-CNN 네트워크의 트레이닝 명령은 다음과 같습니다.

1. OMP_NUM_THREADS=10 mpirun -np 32 -cpus-per-proc 10
2. --map-by socket -hostfile HOSTFILE
3. --report-bindings
4. --oversubscribe -x LD_LIBRARY_PATH -x
5. PATH -x OMP_NUM_THREADS -x HOROVOD_FUSION_THRESHOLD numactl -l

```

6.
7. python tf_cnn_benchmarks.py
8. --model=mconv
9. --batch_size=8
10. --data_format=NCHW
11. --data_dir=INPUT_DATA_DIR
12. --data_name=mconv
13. --num_intra_threads=10
14. --num_inter_threads=2
15. --num_batches=2000
16. --num_warmup_batches=70
17. --display_every=5
18. --momentum=0.9
19. --weight_decay=0.00005
20. --optimizer=momentum
21. --resize_method=bilinear
22. --distortions=False
23. --sync_on_finish=True
24. --device=cpu
25. --mkldn=True
26. --kmp_affinity="--granularity=fine,compact,1,0"
27. --variable_update=horovod
28. --local_parameter_device=cpu
29. --kmp_blocktime=1
30. --horovod_device=cpu
31. --piecewise_learning_rate_schedule=0.008;0.032;5;0.029;10;0.026;15;0.001;20;0.0001
32. --train_dir=TRAIN_DATAWRITE_DIR
33. --save_summaries_steps=1
34. --summary_verbosity=1

```

딥러닝을 이용해 약물 연구개발 효율을 향상시킨 노바티스

배경

세계를 선두하는 제약회사 노바티스는 디지털화 전환 흐름에 힘입어 약물 혁신과 질병 진단, 약물 연구 등 방면에서 우위를 점하고 있습니다. "AI+약물 발견"은 가까운 미래의 약물 연구 개발에 있어 아주 중요한 부분입니다.

현재 노바티스는 인텔과 함께 딥러닝 방식을 이용한 HCS 가속화를 연구 중이며, 세포 표현형인 HCS는 노바티스가 진행 중인 초기 약물 발견 사업에 중요한 역할을 하고 있습니다. HCS란 전형적인 이미지 처리 기술을 사용해 이미지에서 추출한 수천 개의 사전 정의 특성(예: 크기, 형태, 무늬 등등)의 집합을 가리킵니다. HCS는 마이크로이미지를 분석하며, 수천 가지의 유전 연구나 화학 처리로 여러 세포 배양물에 영향을 미칩니다. 딥러닝 방식을 이용하면 데이터에서 자동 학습할 수 있고, 하나님의 치료 방식과 또 다른 치료 방식 간의 관련 이미지 특성을 구분할 수 있습니다. 하지만 세포 현미경 이미지의 방대한 정보량 때문에 여전히 긴 시간이 소요된다는 점은 개선되어야 할 부분입니다. 해당 이미지 분석 모델의 트레이닝 시간은 약 11시간³⁶입니다.

³⁵ LR 기술 설정 관련 세부 정보는 다음을 참조하십시오. Yang You et al, 2017, "ImageNet Training in Minutes"

³⁶ 데이터 출처: <https://newsroom.intel.com/news/using-deep-neural-network-acceleration-image-analysis-drug-discovery/#gs.ptk50k>

이제 인텔과 노바티스의 생물학자와 데이터 과학자들은 최적화된 인텔® 제온® 확장성 CPU 플랫폼에 배치한 M-CNN 네트워크로 HCS 분석을 가속화하려고 합니다. 이번 연합 작업에서 팀은 단독 프로세스로 이미지의 개별 세포를 우선 식별하는 방법이 아닌 현미경 이미지 전체에 집중했습니다. 이번에 사용한 데이터셋은 Broad Bioimage Benchmark Collection* 021(BBBC-021)의 현미경 이미지는 흔한 딥러닝 데이터셋의 이미지보다 훨씬 클 수 있습니다.

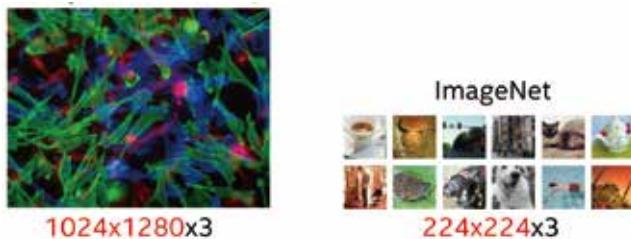


그림 4-3-1 HCS에 사용한 현미경 이미지와 일반 이미지 데이터셋과의 비교

그림 4-3-1을 보면, 왼쪽은 HCS에 사용된 현미경 이미지로 장당 픽셀이 400만에 달합니다. 오른쪽은 유명한 ImageNet 데이터셋³⁷의 이미지이며, 트레이닝 데이터셋의 장당 이미지가 15만 픽셀로 두 이미지의 차이가 26배에 달합니다. 크기가 방대한 현미경 이미지에 수백만 개의 파라미터, 거기에 1회 트레이닝에 수천 개에 달하는 규모까지 덜해지면 시스템 메모리에게는 부담이 될 수 있고, 막대한 컴퓨팅 부하를 가져올 수도 있습니다. 이런 부담을 해결하기 위해 양측은 DNN 최적화와 가속화 기술을 채택해 더 짧은 시간 안에 시스템이 여러 개의 이미지를 처리하고 정확도를 유지할 수 있도록 도왔습니다.

방법과 성과

최적화 방안은 두 가지 측면에서 인텔® 제온® 확장성 CPU 기반 플랫폼으로 배치한 M-CNN 모델의 트레이닝을 가속화했습니다. 첫 번째, 단일 컴퓨팅 노드에서 인텔® 제온® 확장성 CPU 플랫폼을 이용해 대형 메모리를 지원하기 때문에 대형 Batch_Size(방안에서는 32로 설정)를 채택할 수 있고, NUMA 기술로 워커 스레드를 증가해 트레이닝 효율을 올릴 수 있습니다. 두 번째, 멀티플 컴퓨팅 노드에 오픈 소스의 TensorFlow 분산형 딥러닝 아키텍처 Horovod를 도입하고, 인텔® 제온® 확장성 CPU가 지원한 인텔® Omni-Path 아키텍처를 결합해 멀티 노드에서 M-CNN 모델의 트레이닝 효율을 대폭 향상시킵니다. 거기에 최적화 이후 학습률을 수렴하고 조정 방안을 설계 및 채택해 성능³⁸을 향상시킵니다.

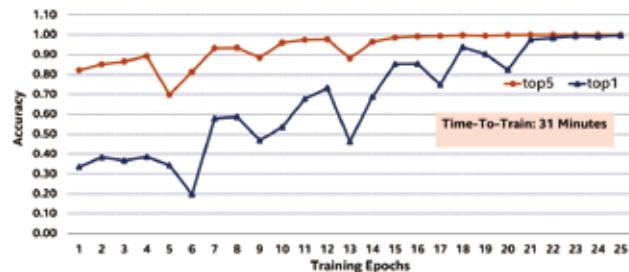


그림 4-3-2 노바티스 최적화 이후 방안의 트레이닝 효과

방안에서는 인텔® 제온® 확장성 CPU 기반 노드 8개를 배치하고, Broad Bioimage Benchmark Collection* 021(BBBC-021) 데이터셋을 사용합니다. 이미지 총량은 1만장이고, 크기는 1024*1280*3입니다. 20회 이상 트레이닝 총시간은 그림 4-3-2처럼 31분이고, 정확도는 99%가 넘습니다. 또한, NUMA 기술을 사용해 TensorFlow WP(노드당 워커 스레드 4개)를 32개 형성하면, 처리능력이 120 다중 이미지/초가 되어 최적화하기 이전과 비교해 성능이 눈에 띄게 향상됩니다.

³⁷ ImageNet: Russakovsky O et al, ImageNet Large Scale Visual Recognition Challenge, IJCV, 2015

³⁸ 데이터에 사용한 테스트 구성: 투웨이 인텔® 제온® 골드 6148 CPU, 2.40GHz 코어/스레드: 20/40, HT: ON, Turbo: ON, 메모리: 16GB DDR4 2666*12, 하드 디스크: 480GB 인텔® SSD OS drive*1, 1.6TB 인텔® SSD data drive*1 네트워크 어댑터: 인텔® Omni-Path 호스트 구조 인터페이스(HFI), BIOS: SE5C620.8 6B.02.01.0008.031920191559, 운영 체제: CentOS Linux 7.3, gcc 버전: 6.2, Tensorflow 버전: 인텔® 아키텍처 최적화 지향 Tensorflow v1.7.0 Horovod 버전: 0.12.1, OpenMPI: 3.0.0, ToRSwitch: 인텔® Omni-Path 아키텍처 워크로드: Broad Bioimage Benchmark Collection* 021(BBBC-021) 데이터셋으로 이미지는 1만장이며, 크기는 1024*1280*3입니다.

요약

신약 개발은 발견부터 테스트, 생산까지 보통 수년의 시간이 소요되다 보니 결과를 기다리는 환자와 가족은 항상 애가 탑니다. 하지만 AI 기술을 이용해 약물 연구 개발을 가속화하면 여러 제약 회사가 혁신을 가속화할 수 있고, 코어 경쟁력의 보편적 선택을 유지할 수 있을 뿐만 아니라 과학 기술이 인류에게 행복을 선사해 건강한 생활을 유지할 수 있도록 도와줍니다. 약물 연구 개발에 AI 방안이 빠르게 응용될 수 있도록 인텔 역시 여러 제약회사와 함께 노력 중입니다.

합리적인 최적화 방안과 인텔® 제온® 확장성 CPU, 인텔® Omni-Path 아키텍처 등 첨단 기술과 제품을 더해지면, 딥러닝에 기초한 HCS 등 AI 응용에 대형 메모리와 대형 Batch_Size, 다중 TensorFlow WP를 지원할 수 있기 때문에 단일 노드 또는 멀티 노드의 트레이닝 효율을 가속화할 수 있습니다. 또한 대역폭이 높고 지연 시간이 낮은 첨단 접속 랙으로 Horovod 분산형 트레이닝 아키텍처를 지원해 노바티스 등 제약회사의 약물 연구 개발 과정을 대폭 가속화할 수 있습니다.

현재 인텔® 제온® 확장성 CPU 플랫폼 기반 AI 응용은 이미 여러 제약 회사에 배치되었으며 뛰어난 성과를 거두었습니다. 본문의 테스트는 인텔® 제온® 골드 6148 CPU 플랫폼에서 진행했지만, 2세대 인텔® 제온® 확장성 CPU와 인텔® 옵테인™ 데이터 센터급 영구 메모리 등 차세대 인텔 하드웨어와 기술을 출시하고 응용했습니다. 그렇기 때문에 사용자는 향후 실제로 배치하는 과정에서 업데이트된 인텔 하드웨어 플랫폼과 소프트웨어 최적화 방안을 이용해 더 강력한 성능의 딥러닝 방안을 구축할 수 있게 되었습니다. 덕분에 더 뛰어난 트레이닝 및 추론 효과를 얻고, 약물 개발 과정을 한 단계 더 가속화할 수 있어서 환자의 치료와 회복에 도움이 됩니다.

