英特尔中国
医疗健康行业AI实战手册

英特尔® 至强® 可扩展平台
高效实践人工智能

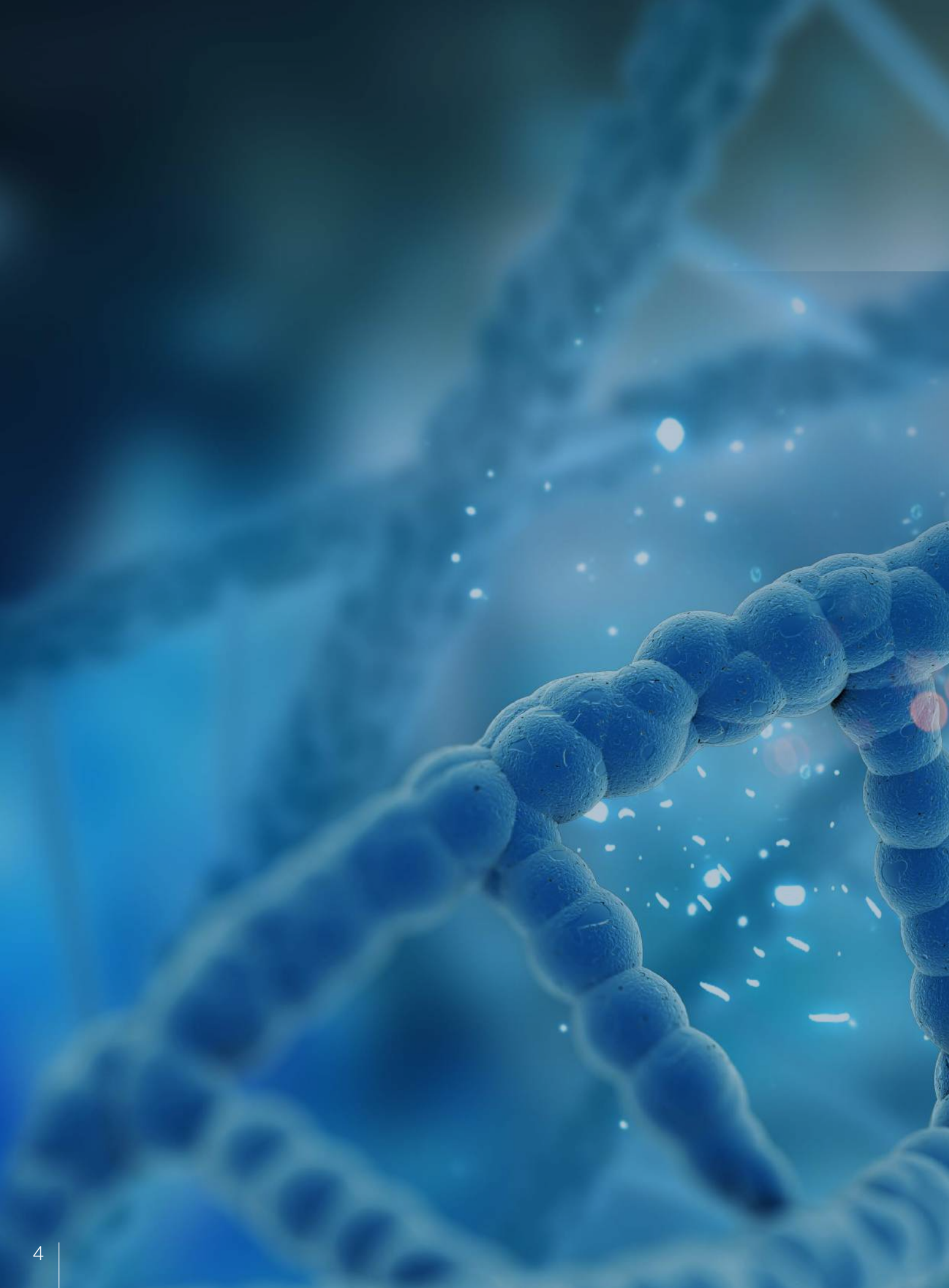
加速AI实践，请访问www.intel.cn/ai

目录 CONTENTS

趋势篇

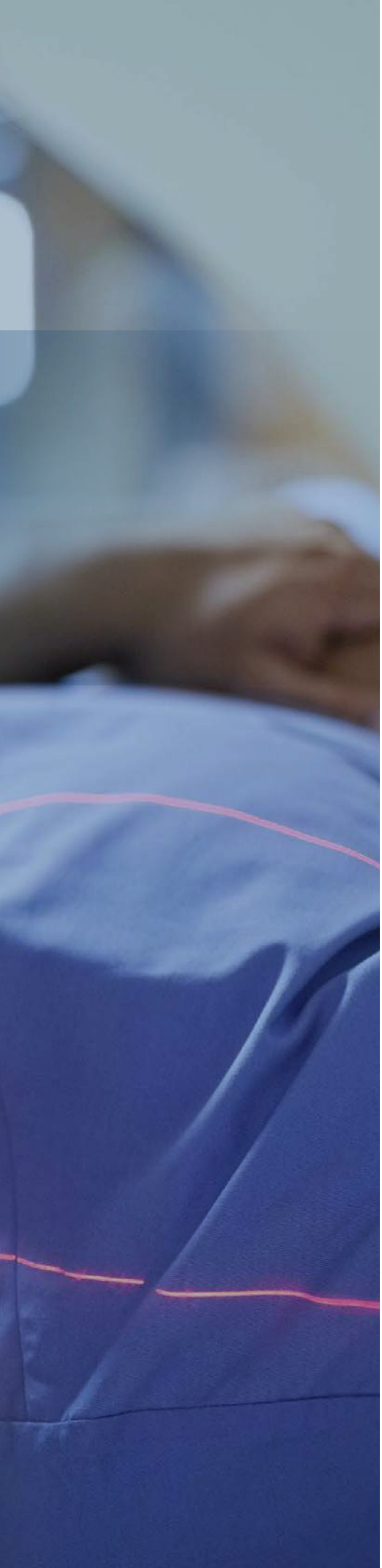
实战篇

技术篇



趋势篇

人工智能在医疗健康领域的发展与应用



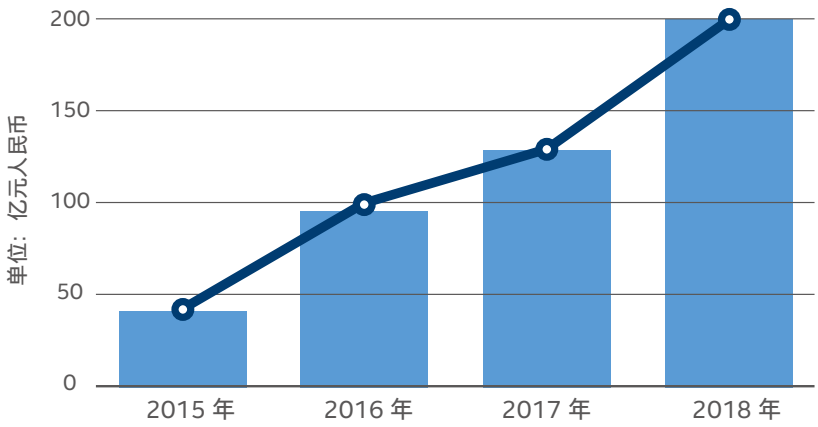
人工智能在医疗健康领域的发展

医疗人工智能的市场趋势

随着算法的进一步成熟、算力的提高以及数据的持续积累，人工智能（Artificial Intelligence，AI）得到迅猛发展，深度学习成为其代表，并呈现出应用领域日益集中的趋势。2018 年 9 月中国信息通信研究院发布的《2018 世界人工智能产业蓝皮书》指出，在中国各类垂直行业中，人工智能渗透较多的领域包括了医疗健康、金融、商业、教育和安防等，其中与医疗健康相关的 AI 企业在所有 AI 企业中占比最高，达到 22%[1]。

作为主要应用领域之一，医疗健康行业对人工智能技术的投资也在快速增长。前瞻产业研究院发布的《2018-2023 年中国人工智能行业市场前瞻与投资战略规划分析报告》显示，2016 年中国医疗人工智能市场规模达到 96.61 亿元人民币，增长 37.9%；2017 年超过 130 亿元人民币，增长 40.7%；2018 年有望达到 200 亿元人民币[2]。这一高速增长一方面得益于中国医疗市场的迫切需求，另一方面则源于近年来医疗人工智能技术的发展以及相关政策的支持。

从全球来看，医疗人工智能的应用细分领域与中国略有不同。根据 Global Market Insight 的统计数据，药物研发在全球医疗人工智能市场中的占比最大，达到 35%。紧随其后的是医学影像人工智能（占比 25%[3]），并将以超过 40% 的增速发展，预计 2024 年将达到 25 亿美元的规模。



中国医疗人工智能产业市场规模

[1]《2018 世界人工智能产业蓝皮书》：http://www.semi.org.cn/siip/pdf/20180920p2.pdf
[2] 前瞻产业研究院，《2018-2023 年中国人工智能行业市场前瞻与投资战略规划分析报告》，2018 年 https://bg.qianzhan.com/report/detail/300/190314-389cc4a4.html
[3] Global Market Insights report. 2018 年 4 月 www.elecfans.com/rengongzhineng/592041.html

此外，基因组学分析是人工智能应用的又一重要领域。预计到 2022 年，该细分市场的规模仅在中国就将接近 300 亿元人民币左右[4]。基因测序与人工智能进一步结合，势必还会加速其发展，同时随之带来的测序时间缩短以及成本大幅降低，又会为医疗行业创造更大的想象空间。

在中国，政策激励是加速医疗人工智能应用落地的关键因素之一。自 2015 年以来，相关政府部门陆续推出了近 20 项政策，从人才培养、技术创新、标准监管、行业融合、产品落地等方面推动人工智能发展。2017 年 3 月，“人工智能” 首次被写入政府工作报告；同年 7 月，国务院正式印发《新一代人工智能发展规划》，明确指出新一代人工智能发展分三步走的战略目标；同年 10 月，“人工智能” 被写入十九大报告，该报告将 “推动互联网、大数据、人工智能和实体经济深度融合”，确定为中国数字经济发展的方向；同年 12 月，工信部发布《促进新一代人工智能产业发展三年行动计划（2018-2020 年）》，详细规划了人工智能在未来三年的重点发展方向和目标。随后在 2018 年 1 月，国家标准化管理委员会指导下的《人工智能标准化白皮书（2018 版）》发布；在 4 月，国务院印发《关于促进 “互联网 + 医疗健康”发展的意见》，将推进 “互联网 +” 人工智能应用服务作为实施 “健康中国” 战略的重要举措，并表示将重点支持研发医疗健康相关的人工智能技术、医用机器人、大型医疗设备等。

医疗人工智能的应用趋势

人工智能在医疗健康领域的应用非常广泛，从医学影像、辅助诊断、疾病预测，到健康管理、药物研发等诸多环节，都可发挥关键作用。其中，人工智能用于医学影像、辅助诊断、疾病预测，主要服务于医院或其他医疗机构，其应用集中在疾病筛查方面，关注点在于如何提高诊断准确率。但囿于存在假阴性的情况，还需要医生审阅所有片子以防漏诊，致使此类应用在减轻医生工作量方面的效果并不显著。

未来，人工智能在不同层级的医疗机构的应用方向可能会呈现出更加多元化的趋势，即在基层医院或第三方体检中心，其应用以辅助筛查和辅助诊断为主；在三甲医院，则以提高医生工作效率为主。在健康管理方面，人工智能以支持单位和个人支付的健康体检为主要方向；在药物研发这一细分领域，人工智能应用又表现出不同特点，需要人工智能技术公司与大型药企、医药研究机构通力合作来推进。

虽然人工智能在医疗健康领域迅速得以应用，但由于数据、模型等方面的原因，仍然面临诸多挑战：

- 数据量。模型越复杂，参数越多，所需要的训练样本量就越大。但是对许多复杂的临床场景而言，所需要的大量可靠数据却并不容易获得；
- 数据质量。一般而言，健康数据的组织化和标准化程度都不高，且数据分散、有噪声。在条件不好的诊所，还存在电子病历信息缺失或有误、多机构间分散存储等问题，同时接口数据可靠性也很差；
- 模型的可解释性。深度学习模型是个黑盒子，对如何得出结论没有明确的解释，其决策模式的权威性也尚待验证；
- 模型的通用性。首先是模型偏差，比如采用白种人患者数据进行训练的模型，可能在其他种族患者中效果不佳；还有就是模型互操作性差，即很难建立一个适用于两种不同电子病历系统中的深度学习模型；
- 模型安全。即便是训练有素的图像处理模型，也有可能因输入图像的扰动而受到不良影响，但这一扰动却无法被人察觉。此外，还存在数据 “差之毫厘” 就可能带来预测结果 “失之千里” 的问题，比如，轻微改变患者电子病历数据中的实验检测值，就会极大影响模型对住院死亡率的预测结果。

针对这些挑战，医疗和人工智能等领域的专家已经提出多项应对措施，来改善应用环境

- 收集大规模和多样化的健康数据。广泛收集来自不同种族、民族、语言和社会经济地位患者的数据，并对其进行标准化和集成；
- 提高数据质量。提供可靠、高质量的数据输入至关重要。可以利用工具提高数据收集的质量，如进行错误纠正、提出关于缺失数据的警告等；
- 融入临床工作流程。将深度学习融入现有电子病历系统的管理，提高临床医生的工作效率；
- 法制化规范化。针对诸如计算机黑客篡改数据，从而影响深度学习模型的结果等数据安全问题，要制定相应法规，保护分析模型的安全。

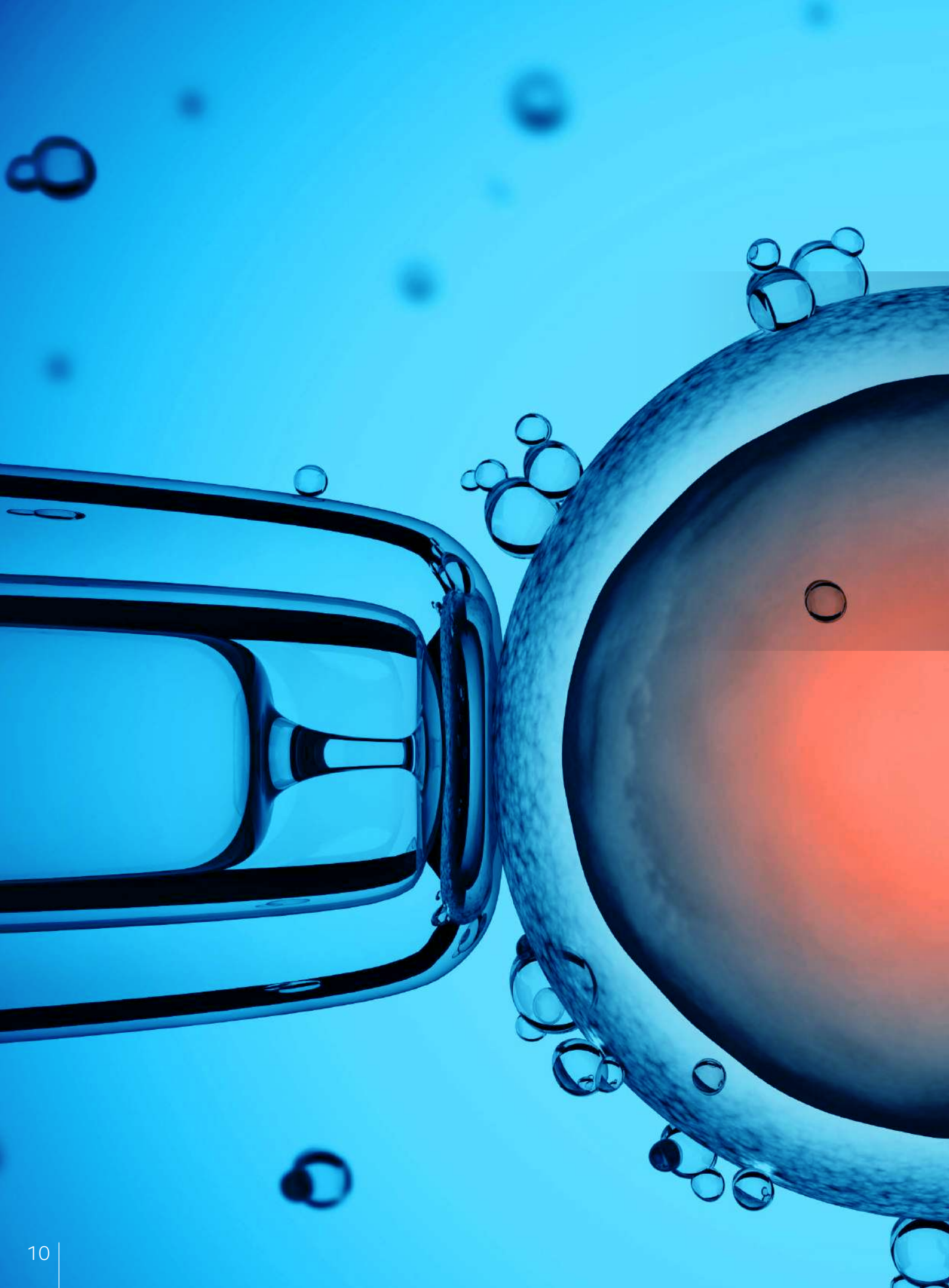
[4] 前瞻产业研究院，2018-2023 年中国基因测序行业市场前瞻与投资战略规划报告，2018 年 https://bg.qianzhan.com/trends/detail/506/180411-e7daa2c4.html

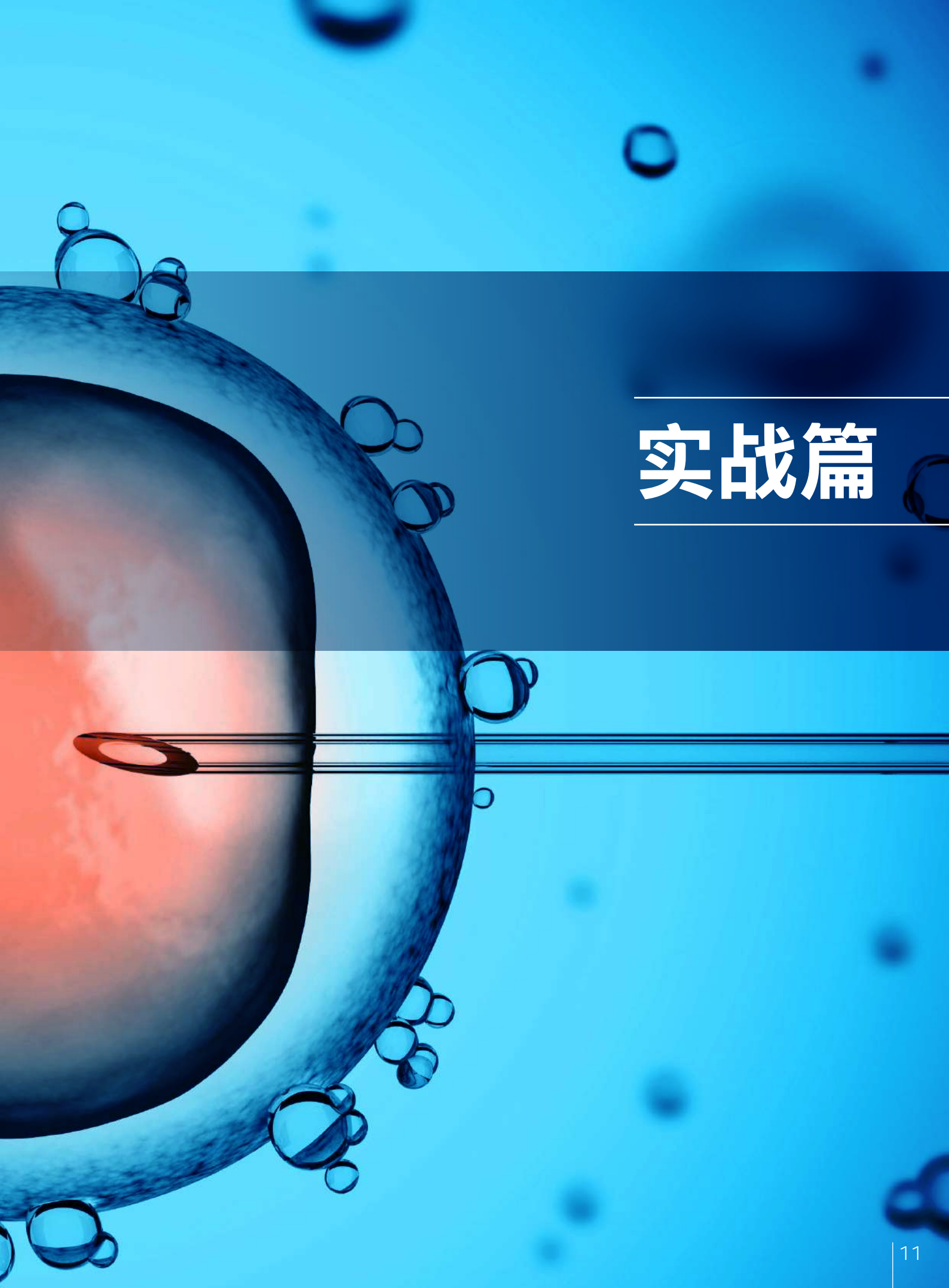
人工智能在医疗健康领域的应用场景

医疗健康是人工智能应用落地最具潜力的领域之一，对此业界已有共识。伴随着其应用的不断深入，人工智能将在以下多种医疗健康应用场景中大显身手：

- 慢病管理与疾病监测：基于患者体征对（潜在）慢性疾病进行风险预估，从而通过早期干预，大大降低患者的医疗费用；
- 临床预测分析：例如基于电子病历数据评估在院内感染疾病（如败血症）的风险，根据运营模型预测患者再入院率，根据财务模型制定捆绑销售服务方案等；
- 病历搜索与质量控制：精准提取医疗文本中的关键信息，进行医学实体识别，进而实现灵活的全量电子病历搜索；
- 虚拟现实助手：通过虚拟现实会话，参与到患教活动中，帮助患者清楚了解其病因，使医患沟通更有效；
- 智能导诊：通过语音、触屏等多种交互方式更好地提供院内导航、导诊、导医，提升精准分诊、健康咨询、健康宣教等服务的水平。
- 图像识别：例如，利用扫描技术、OCR技术或图像处理软件，辨读病历或药品外包装上的信息，快速调取相关资料；
- 影像辅助诊断：帮助放射科医生快速筛除正常影像，提高医生的病例处理量；提高分析影像的准确度，缩短诊断结果报告时间，提升医疗系统的诊断能力；
- 病理分析：例如高效、准确地检测和分类癌细胞，精准勾画癌症放疗靶区等；
- 基因组学分析：用以大幅降低基因测序成本，快速精确实现规模庞大的基因组数据分析，为癌症等疾病的诊断和治疗等提供利器；
- 药物发现：加快药物研发效率，降低成本。