의료 업계에 응용된 AI 기반 화상 인식 기술

스마트 의료와 화상 인식 기술

의료 업계의 화상 인식 기술

병원 정보 시스템(Hospital Information System, HIS)과 진료 정보 시스템(Clinical Information System, CIS), 전자 진료 기록(Electronic Medical Record, EMR) 및 의료영상전송시스템(Picture Archiving and Communication Systems, PACS)등과 같은 규범화된 정보 시스템을 이용해 더 스마트한 의료 정보화 능력을 만들어내는 의료 기관이 점점 더 많아지고 있습니다. 그 덕분에 환자는 의료 종사자, 의료 기관, 의료 설비 관계자와 더욱 효율적으로 소통할 수 있게 되었습니다. 고효율의 식별 기술은 시스템을 관통하고 스마트 의료가 효능을 발휘하도록 다양한 지원을 제공합니다.

예를 들어, 의료 기관은 전통적으로 바코드 식별과 광학적 문자 인식(Optical Character Recognition, OCR), 소프트웨어 식별 등의 기술로 환자의 신원 확인과 약품 지급 등 작업을 실행했는데, AI 기술이 점차 발전하면서 머신러닝과 딥러닝 등 AI 방식을 사용하는 의료 기관이 점점 더 많아졌습니다. AI 방식은 환자의 신원을 실시간으로 확인하고 약품을 더 정확하게 지급하며 의료 검사 프로세스를 완벽하게 연결할 수 있기 때문에 식별 시스템 전체의 효율과 정확도를 향상시키고, 의료 기관의 업무 효율을 강화할 수 있습니다.

■ 바코드 식별

바코드 식별 기술이란 바코드 스캐너 등 광전 변환 기기를 이용해 스캔한 바코드를 식별하는 방식을 가리킵니다. 바코드는 두껍고 얇은 검은색과 흰색의 막대 모양으로 구성된 시퀀스로서 일정 숫자와 문자 부호를 표시합니다. 바코드 식별은 현재 의료 업계에서 흔하게 사용되는 식별 기술 중 하나이며, 바코드를 진료 기록과 검사 보고서 및 기타 물품에 인쇄할 수 있고 정확하고 빠르게 식별하기 때문에 시스템 통합이 간단하다는 점이 가장 큰 장점입니다. 예를 들어, 의사가 치료를 진행하기 전 스캐너로 진료 기록의 바코드를 스캔만 하면 백그라운드의 연관 바코드 라이브러리에서 질환 정보를 추출할 수 있습니다. 약품을 지급할 때에는 스캐너로 약품 겉포장지의 바코드를 스캔만 하면 약품 정보를 바로 얻을 수 있습니다.

하지만 바코드 식별 기술 역시 단점은 있습니다. 우선 작업자가 바코드를 일일이 스캔해야 하다 보니 작업 속도가 너무 느립니다. 또 하나는 모든 프로세스에 바코드가 있지 않다는 점입니다. 주사용 주사약이나 약병에는 바코드가 부착되어 있지 않다 보니 간호사가 주사를 놓을 때마다 일일이 대조해야 하는 번거로움이 있습니다. 게다가 바코드 라이브러리를 유지 보수하려면 상당한 시간과 에너지가 드는 것도 단점 중 하나입니다.

■ OCR 식별

OCR 식별은 일반적으로 스캐너같은 전자 기기로 지면의 문자 부호나 부호를 획득하는 방식을 가리킵니다. 우선 양암측정 모드로 형태를 확인한 뒤 문자 부호 식별 방식으로 컴퓨터가 식별할 수 있는 문자 부호로 전환합니다. 이 방식의 장점은 획득할 수 있는 정보의 범위가 넓어서 1회 스캔만으로도 페이지의 전체 문자 정보를 획득할 수 있다는 점입니다. 그렇기 때문에 의료 업계에서 OCR을 이용하면 진료 기록과 검사 보고서, 약품 겉포장지 등 이미지를 수집할 수 있는 것은 물론, OCR로 내부 정보까지 판독할 수 있습니다.

물론 OCR 식별 방식도 두드러지는 단점이 있습니다. 첫째, OCR은 각도와 광선 등 조건의 영향을 쉽게 받기 때문에 식별 오차가 커서 100% 정확하게 식별하기 힘듭니다. 둘째, OCR 식별은 문자(알파벳, 숫자, 부호 등)만 식별하며, 이미지는 식별할 수 없습니다. 셋째, OCR 식별 효율이 낮아 시간적 제약을 많이 받는 진료 프로세스에 응용할 경우 업무 전체가 지연될 수 있습니다.

■ 소프트웨어 식별

컴퓨터 이미지 기술이 끊임 없이 발전하면서 이미지 처리 소프트웨어와 기술을 의료 업계에 도입해 이미지와 문자 식별을 진행하는 경우가 점점 많아졌습니다. 예를 들어, OpenCV 컴퓨터 비전 라이브러리는 플랫폼 간 물체 식별과 이미지 분할, 언어 인식, 문자 인식 등 일련의 이미지 처리와 분석 업무를 구현할 수 있습니다. 일반적으로 사용자는 Python, Java 같은 프로그래밍 언어를 채택해 OpenCV에 기초한 자신만의 식별 시스템을 개발할 수 있습니다. 이 방식은 식별 확률이 높아 문자와 사진을 동시에 식별할 수 있는 장점을 가지고 있습니다. 하지만 맞춤형으로 개발한 소프트웨어는 업데이트를 반복할 경우 속도가 느려진다는 단점이 있기 때문에, 수요 변화가 빠른 의료 기관에 적용하기에는 어려움이 있습니다.

딥러닝 기반 화상 인식 기술

딥러닝 기반 화상 인식 기술은 전통적인 화상 인식 기술보다 정확도와 업무 효율이 높아 업데이트 메커니즘을 형성하기에 더 유리합니다. 또한 이미지 특성에 기초해 식별을 진행하기 때문에 1회만으로도 다양한 종류와 다수의 이미지를 획득해 특성을 식별할 수 있습니다. 현재 여러 대형 오픈 소스 커뮤니티에서는 비교적 완전한 형태의 화상 인식 알고리즘과 딥러닝 아키텍처를 참고 및 조정용으로 보유하고 있습니다.

■ 콘볼루션 뉴럴 네트워크

콘볼루션 뉴럴 네트워크(CNN)은 딥러닝의 대표 알고리즘 중 하나로서 콘볼루션 컴퓨팅을 포함하고 다층 구조의 순방향 신경망을 갖추고 있습니다. CNN으로

구축한 모델은 이미지의 특성을 손쉽게 식별하고 정렬할 수 있으며,

CNN은 이미지의 변위와 크기 조절, 왜곡에 대해 특성 불변성을 지니고 있습니다. 즉 일반화 성능이 강한 편입니다. 게다가 CNN의 가중치 공유 구조는 뉴럴 네트워크의 파라미터 수량을 대폭 감소시킬 수 있기 때문에 과도한 맞춤을 방지하고 뉴럴 네트워크 모델의 복잡도를 떨어뜨릴 수도 있습니다. 이와 같은 특성 덕분에 CNN을 의료 영역 중 물품 식별 응용 시나리오에서 사용하면 광선과 배열 위치 등의 요소로부터 받는 영향을 없앨 수 있습니다. 그러면 화상 인식 정확도가 향상될 뿐만 아니라 복잡도도 적당해져 사용자가 종복 트레이닝과 학습을 진행하기에 유리해집니다. 데이터에서 볼 수 있듯이, 2016년 뉴럴 네트워크 모델 기반 화상 인식 top5 오류율이 이미 2.991%까지 하락하면서 동일 화상 인식이 식별할 때 오류율이 5.1%³⁹보다 낮아졌음을 알 수 있습니다. 현재 LeNet-5, ZFNet, VGGNet과 ResNet 등, CNN의 변형체는 텍스트와 인물, 동작 등 화상 인식 및 분석에 다양하게 사용되고 있습니다.

■ LeNet-5 CNN 기반 딥러닝 방식

LeNet-5 CNN은 현대 CNN의 기본 구조를 형성했습니다. 교체 출현한 콘볼루션 층과 풀링 계층은 입력한 이미지의 TI 특성을 추출해 수기 서명처럼 변위되고 크기 조절 및 왜곡된 이미지를 식별할 수 있는 능력을 갖추고 있습니다.

그림 5-1-1은 전형적인 7층 LeNet-5 CNN 트레이닝 모델로 입력과 출력 외에도 여려 개의 콘볼루션 층과 풀링 계층, FC를 포함하고 있습니다.

LeNet-5 CNN은 입력 층에서 2차원 픽셀점, RGB 채널 등 다차원의 입력 데이터를 수신하고 처리할 수 있으며, 표준화 처리까지 진행합니다. 표준화 처리는 데이터를 CNN에 입력하기 전에 채널과 시간, 빈도 등 차원에서 입력 데이터를 정규화 컴퓨팅해 모델의 운용과 학습 효과에 유리하게 만들어 줍니다.

입력 층 다음은 몇 개의 콘볼루션 층과 풀링 계층으로 이루어져 있습니다. 콘볼루션 층의 주요 기능은 입력 데이터에 특성 추출을 진행하는 것이며, 이 특성 추출은 층별로 점차 증가합니다. 첫 번째 콘볼루션 층은 보통 비교적 간단한 특성만 추출하지만, 그 다음 콘볼루션 층은 이 간단한 특성을 기초로 더 복잡한 특성을 추출할 수 있습니다. 풀링 계층의 역할은 콘볼루션 층이 출력한 특성을 선택하고 정보를 필터링하는 것입니다. FC는 일반적으로 CNN의 가장 마지막 부분에 구성되며, 특성의 3차원 구조를 벡터로 전환하고 다음 층으로 전달할 수 있습니다. 마지막 출력 층에서는 로직 함수나 정규화 지수 함수를 사용해 정렬 태그를 출력합니다.

모델 구현 및 최적화

■ TensorFlow 구현 및 최적화

TensorFlow로 LeNet-5 CNN을 구현하면 공식 모델 코드를 채택해 직접 모델 트레이닝을 진행할 수 있습니다. Github의 slim 목록에는 CNN 모델을 채택한 다양한 트레이닝 코드가 집약되어 있으며, train_image_classifier.py로 직접 훈련⁴¹할 수 있습니다. 구체적인 트레이닝 명령은 다음과 같습니다.

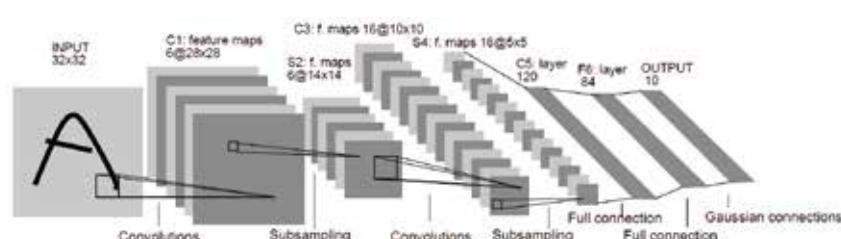


그림 5-1-1 전형적인 7층 LeNet-5 CNN 트레이닝 모델⁴⁰

³⁹ Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition[R/OL]. <https://arxiv.org/abs/1512.03385>, 2015-12-10

⁴⁰ 사진과 LeNet-5 관련 설명 출처: LeCun, Y.; Bottou, L.; Bengio, Y.&Haffner, P.(1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE.86(11): 2278 - 2324.]

⁴¹ Github 관련 주소: <https://github.com/tensorflow/models/research/slim/>

```
1. [root@worker105 slim]# python train_image_classifier.py -model_name lenet -dataset_name mnist -batch_size 32 |
```

또한 인텔®CPU의 컴퓨팅 리소스가 충분히 이용될 수 있도록 트레이닝 코드에서 다음 최적화를 진행할 수도 있습니다.

```
1. config = tf.ConfigProto()
2. config.intra_op_parallelism_threads = 12
3. config.inter_op_parallelism_threads = 1
4. tf.Session(config=config)
```

Intra_op_parallelism_threads 파라미터와 inter_op_parallelism_threads 파라미터는 모든 오퍼레이터 op가 병렬 컴퓨팅한 스레드 개수를 제어하는 데 사용합니다. 전자는 연산자 op 내부의 병렬을, 후자는 여러 개의 연산자 op간의 병렬 컴퓨팅을 제어합니다.

Python 코드를 실행하기 전에는 환경 변수를 설정해 인텔® MKL-DNN을 최적의 상태⁴²로 만들 수도 있습니다. 상세 정보는 다음과 같습니다.

```
1. export KMP_BLOCKTIME=1
2. export KMP_SETTINGS=1
3. export KMP_AFFINITY=granularity=fine,compact,1,0
4. export OMP_NUM_THREADS=12
```

KMP_BLOCKTIME을 1로 설정합니다. 이는 스레드 현재 임무를 실행 완료하고 휴면 상태에 들어가기 전까지 대기하는 시간 설정으로 일반적으로 1 밀리 초로 설정합니다. KMP_SETTINGS을 1로 설정하면 프로그래밍 실행 기간 동안 OpenMP RTL(run-time library) 환경 변수를 출력할 수 있습니다. KMP_AFFINITY를 Compact로 설정합니다. 이는 해당 모드에서 스레드 바인딩이 컴퓨팅 코어의 컴퓨팅에 따라 우선 순위를 정하는 것을 의미합니다. 우선 동일한 코어를 바인딩하고, 그 다음으로 동일 CPU의 다음 코어를 바인딩합니다. 이러한 바인딩 방식은 스레드 사이에 데이터 교류가 이루어지거나 공용 데이터를 갖춘 상황에 적용되며, 다양한 계층의 캐시를 충분히 이용할 수 있다는 점이 장점입니다. OMP_NUM_THREADS는 사용할 스레드 수를 가리킵니다.

* 인텔®아키텍처 최적화를 지향하는 TensorFlow 기술과 관련한 더 자세한 사항은 해당 매뉴얼 기술편 소개 부분을 참조하십시오.

소프트웨어/하드웨어 구성 제안

스마트 의료에 딥러닝 기반 화상 인식 방안을 구축 하려면, 인텔®아키텍처 플랫폼 기반 소프트/하드웨어 구성을 참조하십시오.

명칭	사양
CPU	인텔® 제온®골드 6240 CPU 또는 그 이상
하이퍼스레딩	ON
터보 가속	ON
메모리	16GB DDR4 2666MHz* 12 및 이상
저장	인텔®SSD D5 P4320 시리즈 및 이상
운영 체제	CentOS Linux 7.6 또는 최신 버전
Linux 코어	3.10.0 또는 최신 버전
컴파일러	GCC 4.8.5 또는 최신 버전
Python 버전	Python 3.6 또는 최신 버전
Tensorflow 버전	R1.13.1 또는 최신 버전

외래 진찰 약품 지급에 딥러닝 기술을 이용한 해방군총병원

배경

처방에 따른 약품 지급과 약품 사용은 병원의 약품 유통 사이클 중 "마지막 1미터"에 해당하는 중요한 단계일 뿐만 아니라 질환 치료에 있어 가장 중요한 단계 중 하나이기도 합니다. 병원은 오랫동안 여러 제도를 통해 약품 지급과 약품 사용 단계를 엄격하게 관리해 왔습니다. 외래진료 창구에서 약을 발급할 때에는 "4가지 검사와 10가지 대조"를, 병실에서 의사의 지시에 따라 약물을 치료할 때에는 "3가지 검사와 7가지 대조"를 진행해야 했습니다. 병실에는 약품이 혼용되는 것을 막기 위해 전용 라벨을 붙였고, 약품 지급이나 사용이 잘못되었을 경우 의료 사고로 기록해 상부 보고 처리됩니다.

그런데 관리 규정이 이렇게 엄격한데도 불구하고 여러 요인으로 약물 지급 및 사용

오류로 인한 의료 분쟁과 사고는 여전히 빈번하게 일어나고 있습니다. 약품 지급 및 사용 오류의 주요 원인은 다음과 같이 분석할 수 있습니다.

- 관리 제도가 실제 상황에 적절히 적용되지 않았습니다. 제도 제정은 완전한 편이지만, 실행력은 여전히 부족합니다.
- 약사나 의료인의 업무 시간이 길고 업무량이 많다 보니 스트레스로 인해 실수가 자주 일어납니다.
- 약품 자체의 문제도 있습니다. 일부 약품은 이름과 걸포장지가 쉽게 혼용되며 디자인되어 있습니다.

유명한 3등급 의원인 해방군총병원 역시 이런 문제에 직면해 있습니다. 2010년부터 2017년 까지 이 병의 외래 진찰 양은 연평균 5.71%, 처방약 지급양은 연평균 7.63% 급증했고, 외래 진찰 약국 약품 중 15.4%가 혼용⁴³하기 쉬운 라벨이 부착되어 있습니다. 하지만 약국 업무 인원수는 여전히 요지부동이고, 노동 강도는 반대로 증가하다 보니 약품 지급 오류 리스크는 오히려 높아진 상황입니다.

이러한 문제를 해결하기 위해 해방군총병원은 정보화 수단을 시범적으로 사용해 약품 지급 오류를 줄이려 하고 있습니다. 우선, 컴퓨터 비전 기술을 이용해 외래 진찰 약품 지급 창구에 약품 분류와 수량을 표시해 식별합니다. 그 다음으로 이 식별 시스템을 HIS 시스템의 처방 데이터와 자동 연결 및 매칭해 정보 비교를 통해 지급 예정 약품이 처방 정보와 일치하는지 판단하고, 그 결과를 약사에게 실시간으로 피드백해 약품 지급 오류 감소 목표를 달성합니다.

방법과 성과

해방군총병원이 딥러닝 기술을 이용해 외래 진찰 약품 지급 문제를 해결하는 방안의 기본 과정은 그림 5-2-1과 같습니다.

단계 1: 약사가 지급 예정인 약품을 약품 지급 창구 작업대에 올려 놓습니다.

단계 2: 작업대 상단의 이미지 수집 장치에서 약품 이미지를 자동 캡처해 백그라운드 시스템에 전송합니다.

⁴² 구체적인 내용은 TensorFlow 공식 홈페이지 참조: <https://tensorflow.google.cn/guide/performance/overview?hl=zh-cn>

⁴³ 데이터 출처: 장전지양, 스화위, 신하이리, 리황, 류민차오의 《딥러닝 기술 외래 진찰 약품 지급 실제 사례》

단계 3: 딥러닝에 기초한 약품 걸포장지 식별 모듈은 약품 걸포장지를 빠르게 식별하고 그 결과를 스크린에 표시합니다.

시스템은 HIS 처방 정보와 자동 연결해 약품 이름과 사양, 제조업체, 수량 등 파라미터를 각각 매칭한 뒤 오류 정보가 있을 경우 다른 색으로 표시해 알립니다.



그림 5-2-1 약품 지급 창구 응용 시나리오 설명도

해방군총병원은 이 방안을 실행하면서 CNN을 이용해 약품 걸포장지 식별 모듈을 구축하고, 딥러닝 방식으로 약품 걸포장지의 특성을 식별했습니다. 식별 목표는 두 가지입니다. 하나는 약품, 특히 혼용하기 쉬운 약품을 효과적이고 정확하게 식별하는 것입니다. 또 다른 하나는 약품 수량을 집계하는 것입니다.

그림 5-2-2처럼 뉴럴 네트워크 기반 약품 걸포장지 식별 모듈 작업 프로세스는 주로 이미지 데이터 전처리와 모델 트레이닝, 반복 최적화, 추론 예측 등 단계를 포함하고 있습니다. 이미지 데이터 전처리 단계에서는 다양한 라벨 데이터가 필요한 뉴럴 네트워크의 모델 트레이닝 특성에 따라, 소량의 초기 사진을 우선 수집한 뒤 다양한 트레이닝용 이미지를 자동 생성하는 모드를 채택했습니다.

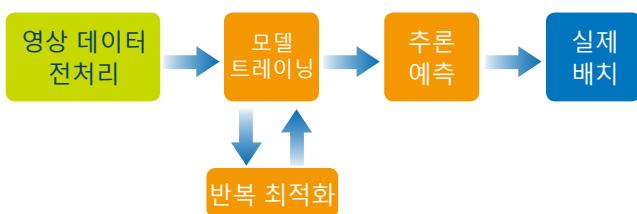


그림 5-2-2 뉴럴 네트워크 기반 약품 걸포장지 식별 모듈

우선, 약품 걸포장지 6개의 초기 사진을 각각 수집해 유통기한과 전자 감독 코드를 포함한 이미지 정보를 삭제합니다. 그런 다음, 등고선 알고리즘으로 약품 사진을 획득해 간섭을 줄입니다. 그리고는 드로잉과 왜곡, 크기 조절, 회전, 랜덤 변위 등 이미지 처리 방식으로 새로운 이미지를 획득하고, 모든 약품의 사진을 단독 목록에 저장합니다. 해방군총병원은 56개종 약품에서 279개의 초기 이미지를 수집했고, 전처리로 생성한 이미지가 467,752장에 달합니다. 그 중 랜덤으로 289,448장을 선택해 트레이닝에, 178,304장을 선택해 모델 검증에 사용했습니다.



그림 5-2-3 약품 걸포장지 화상 인식에 사용되는 딥러닝 트레이닝 모델

실제 수요를 만족시키기 위해 해방군총병원은 LeNet-5 CNN을 기초로 해 그림 5-2-4처럼 약품 걸포장지 화상 인식에 사용하는 7층 CNN 트레이닝 모델(콘볼루션 층 3개, 풀링 계층 3개, FC 1개 포함)을 설계해 56개종 약품의 279개 유형에 모델 트레이닝을 진행했습니다. 트레이닝 과정에서 시스템은 약 상자별로 예측 정밀률 값을 1개씩 컴퓨팅했으며, 96% 이상의 최대 예측 확률 값에 대응하는 최종 예측 정밀률을 진행했습니다. 전체 예측 확률 값이 모두 96% 미만이면 식별 실패로 판단합니다. 전체 트레이닝 반복 횟수는 3,000회이며, 총시간은 약 12시간입니다. 그림 5-2-4처럼 트레이닝 횟수가 증가함에 따라 예측 정확도도 점차 향상되며, 마지막에는 식별 정확도⁴⁴가 95.6%까지 다다랐습니다.