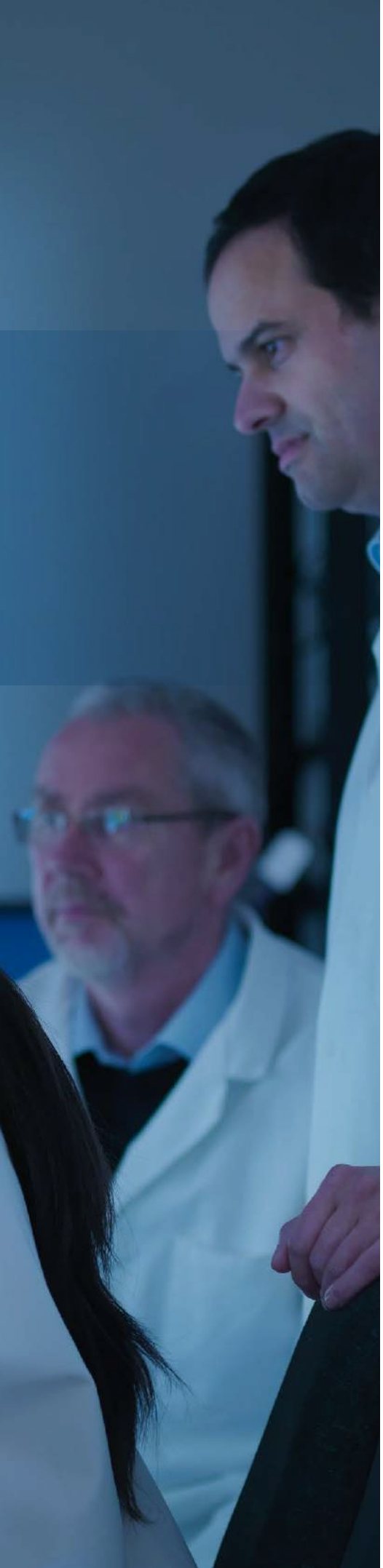
在下一章“实战篇”中，我们将结合英特尔与东软、西门子、解放军总医院、盈谷以及江丰生物等产业伙伴及客户在医疗人工智能领域的实战案例，详细介绍项目的背景、实施过程以及取得的经验与成果，还将结合各应用场景提供相对应的软、硬件配置推荐。





实战篇

加速应用于医疗行业的图像分割 AI 推理



医学影像处理中的图像分割

传统医学影像图像分割方法

计算机视觉中的图像分割[5]是指以图像中的自然边界，例如物体轮廓、线条等，将图像切分为多个区域，其目的是用于简化或改变图像的表现形式，使之更易解读和分析。在计算机方法中，这一过程通常会被解构为将图像中的每个像素加上标签，使具有相同标签的像素有着某种共同视觉特性，例如颜色、亮度、纹理等，由此进行度量或计算得出的一定区域的像素特性都是相似的，而邻接区域则有着很大的不同。

作为计算机视觉技术的重要分支，图像分割已在医学影像处理、人脸识别、工业机器人、智能交通、指纹识别以及卫星图像定位等多个行业和领域获得广泛应用。在医学影像处理领域，图像分割也已在肿瘤和其他病理位置定位、组织体积测量、解剖学研究、计算机辅助手术、治疗方案制定以及临床辅助诊断等多个场景证明了其价值。

传统的图像分割方法主要有以下几种常见方法：

- 基于聚类的方法：聚类法是基于 K- 均值算法，将图像迭代分割成 K 个聚类。该算法中，分割图像中像素与聚类中心之间都有着相似的距离偏差，距离偏差通常采用颜色、亮度、纹理、位置等指标。该算法具有良好的收敛性；
- 基于阈值的方法：该方法是通过计算图像的一个或多个灰度阈值后，将每个像素的灰度值与阈值相比较，最后进行归类的方法；
- 基于边缘的方法：该方法是根据图像中自然边缘的灰度、颜色、纹理等特性的突变性来对图像进行分割。一般来说，基于边缘的分割方法依赖于灰度值边缘检测，当边缘灰度值呈现阶跃型等变化时，判断为图像边缘；
- 基于区域的方法：该方法是根据图像的相似性来对图像进行分割，其判断原则是根据相邻像素点的灰度、颜色、纹理等特性是否存在相似性，如有相似，则扩大像素点的集合。

基于深度学习的图像分割方法

随着近年来 AI 技术的飞速发展，尤其是在图像领域，基于 AI 技术的图像识别、图像处理应用已经被用在很多场景中，其对各类医学影像的分析已经超过人类的识别能力。与卷积神经网络 (Convolutional Neural Network, CNN) 类似的模型，是目前基于 AI 的图像分割技术中常见的网络模型。这其中，全卷积网络 (Fully Convolutional Network, FCN)、U-net 和 V-net 是常见的几种基于深度学习的图像分割方法。

[5] 关于图像分割的描述，部分参考：Linda G. Shapiro and George C. Stockman (2001): “Computer Vision” , pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13-030796-3

■ FCN

CNN 的典型用途是对任务进行分类。对图像处理而言，它的输出是单个类别标签。在生物医学的图像分割处理中，期望的输出应该包括定位，即应该将类标签分配给每个像素。作为卷积神经网络的升级扩展版本，如图 1-1-1 所示，FCN[6] 遵循编码、解码的网络结构模式，级联了卷积层和池化层。卷积层和最大池化层有效降低了原始图像的空间维度。同时，FCN 使用 AlexNet 作为网络的编码器，采用多重转置卷积重复扩展的方式，对编码器最后一个卷积层输出的特征图进行上采样，直到特征图恢复到输入图像的分辨率，因而，可以实现像素级别的图像分割。

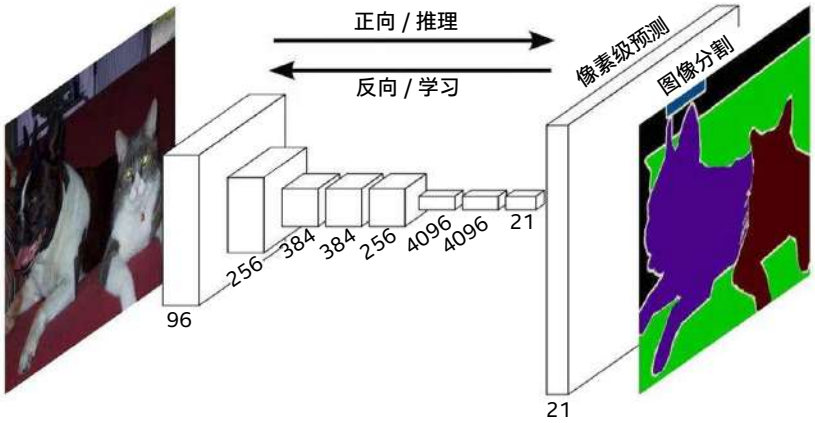


图 1-1-1 FCN 方法原理图

■ U-net

作为 FCN 网络的一个改进版本，U-net 具有一个鲜明的 U 型结构，其拓扑图如 1-1-2 所示，其在每个 Encoder 上都会进行 4 次上采样，这使得分割图恢复边缘等信息会更为精细。同时，在同一个 stage 上，U-net 都采用了跳跃连接（skip connection），而不是直接在高级语义特征上进行监督和 loss 反传，这样就可以保证最后得到的特征图融合了更多的低层级（low-level）的特征，也使得不同尺度的特征得到了融合，从而可以进行多尺度预测（Multi-Scale Prediction）和深度监督（Deep Supervision）。另外，U-net 在网络后部补充了一个与前面类似的网络，形成 U 性结构。其中池化运算符由上采样运算符替换，因此增加了输出的分辨率。同时，为了定位，模型从收缩路径的高分辨率特征与上采样输出相结合。连续卷积层可以采用 relu 激活函数来对原始图片进行降采样操作，从而获得更精确的输出。

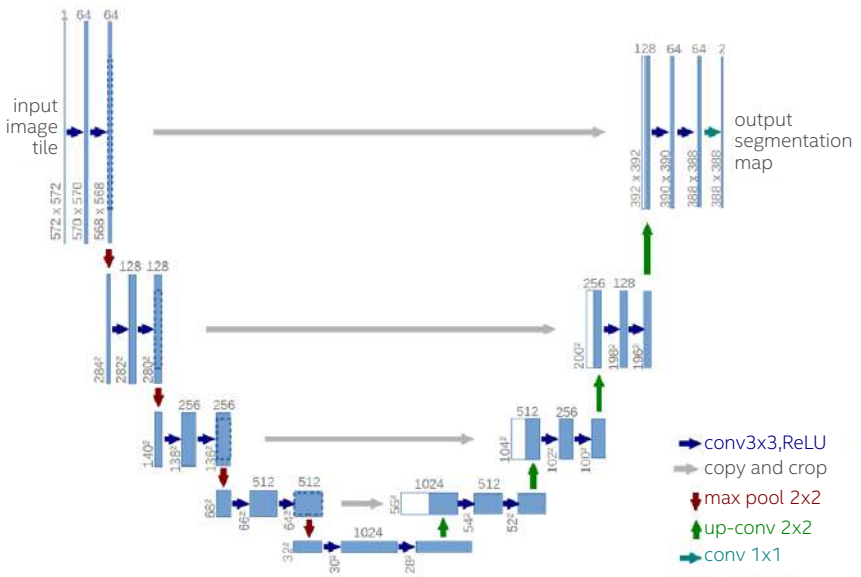


图 1-1-2 U-Net 拓扑

医学影像在实际应用中也有其独有的特性。我们可以看到，一般胸片影像是胸片 CT，而眼底检查则是眼底 OCT，均为针对一个指定器官的成像，而非全身。而器官本身结构比较固定，语义信息并非特别丰富。所以高级语义信息和低层级特征就显得非常重要，而 U-net 的 U 型结构和跳跃连接在这种场景中，可以发挥出更大作用。近年来，U-net 在医学影像分割领域良好的应用效果，已在很多部署中得到了充分证明。

[6] 关于 FCN 的相关技术描述，摘选自 UC Berkeley jonlong、shelhamer 和 trevor 的论文《Fully Convolutional Networks for Semantic Segmentation》：https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf

■ V-net

V-net 可以视为 3D 版本的 U-net，如图 1-1-3 所示，它与 U-net 有着类似的拓扑形态，适用于三维结构的医学影像分割。V-net 能够实现基于 3D 图像的端到端图像语义分割，并通过类似于残差学习的 trick 来对网络进行改进。

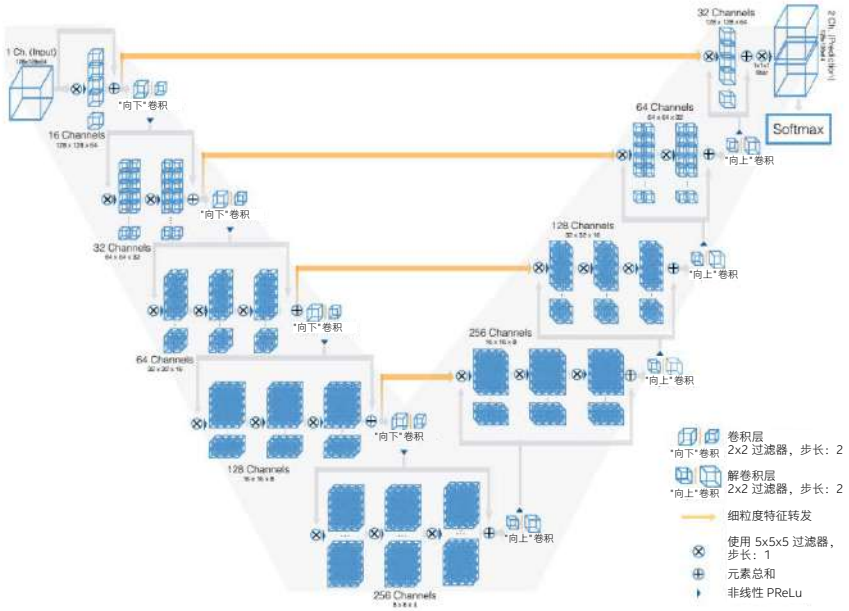


图 1-1-3 V-Net 拓扑思想

软硬件配置建议

对于在医疗行业中构建基于深度学习的图像分割方法，可以参考以下基于英特尔® 架构平台的软硬件配置来完成。

名称	规格
处理器	英特尔® 至强® 金牌 6240 处理器或更高
超线程	ON
睿频加速	ON
内存	16GB DDR4 2666MHz* 12及以上
存储	英特尔® 固态盘 D5 P4320系列及以上
操作系统	CentOS Linux 7.6或最新版本
Linux核心	3.10.0或最新版本
编译器	GCC 4.8.5或最新版本
Python版本	Python 3.6或最新版本
Tensorflow版本	R1.13.1或最新版本
OpenVINO™ 工具套件	2019 R1或最新版本
Keras版本	2.1.3或最新版本

U-net 分割网络的优化方法

基于英特尔® 架构的优化方法

将传统的 CNN 图像分割方法用于医学图像时，往往存在以下困难：

- CNN 通常都是应用于分类，生物医学图像则更关注分割以及定位的任务；
- CNN 需要获取大量的训练数据，而医学图像很难获得相应较大规模的数据。

以往在应对上述困难时，通常采用滑窗的方法，即为每一个待分类的像素点取周围的一部分邻域输入。这种方法好处有两点：首先，这一方法能够在滑窗的同时完成定位工作；其次，每次动作都会取一个像素点周围的邻域，可以大大增加训练的数据量。但是，这一方法也有两个缺点：一是通过滑窗所取的块之间有较大的重叠，会导致训练和推理速度变慢；二是网络需要在局部准确性和获取上下文之间进行取舍，因为如果滑窗取的块过大，就需要更多的池化层，定位准确率会降低，而取的块过小，则网络只能看到很小的一部分上下文。

基于英特尔® 架构平台开展的一系列优化，可以从另一个层面帮助用户解决以上问题。这些优化方法包括：调整处理器核心数量、引入非统一内存访问架构 (Non- Uniform Memory Access Architecture，NUMA) 技术以及面向深度神经网络的英特尔® 数学核心函数库 (Intel® Math Kernel Library for Deep Neural Networks, 英特尔® MKL-DNN)，从而为 U-net 图像分割法提供多层次的优化。优化步骤如下：

■ 环境变量设置

首先，需要对环境变量进行设置，如以下所示，命令包括：清空系统的缓存（cache），将处理器设置为性能优先的模式，即运行在最高频率，打开处理器的睿频加速。

```
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_OMP_NUM_THREADS=20
```

- KMP_BLOCKTIME 设置为 1，是设置某个线程在执行完当前任务并进入休眠之前需要等待的时间，通常设置为 1 毫秒；
- KMP_AFFINITY 设置为 Compact，是表示在该模式下，线程绑定按计算核心的计算要求优先，先绑定同一个核心，再依次绑定同一个处理器上的下一个核心。此种绑定适用于线程之间具有数据交换或有公共数据的计算情况，优势在于可以充分利用多级缓存的特性；
- OMP_NUM_THREADS 设置为 20，是将并行执行线程的数量设定为所需的物理核心数。

■ 测试代码中添加线程控制

```
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
```

如上述设置命令所示，在进行 tf.ConfigProto() 初始化时，我们也可以通过设置 intra_op_parallelism_threads 参数和 inter_op_parallelism_threads 参数，来控制每个操作符 op 并行计算的线程个数。二者的区别在于：

- intra_op_parallelism_threads 控制运算符 op 内部的并行，当运算符 op 为单一运算符，并且内部可以实现并行时，如矩阵乘法、reduce_sum 之类的操作，可以通过设置 intra_op_parallelism_threads 参数来并行，intra 代表内部。
- inter_op_parallelism_threads 控制多个运算符 op 之间的并行计算，当有多个运算符 op，并且它们之间比较独立，运算符和运算符之间没有直接的路径 Path 相连时，Tensorflow 会尝试并行地对其进行计算，并使用由 inter_op_parallelism_threads 参数来控制数量的一个线程池。

通常而言，intra_op_parallelism_threads 设置为单个处理器的物理核心数量，而 inter_op_parallelism_threads 则设置为 1 或者 2。

■ 利用NUMA特征来控制处理器计算资源的使用

数据中心使用的服务器，通常都是配置两颗或更多的处理器，多数都采用 NUMA 技术，使众多服务器像单一系统那样运转。处理器访问它自己的本地存储器的速度比非本地存储器更快一些。为了在这样的系统上获取最好的计算性能，需要通过一些特定指令来加以控制。Numactl 就是用于控制进程与共享存储的一种技术机制，也是在 Linux 系统中广泛使用的计算资源控制方法。具体使用方法如下所示：

图 1-2-1 用 NUMA 特征来控制处理器计算资源的使用

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

上述指令表示的是 test.py 在执行的时候只使用了处理器 #CPU0 中的 0-19 和 40-59 核，以及处理器 #CPU0 对应的近端内存。

■ 采用面向英特尔® MKL-DNN 优化的 TensorFlow

为了使用户在通用处理器平台上进行高效的 AI 计算，英特尔针对众多主流的深度学习开源框架进行了大量的优化，包括目前在工业界和学术界使用十分广泛的 TensorFlow。

通过使用英特尔® MKL-DNN 优化的多种原语（Primitive），英特尔对 TensorFlow 进行了优化。英特尔® MKL-DNN 是从 TensorFlow

1.2 开始添加的。除了在训练基于 CNN 的模型时能显著提升性能之外，使用英特尔 ® MKL-DNN 进行编译还可以创建针对英特尔 ® 高级矢量扩展指令集 (Intel® Advanced Vector Extensions，英特尔 ® AVX)、英特尔® AVX 2 和英特尔 ® AVX-512 进行优化的二进制文件，从而得到一个经过优化且与大多数现代 (2011 年后) 处理器兼容的文件。

参考文献:

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

*** 更多英特尔 ® MKL-DNN 的技术细节，请参阅本手册技术篇相关介绍。**

U-net基于英特尔® 架构优化后的测试及结果

通过以上四个方面的优化，U-net 在基于英特尔 ® 架构的处理器平台上的性能得到了显著提升，测试结果如下图所示 [7]:

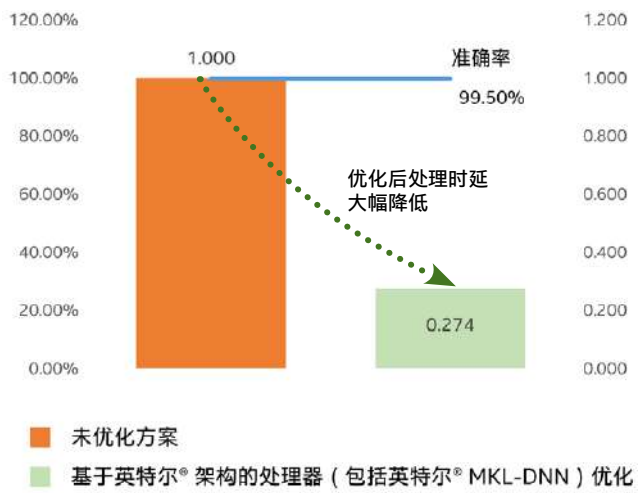


图 1-2-2 基于英特尔 ® 架构优化前后性能对比

基于 OpenVINO™ 工具套件英特尔® 发行版对 U-net 进一步优化

为满足客户在实际应用场景中的需求，在上述结果的基础上，英特尔又基于 OpenVINO™ 工具套件英特尔® 发行版(以下简称"OpenVINO™ 工具套件") 对 U-net 图像切割方法实施了进一步的优化，具体优化步骤如下:

■ 模型转换

由于原有的模型是基于 Keras 进行训练，生成的模型为 hdf5 格式，这种格式的模型无法直接作为 OpenVINO ™ 工具套件的输入，需要先进行格式转换，操作命令如下:

```
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 ##做模型转换
```

■ 将模型通过OpenVINO™ 工具套件的mo.py转换成xml文件和bin文件

命令如下:

```
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]
```

■ 通过Inference Engine 来进行模型推理

命令如下:

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

其中，做推理的代码包含如下逻辑模块:

```
1. #Load input data ##数据数据预处理及导入，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_bi
   n) ##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] 测试配置为: 处理器: 双路英特尔 ® 至强 ® 金牌 6148 处理器，2.40GHz; 核心 / 线程: 20/40; 内存: 16GB DDR4 2666MHz * 12; 硬盘: 英特尔 ® 固态盘 SC2BB480G7; BIOS: SE5C620.86B.02.01.0008.031920191559; 操作系统: CentOS Linux 7.6; Linux 内核: 3.10.0-957.21.3.el7.x86_64; gcc 版本: 7.2; Python 版本: Python 3.6; Tensorflow 版本: R1.13.1。

基于 OpenVINO™ 工具套件的优化结果

优化结果如图 1-2-3 所示，最左列为脑部 CT 原图，中间列是未优化时的图像分割结果，最右列是通过 OpenVINO™ 工具套件优化之后生成的图像分割结果。可以看出，通过 OpenVINO™ 工具套件优化后生成的图像分割结果，在准确率上与未优化时基本保持一致，但在推理速度上却远高于未优化时[8]。

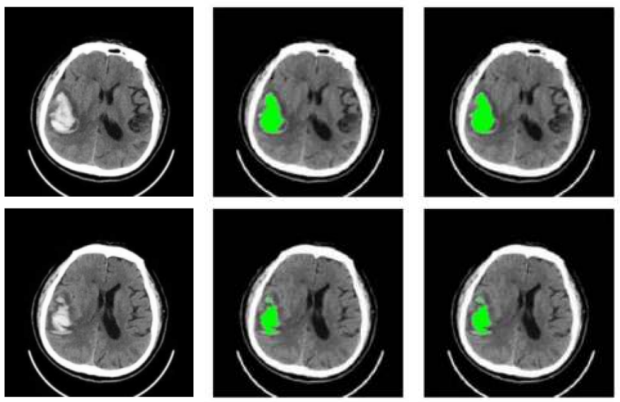


图 1-2-3 基于 OpenVINO ™ 工具套件对 U-net 的优化结果

***更多OpenVINO™ 工具套件的技术细节，请参阅本手册技术篇相关介绍。**

基于第二代英特尔® 至强® 可扩展处理器构建的 Dense U-net 图像分割方法

英特尔® 深度学习加速 (Intel® Deep Learning Boost, 英特尔® DL Boost) 技术

第二代英特尔® 至强® 可扩展处理器，不仅以优化的微架构、更多的内核及更快的内存通道带来了计算性能的提升，更面向 AI 应用提供了更为全面的加速能力，尤其是在其集成的英特尔® 深度学习加速技术（VNNI 指令集）中，加入了对 INT8 的支持，为用户提供了高效的 INT8 深度学习推理加速能力，这一能力将有效提升 U-net 图像分割方法的执行效率。

英特尔® 深度学习加速技术通过 VNNI 指令集来支持 8 位或 16 位低精度数值相乘，这对于需要执行大量矩阵乘法的深度学习计算而言尤为重要。它的导入，使得用户在执行 INT8 推理时，对系统内存的要求最大可减少 75%[9]，而对内存和所需带宽的减少，也加快了低数值精度运算的速度，从而使系统整体性能获得大幅提升。

与以往的 FP32 模型相比，INT8 模型具有更小的数值精度和动态范围，因此在图像切割等深度学习中采用 INT8 推理方式，需要着重解决计算执行时的信息损失问题。一般地来讲，INT8 推理功能可以通过量化校准的方式来形成待推理的 INT8 模型，进而实现将 FP32 在信息损失最小化的前提下转换为 INT8 的目标。

以图像分析应用为例，从高精度数值向低精度数据转换，实际是一个边计算边缩减的过程。换言之，如何确认缩减的范围是实现信息损失最小化的关键。在 FP32 向 INT8 映射的过程中，采用根据数据集校准的方式，来确定映射缩减的参数。在确定参数后，平台再根据所支持的 INT8 操作列表，对图形进行分析并执行量化 / 反量化等操作。量化操作用于 FP32 向 S8（有符号 INT8）或 U8（无符号 INT8）的量化，反量化操作则执行反向操作。

基于 OpenVINO™ 工具套件进行 FP32 模型到 INT8 模型的转换

通常地，通过神经网络训练好的模型是单精浮点精度的，即 FP32，用户可以将这样的模型直接部署在实际应用场景中，并通过量化技术得到低精度模型，比如 INT8 模型在保证模型精度的基础之上可以提供效率更高的模型推理应用，通常情况下模型精度的损失小于 1%。

OpenVINO™ 工具套件从 2018 R4 版本开始提供 FP32 模型到 INT8 模型的转换功能，并且从 2019 R1 版本开始，支持基于第二代英特尔® 至强® 可扩展处理器所集成的英特尔 ® 深度学习加速技术。

首先，OpenVINO™ 工具套件会将训练好的、基于开放神经网络交换（Open Neural Network Exchange，ONNX）训练的模型进行转换和优化，生成 FP32 格式的 xml 文件和 bin 文件，其中的优化包含节点融合、批量归一化的去除和常量折叠等方法；然后，通过 OpenVINO™ 工具套件中的转换工具（Calibration Tool）将 FP32 格式的文件转换为 INT8 格式的 xml 文件和 bin 文件，在转换的过程中需要用到一个小批量的验证数据集，并且会将转换量化过程中

[8] 相关验证测试配置为：处理器：双路英特尔® 至强® 金牌 6148 处理器，2.40GHz；核心/线程：20/40；内存：16GB DDR4 2666MHz* 12；硬盘：英特尔® 固态盘SC2BB480G7；BIOS：SE5C620.86B.02.01.0008.031920191559；操作系统：CentOS Linux 7.6；Linux内核：3.10.0-957.21.3.el7.x86_64；gcc版本：4.8.5；Python版本：Python 3.6；OpenVINO™ 工具套件：2019 R1；Keras：2.1.3。

[9] 数据源引自https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

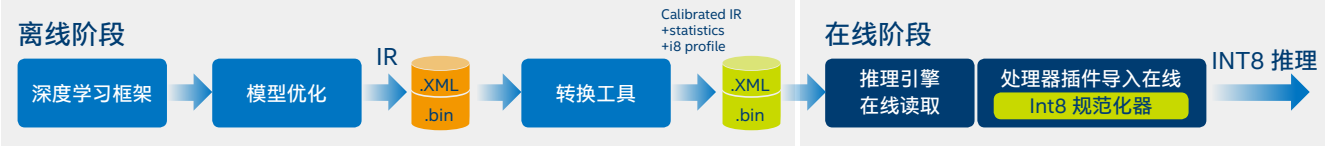


图 1-3-1 基于 OpenVINO ™ 工具套件的 FP32 模型到 INT8 模型的转换

的统计数据存储下来，以便在后续的推理时确保精度不受影响。上述的转换流程是离线运行的，也就是只要转换一次即可，详细做法如图 1-3-1 所示：

按照上述模型转换之后得到的初步模型的性能如下图所示：

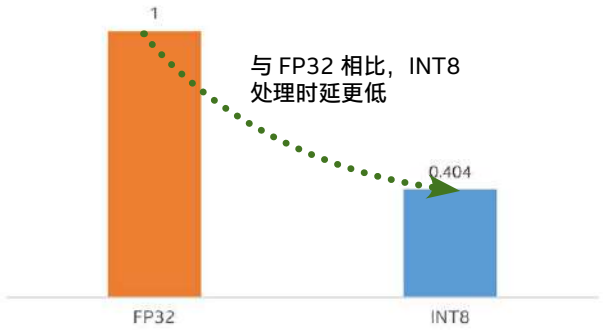


图 1-3-2 FP32 与 INT8 的时延性能对比

通过对两种模型进行性能分析可以看出，FP32模型中的重排序操作（Reorder Ops）占据了大量的执行时间，在INT8模型中，重采样（Resample Ops）只支持FP32的操作，连接操作（concat Ops）执行时间过长，而本来占比最高的卷积操作（Convolution Ops）在整个模型运行中占据的时间比例反而少。因此，需可以对其进行进一步的优化。

如图1-3-3所示，经过优化，模型的延迟有了大幅度的降低。

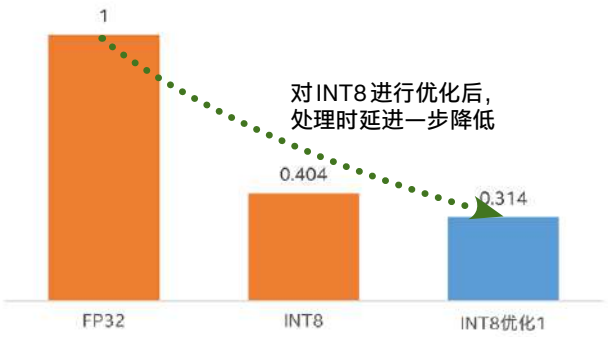


图 1-3-3 优化后的 INT8 模型时延性能对比

此时再将INT8模型进行逐层分析，可以看到相比之前已经有了很明显的提升。但此时可以看到在优化之后的模型中，Concat Ops所占据的执行时间还是较长，为了进一步提升模型的吞吐量，需对Concat Ops进行了特定优化，并且不再使用英特尔® MKL-DNN中的原语，而是要进行定制化，详细代码如下所示：

```
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[j] + (i * channels_size), src_ptrs[j] + i * channels[j], channels[j]);
12.     }
13. });
```

上述优化主要的目的是实现并行化地批量拷贝数据到指定位置，通过此类型的优化，模型性能有了进一步的提升。此时的模型执行时间基本达到了理想状况，最终优化结果如图1-3-4所示：

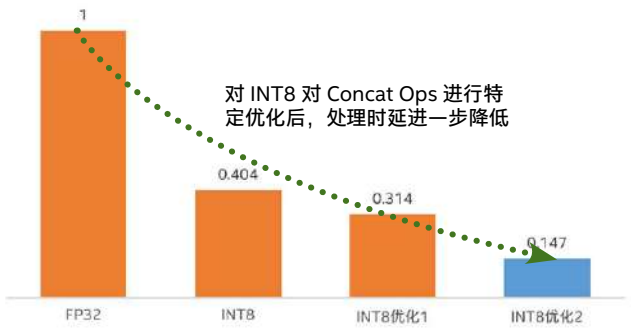


图 1-3-4 进一步优化后的 INT8 模型时延性能对比

从性能分析可以获知，此时模型运行占比最高的原语成了卷积操作，完全符合本实例中Dense U-net这个模型本应有的效果。

应用案例

东软 eStroke 溶栓取栓影像平台

■ 背景

脑卒中一直是危害公众健康的主要 “杀手”。据估算，全国每年新发脑卒中约 200 万人，65 岁以下人群约占 50%。这表明，我国脑卒中年轻化趋势严重，且每年仍以 13% 的速率在上升，复发率高达 17.7%[10]，给患者及社会带来了沉重的负担。脑卒中的首选有效治疗手段为溶栓和取栓治疗，但这一方法有赖于对脑部医疗影像的快速和准确判读。脑卒中救治关键时间只有 30 分钟，基本没有时间转诊，但是施救的关键点往往在基层区县医院。一方面，囿于基层医院技术能力不足，溶栓、取栓比例较低；另一方面，医生判读水平参差不齐，专业影像医生资源不足，中心医院影像专家也分身乏术。这些都导致脑卒中溶栓、取栓缺乏有效的影像学指导，无法有效识别出可挽救的组织，使患者失去了宝贵的抢救窗口。

为应对这一挑战，医疗行业需要一种即便在基层医院医生判断水平不足的情况下，仍然可以快速准确地对相关医学影像进行分析的工具。现在，基于深度学习的医学影像判读已经逐步走入医疗机构，帮助应对以上问题。东软智能医疗研究院、沈阳东软医疗系统有限公司（以下简称 “东软”）联合众多合作伙伴一同，打造的高质量 eStroke 溶栓取栓影像平台，就能够为急性脑卒中静脉溶栓和动脉取栓治疗提供更精准指导。

■ 方案与成效

eStroke 溶栓取栓影像平台是基于缺血性脑卒中半暗带、脑微出血、脑侧支循环做出定量评价的云服务平台，可以实现对溶栓、取栓多模态影像做出精准评价，具有以下优势：

- 支持多模态影像学设备，其中包括电子计算机断层扫描（Computed Tomography，CT）、核磁共振成像（Magnetic Resonance Imaging，MRI）图像等 16 排以上多层螺旋 CT 以及 1.5T 以上 MRI；
- 实现全流程自动化，从医院设备扫描序列开始到影像后处理分析，一直到输出影像诊断报告，均无需人工干预；
- 能够接入互联网医疗诊治技术应用研究平台等外部诊疗系统，支撑开展心脑血管病远程急救、移动急救、高危人群智能预警及干预、心脑血管病联合救治、虚拟手术等技术研发和工程化。

以 eStroke 溶栓取栓影像平台为载体，东软与英特尔携手，基于 U-net 模型对平台中的脑卒中医学影像进行图像分割处理，根据 eStroke 平台对灌注成像的各个参数，包括 CBF、CBV、MTT 和 TMAX（分别对应脑血流量、血脑容量、平均通过时间和残留函数的达峰时间）的计算，并结合以上参数通过左右脑循环的对称性，如图 1-4-1 所示，进一步推理出用于医学诊断的缺血半影带和梗死核心的所在区域。

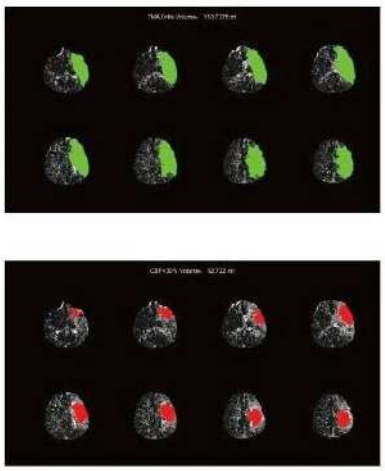


图 1-4-1 通过 TMAX&CBF 异常区域计算出缺血半影带和梗死核心区域

该方案采用面向英特尔® 架构优化的 TensorFlow（基于英特尔® MKL-DNN 优化）以及 OpenVINO™ 工具套件进行了优化，使基于 U-net 模型的深度学习推理在保证准确性的同时，推理时间得以大幅减少。这对于争分夺秒的脑卒中诊治而言，无疑有着重大的实践意义。如图 1-4-2 所示，在推理准确性基本一致的情况下，采用两个工具优化后的方案与未经优化的方案对比，推理延迟分别降低 72.6% 和 85.4%[11]。

图 1-4-2 东软 U-net 图像分割各方案性能对比

[10] 数据援引自《安徽省脑卒中分级诊疗指南（2015版）》

[11] 该数据所使用的测试配置为：处理器：双路英特尔® 至强® 金牌 6148 处理器，2.40GHz；核心/线程：20/40；内存：16GB DDR4 2666MHz * 12；硬盘：英特尔® 固态盘SC2BB480G7；BIOS：SE5C620.86B.02.01.0008.031920191559；操作系统：CentOS Linux 7.6；Linux内核：3.10.0-957.21.3.el7.x86_64；gcc版本：7.2(Tensorflow) & 4.8.5(OpenVINO)；Python版本：Python 3.6；Tensorflow版本：R1.13.1；OpenVINO™ 工具套件：2019 R1；Keras：2.1.3。

西门子医疗利用英特尔® 深度学习加速技术，加速心血管疾病治疗中的 AI 应用

■ 背景

心血管疾病一直是危害人类健康的大敌。据统计，心血管疾病每年导致约 1,800 万人失去生命[12]。采用心脏磁共振成像检查（MRI），通过对心脏磁共振成像（Cardiac Magnetic Resonance，CMR）图像的定量测量，一直是评估心脏功能、心室容量和心肌组织状况的金标准。传统上，心血管专家需要凭借经验来对 MRI 影像进行判读，不仅费时费力，且错误率较高，在解释图像时也容易受到主观因素的影响，导致漏诊和误诊。

现在，西门子医疗正在开展一系列创新医疗 AI 应用的研究，并将成果纳入心脏病学与放射性影像分析的实际应用中。但要将这些 AI 能力真正应用到医疗实践中，还面临着一系列的挑战。

首先，AI 应用不能对临床诊疗造成延迟，AI 应用需要与各类检查仪器生成的数据保持同步，并保证 AI 推理具备高吞吐、低延迟的特性，才能让基于 AI 的医疗系统服务更多患者。其次，AI 应用应当尽可能与临床诊疗流程进行融合，以便节省时间，并提高测量和诊断之间的一致性和准确性。

为此，西门子医疗与英特尔一起，基于通用处理器平台来开展针对 MRI 影像的判读和测量，实施高效的 AI 推理工作。双方不仅利用深度学习的方法对来自 MRI 的心血管医学影像进行了 AI 判读研究，同时基于全新的第二代英特尔® 至强® 可扩展处理器平台以及 OpenVINO™ 工具套件等，进行了优化工作，使推理速度大幅提升，为临床医学诊疗提供了强有力的支撑。

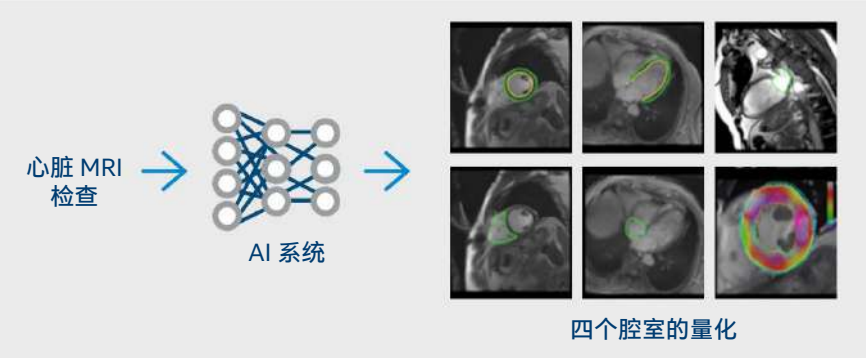


图 1-4-3 西门子医疗与英特尔一起构建心脏 MRI 的 AI 分析能力

■ 方案与成效

在本案例中，西门子医疗与英特尔一起合作，优化了基于全新的第二代英特尔® 至强® 可扩展处理器构建的心腔检测和量化模型。该 AI 模型基于 Dense U-net，可对心脏的左右心室进行语义分割，并可扩展到所有四个腔室。AI 模型的输入是跳动心脏的 MRI 图像的堆叠，输出则是识别心脏的区域以及结构，其中每个结构都会被颜色编码。这样可以实现原先需要人工识别标注的过程，从而加快影像判读速度，其整体工作流程见图 1-4-3 所示。

第二代英特尔® 至强® 可扩展处理器为该 AI 模型的推理提供了高效、灵活和可扩展的平台，特别是与 OpenVINO™ 工具套件的紧密结合，有效地加速了针对视觉应用的深度学习推理，提高了诊疗过程中宝贵的诊断、决策速度和准确性。同时，处理器集成的英特尔® 深度学习加速技术，具有全新的矢量神经网络指令（VNNI），能够进一步加速深度学习中的各种计算密集型操作，让图像分类、图像分割、目标检测等 AI 应用在英特尔® 处理器平台上的推理效率变得更高。英特尔® 深度学习加速技术对 INT8 良好的支持能力，使其可以将 FP32 训练模型转化为 INT8，在保持准确性的同时大幅提升推理速度。

在本案例中，深度神经网络（例如 Dense U-net）经过训练后被用以识别心脏区域，神经网络的权值通常采用浮点数值（FP32）来表示，因此模型通常情况下会通过 FP32 精度来进行训练和推理。但 INT8 同样可以在损失很小的准确率（通常 <0.5%，本案例中可达到 <0.001%）情况下来提升推理速度[13]。

通过英特尔® 深度学习加速技术和 OpenVINO™ 工具套件提供的 FP32 到 INT8 的转换工具（Calibration Tool），英

[12, 13] 该数据援引自 Journal of the American College of Cardiology, 2017.

特尔帮助西门子医疗实现了在保持准确率的情况下以更高的速度进行推理运算的能力。图 1-4-4 显示了利用 AI 进行心脏图像分割，左图显示 AI 模型分割了心脏的各种结构，右图上部是未使用 INT8 模型的传统 ONNX 输出图像，而右图下部是使用 INT8 模型的输出图像，可以直观地看到，两者的精度基本一致。

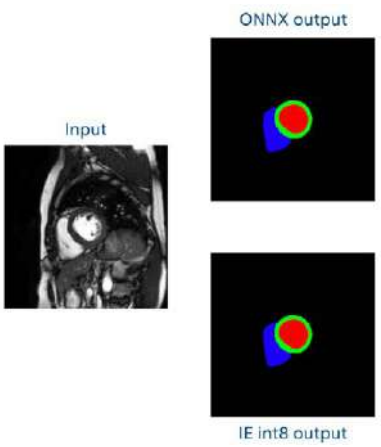


图 1-4-4 使用 INT8 模型前后的输出精度对比

从推理速度来看，基于第二代英特尔® 至强® 可扩展处理器、英特尔® 深度学习加速技术以及 OpenVINO ™ 工具套件进行优化后，心脏 MRI 的 AI 分析能力得以大幅增强。一方面，心脏 MRI 影像的处理速度获得了显著增强，达到了 200 FPS (帧每秒) ，这意味着，一次完整的心脏 MRI 检查数据可以在不到 1 秒的时间内就分析完毕，这为心脏 MRI 在临床上的近实时应用开辟了可能；另一方面，优化后的解决方案，在量化和执行模型时，在几乎没有降低精度的情况下，性能可以提升到未优化方案的 5.5 倍[14]。

通用电气医疗集团利用英特尔技术与产品，优化深度学习模型，提升 CT 图像推理性能

■ 背景

电子计算机断层扫描 (Computed Tomography，CT) 检查是现代医学中最常用的检查手段之一。其通过 X 射线束对人体层面进行扫描，并得到相关部位的断面或立体图像，从而发现人体的病变情况。CT 检查虽然有着极为重要的临床意义，但 CT 切片图像的检查在传统上往往依赖经验丰富的医生进行人工读片，不仅效率较低，其固有的主观性也可能带来误诊、漏诊。

现在，通用电气医疗集团 (以下简称 “GE 医疗”) 正利用深度学习的方法，对 CT 切片图像进行分类和标记，这更便于医生寻找到微小病灶，并将其用于研究或临床比较。在 2018 年的医学成像光学会议 (SPIE) 上，GE 医疗发表了一篇关于基于 AI 的结构分类器的论文，其中 GE 医疗的 CT 成像专家使用 Python 语言、TensorFlow 框架以及 Keras 库构建并训练了新的 AI 模型。通过与英特尔开展的深入技术合作，双方正利用英特尔 ® 至强 ® 处理器、英特尔 ® 深度学习部署工具 (Intel® Deep Learning Deployment Toolkit, 英特尔 ® DLDT) 等产品与技术，来优化其面向 CT 推理的解决方案。

■ 方案与成效

方案中引入了英特尔 ® DLDT 来优化深度学习模型，并在英特尔 ® 至强 ® 处理器平台展现出更好的推理性能。

英特尔 ® DLDT 是 OpenVINO ™ 工具套件中，专门用于深度学习模型的推理加速部件。通过该工具，训练收敛的模型可以在多种英特尔 ® 处理器平台上获得更高的数据处理能力，以及更低的数据处理延时。其可以对多种主流深度学习开源框架训练好的模型进行转换和优化，生成独立于深度学习框架的 bin 文件和 xml 文件。其中 bin 文件用于存放深度学习模型的权重，以二进制形式存储，而 xml 文件则描述深度学习模型的网络结构，二者结合起来共同解析模型。这使得模型的表征文件不依赖于任何深度学习框架，可以更方便地进行部署。同时，在生成这两个文件的过程中，还会对模型进行常量折叠、Batch 层融合、水平方向层融合、无效节点消除等模型优化操作。

如图 1-4-5 所示，英特尔 ® DLDT 可以轻松地导入 GE 医疗基于 TensorFlow 等框架训练得到的模型。

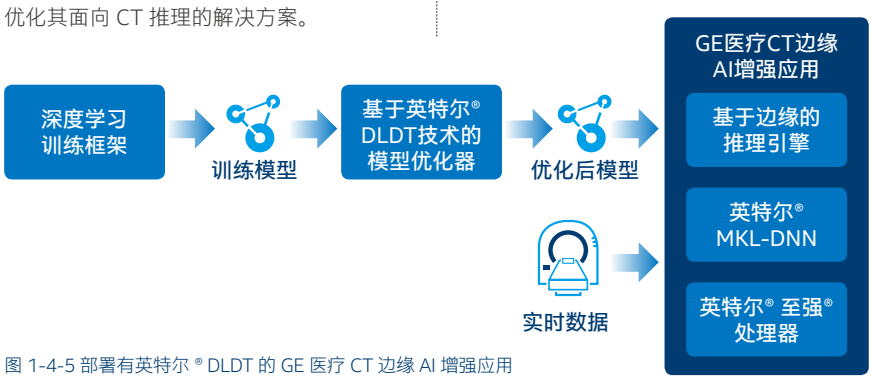


图 1-4-5 部署有英特尔 ® DLDT 的 GE 医疗 CT 边缘 AI 增强应用

[14] 该数据所使用的测试配置为：处理器：双路英特尔® 至强® 铂金 8280 处理器，2.70GHz；核心/线程：28/56；HT：ON；Turbo：ON；内存：192GB DDR4 2933；硬盘：英特尔® 固态盘SC2KG48；BIOS：SE5C620.86B.02.01.0008.031920191559；操作系统：CentOS Linux 7.6.1810；Linux内核：4.19.5-1.el7.elrepo.x86_64；gcc版本：4.8.5；OpenVINO™ 工具套件：2019 R1；工作负载：Dense U-Net。

利用英特尔® DLDT 对模型进行转换和优化后，可将优化后的模型导入 GE 医疗 CT 边缘 AI 增强应用中，该应用在英特尔® 至强® 处理器平台和英特尔® MKL-DNN 的基础上，构建了基于边缘的强大推理引擎。

为了验证这一优化方案的实际效能，双方进行了一系列的性能测试，该数据集具有 8,834 个 CT 扫描图像。GE 医疗希望在对模型实施优化后，能够在使用小于 4 个处理器核心的情况下，使推理引擎每秒可处理的图像数量达到 100 张[15]。

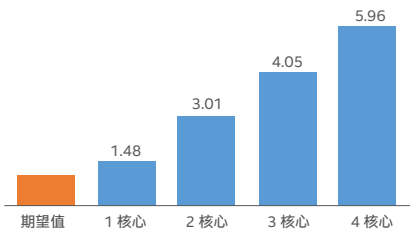


图 1-4-6 多核心带来了推理性能的稳步扩展

测试结果显示，在只启动单核心的英特尔® 至强® 处理器 E5-2650 v4 上，优化后的模型可使推理吞吐量提高到优化前的 14 倍。同时，英特尔® 至强® 处理器的多核心性能，使得 GE 医疗推理引擎的效率获得大幅提升，如图 1-4-6 所示，在使用了 4 个处理器核心后，推理引擎每秒可处理的图像数量提升到了 596 张，近 6 倍于最初的期望值。

小结

医疗图像分割、目标检测是 AI 应用于医疗领域的重要分支。良好的图像分割模型，能有效帮助医疗机构提高医学影像判读效率，进而增强临床诊疗能力、提升疾病治愈率以及减少病患等待时间，弥补因医疗机构影像科资源缺乏带来的多种问题。

与基于 AI 的、在其他图像处理领域的应用不同，医疗领域的图像分割对时效性要求更高，留给病患的黄金诊疗窗口往往只有数十分钟。因此，如果图像分割 AI 应用的推理效率不够高，就有可能延误宝贵的抢救时间。来自多个行业、多个场景的案例显示，英特尔® 至强® 可扩展处理器、第二代英特尔® 至强® 可扩展处理器、英特尔® 深度学习加速技术以及 OpenVINO™ 工具套件等产品和技术，可以有效提升深度学习模型的推理效率。

在东软的应用中，OpenVINO™ 工具套件在保证 U-net 模型高准确率的同时，大幅降低了推理时间，为脑卒中抢救争取宝贵时间。在西门子的应用中，英特尔® 深度学习加速技术对 INT8 良好的支持能力，使推理性能获得进一步加强，满足用户对实时性的要求；而在 GE 医疗的案例中，来自 OpenVINO™ 工具套件的英特尔® DLDT 工具，提供了对多种主流深度学习框架的优化能力，让用户拥有了更多优化方案选择。

值得一提的是，在现有案例的方案中，多是基于既有的英特尔® 至强® 处理器、英特尔® 至强® 可扩展处理器等硬件产品展开。未来，用户可选择性能更强、在 AI 领域有着更突出表现的新一代英特尔产品与技术，例如第二代英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 数据中心级持久内存等，来构建其解决方案。基于这些先进的产品与技术，英特尔将一如既往地推动和探索医疗行业中 AI 应用的创新和落地，使科技更好地服务于人们的健康生活。

[15] 该数据所使用的测试配置为：处理器：英特尔® 至强® 处理器 E5-2650 v4，2.20GHz；核心 / 线程：12/24；HT：ON；Turbo：ON；内存：264GB；硬盘：480GB；操作系统：CentOS Linux 7.4.1708；Linux 内核：3.10.0-693.el7.x86_64；gcc 版本：4.8.5；工作负载：包含了 8,834 个 CT 扫描图像的数据集。

AI + Cloud，协力共建高效医学影像分析能力

医疗领域中的医学影像分析

医学影像分析面临挑战

众所周知，高水平诊疗的前提，是对病情的准确掌握和精准分析。古时，医技高明的大夫以望、闻、问、切来获取和推断病情。今天，通过各类医疗设备和信息系统，尤其是医学影像设备的辅助，医生更能驾驭诊疗过程，为病人提供优质医疗服务。目前，在大中型医疗机构中，X 光机、CT 机、核磁共振等设备已逐渐普及，即便在基层医疗机构，患者也能进行各类医学影像检查。

医学影像设备和系统虽然可以迅速到位，但 “软实力” 却无法一蹴而就。如医学影像分析需要影像科医生拥有较高的专业素养，不仅具备临床医学、医学影像学等方面的专业知识，还必须熟练掌握放射学、CT、核磁共振、超声学等相关技能，同时，还需具备运用各种影像分析技术进行疾病诊断的能力。

因此，虽然医学影像设备在医疗机构已相当普及，但在一些边远地区或基层医疗机构，却常常面临空有设备却无人有能力 “看片” 的尴尬境地。以一些省份为例，很多医学影像设备已部署到县、社区一级的医疗机构，但病人接受检查后，当地医院却依然无法做出精准的判断和分析，需要将影像文件通过拍照、扫描等方式传给上一级医疗机构。有时会因为影像文件的质量得不到保障乃至失真，造成病情的延误或误判。

不仅如此，由于各医疗机构的信息化系统彼此独立，且数据标准未完全统一。例如各个影像归档和通信系统 (Picture Archiving and Communication Systems，PACS) 上存储的医学影像数据几乎没有连通，形成了一个个信息 “孤岛”，这些都会造成偏远地区患者在基层医疗机构得不到有效的病情分析，长途奔波到大医院后，却还需要接受重复检查的怪现象，存在引发医患矛盾的风险。

“云技术 + 大数据” 在医学影像分析中的应用

云计算技术的快速发展，让信息孤岛问题逐渐得以解决，如图 2-1-1 所示，越来越多的医疗机构开始将相关医技设备及医疗服务过程都通过云的方式链接起来，并在其上构建全医技协同平台、影像协同平台等能力和应用，以平台即服务（Platform as a Service，PaaS）或软件即服务（Software as a service，SaaS）的方式满足各层级医疗机构的不同需求。

图 2-1-1 云服务将医技设备聚合起来

以全医技协同服务平台为例，通过接入云服务，各级医疗机构能够获得跨终端、跨平台的全医技功能应用。而影像协同平台则能够让来自大、中型医疗机构的医学影像专家随时随地处理从不同地区传来的影像数据，并对疑难杂症进行协同会诊，来实现医疗资源的高效共享。

以医学影像数据为例，基于云计算和大数据技术的互联互通，不仅让各医疗机构可以规避过度检查、重复治疗等问题，还有力地打破了数据孤岛现象，建立起无边界医疗全连接，提高了医疗服务质量。同时，通过影像数据的积累和分析，也让基于 AI 的医学影像分析应用日趋走向成熟。现在，基于云技术 +AI 的医学影像分析已逐渐在各个医疗机构获得部署，并获得良好反馈。

基于 AI 的医学影像分析

通过云服务和大数据系统汇集的海量数据，让目标侦测神经网络等 AI 模型获得大量的训练样本，令基于 AI 的智能化辅助诊断系统能够更有效地帮助医疗机构提升诊疗能力。

以肺癌早期发现为例，肺癌是令人生畏的恶性肿瘤，而早期肺癌常表现为无症状、易被忽视的肺结节。肺结节的早期确认（良性或恶性）能有效降低肺癌的死亡率。由于微小的肺结节往往难以被人眼及时、准确地发现，因此肺癌一旦发现，往往已是中晚期，导致患者失去了最佳治疗窗口期。

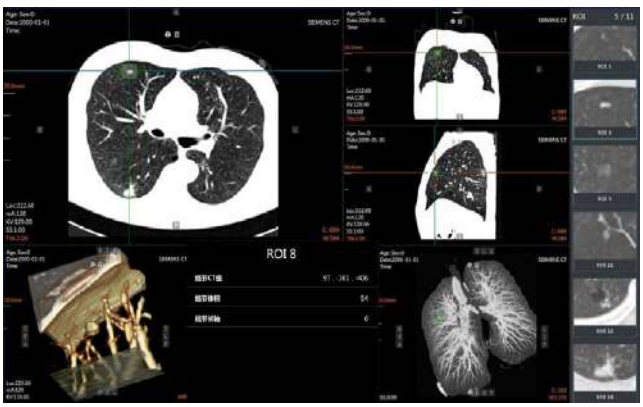


图 2-1-2 利用低剂量 CT 对肺小结节进行的智能化辅助诊断

现在，在医学影像 AI 分析应用中，如图 2-1-2 所示，部分医疗机构正利用低剂量 CT 对肺小结节进行智能化辅助诊断。实践数据显示，其定量的监测敏感度（探测率）已达到 95%，筛查时间也由人工所需的 10 多分钟缩短到秒级[16]。通过 AI 模型识别出肺结节后，再交由医生执行进一步诊断，效率和精准度都获得了大幅提升。

目前，在医学影像 AI 分析应用中，目标侦测神经网络正被广泛地运用，其通过深度学习的方法，能够对 X 光片、CT 成像等医学影像进行高效、准确的病灶检测。

目标侦测神经网络

典型的目标侦测神经网络有 R-CNN、Fast R-CNN、SPP-NET、R-FCN[17] 等。R-FCN 是近年来在医学影像分析领域常见的目标侦测神经网络模型。

一个典型的 R-FCN 结构，如图 2-1-3 所示，首先，对需要处理的影像图片进行预处理操作后，送入一个预先训练好的卷积神经网络（CNN）中，例如 ResNet-101 网络。在该网络最后一个卷积层

[16] 数据援引自盈谷内部测试数据：https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yinggu-case-study-medical.html
[17] R-FCN 相关技术描述，援引自 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

获得的特征地图（feature map）上，会引出 3 个分支。第 1 个分支是将特征地图导入区域生成网络（ Region Proposal Network，RPN），并获得相应的兴趣区域（(Region Of Interest，ROI）。第 2 个分支是在该特征地图上获得一个用于分类的多维位置敏感得分映射（position-sensitive score map）；第 3 个分支就是在该特征地图上获得一个用于回归的多维位置敏感得分映射。最后，在两个位置敏感得分映射上，分别执行位置敏感的 ROI 池化操作（Position-Sensitive Rol Pooling），由此获得对应的类别和位置信息。

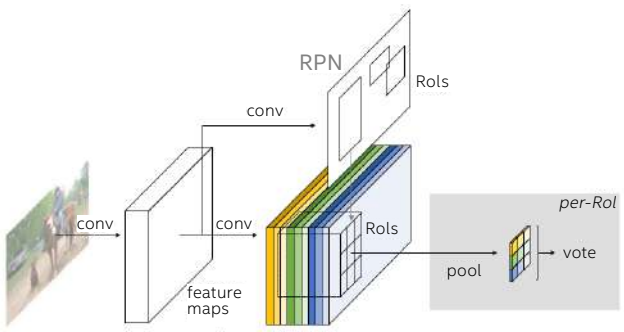


图 2-1-3 典型的 R-FCN 结构

与其他目标侦测神经网络模型，例如 Faster R-CNN 相比，R-FCN 具有检测速度更快，检测精度也更高等特点 [18]。

软硬件配置建议

对于基于 AI 的医疗影像分析方案构建，可以参考以下基于英特尔® 架构平台的软硬件配置来完成。

名称	规格
处理器	英特尔® 至强® 金牌 6240 处理器或更高
超线程	ON
睿频加速	ON
内存	16GB DDR4 2666MHz* 12 及以上
存储	英特尔® 固态盘 D5 P4320 系列及以上
操作系统	CentOS Linux 7.6 或最新版本
Linux核心	3.10.0 或最新版本
编译器	GCC 4.8.5 或最新版本
Caffe版本	面向英特尔® 架构优化的 Caffe 1.1.6 或最新版本

优化 AI 模型效率

基于英特尔® 架构处理器平台的优化

包括英特尔® 至强® 可扩展处理器、第二代英特尔® 至强® 可扩展处理器等在内的英特尔® 架构处理器平台，不仅可为基于 AI + Cloud 的智能医疗影像分析系统带来强大的通用计算能力，更可为其提供亟需的并行计算能力。在深度学习模型的推理过程中，往往对并行计算能力有着较高要求，而英特尔® 至强® 可扩展处理器通过引入英特尔® AVX -512，提供了更高效的单指令多数据流（Single Instruction Multiple Data，SIMD）执行效率，让系统获得了更强大的并行计算加速能力。

同时，英特尔® 数学核心函数库（Intel® Math Kernel Library，英特尔® MKL）、英特尔® MKL-DNN 的加入，可以进一步提升 AI 模型的工作效率，其主要通过以下三个方面来提升人工智能模型性能：

- 使用 Cache Blocking 技术优化数据缓存，提高数据命中率；
- 对神经网络中的常用算子进行并行化与向量化优化；
- 使用 Winograd 算法级优化。

而全新的第二代英特尔® 至强® 可扩展处理器中加入的英特尔® 深度学习加速技术，让深度学习推理可以使用 INT8 来获得更佳的性能表现。

在英特尔® 至强® 可扩展处理器平台上，以单幅胸部 Dicom 数据执行 R-FCN 模型为例，来自某应用的数据表明，英特尔® 至强® 金牌 6148 处理器经过优化，可以把性能提升近 5 倍 [19]。

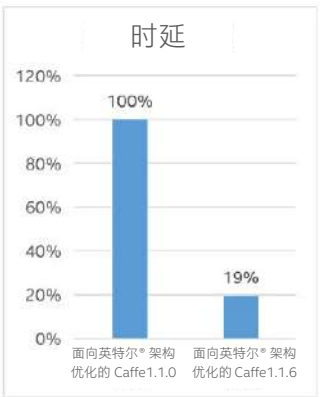


图 2-2-1 单幅胸部 Dicom 数据执行 R-FCN 模型处理延时比较

[18] R-FCN 性能数据，请参阅 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] 性能测试结果基于【2019 年 4 月 10 日】进行的测试，测试配置为：2 路英特尔® 至强® 金牌 6148 处理器，20 核心 /40 线程，启用 HT/Turbo, 搭载 192GB 内存 (12 slots / 16GB / 2666MHz)，CentOS 7.6, BIOS:SE5C620.86B.02.01.0008.031920191559 (uncode:0x200005e)，Kurnel 版本：3.10.0-957.21.3.el7.x86_64，编译器 GCC 4.8.5。测试组使用英特尔® MKL-DNN 0.12 版本，对比组使用英特尔® MKL-DNN 0.18 版本，框架：面向英特尔® 架构优化的 Caffe1.1.0，对比组使用面向英特尔® 架构优化的 Caffe 1.1.6. Minibatch=1 配置下完成。

面向英特尔® 架构优化的Caffe

与伯克利视觉和学习中心 (Berkeley Vision and Learning Center，BLVC) 版本的 Caffe[20] 相比，面向英特尔 ® 架构优化的 Caffe[21] 专门面向英特尔 ® 架构进行了大量优化，并加入了对英特尔 ® MKL、英特尔 ® MKL-DNN 以及英特尔 ® AVX-512 的支持，在各个深度学习模型上都有着更好的性能表现，推理效率也更高。

为了使英特尔® 架构处理器的计算资源得以充分利用，一般在执行推理之前还可以进行一些环境变量的设置，例如：

```
1. export OMP_NUM_THREADS=36
```

这里 OMP_NUM_THREADS 是指定要使用的线程数。

通过对 BLVC Caffe 实施的性能分析，面向英特尔 ® 架构优化的 Caffe 进行了以下几个方面的优化。

■ 代码矢量化优化

优化内容包括：

- 将基本线性代数子程序 (BLAS) 库从自动调优线性代数系统 (ATLAS) 切换至英特尔® MKL-DNN，从而使通用矩阵乘法 (GEMM) 等优化后，更适用于矢量化、多线程化的工作负载，并提高缓存量；
- 使用 Xbyak just-in-time (JIT) 汇编程序执行编译过程。作为一种 x86/x64 JIT 汇编程序，Xbyak 对英特尔® 架构下的指令集，例如 MMX ™ 技术、英特尔® 流式单指令多数据扩展 (Intel® Streaming SIMD Extensions, 英特尔® SSE) 、英特尔® AVX 系列技术等有着更好的支持；同时，还可帮助面向英特尔 ® 架构优化的 Caffe 在代码实施过程中提高矢量化率。
- 对 GNU Compiler Collection (GCC) 和 Open Multi-Processing (OpenMP) 进行代码矢量化。矢量化率的提高，有利于 SIMD 指令同时处理更多数据，提高数据并行利用率。同时，对代码进行矢量化处理，也能有效提升深度学习模型中池化层的性能。

■ 常规代码优化

优化内容包括：

- 降低编程复杂性；
- 减少计算数量；
- 展开循环。

例如在代码优化过程中采用一些标量优化技巧，代码如下：

```
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;}}
```

其代码片段的第三行，关于 h_im 计算，可以将其移出最内层，如下所示：

```
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;}}
```

■ 基于英特尔 ® 架构处理器的其他优化措施

优化内容包括：

- 改进 im2col_cpu/col2im_cpu 执行效率，im2col_cpu 函数是深度学习计算中的常用函数，其能使用优化后的 BLAS 库，以 GEMM 方式执行直接卷积。可对 im2col_cpu 实施以下优化：在 BLVC Caffe 代码中

```
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] = //...
```

其中的四次算术运算 (两次加法和两次乘法)，可替换为单次索引递增运算来提升运算效率；

- 降低归一化批处理的复杂性；
- 特定的处理器 / 系统的优化方法；
- 每个计算线程锁定一个核心，避免线程移动，可设置如下环境变量来实现。

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

通过紧密设置相邻线程，可提高 GEMM 操作性能，因为所有线程都可共享相同的末级高速缓存 (LLC)，从而可将之前预取的缓存行重复用于数据，提高效率。

■ 借助 OpenMP 实现代码并行化

采用 OpenMP 多线程并行处理方法，可以有效提升神经网络的推理效率，例如在池化层中，单一池化层适用于处理单张特征图，但如果池化层与 OpenMP 多线程并行执行，由于图像相互独立，因此多个线程可并行同时处理多个图像，提升效率。代码如下：