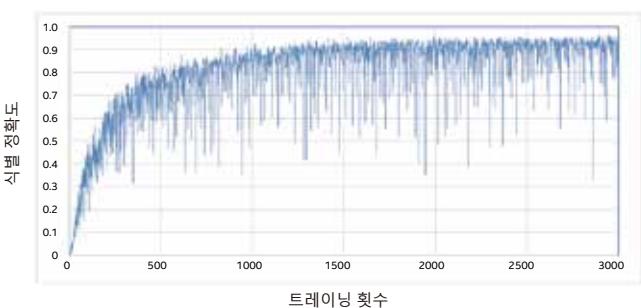


그림 5-2-4 식별 모델 트레이닝 횟수와 식별 확률 변화 차트

이어지는 실제 시나리오 검증 테스트에서 해방군총병원은 일부 혼용되기 쉬운 약품에 식별 테스트를 진행했습니다. 그 결과, 약품 걸포장지 식별 모듈은 특성에 따라 정확히 구분하는 것으로 밝혀졌습니다. 그림 5-2-5에서 왼쪽 그림은 샤메이터루어티카 흡입제(가루형) 50마이크로그램/250마이크로그램과 50마이크로그램/100마이크로그램 두 사양의 약품 식별 효과고, 오른쪽 그림은 토브라마이신 점안액과 토브라마이신 텍사메타손 점안액의 식별 효과(초록색은 시스템 판별 결과임)입니다.

⁴⁴ 데이터 출처: 장전지양, 스학위, 신하이리, 리황, 류민차오의《딥러닝 기술 외래 진찰 약품 지급 실제 사례》, 사용한 테스트 구성: 투웨이 인텔® 제온®골드 6126 CPU, 2.60GHz 코어/스레드: 12/24, HT: ON, Turbo: ON, 메모리: 16GB DDR4 2666*12,

하드 디스크: 인텔® DC S3320 데이터 센터급 SSD BIOS: SE5C620.86B.00.01.0013.030920180427, 운영 체제: CentOS Linux release 7.4.1708(Core). Linux 커널: 3.10.0-957.12.1.el7.x86_64, gcc 버전: 7.1, Tensorflow 버전: R1.10.0, Python 버전: Python 2.7, 워크로드: Lenet



그림 5-2-5 약품 걸포장지 식별 모듈 식별 테스트 결과

약품 걸포장지 식별 모듈은 이미 해방군총병원 외래 진찰 약품 지급 시스템에 배치되었으며, 약국 업무자의 업무 강도를 효과적으로 낮추었을 뿐만 아니라 약품 지급 정확도를 높인 덕분에 환자의 만족도까지 대폭 향상되었습니다.

앞으로 해방군총병원은 외래 진찰 약품 지급 시스템을 다음 측면에서 최적화하고 개선하려고 계획하고 있습니다.

- 업무 프로세스와 방안을 한 단계 더 융합: 현재 방안에서는 약사가 약품을 일일이 작업대에 놓아야 하다 보니 작업 소요 시간이 긴 편입니다. 약 바구니를 통째로 작업대에 올리면 시스템은 즉시 수집한 약품 이미지를 분류하고 수량을 집계하기 때문에 효율을 한 단계 더 최적화할 수 있습니다. 하지만 등고선 식별 알고리즘에 대한 개선이 필요합니다.
- 다양한 형태의 약품 식별: 유리병에 담긴 주사약제나 가소성 고분자 화합물에 담긴 정제와 캡슐은 6면으로 된 박스와 형태가 다르기 때문에 기존 방안에서 특성으로 식별하기에 어려움이 있습니다. 향후 해방군총병원은 다양한 형태의 약품의 특성을 더 세밀하게 식별할 예정입니다.
- 모델 트레이닝 효율 향상: 현재 방안의 식별 모델은 56 가지 약품의 분류 식별만 구현합니다. 향후 해방군총병원은 2,000가지에 달하는 약품을 정확히 분류할 예정이며, 그에 따라 모델의 복잡도 역시 증가할 것입니다. 따라서 트레이닝 효율을 최적화하고 트레이닝 시간을 단축할 필요가 있습니다.
- 모델 업데이트 메커니즘 구축: 약품의 종류와 사양이 변경되거나 새로운 약품이 입고되면 병원은 모델을 다시 트레이닝해야 합니다. 이로 인해 효율을 높일 수 있는 더 빠르고 새로운 메커니즘이 필요합니다.

요약

화상 인식은 의료 업계를 디지털화 및 스마트화로 전환하는 중요한 기술 중 하나입니다. 우수한 화상 인식 솔루션은 의료 기관별 정보화 시스템의 정보 순환을 대폭 향상하고 스마트화 수준을 높여 환자에게 더 훌륭한 진료 서비스를 제공할 수 있습니다. 딥러닝 기반 식별 모듈을 이용한 약품 지급 방식은 사람이 직접 작업하면서 발생하는 오류율을 낮출 수 있음을 이미 증명했으며 앞으로 충분히 응용될 가능성을 갖고 있습니다. 앞으로 이와 유사한 AI 기술은 약국의 약품 지급 등 시나리오에 응용될 뿐만 아니라, 수술 관련 소모품 지급 관리나 환자의 의료 정보 관리 등 다른 의료 시나리오에서도 큰 역할을 할 수 있습니다.

인텔® 제온® 확장성 CPU와 인텔® MKL-DNN 등이 포함된 인텔의 첨단 제품과 기술은 위에서 설명한 응용 시나리오에서도 다양한 딥러닝 모델이 잠재력을 효과적으로 발휘할 수 있음을 증명했습니다. 인텔® MKL-DNN으로 최적화한 TensorFlow 아키텍처는 LeNet-5 CNN의 효율 향상 뿐만 아니라 트레이닝과 추론 효율 향상도 도와주기 때문에 AI 응용 처리 효율이 대폭 향상됩니다.

해방군총병원에서 채택한 방안에서는 인텔® 제온® 확장성 CPU와 DRAM 메모리, 그리고 전통 방식의 인텔® NAND SSD의 하드웨어 플랫폼을 사용했습니다. 앞으로 이용자에게는 더 넓은 폭의 선택권이 주어질 것이고, AI 업계에는 더 다양한 조치가 취해진 2세대 인텔® 제온® 확장성 CPU와 인텔® 옵테인™ 데이터 센터급 영구 메모리가 제공하는 대용량 메모리 방안이 주어질 것입니다. 그리고 이를 고성능 인텔® 옵테인™ SSD와 결합해 더 탁월한 성능을 자닌 솔루션을 구축할 것입니다.

인텔 역시 여러 유형의 AI 식별 기술을 의료 업계에 지속적으로 응용하고 스마트화 의료 체계를 구축하기 위해 실력을 선보일 것입니다.



기술편

2 세대 인텔® 제온® 확장성 CPU



2세대 인텔® 제온® 확장성 CPU는 데이터 센터 현대화를 위해 전문 설계되었습니다. 이전 제품보다 성능¹이 25%-35% 높고, 다양한 옵션과 새로운 특성을 갖추었습니다. 유연성과 안전성을 향상시키고 메모리 성능을 강화해 사용자가 인프라, 기업 앱과 기술 컴퓨팅 운용 효율을 향상시킬 수 있도록 도와 성능이 더 강력하고 서비스가 더 민첩해 가치가 높은 기능을 형성합니다. 더 나아가 총소유비용(Total Cost of Ownership, TCO)을 개선해 생산력을 높입니다.

인텔® 제온® 골드 CPU 6200 시리즈, 특히 주류의 인텔® 제온® 골드 6248 CPU, 인텔® 제온® 골드 6240 CPU, 인텔® 제온® 골드 6230 CPU는 인텔® 제온® 확장성 CPU 플랫폼의 기동으로서 더 빠른 메모리와 더 강한 메모리 용량과 포웨이 확장 가능성을 지원합니다. 성능과 고급 신뢰성, 하드웨어 증강형 보안 기술 측면에서도 눈에 띄게 개선되었으며, 조건 이 까다로운 주류 데이터 센터, 클라우드, 네트워크와 스토리지 등 워크로드에 최적화를 진행했기 때문에 더 복잡하고 다양화된 응용 시나리오에 적응할 수 있게 되었습니다. 또한 인텔® 제온® 골드 6200 시리즈는 듀얼 FMA 채널을 최초로 지원하며, 이는 FMA 성능이 2배로 향상되었음을 의미합니다.

2세대 인텔® 제온® 확장성 CPU는 딥러닝 가속 기술(VNNI)을 집성하고 인텔® AVX-512를 확장해 플랫폼에 더 다양하고 더 강한 AI 기능을 부여함으로써 인공지능과 딥러닝 추론을 가속화하고 워크로드를 최적화할 수 있었고, 그 결과

AI 가속 능력을 보유한 CPU 아키텍처가 되었습니다. 이 아키텍처를 기반으로 대다수의 주론 업무가 워크로드 또는 애플리케이션에 집성되어 있기 때문에 사용자는 가속화된 성능과 높아진 유연성 등의 장점을 누림으로써 데이터가 중심인 현대 사회에서 클라우드와 AI 엣지 사이에서 성능 전환과 AI 개발, 응용을 방해 받지 않고 고효율적으로 진행할 수 있습니다.

차세대 제온 확장성 플랫폼의 "핵심"으로서 2세대 인텔® 제온® 확장성 CPU는 인텔® 옵테인™ 데이터 센터급 영구 메모리를 새로운 제품을 지원합니다. 하지만 인텔® 옵테인™ 데이터 센터급 영구 메모리는 2세대 인텔® 제온® 골드 및 실버 CPU와 결합해 DRAM 메모리의 강력한 보완 조치로 시스템 성능, 가속 워크로드 처리와 서비스 전달(구체적인 내용은 《인텔® 옵테인™ 데이터 센터급 영구 메모리》 부분 참조)을 뚜렷하게 향상시킬 수 있습니다.

기능 특성:

- 더 높아진 커널별 성능: 많게는 56커널(9200 시리즈)과 28커널(8200 시리즈)에 다르며, 컴퓨팅과 스토리지, 네트워크 응용에서 컴퓨팅 밀집형 워크로드에 더 높은 성능과 확장성을 제공합니다.
- VNNI 기반 인텔® 딥러닝 가속(인텔® DL Boost) 기술: CPU에서 인공지능 추론의 표현을 향상시켰습니다. 이전 세대 제품과 비교하면 성능이 30배까지 향상되면서 데이터 센터에서 엣지까지

AI를 배치 및 응용하는데 도움이 됩니다.

- 업계를 선도하는 메모리와 스토리지, 더 커진 메모리 대역폭 / 용량: 인텔® 옵테인™ 데이터 센터급 영구 메모리를 지원하며, 전통적인 DRAM과 결합해 사용하면 최고 36TB 시스템급 메모리 용량까지 지원할 수 있습니다. 메모리 대역폭과 용량이 50%² 향상되어, 각각 메모리 채널 6개와 최대 4TB DDR4 메모리, 최고 2933 MT/s(1 DPC) 속도를 지원합니다. 그 밖에도 인텔® 옵테인™ 데이터 센터급 SSD와 인텔® QLC 3D NAND SSD를 지원해 데이터 밀집형 워크로드와 획기적인 메모리와 스토리지 메모리 혁신 효율과 성능을 두드러지게 향상시켰습니다.
- 인텔® Infrastructure Management 기술(인텔® IMT): 이 자원 관리 아키텍처는 인텔의 여러 능력을 결합해 플랫폼급 탐지와 보고, 구성을 효과적으로 지원할 수 있습니다.
- 데이터 센터 지향의 인텔® Security Libraries(인텔® SecL-DC): 이 소프트웨어 라이브러리와 모듈은 인텔 하드웨어를 기반으로 한 보안 기능입니다.

제온® 플랫폼의 혁신작으로서, 2세대 인텔® 제온® 확장성 CPU는 획기적인 설계를 기반으로 플랫폼에서 컴퓨팅, 메모리, 스토리지, 네트워크 및 보안 등 기능을 융합하면서 새로운 수준으로 올라설 수 있었습니다.

¹ <https://www.intel.cn/content/www/cn/zh/technology-provider/products-and-solutions/xeon-scalable-family/2gen-data-centric-computing-article.html>

² <https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html>

스마트앱에 적용되는 2 세대 인텔® 제온® 확장성 CPU

* 특정 CPU에서만 지원을 받습니다.

	인텔® 제온® 골드 CPU (6200 시리즈)	인텔® 제온® 골드 CPU (6200 시리즈)			인텔® 제온® 골드 CPU (8200 시리즈)	인텔® 제온® 골드 CPU(9200 시리즈)
		인텔® 제온®골드 6230 CPU	인텔® 제온®골드 6240 CPU	인텔® 제온®골드 6248 CPU		
유비쿼터스의 성능과 안전성						
지원하는 최대 커널 수	24	20	18	20	28	56
지원하는 최고 주파수	4.4 GHz	3.90 GHz	3.90 GHz		4.0 GHz	3.8 GHz
지원하는 CPU 수	최대 4개				최대 8개	최대 2개
인텔®UPI(Ultra Path Interconnect)	3	3	3	3	3	4
인텔®UPI Speed	10.4 GT/s				10.4 GT/s	10.4 GT/s
인텔®고급 벡터 확장 512(인텔® AVX-512)	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA
지원한 최고 메모리 속도(DDR4)	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s
지원하는 최고 메모리 용량*	1 TB ~ 2 TB ~ 4.5 TB	1 TB	1 TB		1 TB ~ 2 TB ~ 4.5 TB	3.0 TB
16 Gb DDR4 DIMM 지원	●	●	●	●	●	●
VNNI를 채택한 인텔®딥러닝 가속(인텔 Boost)	●	●	●	●	●	●
인텔®옵테인™ 데이터 센터급 영구 메모리 지원*	●	●	●	●	●	
인텔®Omni-Path 아키텍처(독립식 PCIe*)	●	●	●	●	●	●
인텔®QuickAssist 기술(칩셋에 집성)	●	●	●	●	●	
인텔®QuickAssist 기술(독립식 PCIe* 카드)	●	●	●	●	●	●
인텔®옵테인™ 데이터 센터급 SSD	●	●	●	●	●	●
인텔®SSD 데이터 센터 제품군(3D NAND)	●	●	●	●	●	●
PCIe 3.0	●	●	●	●	●	●
인텔®QuickData 기술(CBDMA)	●	●	●	●	●	●
NTB	●	●	●	●	●	●
인텔®터보 가속 2.0	●	●	●	●	●	●
인텔®하이퍼스레딩 기술(인텔® HT 기술)	●	●	●	●	●	●
노드 제어 장치 지원	●	●	●	●	●	●

* 특정 CPU에서만 지원을 받습니다.

2세대 인텔® 제온®확장성 CPU 정보는 다음을 참조하십시오.

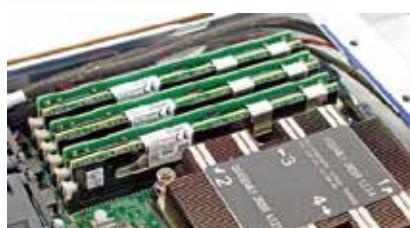
<https://www.intel.cn/content/www/cn/zh/products/processors/xeon/scalable.html>

인텔® 옵테인™ 데이터 센터급 영구 메모리



인텔® 옵테인™ 데이터 센터급 영구 메모리는 인텔의 혁명적인 상품입니다. 새로운 스토리지 레이어를 만들어 메모리와 스토리지 간의 성능 차이를 메워 전통적인 메모리-스토리지 아키텍처를 뒤엎고, 합리적인 가격으로 영구적인 대형 메모리 레이어를 제공하기 때문에 메모리 밀집형 워크로드, 가상 머신 밀도와 라피드 스토리지에 더 뛰어난 성능과 효율, 경제성을 가져다 줄 수 있습니다. 더 나아가 사용자가 이전보다 빨라진 분석과 클라우드 서비스, 인공지능 트레이닝과 추론 및 차세대 통신 서비스를 통해 IT 변화를 가속화함으로써 데이터 시대의 수요를 만족시킬 수 있습니다.

인텔® 옵테인™ 데이터 센터급 영구 메모리는 비용과 비휘발성, 성능 등 특성을 고루 고려하기 때문에 단일 모듈은 128/256/512 GiB 세 가지 중 선택할 수 있으며, DDR4 슬롯과는 호환할 수 있습니다. 전통적인 DDR4 DRAM 메모리와 함께 2세대 인텔® 제온® 확장성 CPU 기반 플랫폼에 응용하면, 적은 비용으로 Eight way 시스템에서 최고 24TiB 용량(way당 최고 구성 가능 용량이 3TiB인 옵테인 데이터 센터급 영구 메모리)을 실현할 수 있습니다. 더 나아가 사용자가 CPU의 메모리 시스템에서 규모가 이전을 훨씬 초과하는 데이터 세트를 로딩하도록 와주기 때문에 메모리 데이터 세트 분석, AI 추론과 반복 등 대형 메모리에 조건이 까다로운 앱 로딩 수요를 만족시킴으로써 더 많은 데이터 처리와 분석을 실시간 처리할 수 있게 됩니다.



메모리와 스토리지 특성을 기반으로 융합한 옵테인™ 데이터 센터급 영구 메모리는: 메모리 모드와 App Direct 모드, 두 가지 업무 모드가 있습니다. 서로 다른 이 두 가지 업무 모드를 통해, 고객은 여러 개의 워크로드를 유연하게 넘나들 수 있기 때문에 시스템 성능이 현저히 향상됩니다.

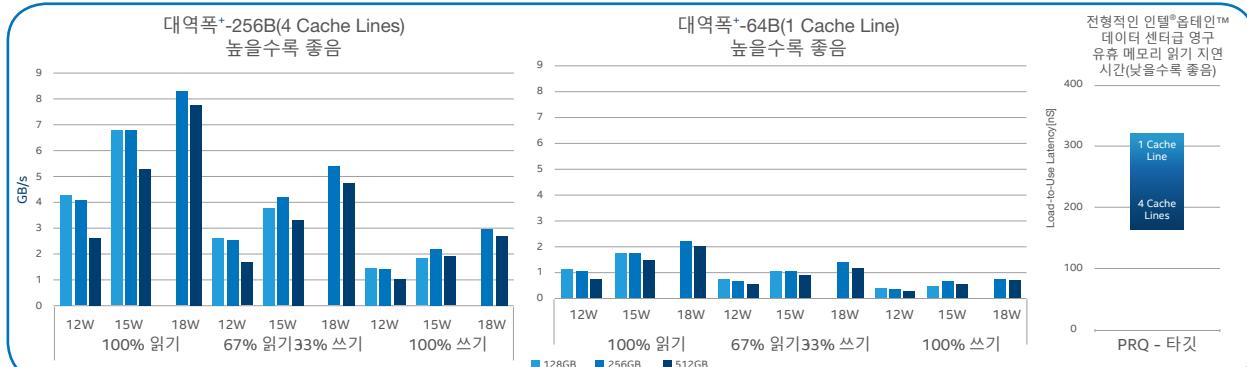
메모리 모드는, 대용량 메모리에 매우 적합하며, 애플리케이션을 변경할 필요가 없습니다. 메모리 모드에서는 CPU 메모리 제어 장치가 인텔® 옵테인™ 데이터 센터급 영구 메모리를 주소지정 방식의 주기억장치로 삼아 운영 체제가 지원하는 사용 가능 휘발성 메모 용량을 확장하고 DRAM을 캐시로 삼습니다. 이는 가상화와 컨테이너 기술이 직접적인 이익을 얻을 수 있게 해주며, 저렴한 비용으로 단일 물리적 서버에서 가상 컴퓨터나 컨테이너의 밀도를 향상시키거나 모든 가상 컴퓨터와 컨테이너에 더 큰 용량의 메모리를 제공할 수 있기 때문에 소프트웨어를 새로 프로그래밍할 필요가 없습니다.

App Direct 모드에서는 운영 체제가 DRAM 메모리와 인텔® 옵테인™ 데이터 센터급 영구 메모리를 두 개의 독립적인 메모리

풀로 간주하기 때문에 인텔® 옵테인™ 데이터 센터급 영구 메모리처럼 주소를 지정하고 스토리지처럼 영구적인 데이터를 보유할 수 있도록 해줍니다. 이러한 데이터의 영구적 특성은 시스템이 유지 또는 재부팅 기간일 때에도 영구 메모리가 데이터를 유지함으로써 시스템의 업무 탄력성을 증가시키고, 재부팅 시간을 단축하며, 그로 인한 비용을 절감할 수 있습니다.

이중 모드는 App Direct의 부분 집합으로 프로비저닝을 통해 인텔® 옵테인™ 데이터 센터급 영구 메모리 부분을 메모리 모드에, 나머지 부분은 App Direct 모드에 놓이게 할 뿐만 아니라 이중 조건을 갖춘 워크로드 또는 응용 시나리오의 수요를 만족시킵니다.

인텔® 옵테인™ 데이터 센터급 영구 메모리는 다른 인텔® 제온® 확장성 CPU 플랫폼 상품과 함께 시장에 출시되었으며, 2세대 인텔® 제온® 확장성 CPU를 최적화했습니다. 메모리만의 독특한 장점 덕분에 출시 초기부터 관심을 끌었고, 수많은 ISV 파트너들이 관련 소프트웨어를 최적화하게 만들었을 뿐만 아니라. 클라우드 인프라 및 데이터 분석 사용자를 보유하고, 인텔® 옵테인™ 데이터 센터급 영구 메모리를 도입하게 만들었습니다. 협력 파트너와 사용자는 앱 투명화와 앱 실전 과정에서 인텔® 옵테인™ 데이터 센터급 영구 메모리의 응용 가치를 일찍이 증명했습니다.