```
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_func(bottom_data, top_data, top_count, image, image+1,
7.                      mask, channel, channel+1, this, use_top_mask);
8. }
```

可以看出，借助 collapse(2) clause，OpenMP #pragma omp parallel 可以扩展到两个 for-loop 嵌套语句，再将批量迭代图像和图像通道两个循环合并成一个循环，并对该循环进行并行化处理。

通过一系列的优化方法和技巧，面向英特尔 ® 架构优化的 Caffe 在性能上相较 BLVC Caffe 有了长足的提升。一项测试表明，面向英特尔 ® 架构优化的 Caffe，工作负载执行时间可缩短至原来的 10%，而整体执行性能则提升到原来的 10 倍以上 [22]。

*** 更多面向英特尔 ® 架构优化的 Caffe 的技术细节，请参阅本手册技术篇相关介绍。**

[20] 该版本源代码请详见https://github.com/BVLC/caffe
[21] 该版本源代码请详见https://github.com/intel/caffe
[22] 相关测试数据，以及更多面向英特尔® 架构优化的Caffe的优化方法，请参阅《 Caffe* Optimized for Intel® Architecture: Applying Modern Code Techniques 》： https://software.intel.com/en-us/articles/caffe-optimized-for-intel-architecture-applying-modern-code-techniques。

西安盈谷利用 AI 技术和云服务，提升医学诊疗辅助能力

背景

医疗资源配置的不均衡，使各个医疗机构在医疗影像的后处理、后分析能力上也参差不齐。同时，数据没有互联互通，也使医疗资源的利用效率难以通过资源共享得到有效提升。专注医学影像核心技术近 20 年的西安盈谷网络科技有限公司（以下简称“西安盈谷”），正致力于将其专业医学影像核心技术和产品，与最新的云计算、大数据和 AI 技术结合起来，形成高效、智能的医疗智能化辅助诊断能力，助力广大医疗机构提升诊疗效率及质量。

在西安盈谷看来，要解决医学影像分析处理能力发展不均衡的问题，就必须通过云计算等方式将医学影像数据有效聚合起来，并在其上形成基于 AI 的数据分析能力，进而以资源共享和 AI 两大能力，来逐渐消除各级医疗机构在医学影像分析能力上的差异。

为此，西安盈谷通过医真云的部署，利用创新的医技设备物联网技术 AMOL，将源自不同设备的海量医学影像数据链接起来。同时，西安盈谷还将深度学习引入医学影像处理中，基于目标侦测神经网络模型构建了全新的 Cloud IDT 服务，在提高检出率、降低决策时间、提高工作效率等多个方面都收效显著。

为帮助西安盈谷更好地推动这一系统的部署落地，英特尔为其提供了英特尔® 至强® 可扩展处理器等最新一代平台产品与技术，助其完成了 Cloud IDT 服务向英特尔® 架构平台的迁移，以及对于 Caffe、TensorFlow 等深度学习框架的部署和优化。通过双方的协作和努力，全新的医疗智能化辅助诊断系统已经在筛查时间、准确率等多个指标维度上获得了用户的一致好评。

方案与成效

在新方案中，一方面，西安盈谷基于目标侦测神经网络模型构建了一系列医学影像分析处理应用，并采用英特尔® 架构处理器执行高效率的模型推理；另一方面，西安盈谷也将其 Cloud IDT 智能应用与医学影像处理及分析云计算 @ iMAGES 核心引擎等结合起来，提供了强劲的影像大数据在线智能处理能力。

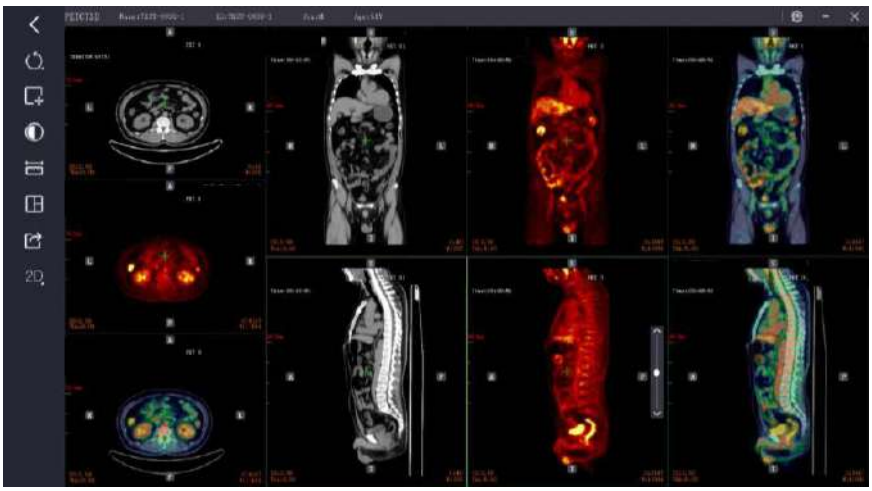


图 2-3-1：云端 PET-CT 融合

如图 2-3-1 所示，结合基于英特尔® 架构的处理器提供的强劲算力，以及 @iMAGES 核心引擎提供的基于云端的强大正电子发射计算机断层显像（Positron Emission Tomography CT，PET-CT）融合能力，不仅能够提供基于形态学和功能的“热力图”，还可以对影像做出半定量化的标准化摄取值（Standard Uptake Value, SUV）分析，而这些影像又可通过 Cloud IDT 智能系统中的 R-FCN 目标侦测神经网络，进一步执行肿瘤等疾病的鉴别和定量分析。

在出色的硬件性能基础上，英特尔还通过对 Caffe、TensorFlow 等人工智能框架的优化，进一步提升了西安盈谷 Cloud IDT 智能系统的执行效率。通过对 R-FCN 模型的优化，模型裁剪融合带来了近 30% 的性能提升，而进一步优化 OpenMP 多线程实现方案后，性能再度提升 40-50%[23]。

此外，英特尔® 至强® 可扩展处理器在通用计算能力和并行计算能力两方面的算力支撑，也可助力智能系统将原先分散在不同平台的任务处理，例如数据统计与模型推理，合并到一起，

[23] 数据援引自盈谷内部测试数据：https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yinggu-case-study-medical.html，所使用的测试配置为：处理器：双路英特尔® 至强® 金牌 6148 处理器，2.40GHz；核心/线程：20/40; HT：ON；Turbo：ON；内存：192GB DDR4 2666；硬盘：英特尔® 固态盘 SC2KB48；网络适配器：英特尔® 以太网聚合网络适配器 XC710；BIOS：SE5C620.86B.02.01.0008.031920191559；操作系统：CentOS Linux 7.6；Linux内核：3.10.0-957.21.3.el7.x86_64；gcc版本：4.8.5；Caffe版本：面向英特尔® 架构优化的Caffe 1.1.6；工作负载：R-FCN。

进而让用户不仅能在其私有云中部署更多的虚拟机，还能带来拥有总成本（Total Cost of Ownership, TCO）的降低。

现在，西安盈谷已基于 AI + Cloud 的模式，构建起肺结节诊断、肋骨骨折诊断、肺结核诊断等一系列智能辅助诊断能力，部分能力如下表所示：

西安盈谷基于AI+Cloud 构建的智能辅助诊断系统能力[24]	
肺结节诊断	在基于大量专家医生标注的胸部CT数据基础上，利用深度学习技术和3D立体图像处理技术，设计特定的深度神经网络和图像算法，可以从胸部CT数据中定位出3mm以上的肺结节，并计算结节大小和结节恶性指标，检测准确率为95%。
肋骨骨折诊断	基于X光胸片数据的全自动智能检测系统，主要方向是肋骨骨折的检测，利用深度学习技术和图像处理技术，自动识别定位骨折并自动将其标记在图像上，检测准确率达90%以上，能够帮助医生快速发现并诊断。
肺结核诊断	自动胸片肺结核检测系统是基于先进的图像处理和人工智能机器学习算法，通过扫描胸片的高分辨率数字影像，自动对其中的可疑病灶点进行检测评分，并简单快速地将检测结果展现出来，敏感度高达 86%，为医疗人员提供了有力的诊断参考。
肺炎AI检测	肺炎AI检测是针对胸部X光片，可以检测出疑似肺炎病灶位置，主要应用于医生日常比较常见的肺部炎症疾病筛选，可帮助医生提高诊断效率，其肺炎检测指标敏感度为82%。
胸部健康片筛查	胸部健康片筛查针对胸部X光片，将正常胸片从所有胸片中挑选出来，可降低筛查工作量，帮助医生只需要把主要精力放到异常数据的明确诊断上即可。此服务主要用于筛查场景（如体检），且胸部健康片筛查指标可以达到敏感性99%、特异性22%。
尘肺病智能筛查	自动胸片肺尘肺检测系统是基于先进的图像处理和人工智能机器学习算法，设计特定的深度神经网络和图像算法，扫描胸片的高分辨率数字影像，自动对其中的可疑病灶点进行检测，并简单快速地将检测结果展现出来，敏感度高达92%。
乳腺癌智能早筛	在基于深度学习大量专家标注的乳腺钼靶摄片数据上，设计特定的深度神经网络，自动识别乳腺钼靶摄片中的异常肿块、钙化灶等病变，钙化斑点、肿块的敏感度大于95%，钙化识别敏感度92%。
糖网增生病变筛查	眼底摄影是基于眼底数据的全自动智能检测系统，主要方向是糖网的检测及分级预测，通过构建数据核查和医学逻辑模块，实现了在眼底图像中识别多种眼底病变和疾病，能够辅助医生有效筛查出早期患者，减少误诊漏诊。
小肠梗阻智能识别	普放肠梗阻数据的全自动智能检测，主要方向是对肠腔内积液面检测，通过设计特定的AI智能算法和图像算法，实现了在腹部立位图像中识别多种梗阻病变和疾病，提供多种形式的筛查，以辅助医生有效甄别急腹症患者，减少误诊漏诊。
CTA冠脉智能诊断	CTA冠脉全自动诊断是针对CTA薄层图像进行智能化处理分析，实现全程自动化的冠脉的分割、分段检测、斑块分类检测、狭窄分析、钙化积分，并结合VR及剖面的可视化辅助，最终实现自动化、结构化报告输出，能够将现有CTA冠脉诊断效率提升近6倍以上。

[24] 盈谷AI应用介绍以及相关数据，源引自盈谷医真云AI官网：http://ai.yizhen.cn/#Page03

小结

以数据驱动医疗信息化的美好明天，是英特尔与西安盈谷等合作伙伴的共同心愿。基于云计算、物联网、大数据以及 AI 等技术领域，针对医疗信息化、智能化的应用目前已经得到了广泛的开展和探索，并在医学影像数据实时计算展现、医学视觉类数据人工智能研究等多个方面都获得了突破，在各个医疗机构的实际部署和实施中都获得了良好的反馈。

为不断挖掘目前主流 AI 框架在基于英特尔® 架构的平台上的潜力，英特尔对这些框架开展了多方位的优化工作。面向英特尔 ® 架构优化的 Caffe 框架通过代码矢量化、借助 OpenMP 并行化等优化手段，使模型整体性能相较 BLVC Caffe 获得巨大提升，在与西安盈谷 Cloud IDT 智能应用、医学影像处理及分析云计算 @ iMAGES 核心引擎等应用结合后，已在肺结节诊断等一大批关键场景中建立起 “AI+Cloud” 的智能辅助诊断系统能力。

随着第二代英特尔 ® 至强 ® 可扩展处理器、英特尔 ® 傲腾™ 数据中心级持久内存等更新一代英特尔技术与产品的涌现，相信基于英特尔 ® 架构平台构建的医疗影像分析解决方案将会输出更强大的性能表现以及更高超的 AI 能力。未来，英特尔还计划与更多合作伙伴继续深入开展合作，将更多、更先进的产品与技术与医疗信息化进程结合起来，推动精准医疗、智慧医疗的前行，让信息化、数字化和智能化更有效地提升医疗服务水平，为患者带去更舒心和贴心的医疗健康服务。

AI技术加速病理切片分析

医疗领域中的病理切片分析

传统病理切片分析方法面临挑战

病理切片是将部分病变组织或脏器，经过一系列处理后形成微米级薄片，粘附在玻片上并进行染色处理，然后再交至病理科，病理科医生通过显微镜对病理切片进行镜检，观察病理变化，并作出病理诊断和预后评估。病理切片检查是一项非常复杂和具有挑战性的工作，而想要成为病理学方面的专家，更是需要具备多年的读片经验与数万张切片的阅片积累以及具有丰富专业知识。然而，据统计，目前全国病理科医生还不足万人[25]。

此外，人工检查不免带有较大主观性，由不同病理科医生对同一患者的病理切片作出的诊断，也经常会存在差异，这可能导致误诊、漏诊等现象产生。同时，在实际的病理切片检查中，患者的病理切片以40倍的放大倍数进行数字化后，单个病理切片的像素点可能超过百万像素。病理科医生需要连续观察多张百万像素级的图片，并且需要注意到图片里微观区域的异常，不仅费时费力，还容易出现错漏。且较长的阅片时间也会导致病患等待时间长，有可能会造成病情的延误。

基于 AI 的病理切片分析方法

随着基于 AI 的图像处理与分析技术获得巨大进步，各个医疗机构均不遗余力地开展了基于深度学习或机器学习的病理切片分析方法，并取得了良好的成效。例如通过 ResNet50 网络进行的深度学习模型训练，可用于执行肿瘤病理组织的辨识工作。尽管其得到的肿瘤预测热学图依然存在噪声等问题，但已经可以像病理科医生一样，以不同的放大倍数来检查病理切片图像。实验表明，医疗机构有可能通过训练一个深度网络模型，使其不仅能够具备专业的检测技术，还能有超快的检测速度和无限的工作时间。

来自纽约大学的一项最新研究成果表明，利用大量数字化病理切片图像训练的 Inception v3 深度学习模型，识别癌组织和正常组织的准确率已达到 99%，区分腺癌和鳞癌的准确率已达到 97%[26]。

现在，基于 CNN 的分类算法以及目标侦测算法都已经获得了长足的发展。作为深度学习的代表方法之一，CNN 的典型代表，例如 LeNet、ZFNet、VGGNet 和 ResNet 等，已经被广泛地运用于图像分类、人像识别、目标定位和图像分析等领域。

[25] 该数据援引自媒体报道：https://www.cn-healthcare.com/article/20141118/content-463705.html
[26] 数据源引自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.

■ 分类卷积神经网络

在医疗图像的检测结果中，往往会出现明显的分类情况，例如阴性为正常，阳性为非正常。可以看出，此时检测所期望的结果，是一系列的离散数字，例如 0 或 1，这就构成了一个典型的分类问题。据此可以认为，利用类似二分类的分类算法，CNN 能够有效帮助医疗机构先初步、定性地筛选出有问题的区域或组织，然后再进行定量的分析和判读。

典型的二分类算法，如逻辑回归，是一种广义的线性回归分析模型。以根据病理切片图片来预测患者是否患有癌症为例，假设随着患者年龄的增加，当发现某种细胞超过 x 个即可判定患有癌症，此时，其在数学上就表现为一个阈值为 x 的线性函数，即 y= 年龄 (n)*a+ 初始值 (b)，当 y>=x 时，判为癌症。

而在实际场景中，这一函数会复杂得多，例如除了年龄以外，异常细胞的大小、状态等也可能成为判断依据，此时，线性函数就会变成一个多元线性函数，例如

y=n*a+m*c+o*d……+b

如前所述，分类问题需要输出一系列离散的结果，因此需要在线性函数上加上一个激活函数，使其输出结果呈离散化。而对于神经网络而言，激活函数的作用是能够给神经网络加入一些非线性因素，使神经网络可以更好地解决较为复杂的问题。常见的激活函数有 Sigmoid 函数、tanh 函数、ReLU 函数等。另外，逻辑回归会采用梯度下降迭代求解的方法，来获取最小化的损失函数。

通常，基于二分类算法的 CNN 图像分类具有以下几个主要模块，如图 3-1-1 所示，包括图像读取与预处理、图像训练、迭代优化和图像预测。其中基于 CNN 的模型训练，由卷积层、池化层以及全连接层等构成，可采用交叉熵损失函数，以及 MBGD 梯度下降算法或 BGD 梯度下降算法。

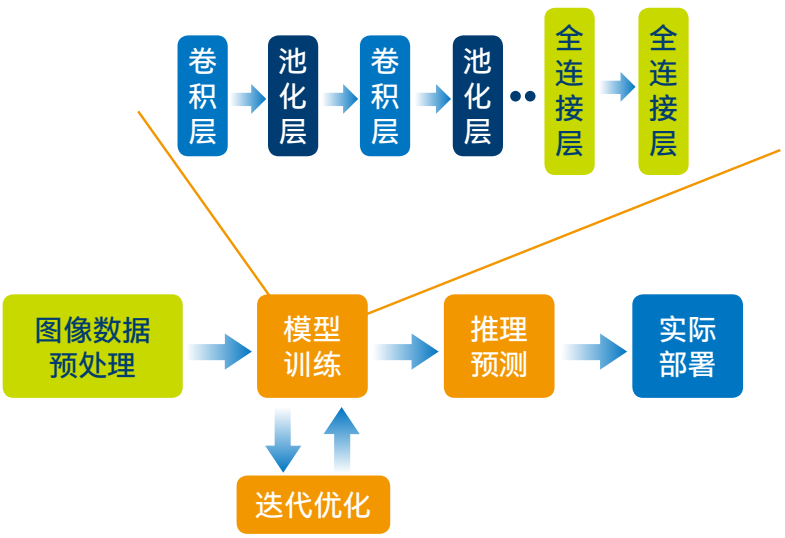


图 3-1-1 基于二分类算法的 CNN 图像分类组成模块

在实际应用中，残差网络 (Residual Net，ResNet) 也是常见的分类卷积神经网络之一，其在 2D 图像分类、检测及定位上有着非常优异的特性。与其他 CNN 相比，ResNet 在网络中增加了直连通道，允许输入信息直接传到后面的层中，如图 3-1-2 所示：

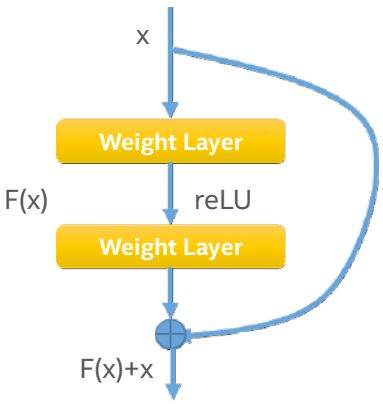


图 3-1-2 ResNet 的残差结构

这一结构 (残差结构) 在一定程度上解决了经典 CNN 网络结构在信息传递时可能存在的信息丢失、损耗，乃至梯度消失等问题，这些问题是深度模型的层数无法变得太多的原因之一。采用 ResNet 后，训练模型的层数可以大幅增加，也由此提高了分类准确率。

■ 目标侦测神经网络

目标侦测神经网络是指在给定的图片中精确找到物体所在位置，并标注出物体的类别。常见的目标侦测神经网络有 R-CNN、Fast R-CNN、SPP-NET、R-FCN 等。

R-CNN 是经典的深度学习目标检测算法，其基本工作流程如下：首先，R-CNN 会基于 selective search 方法在原始图上生成数

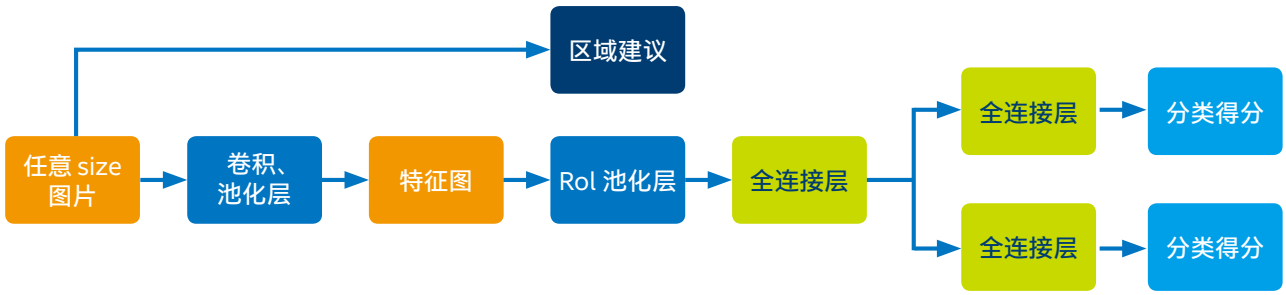


图 3-1-3 Fast R-CNN 网络结构

千个大小一致的候选区域，并输入 CNN 网络。由该网络模型得到的特征向量将通过多类别的支持向量机（Support Vector Machines，SVM）分类器，每个目标都会训练一个 SVM 分类器，并从特征向量中推断其属于该目标的概率。同时，R-CNN 还设置了一个边界框的回归模型来提升定位准确性，通过边框回归模型对边界框的准确位置进行了优化。

为了解决 R-CNN 在实际应用中训练、推理和测试速度较慢，训练所需空间大等问题，Fast R-CNN 采用了以下方法来应对，并获得了比 R-CNN 更好的应用效果。方法为：

- 将整个图像先进行归一化后再送入 CNN 网络；
- 在卷积层不进行候选区域的特征提取，而是在最后一个池化层加入候选区域坐标信息进行特征提取的计算；
- 在 CNN 网络中统一做目标与候选框回归。

而后续的 Faster R-CNN 又将特征抽取 (feature extraction)、proposal 提取，bounding box regression (rect refine)、classification 都整合在了一个网络中，使得综合性能有较大提高，在检测速度方面尤为明显。

软硬件配置建议

对于基于 AI 的病理切片分析方案构建，可以参考以下基于英特尔® 架构平台的软硬件配置来完成。

名称	规格
处理器	英特尔® 至强® 金牌 6240 处理器或更高
超线程	ON
睿频加速	ON
内存	16GB DDR4 2666MHz* 12及以上
存储	英特尔® 固态盘D5 P4320系列及以上
操作系统	CentOS Linux 7.6或最新版本
Linux核心	3.10.0或最新版本
编译器	GCC 4.8.5或最新版本
Caffe版本	面向英特尔® 架构优化的Caffe 1.1.6或最新版本

基于深度学习的病理切片分析方法的优化

基于英特尔® 架构处理器的优化方法

在英特尔® 处理器平台上进行基于深度学习的病理切片分析方法的构建和优化，可以为用户带来以下几个方面的收益：

- 病理切片图像每个文件容量都动辄有数十、上百 MB。传统上，由于存储空间的限制，训练中设定的 Batch Size 都偏小，由此会带来训练时间的增加。而采用基于英特尔® 架构处理器平台，服务器具备了大内存（普遍具备数 TB 乃至数十 TB），可以让 Batch_Size 轻松设置至 100 以上，能够加快训练速度；
- 基于 3D XPoint™ 存储介质构建的英特尔® 傲腾™ 数据中心级持久内存的引入，让至强可扩展平台的优势得到进一步加强。与昂贵的动态随机存取存储器（Dynamic Random-Access Memory，

DRAM）内存相比，英特尔 ® 傲腾™ 数据中心级持久内存大容量和非易失性的特性，及其在实现容量扩展时更低的成本优势，可以有效提升执行模型训练和推理的服务器的内存密度以及计算效率，并大幅降低 TCO;

- 英特尔 ® 至强 ® 可扩展处理器创新的微架构，包括更多数量的核心、更高并发度的线程和更充沛的高速缓存，配合它集成的大量硬件增强技术，特别是英特尔 ® AVX-512 等，都能为 AI 应用提供更强的算力。

*** 更多英特尔 ® 傲腾™ 数据中心级持久内存的技术细节，请参阅本手册技术篇相关介绍。**

面向英特尔® 架构优化的 Caffe

Caffe 是一种常用的深度学习框架，其在视频、图像处理等领域的 AI 训练和推理上有着广泛的运用。为了进一步提升和优化基于 Caffe 的深度学习模型的工作效率，基于英特尔 ® 架构特性，英特尔对 Caffe 进行了大量优化。

这些优化工作包括：

■ 针对典型 ResNet 网络开展的优化

面向英特尔 ® 架构优化的 Caffe 利用 ResNet 系列模型特性，来减少计算和内存访问带来的开销。图 3-2-1 是一种典型的 ResNet 的残差结构，从图左半部可以看出，其底部的 2 个 1*1 卷 Stride-2 卷积层只消耗了一半激活操作。优化方案更改了绑定层设置，如图右半部所示，其将一个 1*1 的池化层加入直连通道，减少了一半的计算量。

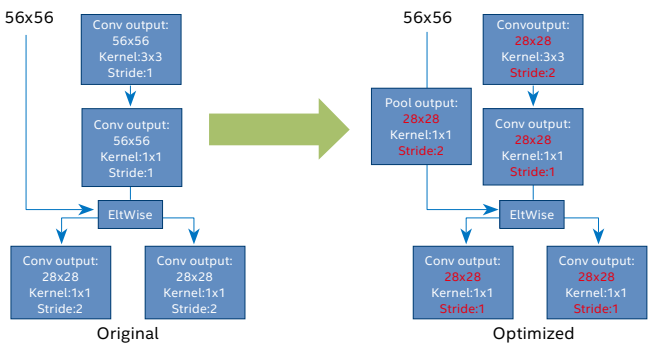


图 3-2-1 面向英特尔® 架构优化的 Caffe 对 ResNet 网络的优化方案

■ 层融合技术

面向英特尔 ® 架构优化的 Caffe 除了针对指令集的向量化、线程级并行进行优化外，还在 Caffe 框架中引入了更为有效的层融合 (Layer Fusion) 优化手段，如 BN+Scale、Conv+Sum、Conv+Relu、BN inplace 以及 sparse fusion, 这些手段使得神经网络，如 ResNet50 的性能获得了极大提升。如图 3-2-2 所示，这是一种残差结构的 Conv 层与 Eltwise 层的融合，图左半部中的卷积层（Conv）res2a_branch2c 和 Eltwise 层 res2a_relu 被融合到一个新的卷积层 res2a_branch2c 中（图右半部所示），有效地提升了 ResNet 类网络模型的性能表现。

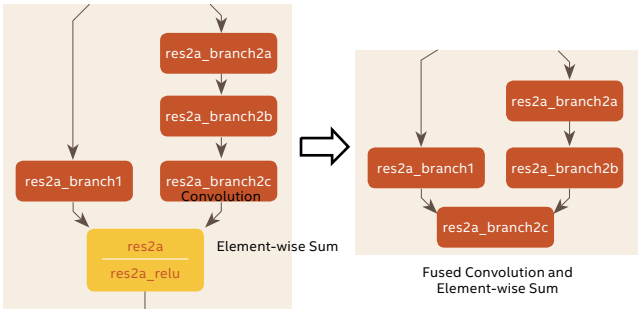


图 3-2-2 Conv 层与 Eltwise 层融合

同时，面向英特尔® 架构优化的 Caffe 还对 INT8 有着良好支持，并提供了 calibration 工具，可帮助用户将神经网络无缝转换到 INT8，以大幅提升性能。

一项测试表明，与使用 BVLC Caffe 相比，面向英特尔® 架构优化的 Caffe 在英特尔® 至强 ® 可扩展处理器上，通过加入层融合技术，使用 ResNet50 卷积神经网络在同等测评环境中执行 AI 推理，如图 3-2-3 所示，单位时间推理性能可提升达前者的 51 倍之多，推理时长则缩短至前者的 4.7%[27]。

[27] 该数据援引自《Highly Efficient 8-bit Low Precision Inference of Convolutional Neural Networks with IntelCaffe》一文：https://arxiv.org/pdf/1805.08691.pdf，测试配置如下：卷积模型：ResNet50，硬件：AWS single-socket c5.18xlarge。

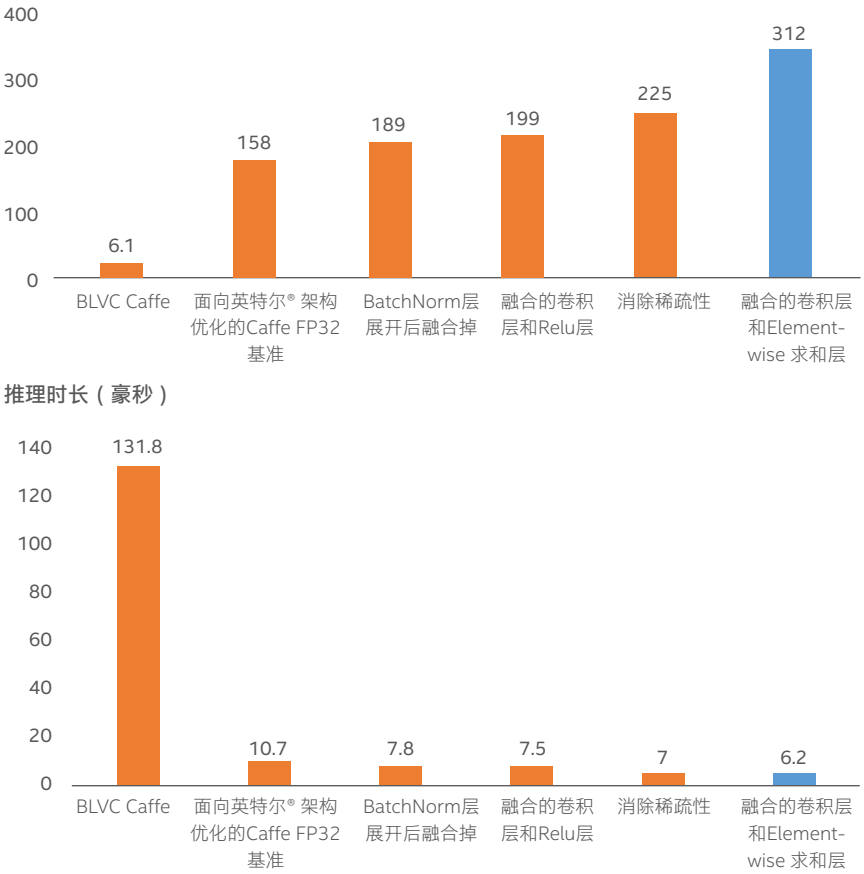


图3-2-3 面向英特尔® 架构优化的Caffe在英特尔® 至强® 可扩展处理器上加入优化方案后，在推理吞吐量和推理时长性能上与BLVC Caffe对比

英特尔® 深度学习加速技术

在全新的第二代英特尔® 至强® 可扩展处理器中，加入了对 INT8 有着良好优化支持的英特尔® 深度学习加速技术，它能够在不影响预测准确率的情况下加速多种深度学习模型在使用 INT8 时的推理速度，有效提升用户深度学习应用的工作效能。

在 图像分类、目标检测等深度学习场景中，采用 INT8 等较低精度的数值替代 FP32 是一种良好的性能优化方案。低精度数值可以更好地使用高速缓存，增加内存数据传输效率，减少带宽瓶颈，且在充分利用计算和存储资源的同时，还能有效降低系统功率。另外，在同样的资源支持下，INT8 还可为深度学习的推理带来更多的每秒操作数（Operations Per Second，OPS）。

英特尔®深度学习加速技术通过 VNNI 指令集，提供了多条全新的 FMA 内核指令，用于支持 8 位或 16 位低精度数值相乘，这对于需要执行大量矩阵乘法的深度学习计算而言尤为重要。它使用户在执行 INT8 推理时，对系统内存的要求最大可减少 75%[28]，而对内存和所需带宽的减少，也加快了低数值精度运算的速度，从而使系统整体性能获得大幅提升。

*** 更多有关英特尔® 至强® 可扩展处理器以及英特尔® 深度学习加速技术的技术细节，请参阅本手册技术篇相关介绍。**

利用工具进行模型准确率优化方法

■ 相似性度量工具

在深度学习中，可以使用相似性度量（Similarity）工具来判断两个特征值之间的相似度。不同的工具可以从不同维度来进行相似性度量，比较常见的有以下几种：

- 欧氏距离 (Euclidean Distance)：是最常见的距离度量，通过对坐标系中的两个点来计算两点之间的绝对距离，距离越大，相似度越低。
- 向量空间余弦相似度 (Cosine Similarity)：使用向量空间中两个向量夹角的余弦值，来衡量两个个体间的差异。与距离度量相比，余弦相似度更加注重两个向量在方向上的差异，夹角越小，相似度越高。
- 标准化欧氏距离 (Standardized Euclidean distance)：是欧氏距离改进版，在计算各个特征的距离之前，需要先将各个分量进行标准化计算。
- 马氏距离 (Mahalanobis distance)：用来表示点与一个分布之间的距离，简单而言，单一样本和哪个样本集距离最近，就属于该样本集。

假阳性预测结果　　真阳性标注

图 3-2-4 利用相似性度量工具分析预测失败原因

[28] 数据源引自 https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

利用相似性度量工具，可以灵活地设计和组合出一系列提升模型训练准确率的方法。例如，通过计算两个特征之间的欧氏距离，来分析预测失败的原因。如图 3-2-4 所示，通过测量假阳性样本在特征抽取层和哪个阳性标注最为接近，可以推导出导致误判的主要原因。

■ 层级相关性传播工具

传统上，深度学习模型各层之间的信息传递和逻辑，一直像一个黑盒一样难以回溯，利用层级相关性传播（Layer-wise Relevance Propagation，LRP）工具可以在一定程度上帮助用户解决这一困惑。LRP 工具是利用计算相关性，将相关性逐层向后传播，具有较好的回溯性。同时，利用这一机制，系统也可以推导出哪些因素对预测结果起到的作用更大，从而提升模型准确率。

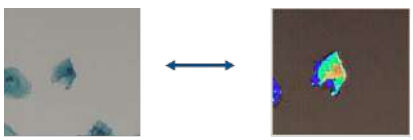


图 3-2-5 利用 LRP 检测不同像素点对于推理效果的作用

如图 3-2-5 所示，在医疗图像分析预测的 AI 应用中，利用 LRP 工具，可以看到不同像素点对于推理结果的效果，并形成热力图，从而帮助方案推导出哪个像素点对最终的预测结果起的作用更大。

江丰生物利用 AI 技术提升宫颈癌筛查效率

背景

宫颈癌是目前严重危害女性健康的恶性肿瘤之一。据统计，在 2018 年的 570,000 例女性癌症患者中，宫颈癌占 6.6%，已经成为女性癌症患者中排名第四的致命疾病[29]。但与此同时，宫颈癌也是唯一一种可以确认致病原因，能够被早期发现并有效预防的癌症病种。宫颈液基细胞学制片（Liquid-Based Cytologic Preparation，LBP）筛查简单易操作，准确率高，可以有效地检测早期癌症病变，帮助早期确诊、及时治疗并阻止癌细胞的进一步扩散。

目前，中国每年都会产生数千万新的宫颈 LBP 涂片，这对医疗机构的病理分析能力构成巨大的挑战。为此，江丰生物与英特尔一起，开始利用先进的 AI 技术，构建和优化基于宫颈 LBP 切片的宫颈癌筛查 AI 解决方案，致力于推动宫颈癌的有效防范与治疗。

目前，有几个因素制约着方案的筛查效率和准确率，使其无法进一步提高。首先是数据标注问题：与其他的医疗数据相比，病理切片的分析数据有其独特之处。如图 3-3-1 所示，病理切片图片会有 1 到 40 倍的不同缩放尺度，缩放尺度较小时，图片基本无法进行标注，而当把图片放到到 20 倍甚至 40 倍时候，只能对整张图片中的很小一部分区域进行人工标注，无法覆盖该切片中的所有问题细胞。

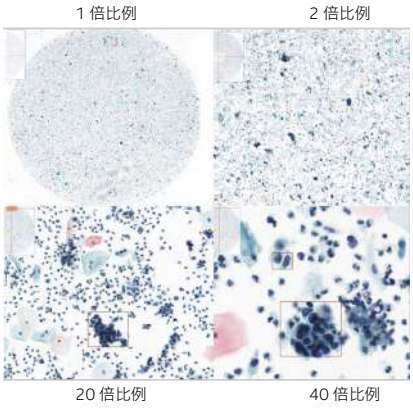


图 3-3-1 不同尺寸的病理切片

此外，在标注过程中，也存在着标注不完整的问题。有时，标注人员只会标注视野中最严重的问题细胞。如图 3-3-2 左侧所示，右下角蓝框中的恶性肿瘤被标注了出来，但未标注左上角的红框中的弱阳性细胞；而图 3-3-2 右侧，则出现了标注位置不够精准的情况。

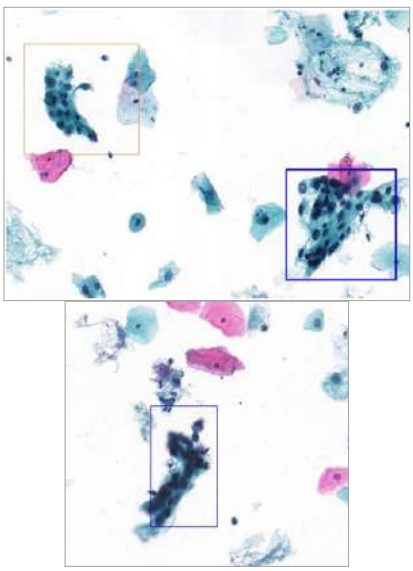


图 3-3-2 标注不够完整的病理切片图片

[29] 引自 WTO 官网 http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

同时，在目前的标注方案中，通常只关注阳性细胞，对于阴性细胞不够重视。即便对阴性细胞进行标注，也只覆盖到切片级别。对于占总量大多数的阴性细胞，没有有效的利用方案。另外，现有的标注样本严重不均衡，非典型鳞状上皮细胞（ASC-US）占绝大部分，而鳞状细胞癌（SCC）、宫内膜、滴虫等样本较少，不利于学习效率的提高。

另一个需要关注的问题是神经网络的选择。从实践的效果来看，目前常用的细胞病变目标侦测网络可以输出病变细胞所在位置矩形坐标以及病变细胞具体的描述性（The Bethesda system，TBS）分级，但单独的目标侦测网络并不能很好地解决标注完整性问题。为解决以上这些问题，江丰生物与英特尔一起，从以下几个维度展开优化，以提升筛查深度学习模型的工作效率：

- 优化数据清理和预处理流程；
- 构建两阶段端到端神经网络；
- 引入模型准确率优化工具。

方案与成效

江丰生物联合英特尔构建的基于宫颈 LBP 切片的宫颈癌筛查 AI 解决方案，主要工作流程如图 3-3-3 所示，系统在输入图片后，经由数据预处理、分类卷积神经网络和后处理阶段，分别得到阳性预测和阴性预测。对于阳性预测，方案则进行第二阶段的目标侦测网络（基于 Resnet50）模型的训练，然后进行阳性识别的推理过程，并交由医生做最终审查。

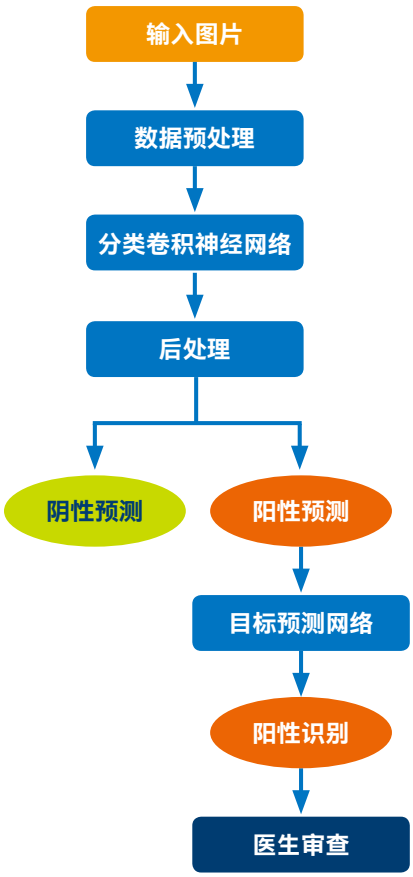


图 3-3-3 优化后的方案流程

在优化数据清理和预处理流程中，针对切片图像的不同缩放尺度问题，方案将切片缩放尺度较大、且阳性标注为细胞 / 细胞块级的病理切片图像，采用从大切片图像上裁剪小图的方式来得到训练数据。而针对切片中样本不均衡的问题，训练集采用了阳性：阴性 =1:5 这一比例，同时，由于阳性标注样本相对较少，方案也对样本进行了旋转，以扩大样本的多样性。

同时，为了提升阴性细胞样本的利用效率，方案假设阴性切片中所有细胞均为阴性细胞，阴性切片的训练集从每一张阴性切片上按比例随机裁剪（目的是除去切片边缘干扰）。而对阳性切片的训练集，则直接根据在阳性切片上标注的坐标中心点，加上合理的随机偏移量裁剪为 512*512 的子图。

为提升识别准确率和效率，方案创新地构建了两阶段端到端神经网络。其中，阶段一为分类卷积神经网络，阶段二为目标侦测神经网络。如图 3-3-4 所示，分类卷积神经网络的主要作用是在每张切片产生的滑动窗上进行二分类推理，并对该切片所有的滑动窗结果进行融合处理，从而得到切片级推理结果。目标侦测网络则是用于对上一阶段确定为阳性的切片进行进一步的阳性区域侦测。

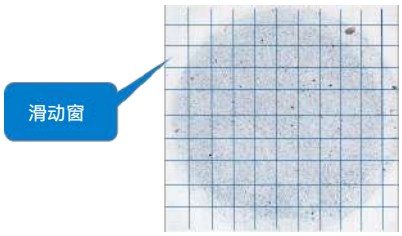


图 3-3-4 基于滑动窗操作的分类卷积神经网络

在模型训练的过程中，方案采用了以下的优化方案来提升训练效果：

- 模型采用了在 Imagenet 数据集上具备优异性能的 ResNet50 来进行训练；
- 训练集准备好后会对其进行旋转，然后按中心点裁剪到 224*224 做均值（Normalize）和归一化（Scale）处理，接下来开始模型训练；
- 鉴于训练集中的正负样本数量较为悬殊，方案将训练好的部分阴性切片和部分阳性切片的子图做集合，递增地加入到训练集中，形成迭代训练。训练集阳性：阴性比为 1:5，从而进一步提升模型的准确率；

- 方案中也加入了相似性度量（Similarity）工具和层级相关性传播（LRP）工具来提升模型准确率。

江丰生物和英特尔一同测评了优化后的基于宫颈 LBP 切片的宫颈癌筛查 AI 解决方案，基于 5,961 张精准标注样本进行了训练，并在 246 张测试集上评估了不同的模型。

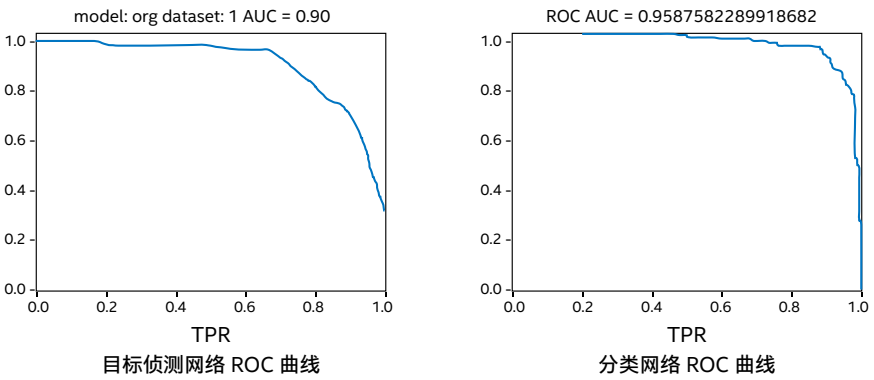


图 3-3-5 优化方案与传统方案准确性对比

评估结果表明，加入分类网络后的优化方案，其准确性比单独的目标侦测网络方案有了大幅提升。如图 3-3-5 所示，可以看出，加入分类网络后，当其敏感度（真阳性率，TPR）为 96% 时，特异度（真阴性率，TNR）接近 70%；而在单独目标侦测网络方案中，特异度仅为 40% 左右[30]，这意味着准确性获得了大幅度的提升[31]。

[30] 该数据援引自江丰生物与英特尔发布的《基于深度学习的病理图像分析》

[31] 数据所使用的测试配置为：双路英特尔® 至强® 铂金 8280 处理器，2.70GHz；核心 / 线程：28/56; HT：ON；Turbo：ON；内存：192GB DDR4 2933；硬盘：英特尔® 固态盘 SC2KG48；网络适配器：英特尔® 以太网网络适配器 X722 for 10GBASE-T ; BIOS：SE5C620.86B.02.01.0003.020220190234；操作系统：CentOS Linux 7.6；Linux 内核：3.10.0-957.el7.x86_64；编译器版本：ICC 18.0.1 20171018；Caffe 版本：面向英特尔® 架构优化的 Caffe 1.1.0；工作负载：Resnet50 with 2 classes，130 张图像每秒。

小结

利用深度学习的方法来对病理切片图像等做出快速检测，不仅可以大大提升医疗机构病理检测的生产力，消弭因专业病理科医生不足带来的一系列问题，也能为病患带去更精确、更及时的治疗方案。目前，基于图像分类和目标检测的病理切片检测 AI 应用，已在众多医疗机构进行了落地部署，并获得良好的反馈。

英特尔 ® 架构处理器平台、面向英特尔 ® 架构优化的 Caffe、英特尔 ® 深度学习加速技术等在内的一系列英特尔先进产品和技术，已在众多应用场景中，助力基于深度学习的病理切片检测应用大幅提升其工作效率。例如英特尔 ® 架构处理器平台对大内存的良好支持，使得在模型训练中可以设定更大的 Batch_Size，从而大幅提升训练效率；再如面向英特尔 ® 架构优化的 Caffe，以及英特尔 ® 深度学习加速技术对 INT8 的良好支持，可以有效提升推理效率，提升病理切片分析的实时性。

虽然本案例涉及的处理器平台为第一代英特尔® 至强® 可扩展处理器，但随着全新的第二代英特尔 ® 至强 ® 可扩展处理器以及其他英特尔新产品、新技术的到来，用户可以基于这些更新的软硬件，来构建训练和推理性能更为强大的 AI 应用。同时，英特尔还计划针对更多的深度学习模型开展推理优化研究，以帮助更多的病患赢得宝贵的治疗时间和效率。

AI 技术助力药物研发

深度学习方法加速药物筛选

基于 HCS 的表型分类

越来越多的新技术正被运用于加速药物研发进程。基于细胞图像的高内涵筛选（High Content Screening，HCS）方法是目前在系统生物学和药物研发领域常用的自动化分析方法之一，也是 AI 技术在药物发现早期环节的重要应用。其通过显微成像法获得的图像信息，来分析和获得由遗传或化学处理诱导的细胞表型特征。

在这一流程中，对细胞图像的表型检测、分析和分类是最重要的几个环节。但生物学分析过程的固有复杂性和细胞测定的固有可变性，对细胞图像中的表型进行分析带来了严峻挑战。传统细胞表型特征提取的图像分析方法主要由一系列独立的数据分析步骤组成。如图 4-1-1 所示，在输入原始图像后，首先利用目标检测（Object Detection）方法，在细胞层级或图像层级上提取特征（Feature Extraction），随后对这些特性进行转换（选择、标准化等），最后是总结归纳相关特征，并作为预测表型的分类算法的输入。

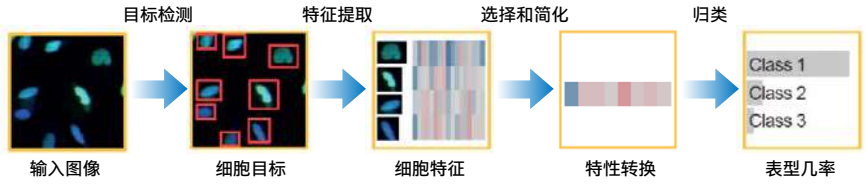


图 4-1-1 传统的 HCS 方法

尽管以上的特征检测、分析和分类方法已经在大量药物研发过程中获得成功应用，但其仍存在许多局限性。例如对于对象分割、降维和表型分类，通常需要大量的先验知识，（例如所预期的表型几何形态（The geometric properties of the expected phenotypes）要对每个测定流程进行定制。同时，采用传统的 HCS 方法，执行每一个步骤，都涉及方法的定制以及参数的调整。而在对整个分析流程的性能调优过程中，如何对所有参数进行联合优化，以达到性能最优化，目前仍面临挑战，因此整体效率还有待提高。为此，更多基于深度学习的 AI 方法正逐渐被引入基于细胞图像的 HCS 表型分类工作。

基于深度学习的 HCS 方法[32]

背景

在传统的 HCS 图像分析方法中，会将图像数据转换为不同的抽象级别，例如像素亮度（Pixel Intensity）等。在深度神经网络等深度学习方法中，可以通过一个框架来对这些图像数据中的分层抽象进行计算和分析，但这些方法在很大程度依赖手动定义的特征。与之相比，CNN

[32] 本节中有关基于 CNN 及 M-CNN 的 HCS 的技术描述，详情请参阅：Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

能够自动地从图像中学习和提取特征，因此在对细胞图像的表型预测中具有更好的效率。

CNN 网络通常包括了输入层、卷积层、ReLU 层、池化层、全连接层等。其中卷积层通过计算层输入（例如原始图像或前一卷积层的输出）和多个二维卷积核之间的卷积，来获得图像中的二维几何信息。每个卷积核都可编码一个几何特征（Geometric Pattern），并可卷积得到一个卷积核映射（或特征映射），该映射是一个基于像素的非线性激活函数，并会被传递到后续的卷积层，获得更复杂的模式。最后，卷积层的输出被送至全连接层，并以前反馈的方式对给定的输入生成预测。

假设CNN的输出层有N_p个待分类的表型，那么对于给定的输入图像 x，网络将在输出层为计算每一路 j 单元的激活函数 a_j (x)，并基于此计算一个向量ρ，ρ _k 可以构成一个概率质量函数，用于覆盖 N_p 个待分类的表型：

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

其中，k 为表型的序号，根据这些概率，可以得到表型的预测值为：

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

由此可知，诸如层数、卷积层内单元数量，以及卷积核和池化因子的大小选择，都会对预测性能带来影响。而在细胞表型分类中，存在着另外一个问题，即由于细胞本身大小不同，显微成像大小不同，导致在图像数据中往往存在着较大的空间差异，此时如果仍沿用经典的 CNN 网络结构，可能会造成准确率的下降。

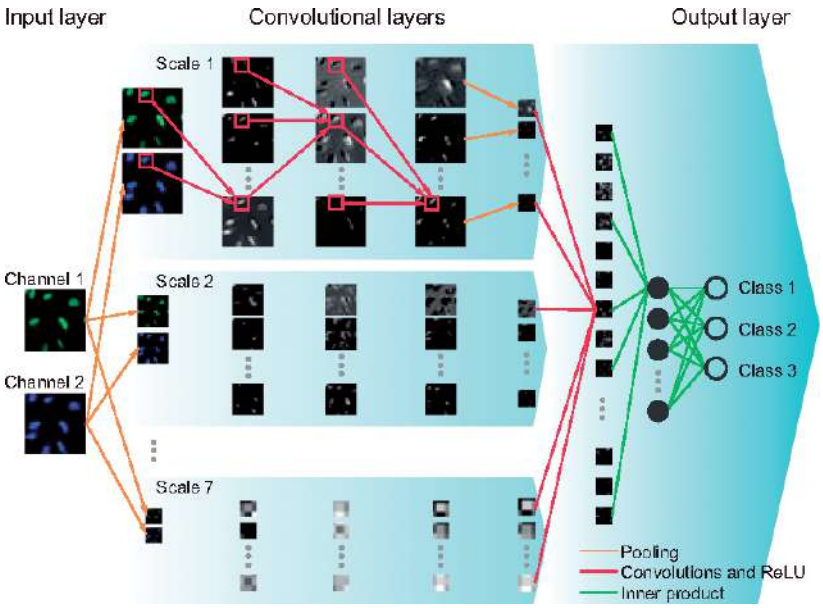


图 4-1-2 M-CNN 架构示意图

多尺度卷积神经网络（Multi-scale Convolutional Neural Networks，M-CNN）可以较好地解决这一问题。与经典 CNN 网络结构相比，其加入了并行的多尺度分析，对于不同尺度上的图像，可以使用不同的 CNN 网络，以独立的方法进行训练。

图 4-1-2 展示了一种具有 7 个尺度的 M-CNN 网络结构，缩放尺寸自上而下逐渐变化。网络在其输入层将输入图像的七个不同尺度的缩放版本，并使用三个卷积层的序列，处理每一个尺度的缩放图像。每个尺度的卷积路径均独立于其他尺度，而在每个尺度的最后一层，都通过汇集方法将得到的卷积核映射缩放到最粗的尺度，并链接起来，用作最终卷积层的输入，最终的输出层将会输出每个表型的生成概率值。

软硬件配置建议

对于利用AI技术来加速药物研发，可以参考以下基于英特尔® 架构平台的软硬件配置，来进行系统部署。

名称	规格
处理器	英特尔® 至强® 金牌 6240 处理器或更高
超线程	ON
睿频加速	ON
内存	16GB DDR4 2666MHz* 12 及以上
存储	英特尔® 固态盘 D5 P4320系列及以上
操作系统	CentOS Linux 7.6或最新版本
Linux核心	3.10.0 或最新版本
编译器	GCC 4.8.5 或最新版本
Tensorflow 版本	面向英特尔® 架构优化的 TensorFlow v1.7.0或最新版本
Horovod	0.12.1或最新版本
OpenMPI	3.0.0或最新版本
ToRSwitch	英特尔® Omni-Path 架构

基于英特尔® 至强® 可扩展处理器的优化

提升单计算节点训练效率

一款新药的研发时间往往长达数年，而其背后常常伴随着患者焦急的等待。为了进一步提升基于 M-CNN 网络模型的 HCS 方法在药物发现工作中的效率，进而让研发得以加速，已经推出了一系列针对英特尔® 至强® 可扩展处理器的优化方案，其包括提升单计算节点吞吐量、提升多计算节点效率等多种方法。

首先，在单计算节点上启动 M-CNN 模型进行训练代码如下：