제품군		인텔® 옵테인™ 데이터 센터급 영구 메모리										
외형 사양		영구 메모리 모듈 (PMM)										
DCPMM SKU ¹	128 GiB	256 GiB			512 GiB							
사용자 용량	126.4 GiB ⁸	252.4 GiB ⁸			502.5 GiB ⁸							
MOQ	4	50	4	50	4	50						
MM#	999AVV	999AVW	999AVX	999AVZ	999AW1	999AW2						
상품 코드	NMA1XXD128GPSU4	NMA1XXD128GPSUF	NMA1XXD256GPSU4	NMA1XXD256GPSUF	NMA1XXD512GPSU4	NMA1XXD512GPSUF						
모델 문자열	NMA1XXD128GPS		NMA1XXD256GPS		NMA1XXD512GPS							
기술	인텔® 옵테인™ 기술											
수리 기간	5 년											
연평균 고장률 (AFR)	≤ 0.44											
내구성 100% 쓰기 15W 256B	292 PBW	363 PBW			300 PBW							
내구성 100% 쓰기 15W 256B	91 PBW	91 PBW			75 PBW							
내구성 100% 쓰기 15W 64B	6.8 GB/s	6.8 GB/s			5.3 GB/s							
내구성 100% 쓰기 15W 256B	1.85 GB/s	2.3 GB/s			1.89 GB/s							
내구성 100% 쓰기 15W 64B	1.7 GB/s	1.75 GB/s			1.4 GB/s							
내구성 100% 쓰기 15W 64B	0.45 GB/s	0.58 GB/s			0.47 GB/s							
DDR 주파수	2666、2400、2133、1866 MT/s											
최대 열 설계 전력 (TDP)	15W	18W										
온도 (TJMAX)	≤ 84°C(85°C OFF, 83°C 기본) 중간 온도											
온도 (주변 온도)	10W: 54° C @ 2.4m/s											
온도 (주변 온도)	12W: 49° C @ 2.4m/s											
온도 (주변 온도)	15W: 44° C @ 2.7m/s											
온도 (주변 온도)	N/A	18W: 40° C @ 3.7m/s										

주석 : ¹GiB = 2^{30} ; GB = 10^9

더 많은 정보는 다음을 참조하십시오.

<https://www.intel.cn/content/www/cn/zh/products/memory-storage/optane-dc-persistent-memory.html>

Intel® Optane™ 固態硬碟與基於 Intel® QLC 3D NAND 技術的 Intel® 固態硬碟



인텔® 옵테인™ SSD 와 인텔® QLC 3D N A N D 기술을 채택한 인텔® SSD 는 혁신적인 스토리지 아키텍처로 데이터 센터가 미래로 향하고 변혁과 혁신을 가속화하도록 돕습니다.

S S D 상품 중 고급 라인에 속하는 인텔® 옵테인™ SSD 는 혁신적인 3D XPoint™ 스토리지 매개체를 채택하고 일련의 첨단 시스템 메모리 제어 장치와 인터페이스 하드웨어, 소프트웨어 기술을 결합해 낮은 지연과 높은 안전성 부분에서 만족스러운 결과를 보임으로써 데이터 센터 스토리지의 병목 현상을 제거할 수 있었습니다. 더 크고, 경제적이며 실용적인 데이터 세트를 사용하기 때문에 애플리케이션 속도를 가속화하고, 랙 민감형 워크로드 사무처리 비용을 절감하며 데이터 센터의 TCO를 개선할 수 있었습니다. 인텔® 옵테인™ SSD는 더 전면적이고 우수하며 균형적인 I T 인프라 기능을 지닌 덕분에 데이터 밀집형 AI 모델 트레이닝과 주론 효율 역시 더 높아졌습니다. 인텔® 옵테인™ SSD DC P4800X를 예로 들면, 최고 55만 IOPS의 랜덤 읽기 쓰기 기능을 갖추었으며, 읽기쓰기 랙 시간이 10마이크로초 밖에 되지 않기 때문에 사용자가 많고 여리 번 교부해야 하는 시나리오에서 더 뛰어난 성능을 보일 수 있게 도와줍니다. 또

인텔® SSD는 획기적인 의미를 지니고, 신뢰할 수 있는 3D NAND 기술을 채택해 스토리지 경제성을 향상시키고, 더 나아가 전통적인 하드 디스크 (HDD) 를 대체하기 때문에 높은 가성비를 선택할 수 있게 합니다. 그 덕분에 사용자는 경험을 개선하고 응용과 서비스 성능을 향상시키며 TCO 를 절감할 수 있습니다. SSD 의 신예 부대라 할 수 있는 인텔® SSD D5-P4320 시리즈는 인텔의 첨단 64 레이어 3D NAND 기술을 채택해 QLC SSD 단일 디스크 용량이 7.68TB (TeraByte) 에 달하기 때문에 데이터 센터 등 인프라 사용자의 "대용량" 스토리지에 대한 수요를 만족시킬 수 있습니다. 게다가 랜덤으로 읽은 IOPS가 42.7만에 달하며, 2 세대 인텔® 제온® 확장성 CPU와 매칭을 통해 AI 트레이닝 등 응용 시나리오에 적용되어 "한 번 기록 가능 여러 번 판독 가능" 성능에 대한 수요와 복잡하고 다양화된 워크로드에 고효율, 고안전성을 지닌 저에너지 스토리지 아키텍처를 제공했습니다.

더 많은 정보는 다음을 참조하십시오.

- <https://www.intel.cn/content/www/cn/zh/products/memory-storage/optanememory/optane-memory-h10-solidstate-storage.html>
- <https://www.intel.cn/content/www/cn/zh/products/memory-storage/solidstate-drives/data-center-ssds.html>

DNN 지향 인텔® MKL

DNN 지향 인텔® MKL (Intel® Math Kernel Library, 인텔® MKL-DNN)은 딥러닝 응용 기반 오픈소스 인핸스먼트 라이브러리이자, 인텔이 개발자가 인텔® 아키텍처를 충분히 이용해 딥러닝 연구와 응용을 추진할 수 있도록 만든 기본 라이브러리입니다. ([소스코드 주소: https://github.com/intel/mkl-dnn](https://github.com/intel/mkl-dnn))

인텔® MKL-DNN은 인텔® 아키텍처에서 딥러닝 아키텍처의 운용 속도를 가속화하기 위해 전문 설계된 하나의 성능 인핸스먼트 라이브러리로서, 고도의 벡터화와 스레드화의 구성 모듈을 포함하고 있으며 C와 C++ 인터페이스 이용 DNN을 지원하고 광범위한 딥러닝 연구와 개발, 응용 생태계를 보유하고 있습니다. Caffe, TensorFlow, PyTorch Apache, Mxnet, BigDL, CNTK, OpenVINO™ 툴킷 등 여러 딥러닝 소프트웨어 상품에 적용됩니다.



인텔® 아키텍처에서 딥러닝 모델이 운용되는 속도를 효과적으로 향상시키고, 유형별 뉴럴 네트워크 중 기타 성능 민감형 애플리케이션의 효율을 향상시키기 위해, 인텔® MKL-DNN은 여러 최적화된 딥러닝 요소를 제공해 범용 구성 모듈이 높은 효율로 실행되도록 다른 딥러닝 아키텍처에서도 응용할 수 있게 만들었습니다. 이 모듈은 행렬 곱셈과 콘볼루션 외에도 다음을 포함합니다.

- 다이렉트 배치 콘볼루션
- 내적
- 풀링: 최대, 최소, 평균
- 표준화: 크로스 채널 부분 응답 정규화(LRN), 대량 정규화
- 활성화: 선형 유닛 수정 (ReLU)
- 데이터 작업: 다차원 전위(전환), 분해, 통합, 합산과 크기 조절.

이 고효율의 함수 모듈은 딥러닝 모델에 광범위하게 응용될 수 있습니다.

응용 유형	위상학적 구조
화상 인식	AlexNet, VGG, GoogleNet, ResNet, MobileNet
영상 분할	FCN, SegNet, MaskRCNN, U-Net
부피 분할	3D-Unet
객체 탐지	SSD, Faster R-CNN, Yolo
기계 번역	GNMT
음성 문자 인식	DeepSpeech, WaveNet
적대 신경망	DCGAN, 3DGAN
강화 학습	A3C

CPU에서 딥러닝의 성능을 대폭 향상시키기 위해 인텔은 여러 오픈소스 공동체와 협력해 인텔® MKL-DNN을 다양한 딥러닝 아키텍처에 집성했습니다. 일찍이 2016년에는 인텔® MKL-DNN 최적화를 거친 Caffe가 인텔® 제온® CPU E5-2697 v3를 이용하면서 기존의 Caffe보다 그 성능이 10이나 높아졌습니다.¹ 최적화를 거친 ResNet-50 역시 인텔® 제온® 실버 9282 CPU에서 초당 7736 장 그림을 선보이는 성능²을 실현했습니다.

인텔® MKL-DNN은 현재 여러 딥러닝 아키텍처를 CPU에서 운용할 때 사용하는 기본 구성이 되어 있기 때문에, 개발자는 딥러닝 아키텍처 설치와 응용에서 인텔® MKL-DNN의 업그레이드된 성능을 바로 활용할 수 있습니다.

더 많은 정보는 다음을 참조하십시오.

- <https://software.intel.com/zh-cn/articles/intel-mkl-dnn-part-1-library-overview-and-installation>
- <https://software.intel.com/zh-cn/articles/introducing-dnnprimitives-in-intelr-mkl>

¹ <https://software.intel.com/es-es/node/604830?language=en>

² <https://www.intel.com/content/dam/www/public/us/en/images/diagrams/rwd/xeon-scalable-max-inference-rwd.png>

適用於 Intel® 架構最佳化的 Caffe

인텔® 아키텍처 최적화 지향 Caffe 는 당시 최신 버전인 인텔® MKL을 집성한 최신 버전에서 당시 제온® CPU 제품이 이미 집성한 인텔® AVX 2 와 인텔® AVX-512 를 전문적으로 최적화한 것입니다. BVLC Caffe와 완전하게 호환될 뿐만 아니라 BVLC Caffe 의 모든 장점까지 갖추고 있습니다. CPU 최적화 기능도 보유하고 있고, 인텔® 아키텍처 CPU 에서 아주 뛰어난 성능을 보였을 뿐만 아니라 여러 노드 분포 프로세스 트레이닝도 지원합니다.

인텔® 아키텍처 최적화 지향 Caffe는 완벽한 Post-training 양자화 방안을 지원하며, 여러 CNN 모델에서 성과를 보았습니다. 특히 2 세대 인텔® 제온® 확장성 CPU 기반 플랫폼 (CPU 소개 부분 참조) 에서 INT8에 대해 지원한 인텔® 딥러닝 가속 기술 (VNNI 명령어 조합)을 최적화해 예측 정확도에 영향을 미치지 않는 상황에서 여러 딥러닝 모델이 INT8을 사용할 때 주론 속도가 FP32를 사용할 때 속보다 2 ~ 4 배 (아래 그림 참조)¹ 가 되도록 했습니다. 덕분에 사용자는 딥러닝 응용 업무 효율을 대대적으로 향상시켰습니다.

더 많은 정보는 다음을 참조하십시오.

- <https://software.intel.com/en-us/articles/caffe-optimized-for-intelarchitecture-applying-modern-codetechniques>
- <https://software.intel.com/zh-cn/videos/what-is-intel-optimization-for-caffe>

최적화된 딥러닝 아키텍처와 툴 패키지

인텔® 딥러닝 가속 기술을 채택한 ResNet-50 개인

2 세대 인텔® 제온® 플래티넘 8280 CPU vs 인텔® 제온® 플래티넘 8180 CPU

인텔® 제온® 확장성 CPU	2세대 인텔® 제온® 확장성 CPU	mxnet	PYTORCH	TensorFlow	Caffe	OpenVINO®
FP32	INT8 W/ Intel® Dl Boost	3.0x	3.7x	3.9x	4.0x	3.9x
INT8	INT8 W/ Intel® Dl Boost	1.8x	2.1x	1.8x	2.3x	1.9x

¹ 구성 정보는 다음을 참조하십시오. <https://www.intel.cn/content/www/cn/zh/benchmarks/server/xeon-scalable/xeon-scalable-artificial-intelligence.html>

인텔® 아키텍처 최적화 지향 TensorFlow

인텔® 아키텍처 최적화 지향 TensorFlow는 인텔이 CPU에서 딥러닝 모델을 운용할 때 맞닥뜨리는 성능 측면의 도전에 대응하기 위해 출시한 최적화 버전으로 딥러닝 유형의 워크로드가 여러 상황에서 인텔® MKL-DNN 기본 알고리즘 유닛을 이용해 효율적으로 운용할 수 있게 했습니다.

성능을 현저히 향상시키기 위해, 인텔은 다른 조치를 취해 TensorFlow를 최적화했습니다.

산출 그래프 최적화

인텔은 CPU 운용 시, 기본 TensorFlow 작업을 인텔 최적화 버전으로 손쉽게 교체해 사용자가 뉴럴 네트워크를 변경하지 않고 기존의 Python 프로그래밍을 사용한 상황에서 성능을 업그레이드할 수 있도록 다양한 산출 그래프 최적화 채널을 출시했습니다. 또한 불필요한 고가의 데이터 구성 전환을 제거하고, 여러 개의 연산을 통합해 CPU에서 고속 캐시를 고효율적으로 중복 사용하고, 빠르게 뒤로 전파되는 중간 상태를 처리합니다.

산출 그래프를 최적화 조치하면 성능이 한 단계 더 향상되기 때문에 TensorFlow 프로 그래밍을 따로 신경 쓸 필요가 없습니다. 그 밖에 데이터 구성 최적화는 성능을 최적화하는 중요한 조치 중 하나입니다. 이전에는 TensorFlow에서 나온 로컬 형식의 데이터 구성은 내부 형식으로 전환해 삽입해 CPU에서 연산을 실행하고, 연산 출력을 TensorFlow로 전환하면 그로 인해 성능 오버헤드 문제가 발생했습니다. 하지만 지금은 인텔® MKL을 이용해 연산 실행한 서브그래프를 최적화하기 때문에 서브그래프 연산 중 전환 문제를 제거할 수 있습니다. 자동 삽입된 전환 노드는 서브그래프 경계에서 데이터 구성은 전환하고, 채널 융합과 같은 또 다른 중요한 최적화를 통해 여러 개의 연산을 고효율적으로 운영되는 단일 인텔® MKL 연산으로 자동 융합합니다.

기타 최적화

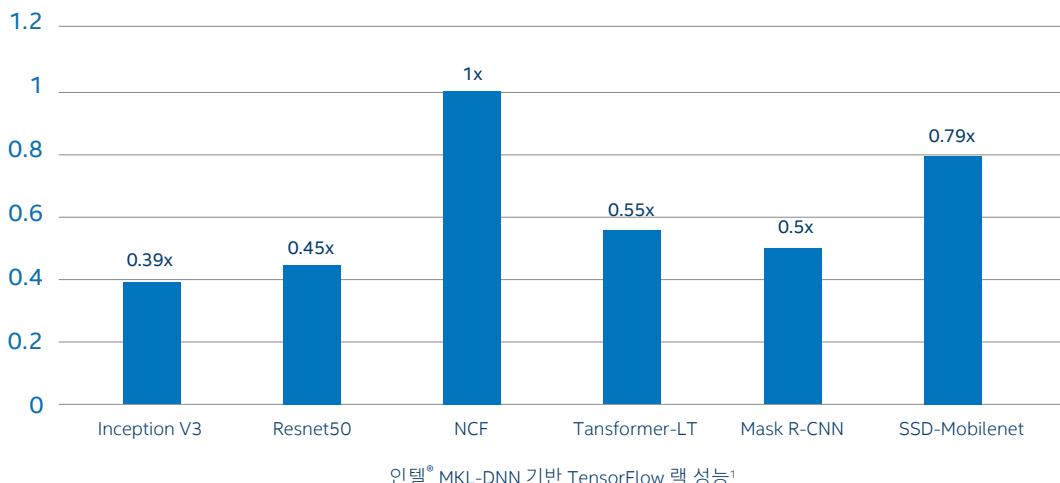
다양한 딥러닝 모델에서 최고의 CPU 성능을 실현할 수 있도록, 인텔은 수많은 TensorFlow를 겨냥해 아키텍처를 조정했습니다. 예를 들어,

TensorFlow에 원래 있는 메모리 풀 스플리터를 사용해 사용자 지정 메모리 풀 스플리터를 개발하고 인텔® MKL과 동일한 메모리 풀(인텔® MKL imalloc 기능 사용)을 공유함으로써 메모리를 너무 일찍 운영 체제로 반환할 필요가 없기 때문에 고비용이 드는 페이지 결실과 페이지 제거 현상을 막을 수 있게 되었습니다. 뿐만 아니라 여러 개의 스레드 라이브러리(TensorFlow가 사용한 pthread와 인텔® MKL이 사용한 OpenMP) 페이지에서 딥 최적화를 진행해 공존할 수 있게 함으로써 서로 CPU 리소스를 생활하는 현상을 방지해 리소스 종합 이용률을 향상시켰습니다.

더 많은 정보는 다음을 참조하십시오.

- https://www.intel.ai/tensorflow/?_ga=2.231295069.330745958.1563951842597697079.1551333838&elq_cid=4287274&erpm_id=7282583
- <https://www.intel.ai/improvingtensorflow-inference-performance-on-intel-xeon-processors/#gs.v0kayg>

지연 시간의 결과(인텔® MKL-DNN/Eigen 라이브러리) — 낮을수록 좋습니다



¹ 시스템 구성: 인텔® 제온® 플래티넘 8180 CPU @2.50GHz, OS: CentOS Linux 7(Core), TensorFlow 소스 코드 : <https://github.com/tensorflow/tensorflow>, TensorFlow 버전 번호 : 355cc566efd2d86fe71fa9d755ceabe546d577a

인텔® 아키텍처 최적화 지향 Python 배포 패키지

인텔® 아키텍처 최적화 지향 Python 배포 패키지는 인텔이 주도적으로 개발하고 강력한 기능을 보유한 소프트웨어 개발 툴킷으로 C 와 Fortran 컴파일러, MKL과 애널라이저 같이 프로그래밍 Python 최초 확장에 필요한 모든 것을 제공합니다. 뿐만 아니라 NumPy, SciPy, Scikit-learn, Pandas, Jupyter, matplotlib, mpi4py 같은 여러 종류의 고성능 데이터 분석과 MKL까지 집성했습니다.

인텔® Python 배포 패키지는 인텔® Parallel Studio XE 의 중요한 툴킷 중 하나로 효율성이 높고 다양한 특성을 지니고 있습니다.

- 숫자와 과학, 데이터 분석, 머신러닝 등이 포함된 계산집약형 응용을 가속화할 수 있게 개봉해 바로 사용하는 uMath, NumPy, SciPy와 Scikit-learn 등 툴을 제공했습니다.

- 인텔® AVX-512 와 다중 스레드 명령, Numba 와 Cython 같은 인텔® 성능 라이브러리(예: 인텔® MKL)를 집성하고, 최신 벡터화 명령을 내장하고, 다중 스레드에 구축한 모듈 라이브러리 인텔® TBB의 커포저블 병렬을 응용해

Python 의 다핵성 CPU 기반 병렬 응용 기능을 해제하고 더 나아가 인텔® 아키텍처에 기반한 플랫폼에서 Python 프로그래밍 운용 성능을 향상 시킴으로써 시스템 호환성을 보장하기 때문에 코드를 변경할 필요가 없습니다.

- Python2.7, Python3.6 과 최신 버전의 인텔® 아키텍처 CPU 를 지원하고, 벡터 컴퓨터 (SVM) 와 K-means 예측, 랜덤 포레스트와 XGBoost 알고리즘 등, 최적화된 TensorFlow, Caffe 등 딥러닝 라이브러리와 머신러닝 라이브러리를 제공하기 때문에 과학 기술 계산과 머신러닝 등 워크로드 구축과 양산 준비 알고리즘 확장이 편리합니다.

벤치마크 테스트를 거쳐 인텔® Python 배포 패키지와 기타 오픈 소스 Python 중 Scikit-learn 툴 패키지의 효율을 비교해보면 인텔® 아키텍처 최적화 지향 Python 지표가 눈에 띄게 향상(그림 참조). 효율이 높을수록 함수 속도가 빠르고 로컬 C언어 속도가 가까워짐)되게 나타납니다. 예를 들어, 인텔® Python 배포 패키지 K-means 클러스터, 선형 회귀 등 알고리즘의 효율은

인텔® 데이터 분석 가속 라이브러리 (Intel® Data Analytics Acceleration Library, 인텔® DAAL) 중 C 언어 효율의 90%를 달성할 수 있습니다.

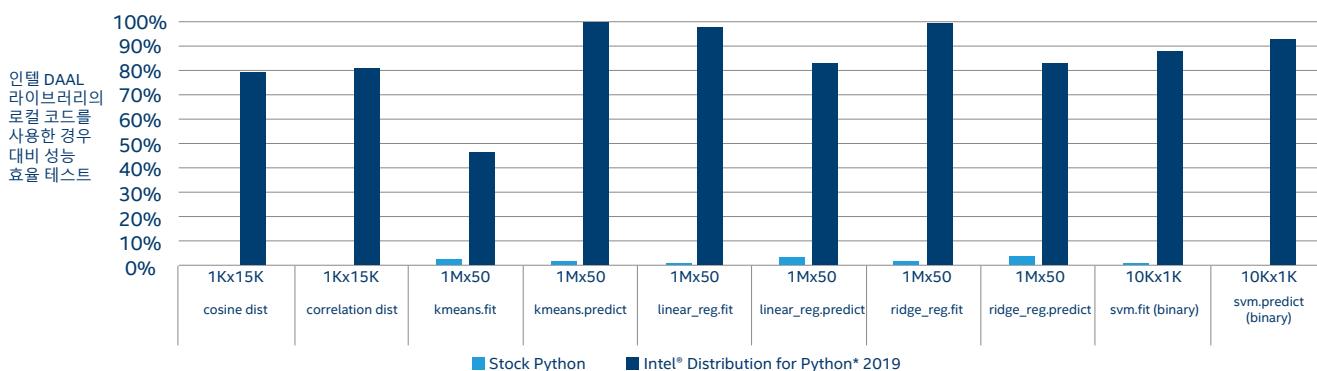
배치가 간단해 사용이 편리하다는 점 역시 인텔® Python 배포 패키지의 특성입니다. Conda 소프트웨어 패키지 관리자와 Anaconda 클라우드를 사용하면 사용자는 명령 하나만 실행하면 코어 Python 환경을 설치할 수 있습니다.

• conda install intelpython3 -c intel

인텔® Python 배포 패키지는 사전에 구축한 이진법 파일로 pip, Docker images, YUM과 APT repos 등 여러 채널을 통해 획득할 수 있습니다.

더 많은 정보는 다음을 참조하십시오.