```
python tf_cnn_benchmarks.py
--model=mcnn
--batch_size=32
--data_format=NCHW
--data_dir=INPUT_DATA_DIR
--data_name=mcnn
--num_intra_threads=40
--num_inter_threads=2
--num_batches=2000
--num_warmup_batches=70
--display_every=5
--momentum=0.9
--weight_decay=0.00005
--optimizer=momentum
--resize_method=bilinear
--distortions=False
--sync_on_finish=True
--device=cpu
--mkl=True
--kmp_affinity=="granularity=fine,compact,1,0"
--variable_update=horovod
```

在单计算节点上，M-CNN 方法遇到的问题之一是内存容量问题。通常而言，深度学习网络的效率可以随着 Batch Size 的增加而有一定程度的提高。用于高内涵筛选的细胞图像通常有着较大尺寸，再加上多尺度联合操作，当 Batch Size 增加到一定量后，所需的内存容量会很大，如图 4-2-1 所示，当 Batch Size 为 32 时，系统所需内存达到了 47.5GB。

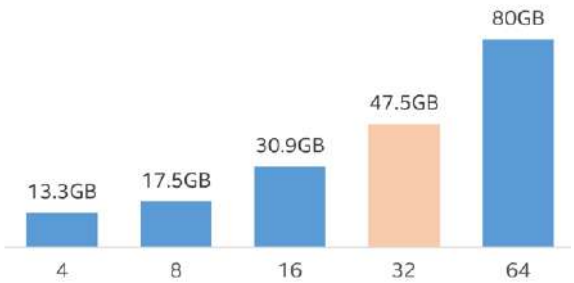


图 4-2-1 不同 Batch Size 下的内存需求量

英特尔® 至强® 可扩展处理器平台对大内存有良好的支持能力，可以有效解决随 Batch Size 增加而带来的大内存需求，其更优化的微架构、更多的核心数量以及对更快、更大容量内存的控制和调度能力，使基于 TensorFlow 框架构建的 M-CNN 方法得以轻松展开。在一项使用 Broad Bioimage Benchmark Collection 021 (BBBC-021) 数据集[33] 所做的测试中，输入的显微镜图像尺寸为 1024*1280*3，在 Batch Size 为 32 时，单一 TensorFlow 工作进程(Worker) 下，处理速度达到 13 张每秒。但这一处理速度对于多达成千上万张图像的数据集而言，整个训练过程仍显漫长，效率亟待提高。

通过非一致性内存访问 (Non Uniform Memory Access Architecture，NUMA) 技术的引入，以及基于分布式深度学习框架Horovod的权重同步技术，可以让用户在TensorFlow框架下，同时使用四个TensorFlow工作进程。如图4-2-2所示，在一个典型的计算节点中部署的双路英特尔® 至强® 可扩展处理器，可以被划分为4个计算区域，每个区域分别执行一个TensorFlow工作进程。

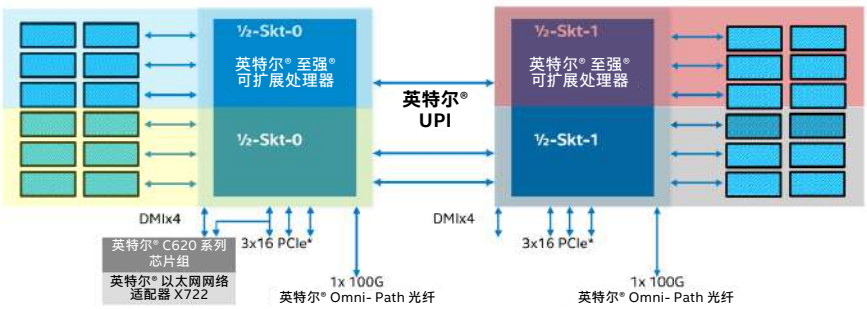


图 4-2-2 典型的计算节点中双路英特尔 ® 至强 ® 可扩展处理器的划分

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

利用 NUMA 的技术特性，可以绑定处理器的不同核心以及不同内存来执行训练，而互相之间不会有计算资源和存储资源的竞争。各个计算区域之间，使用英特尔® 超级通道互联（Intel® Ultra Path Interconnect，英特尔® UPI）技术实现权重同步。通过这种方式，训练模型的吞吐量可获得进一步的提升。如图 4-2-3 所示，右侧使用四个 TensorFlow 工作进程后，在同样 Batch Size 为 32 时，处理速度达到 16.3 张每秒，效率提升达 25.4%。

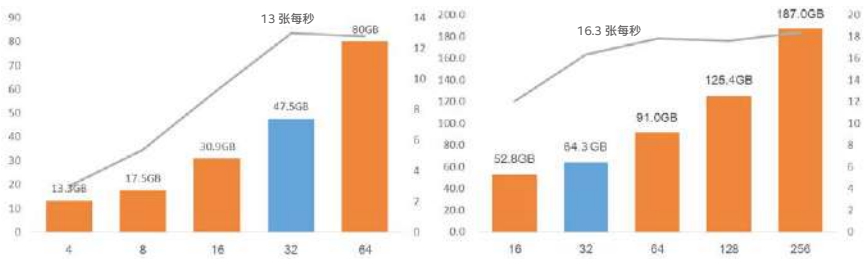


图 4-2-3 TensorFlow 中四个工作线程与单个工作线程性能对比

提升多计算节点训练效率

除了提升单计算节点训练效率之外，利用分布式训练技术方式也可以进一步提升训练效率。在经典的TensorFlow分布式架构中，需要使用参数服务器的方法来平均梯度，每个处理线程都可能作为工作线程或参数服务器。前者用于用户处理和训练数据，计算梯度，并把它们传递到参数服务器上进行平均。

但在这一方法中，如果参数服务器的处理能力不足，可能会造成系统的整体性瓶颈。同时，为了实现最优化性能，使用者在一开始就需要指定合适的初始工作线程和参数服务器，但稍有不慎就会带来性能的下降。新的开源TensorFlow分布式深度学习框架Horovod可以有效解决这一问题。其引入的Ring-allreduce算法构建了新的通信策略，允许工作线程来平均梯度，而无需再加入参数服务器。

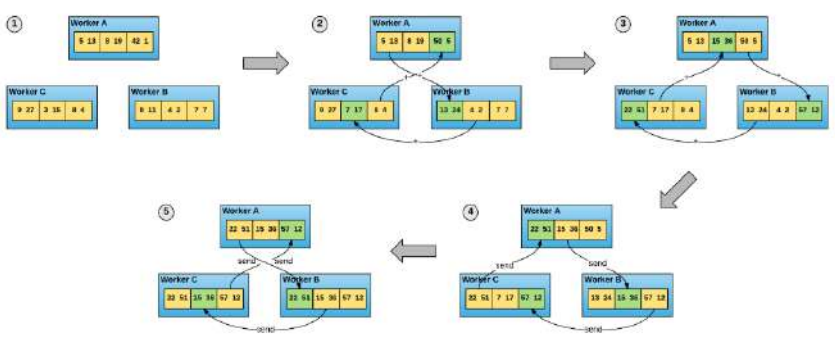


图 4-2-4 Ring-allreduce 算法示意图

如图 4-2-4 所示，在 Ring-allreduce 算法中，每个工作线程首先根据各自的训练数据分别进行梯度计算，得到梯度信息。每个工作线程与其他 N-1 个工作线程进行 2*（N-1）次通信。在这一过程中，一个工作线程发送并接收数据缓冲区传来的梯度信息，每次接收的梯度信息被添加到工作进程缓冲区中，并替代上一次的值。所有的工作线程将在发送和接收 N-1 个梯度消息之后，收到计算更新模型所需的梯度。这一方法可以最大化地利用网络能力，避免计算瓶颈出现 [34]。在此通信策略基础上，Horovod 通过开放消息传递接口（Open Message Passing Interface，OpenMPI）建立基于 TensorFlow 的分布式系统。

即便在采用 Horovod 框架的情况下，所需要传递的梯度信息仍然可观。例如在使用 Broad Bioimage Benchmark Collection* 021 (BBBC-021) 数据集所做的测试中，梯度信息大小为 162.2MB。

英特尔® 至强® 可扩展处理器所支持的英特尔® Omni-Path 架构，可使梯度信息的传递更为迅捷，从而提升 M-CNN 方法的整体训练效率。英特尔® Omni-Path 架构具备 100Gbps 点对点带宽，以及 1us 级的点对点 MPI 通讯延迟；且完全兼容 OFA 软件接口，完全支持 RDMA 以及 PSM 接口，并具有消息包完整性保护、动态链路扩展等革新技术，可为梯度信息的高速传输奠定坚实基础。如图 4-2-5 所示，在 8 个部署了英特尔® 至强® 可扩展处理器的节点中，在使用 Horovod 框架下，同步点传输大于 10Gb。

[34] 相关技术描述详情，请参阅：Alex Sergeev，Mike Del Balso，Meet Horovod: Uber's Open Source Distributed Deep Learning Framework

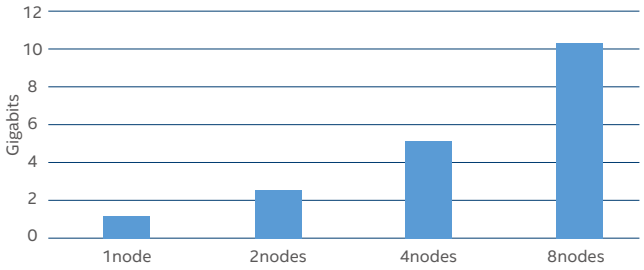


图 4-2-5 使用 Horovod 和英特尔® Omni-Path 架构的同步点传输大于 10Gb

另一个可以对多计算节点训练效率进行优化的方式是收敛和调整学习率（Learning Rate，LR），不同训练阶段的 LR 大小是深度学习中非常重要的设置项，LR 过大会造成振荡，过小则会收敛速度慢且易过拟合。在基于 TensorFlow 框架构建的 M-CNN 模型训练过程中，可以采用如下的 LR 调整方法来获得性能优化。

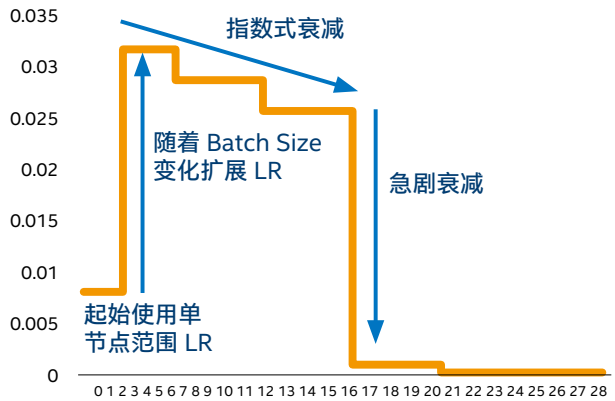


图 4-2-6 M-CNN 网络训练过程中的 LR 调整

如图 4-2-6 所示，在训练之初，首次迭代先使用单节点的 LR，随后将其扩展到全局的 Batch Size 参数。在其后的迭代中，LR 以指数方式衰减，从第 14 次迭代开始，LR 出现一个急剧衰减 [35]。

由此，M-CNN 网络在多计算节点上的训练命令如下：