<https://software.intel.com/en-us/distribution-for-python>



인텔 최적화는 scikit-learn 프로그래밍 패키지의 효율을 향상시켜 인텔® 제온®CPU에서 로컬 코드가 운용되는 속도와 가까워지게 했습니다

OpenVINO™ 툴킷

OpenVINO™ 툴킷은 인텔이 출시한 딥러닝 추론과 배치를 가속화하는 소프트웨어 툴킷으로 고성능 컴퓨터 비전 처리와 응용을 가속화하는데 쓰입니다. 이 툴은 이종 실행을 허용하고, Windows 와 Linux 시스템 및 Python/C++ 언어를 지원하기 때문에 컴퓨터 비전 기술이 스마트 카메라, cctv, 로봇, 지능형 교통, 스마트 의료 등 영역에서 응용될 수 있도록 추진할 수 있습니다.

이 툴킷은 컴퓨터 비전 솔루션의 성능을 향상시키고, 개발 시간을 단축하며, 인텔이 제공한 다양한 하드웨어 옵션에서 효익을 얻는 경로를 간소화했습니다. 이 옵션은 성능을 향상시키고, 전력을 낮추어 하드웨어 이용률을 최대화함으로써 사용자가 낮은 리소스로 높은 수익을 획득할 수 있게 하며, 새로운 제품을 개인 맞춤형으로 디자인할 수 있게 합니다.

딥 콘볼루션 뉴럴 네트워크 (CNN) 를 기반으로 인텔® 하드웨어(가속기 포함)의 워크로드를 확장함으로써 OpenVINO™ 툴킷이 인텔® 아키텍처 CPU가 집성한 그래픽카드 (Integrated GPU), FPGA, 인텔® Movidius™ VPU 등 칩에 따라 비주얼 시스템의 기능과 성능을 강화할 수 있게 합니다. 최신 발표한 OpenVINO™ 버전은 2 세대 인텔® 제온® 확장성 CPU를 지원할 수 있고, 인텔® AVX-512 및 VNNI의 인텔® 딥러닝 가속(인텔® DL Boost) 기술을 채택해 추론 성능을 향상시킴으로써, 고객이 소프트웨어를 변경하지 않는 상황에서 하드웨어 제품 업그레이드와 알고리즘 이식을 빠르게 완료하고 경계에서 고성능 컴퓨터 비전과 딥러닝 응용 개발을 빠르게 실현할 수 있게 해줍니다.

- 인텔® 아키텍처 플랫폼에서 컴퓨터 비전 관련 딥러닝 성능이 19 배 이상 향상됩니다.
- CNN-based 의 네트워크를 종단 장치의 성능 장애에 릴리즈합니다.
- OpenCV, OpenVX 비주얼 라이브러리의 전통적인 API 에 가속화와 최적화를 실현합니다.

- 범용 API 기반 인터페이스를 CPU, GPU, FPGA 등 설비에서 운용합니다.
- 인텔® 플랫폼 기반 최적화 OpenVINO™ 툴킷은 주로 딥러닝 배치 툴 패키지와 전통적인 컴퓨터 비전 툴 패키지 두 부분을 포함합니다. 딥러닝 배치 툴 패키지는 모델 옵티マイ저 (Model Optimizer) 와 추론 엔진 (Inference Engine) 두 개의 핵심 부분을 포함합니다.
- 모델 옵티マイ저(Model Optimizer): 지정 모델을 기본 중간 표현(Intermediate Representation, IR)으로 전환하고, 모델을 최적화하며 ONNX, TensorFlow, Caffe, MXNet, Kaldi* 등 딥러닝 아키텍처를 지원합니다.
- 추론 엔진 (Inference Engine) : 하드웨어 명령어 조합 측면의 딥러닝 모델 가속 운용을 지원하며, 다음의 하드웨어를 지원합니다. CPU, GPU, FPGA, VPU.



크로스 플랫폼 툴을 제공하고, 컴퓨터 비전과 딥러닝 추론 가속을 지원합니다

또한 전통적인 OpenCV 이미지 처리 라이브러리에도 명령어 조합 최적화를 진행해 성능과 속도를 눈에 띄게 향상시킵니다. 컴퓨터 비전 툴 :

- OpenCV (3.3 버전): OpenCV와 최신 인텔 CPU 최적화 비주얼 라이브러리(Intel Photography Vision Library)를 프리컴파일하고, 얼굴 감지/인식, 눈 깜빡임 감지, 미소 감지 기능 등을 보유하고 있습니다.
- OpenVX : 도형 기반 OpenVX 를 실현하고, 전통적인 CV 작업과 CNN 요소를 지원합니다. Khronos OpenVX 뉴럴 네트워크 확장 1.2 를 지원합니다.
- 기타 : OpenCL™ 드라이버와 런타임, 미디어 드라이버, 인텔® Media SDK와 인텔® SDK OpenCL™ 애플리케이션을 함께 작업한 컴퓨터 비전 SDK, 인텔® MKL-DNN, CLDNN 모두 포함되며, 단독 다운로드할 필요가 없습니다.

¹ <https://software.intel.com/en-us/articles/a-guide-for-setting-up-docker-based-openvino-development-environment-with-ubuntu-system>

그 밖에, 여러 애플리케이션에서 컴퓨터 비전과 딥러닝을 빠르고 효율적으로 실행하기 위해 출시된 툴 패키지로, 인텔® 플랫폼 기반 최적화 OpenVINO™ 툴킷은 현재 VGG-16, VGG-19, SqueezeNet, Resnet, Inception, CaffeNet, SSD, Faster-RCNN, FCN8 등과 같은 사전 전환한 Caffe, TensorFlow, Mxnet 모델의 MO 파일을 제공하며, 20 개 이상의 사전 트레이닝 모델을 갖추고 있습니다. 소프트웨어 개발자와 데이터 과학자는 이 툴을 이용해 개인 맞춤형 딥러닝 애플리케이션을 빠르게 실행할 수 있고, OpenCV, OpenVX의 기본 라이브러리를 사용해 특정 알고리즘을 만들고 맞춤형 및 신형 응용 개발을 진행할 수 있습니다.

OpenVINO™ 툴킷은 인텔 플랫폼에서 시각 자료가 현실이 되게 만들어 여러 사용자가 컴퓨터 비전 애플리케이션을 손쉽게 개발하고 빠르게 배치할 수 있도록 도왔습니다. 여러 딥러닝 애플리케이션 시나리오에서는 인공지능 솔루션이 내포한 잠재력을 이미 선보였습니다.

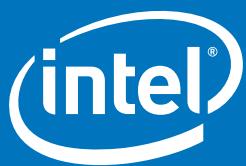
더 많은 정보는 다음을 참조하십시오.

<https://software.intel.com/zh-cn/openvino-toolkit>

해당 매뉴얼 관련 전문 어휘집

영문 풀네임	영문 약자	영문 풀네임
Automatically Tuned Linear Algebra Software	ATLAS	아틀라스
Basic Linear Algebra Subroutine	BLAS	블라스
Batch Size		배치 처리 파라미터
Cardiac Magnetic Resonance	CMR	자기공명영상
Clinical Information System	CIS	진료 정보 시스템
Computed Tomography	CT	전산화 단층촬영
Concat Ops		연속 작업
constant folding		상수 폴딩
Convolution Ops		콘볼루션 작업
Convolutional Layer		콘볼루션 레이어
Convolutional Neural Network	CNN	콘볼루션 뉴럴 네트워크
Cosine Similarity		코사인 유사도
Deep Supervision		딥 감독
Deeping Learning	DL	딥러닝
Double Data Rate	DDR	DDR
Dynamic Random-Access Memory	DRAM	동적 램
Electronic Medical Record	EMR	전자 진료 기록
Euclidean Distance		유클리디안 거리
Feature extraction		특성 추출
Feature map		특징 맵
Fully Connected Layer		FC
Fully Convolutional Network	FCN	전량 콘볼루션 네트워크
Fused Multiply Add	FMA	와이드 FMA
General matrix multiply	GEMM	범용 행렬 곱셈
Geometric pattern		기하학적 무늬
GNU Compiler Collection	GCC	GNU 컴파일러 세트
High Content screening	HCS	HCS
Hospital Information System	HIS	병원 정보 시스템
Intel® d Vector Extensions	Intel® AVX	인텔® 확장 명령어 조합
Intel® Deep Learning Deployment Toolkit	Intel® DLDT	인텔® DLDT
Intel® Streaming SIMD Extensions	Intel® SSE	인텔® Streaming SIMD Extensions
Intel® Ultra Path Interconnect	Intel® UPI	인텔® UPI
Last Level Cache	LLC	LLC
Layer Fusion		레이어 융합
Layer-wise Relevance Propagation	LRP	LRP
Learning Rate	LR	학습률
Liquid-Based Cytologic Preparation	LBP	LBP
Logistics Regression	LR	로지스틱 회귀

영문 풀네임	영문 약자	중문 풀네임
Magnetic Resonance Imaging	MRI	자기공명영상
Mahalanobis distance		마할라노비스 거리
Math Kernel Library for Deep Neural Networks	MKL-DNN	MKL
Matrix multiplication		행렬 곱셈
Multi-scale Convolutional Neural Networks	M-CNN	멀티 스케일 CNN
Multi-Scale Prediction		멀티스케일 예측
Non-Uniform Memory Access Architecture	NUMA	NUMA
Open Message Passing Interface	OpenMPI	개방형 MPI
Open Multi-Processing	OpenMP	
Open Neural Network Exchange	ONNX	ONNX
Operations Per Second	OPS	초당 작업 수
Optical Character Recognition	OCR	광학적 문자 인식
Picture Archiving and Communication Systems	PACS	의료영상전송시스템
Picture Archiving and Communication Systems	PACS	의료영상전송시스템
Pixel Intensity		화소 밝기
Platform as a Service	PaaS	서비스로서의 플랫폼
Pooling Layer		풀링 계층
Position-Sensitive ROI Pooling		센시티브 ROI 풀링
position-sensitive score map		다차원 포지션 센시티브 스코어 맵
Positron Emission Tomography CT	PET-CT	펫시티
Primitive		프리미티브
Region of Interest	ROI	관심영역
Region Proposal Network	RPN	RPN
Reorder Ops		리오더 작업
Resample Ops		리샘플 작업
Residual Net	ResNet	레지듀얼 네트워크
Single Instruction Multiple Data (SIMD)	SIMD	SIMD
Sliding Window Algorithm		슬라이딩 윈도우 알고리즘
Software as a service	SaaS	서비스로서의 소프트웨어
Standard Uptake Value	SUV	SUV
Standardized Euclidean distance		표준화 유클리디안 거리
Support Vector Machines	SVM	서포트 벡터 머신
Total Cost of Ownership (TCO)	TCO	총소유비용



AI 적용을 가속화하려면 다음 사이트를 방문하십시오.



홈페이지
Intel.com/ai



웨이보
@Intel Business



위챗
인텔 비즈니스 채널

© 2019 인텔 회사 저작권 소유. 인텔, Intel, 제온, 옵테인은 인텔 회사의 미국 또는 다른 국가의 상표입니다.
인텔 상표 또는 상표 및 브랜드 명칭 데이터뱅크의 전체 명단은 intel.com 의 상표를 참조하십시오.
* 다른 명칭과 브랜드는 다른 소유자의 자산일 수 있습니다.