```
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=mcnn
9.  --batch_size=8
10. --data_format=NCHW
11. --data_dir=INPUT_DATA_DIR
12. --data_name=mcnn
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. --mkl=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;2;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
```

诺华利用深度学习提高药物研发效率

背景

作为全球领先的医药企业，诺华正积极借助数字化转型来保持其在药物创新、疾病诊断和药物研究等方面的竞争优势，而“AI+ 药物发现”是其面向未来药物研发进程中的重要一环。

现在，诺华正与英特尔一起，合作研究使用深度学习的方法来加速 HCS 进程。细胞表型的 HCS 是目前诺华进行早期药物发现的重要方法之一。所谓高内涵是指使用经典图像处理技术，从图像中提取的数千个预定义特征（例如大小、形状、纹理等等）的丰富集合。HCS 允许分析显微图像，以研究数千种遗传或化学处理对不同细胞培养物的影响。利用深度学习方法，诺华可以从数据中“自动”学习，并区分一种治疗与另一种治疗的相关图像特征，但细胞显微镜图像巨大的信息量使这一方法仍需耗费大量时间——其图像分析模型的训练时间约为 11 小时 [36]。

[35] 更多 LR 设置技术详情，请参阅：Yang You et el, 2017, “ImageNet Training in Minutes”

[36] 该数据援引自 https://newsroom.intel.com/news/using-deep-neural-network-acceleration-image-analysis-drug-discovery/#gs.ptk50k

现在，英特尔和诺华的生物学家、数据科学家们希望通过基于优化的英特尔® 至强® 可扩展处理器平台上部署的 M-CNN 网络，来加快 HCS 分析。在这项联合工作中，该团队专注于整个显微镜图像，而不是使用单独的流程来首先识别图像中的每个细胞。而且，其使用的数据集 Broad Bioimage Benchmark Collection* 021 (BBBC-021) 中的显微镜图像可能比常见深度学习数据集中的图像大得多。

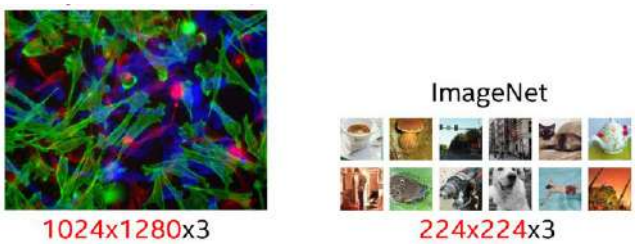


图 4-3-1 用于 HCS 的显微镜图像与常见图像数据集对比

如图 4-3-1 所示，左侧是一个用于 HCS 的显微镜图像，其单张像素接近 400 万，而右侧是来自著名的 ImageNet 数据集 [37] 的图像，其训练数据集单张图像为 15 万像素，双方相差 26 倍。大尺寸的显微镜图像，与其带来的数百万个参数，加之一次训练图像数千个的规模，既对系统内存形成挑战，也带来巨大的计算负荷。为了有效应对这一挑战，双方采用了一系列深度神经网络优化和加速技术，帮助系统能够在更短的时间内处理多个图像，并保持准确率。

方案与成效

优化方案在两个方面对基于英特尔® 至强® 可扩展处理器平台所部署的 M-CNN 模型的训练进行了加速。首先、在单计算节点，充分利用英特尔® 至强® 可扩展处理器平台对大内存的良好支持，使方案可以采用大 Batch_Size(方案中设为 32)，并利用 NUMA 技术增加工作线程来提升训练效率；其次、在多计算节点，引入了开源的 TensorFlow 分布式深度学习框架 Horovod，并结合英特尔® 至强® 可扩展处理器支持的英特尔® Omni-Path 架构，来大幅提升 M-CNN 模型在多节点下的训练效率。同时还设计、采用了优化后的学习率收敛和调整方法来提升性能 [38]。

图 4-3-2 诺华优化后方案的训练效果

方案中部署了 8 个基于英特尔® 至强® 可扩展处理器的节点，使用 Broad Bioimage Benchmark Collection* 021 (BBBC-021) 数据集，图像总量为 1 万张，尺寸为 1024*1280*3。在超过 20 次的训练后，如图 4-3-2 所示，训练时间总长约为 31 分钟，准确率超过 99%。同时，方案在使用 NUMA 技术形成 32 个 TensorFlow 工作进程 (每个节点 4 个工作线程) 后，处理能力达到了每秒 120 多幅图像，与未优化前相比，性能获得了显著提升。

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

[38] 数据所使用的测试配置为：双路英特尔® 至强® 金牌 6148 处理器，2.40GHz；核心/线程：20/40; HT：ON；Turbo：ON；内存：16GB DDR4 2666*12；硬盘：480GB 英特尔® 固态盘 OS drive*1，1.6TB英特尔® 固态盘data drive*1；网络适配器：英特尔®Omni-Path主机结构接口 (HFI)；BIOS：SE5C620.86B.02.01.0008.031920191559；操作系统：CentOS Linux 7.3；gcc版本：6.2；Tensorflow版本：面向英特尔® 架构优化的Tensorflow v1.7.0；Horovod版本：0.12.1；OpenMPI：3.0.0；ToRSwitch：英特尔® Omni-Path架构工作负载：Broad Bioimage Benchmark Collection* 021 (BBBC-021) 数据集，1万张图像，图像尺寸为1024*1280*3。

小结

一款新药从发现、试验到生产，动辄数年，期间伴随着患者及其家属的殷切期待。利用 AI 技术来加速药物研发进程，不仅是众多制药企业加速创新，保持核心竞争力的普遍选择，也是让科技造福人类，助力创造健康生活的重要体现。为此，英特尔也与众多制药企业一起，为加速 AI 方案在药物研发中的应用而努力。

通过合理的优化方案，英特尔® 至强® 可扩展处理器、英特尔® Omni-Path 架构等先进技术与产品，可以为基于深度学习的 HCS 等 AI 应用提供出色且可靠的大内存支持，以及大 Batch_Size 与多 TensorFlow 工作进程支持，来加速单节点或多节点的训练效率，并以高带宽、低延迟的先进互连架色来对 Horovod 分布式训练框架提供支撑，进而大幅加速诺华等药企的药物研发进程。

目前，基于英特尔® 至强® 可扩展处理器平台的一系列 AI 应用，已在众多制药企业获得了落地部署，并产生了良好的效果。值得一提的是，虽然本文中的测试是基于英特尔® 至强® 金牌 6148 处理器平台展开，但随着第二代英特尔® 至强 ® 可扩展处理器、英特尔® 傲腾™ 数据中心级持久内存等新一代英特尔硬件与技术的推出与应用，用户在未来实际部署中可以选用更新的英特尔硬件平台，以及相关软件优化方案来构建性能更强劲的深度学习方案，并获得更佳的训练和推理效果，进而进一步加速药物发现的进程，更好地助力患者治疗与康复。

基于AI的图像识别技术在医疗行业中的应用

智能医疗与图像识别技术

医疗行业中的图像识别技术

越来越多的医疗机构正通过规范的信息系统的建设，例如医院信息管理系统（Hospital Information System，HIS）、临床信息系统（Clinical Information System，CIS）、电子病历（Electronic Medical Record，EMR）以及影像归档和通信系统（Picture Archiving and Communication Systems，PACS）等，来打造更智能的医疗信息化能力，实现患者与医务人员、医疗机构、医疗设备之间的高效互动。高效率的识别技术无疑能够为打通系统，助力智能医疗发挥效能提供更多支持。

比如，传统上，医疗机构使用条码识别、光学字符识别（Optical Character Recognition，OCR）识别以及软件识别等技术来执行对患者身份、药品发放等工作，随着 AI 技术的逐步发展，越来越多的医疗机构开始尝试使用机器学习、深度学习等 AI 方法，来实现患者身份的实时识别，让药品发放更准确，让医疗检查流程实现无缝衔接，进而提升整个系统识别的效率和准确率，增强医疗机构的工作效率。

■ 条码识别

条码识别技术是指利用扫码仪等光电转换设备，对印刷的条形码进行识别。条形码是由一组宽条、窄条和空白组合而成的序列，可用来表示一定的数字和字符。条码识别是目前医疗行业中常见的识别技术之一，条形码可印刷在病历、检测报告和其他物品上，其最主要的优点是可以被准确、快速地识别，系统集成简单。例如，医生在治疗前，用扫描枪扫描病历上贴的条形码，就可通过后台关联的条码库获取病患信息；药房发药时，使用扫描枪扫描药品外包装的条形码，也能马上获悉药品信息。

但条码识别技术也存在缺点。首先，操作者需要对条形码进行逐个扫描，操作速度慢；其次，并不是所有的流程中都有条码，例如一些注射用针剂，瓶身上往往不带条码，护士打针前就需要反复核对。此外，条码库的维护也相当费时费力。

■ OCR 识别

OCR 识别一般是指通过扫描仪等电子设备获取纸面上的字符与符号。通过检测亮暗模式来确定其形状，然后用字符识别的方法将其转化为计算机可识别的字符，其优点在于获取信息范围广泛，可以一次性获取扫描页面上的全部文字信息。在医疗行业，无疑可以采用 OCR 采集病历、检查报告、药品包装等图像，并利用 OCR 组件读取其中的信息。

当然，OCR 识别的缺点也比较明显。首先，OCR 易受角度、光线等条件影响，往往会存在较大识别误差，无法做到 100% 准确识别；其次，OCR 识别只能识别文字（字母、数字、符号等），基本无法识别图像；最后，OCR 识别效率较低，在紧张的诊疗流程中应用，可能会造成一定延误。

■ 软件识别

随着计算机图像技术的不断发展，越来越多的图像处理软件与技术被运用于医疗行业中进行图像和文字识别。例如 OpenCV 计算机视觉库，其可以跨平台实现物体识别、图像分割、人脸识别、文字识别等一系列图像处理与分析工作。通常，用户可以采用 Python、Java 等开发语言，基于 OpenCV 开发自己的识别系统。其优点是识别率高，能够同时识别文字与图片。但定制化开发的软件存在着更新迭代速度慢的弊端，无法针对医疗机构的需求变化迅速做出调整。

基于深度学习的图像识别技术

与传统图像识别技术相比，基于深度学习的图像识别技术准确率和工作效率更高，也更利于形成良好的更新机制。其基于图像特征进行识别，能够一次获取多种类、多数量的图像进行特征识别。目前，各大开源社区都有较成熟的图像识别算法和深度学习框架供参考和调用。

■ 卷积神经网络

卷积神经网络（CNN）是深度学习的代表算法之一，是一种含有卷积计算，且具有多层结构的前馈神经网络。利用卷积神经网络构建的模型，可以方便地对图像进行特征识别并分类。

卷积神经网络对图像的位移、缩放、扭曲具备特征不变性，即泛化性能强劲的。同时。其权值共享结构可以大幅减少神经网络的参数数量，在防止过拟合的同时，又能够降低神经网络模型的复杂度。基于以上特点，在医疗领域物品识别的实际应用场景中，卷积神经网络可以有效规避因为光线、摆放位置等因素造成的影响，提升图像识别准确率，同时复杂度适中，利于用户开展重复训练和学习。有数据显示，2016 年基于神经网络模型的图像识别 top5 错误率已降至 2.991%，低于人类对同类图像识别 5.1% 的错误率 [39]。现在，卷积神经网络的一系列变体，例如 LeNet-5、ZFNet、VGGNet 和 ResNet 等，已经被广泛运用于文本、人像和手势等图像识别和分析领域。

■ 基于 LeNet-5 卷积神经网络的深度学习方法

LeNet-5 卷积神经网络形成了现代卷积神经网络的基本结构，其交替出现的卷积层 - 池化层可以有效提取输入图像的平移不变特征，对位移、缩放和扭曲的图像，例如手写签名，有着良好的特征识别能力。

图 5-1-1 是一个典型 7 层 LeNet-5 卷积神经网络训练模型，除输入输出外，它包含了多个卷积层、池化层和全连接层。

LeNet-5 卷积神经网络在输入层可以接收和处理多维的输入数据，例如二维像素点、RGB 通道等，并进行标准化处理。标准化处理是指将数据输入卷积神经网络前，在通道、时间、频率等维度对输入数据进行归一化计算，这有利于提升模型的运行和学习效率。

在输入层之后，是几个卷积层和池化层。卷积层的主要功能是对输入数据进行特征提取，而且这种特征提取是层层递进的关系。在第一层卷积层，往往只能提取到一些较为简单的特征，而下一个卷积层，则能在这些简单特征的基础上提取更为复杂的特征。池化层的作用是对卷积层输出的特征进行选择和信息过滤。全连接层一般构建在卷积神经网络的最后部分，它可以将特征的 3 维结构转化为向量，并传递至下一层。在最后的输出层，会使用逻辑函数或归一化指数函数输出分类标签。

模型的实现及优化

■ TensorFlow 实现及优化

通过 TensorFlow 来实现 LeNet-5 卷积神经网络，可直接采用官方模型代码来进行模型训练，在 Github 的 slim 目录下已经集成了大量采用 CNN 模型的训练代码，可以直接通过 train_image_classifier.py 来调用 [41]。具体的训练命令如下所示：

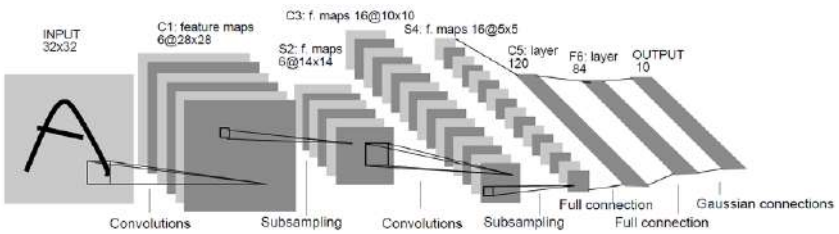


图 5-1-1 典型 7 层 LeNet-5 卷积神经网络训练模型 [40]

[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] 图片以及关于 LeNet-5 的相关描述，援引自 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] 相关 Github 地址为：https://github.com/tensorflow/models/research/slim/

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

同时，为了使英特尔® 处理器的计算资源得以充分利用，还可在训练代码中进行如下优化：

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

其中，intra_op_parallelism_threads 参数和 inter_op_parallelism_threads 参数用来控制每个操作符 op 并行计算的线程个数，前者是控制运算符 op 内部的并行，后者是控制多个运算符 op 之间的并行计算。

同时，在执行 Python 代码之前还可以进行一些环境变量的设置，使英特尔® MKL-DNN 获得最佳状态 [42]，详情如下：

```
export KMP_BLOCKTIME=1
export KMP_SETTINGS=1
export KMP_AFFINITY=granularity=fine,compact,1,0
export OMP_NUM_THREADS=12
```

其中，KMP_BLOCKTIME 设置为 1，是设置某个线程在执行完当前任务并进入休眠之前需要等待的时间，通常设置为 1 毫秒；KMP_SETTINGS 设置为 1，是允许在程序执行期间输出 OpenMP 运行时库环境变量；KMP_AFFINITY 设置为 Compact，是表示在该模式下，线程绑定按计算核心的计算要求优先。先绑定同一个核心，再依次绑定同一个处理器上的下一个核心。此种绑定适用于线程之间具有数据交换或有公共数据的计算情况，优势在于，可以充分利用多级缓存的特性；OMP_NUM_THREADS 是指定要使用的线程数。

*** 更多面向英特尔® 架构优化的 TensorFlow 的技术细节，请参阅本手册技术篇相关介绍。**

软硬件配置建议

对于智能医疗中，基于深度学习的图像识别方案的构建，可以参考以下基于英特尔® 架构平台的软硬件配置来实施。

名称	规格
处理器	英特尔® 至强® 金牌 6240 处理器或更高
超线程	ON
睿频加速	ON
内存	16GB DDR4 2666MHz* 12 及以上
存储	英特尔® 固态盘 D5 P4320 系列及以上
操作系统	CentOS Linux 7.6或最新版本
Linux核心	3.10.0 或最新版本
编译器	GCC 4.8.5 或最新版本
Python版本	Python 3.6 或最新版本
Tensorflow 版本	R1.13.1 或最新版本

解放军总医院利用深度学习技术辅助门诊发药实践

背景

按处方将药品发给患者或用于患者，是药品在医院流通环节的“最后一米”，也是病患获得良好治疗的最重要环节之一。长期以来，医院通过各种制度对最后发药、给药环节进行严格管理，如门诊窗口发药要求“四查十对”，病房药疗医嘱执行要求“三查七对”，药房中还专门设置标牌对易混淆的药品进行标记提醒，如果出现发错药、用错药，都会被记录为医院不良事件，进行上报处理。

管理规定如此严格，却仍然无法杜绝由于各种原因导致的错误，发错或用错药引起的医疗纠纷和事故也屡见不鲜。经分析，错发、错用药品的原因主要包括：

- 管理制度落实不到位。制度制定的比较完善，但是执行力不足；
- 药师或医护人员长时间高负荷工作，工作压力大导致人为出错；
- 药品本身的问题。部分药品名称、外包装设计极易混淆。

解放军总医院虽为知名的三甲医院，但也同样遭遇以上的问题。从 2010 年至 2017 年，该院门诊量年均递增 5.71%，处方发药量年均递增 7.63%，门诊药房有 15.4% 的药品被标记为易混淆药品 [43]，而药房工作人数却基本未变，劳动强度增大提高了发错药的概率和风险。

为应对这一问题，解放军总医院尝试利用信息化手段来辅助减少发药环节的错误。首先，利用计算机视觉技术，在门诊发药窗口对药品的分类和数量进行识别；其次，将该识别系统与 HIS 系统的处方数据进行自动关联和匹配，通过信息比对来判断待发药品实物是否和处方信息一致，并将结果实时反馈给发药的药师，从而达到降低发药出错率的目的。

方案与成效

解放军总医院利用深度学习技术辅助门诊发药解决方案的基本步骤如图 5-2-1 所示：

步骤一：药师会将待发的药品置于发药窗口操作台上；

步骤二：操作台上方的图像采集装置会自动捕获药品图像，并传送到系统后台；

[42] 具体请参见 TensorFlow 官网：https://tensorflow.google.cn/guide/performance/overview?hl=zh-cn

[43] 数据援引自张震江，施华宇，辛海莉，李闯，刘敏超所著《深度学习技术辅助门诊发药实践》一文

步骤三：基于深度学习的药品外包装识别模块，会快速对药品外包装进行识别，并将识别结果显示在电脑屏幕上。

同时，系统会自动与 HIS 处方信息自动关联，对药品名称、规格、厂家和数量等参数分别进行匹配，并将错误信息标记为显著颜色进行提醒。

图 5-2-1 发药窗口应用场景示意图

解放军总医院在方案中，利用 CNN 构建了药品外包装识别模型，并通过深度学习方法对药品外包装进行特征识别。识别包含两个主要目标：一是能够高效准确地识别药品，尤其是易混淆药品；二是能够统计药品数量。

如图 5-2-2 所示，基于神经网络的药品外包装识别模块工作流程主要包含图像数据预处理、模型训练、迭代优化、推理预测等步骤。在图像数据预处理阶段，根据神经网络模型训练需要大量标记数据的特点，方案采用先采集少量原始图片，而后自动生成大量训练用图像数据的模式。

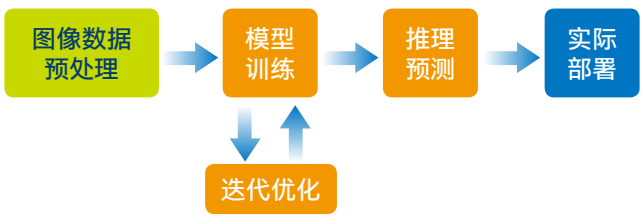


图 5-2-2 基于神经网络的药品外包装识别模块

首先，分别采集药品外包装六个面的原始图片，舍弃含有效期、电子监管码的图像信息；然后，利用轮廓算法获取药品图片以减少干扰；而后，再通过拉伸、扭曲、缩放、旋转、随机位移等图像处理方式得到新图像，并将每个药品的图片以单独目录存储。解放军总医院在方案中针对 56 种药品，采集了 279 个面的原始图像，通过预处理生成了 467,752 张图像，并随机选择 289,448 张用于训练，178,304 张用于模型验证。

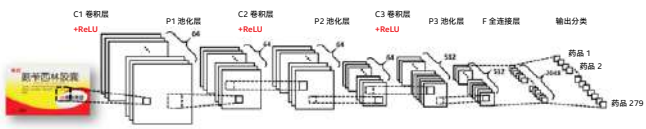


图 5-2-3 用于药品外包装图像识别的深度学习训练模型

根据实际场景需求，解放军总医院在 LeNet-5 卷积神经网络的基础上，如图 5-2-4 所示，设计了一个用于药品外包装图像识别的 7 层卷积神经网络训练模型（包含 3 个卷积层、3 个池化层和 1 个全连接层），对 56 种药品的 279 个分类进行了模型训练。在训练中，系统会对每盒药分别计算出一个预测分类概率值，并设定大于等于 96% 的最大预测概率值所对应的分类为最终预测分类。若全部预测概率值皆小于 96%，则判定为识别失败。整个训练迭代次数为 3,000 次，用时约为 12 个小时，如图 5-2-4 所示，随着训练次数的增加，预测准确率逐渐增高，并最终达到 95.6% 的识别准确率 [44]。

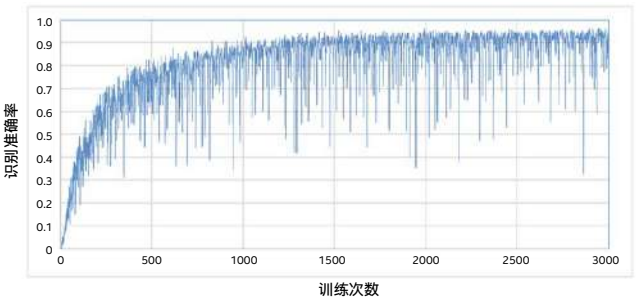


图 5-2-4 识别模型训练次数与识别率变化趋势图

在随后的实际场景验证测试中，解放军总医院对部分极易混淆的药品进行了识别测试。结果显示，药品外包装识别模块均能按照特征

[44] 数据援引自张震江，施华宇，辛海莉，李闯，刘敏超所著《深度学习技术辅助门诊发药实践》一文，所使用的测试配置为：
双路英特尔® 至强® 金牌 6126 处理器，2.60GHz；核心/线程：12/24；HT：ON；Turbo：ON；内存：16GB DDR4 2666*12；
硬盘：英特尔® DC S3320数据中心级固态盘；BIOS：SE5C620.86B.00.01.0013.030920180427；操作系统：CentOS Linux release 7.4.1708 (Core)；
Linux内核：3.10.0-957.12.1.el7.x86_64；gcc版本：7.1；Tensorflow版本：R1.10.0；Python版本：Python 2.7；工作负载：Lenet。

进行准确区分，识别效果如图 5-2-5 所示，左图为沙美特罗替卡松粉吸入剂 50 微克 /250 微克和 50 微克 /100 微克两种规格药品的识别效果，右图为妥布霉素滴眼液和妥布霉素地塞米松滴眼液的识别效果（绿色字体为系统判别结果）。

图 5-2-5 药品外包装识别模块识别测试结果

现在，药品外包装识别模块已在解放军总医院辅助门诊发药系统中得到部署，有效地减轻了药房工作人员的工作强度，同时提升了发药准确率，病患满意度获得大幅提升。

未来，解放军总医院还计划对辅助门诊发药系统进行以下几个方面的优化和改进：

- 业务流程与方案的进一步融合：在目前的方案中，药剂师需要将药品逐一摆放在操作台上，操作耗时较长。如果将整筐药放在操作区，系统即可对采集到的药品图像进行分类和数量统计，无疑可以进一步优化效率，但这需要对轮廓识别算法做进一步的完善；
- 异形药品识别：由于玻璃瓶装的针剂，或圆塑料瓶装的片剂和胶囊不具备盒装药品固定六个平面的特性，在现有方案中进行特征识别有一定难度。下一步，解放军总医院将对异形药品开展更精细的特征识别；
- 提升模型训练效率：目前方案中识别模型仅实现了 56 种药品的分类识别。接下来，解放军总医院计划对全院近 2,000 种药品进行准确分类，模型复杂度也会随之增加。因此需要优化训练效率，缩短训练时间；
- 建立模型更新机制：当医院药品的品类、规格发生更换，或有新进药品时，医院需要对模型重新训练，因此，需要探索一种更快捷的更新机制来提升效率。

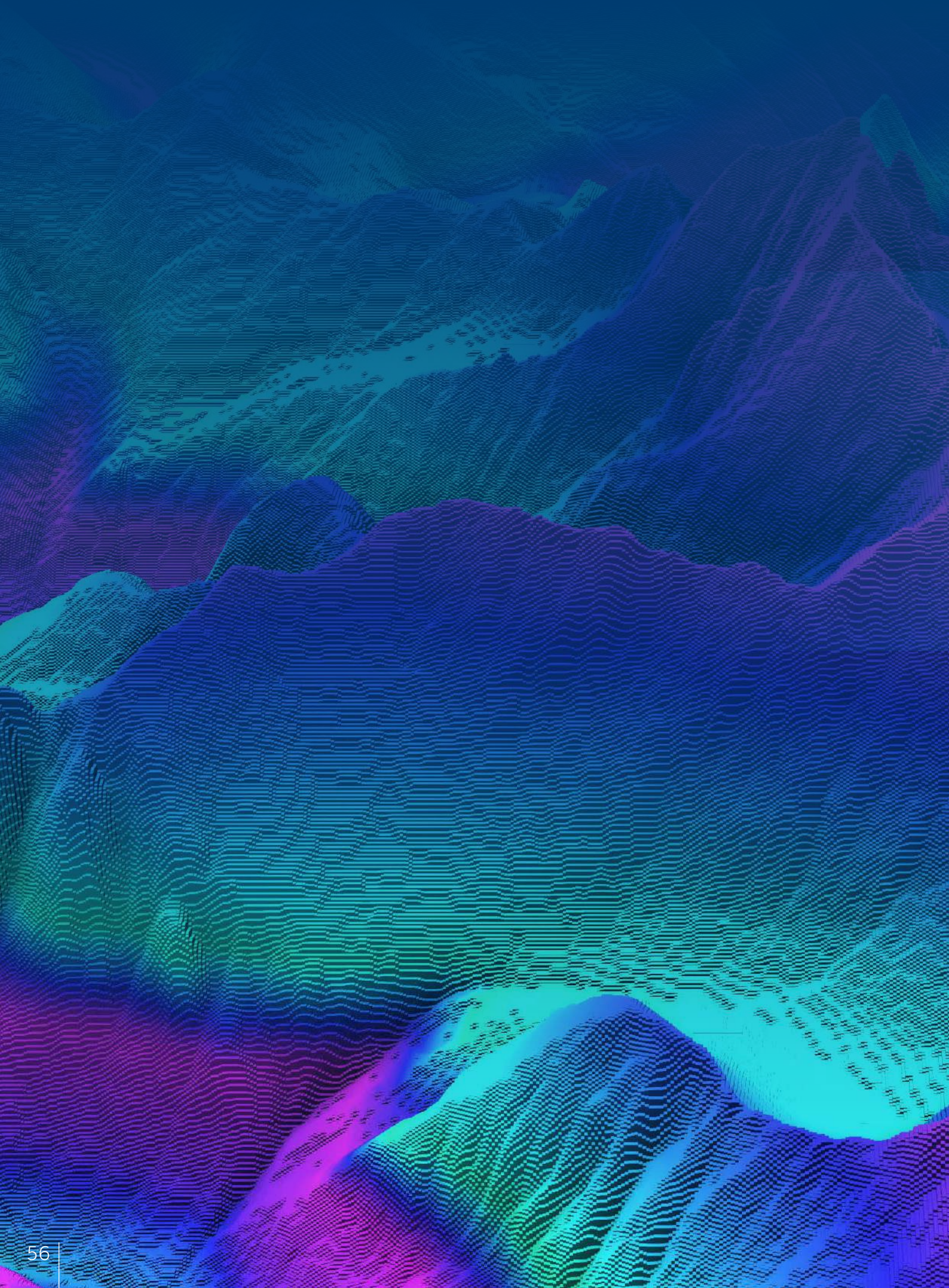
小结

图像识别是支撑医疗行业数字化、智能化转型的关键技术之一，优秀的图像识别解决方案，可以大幅提高医疗机构各信息化系统的信息流转效率，提升智能化水平，为病患创造更好的诊疗体验。目前，利用基于深度学习的识别模块来辅助药品发放等过程，已经被证明可以降低人为因素导致的差错率，具有很好的应用前景。未来，类似的 AI 技术不仅可应用于药房发药等场景，也可在其他医疗场景中发挥巨大作用，例如手术耗材发放管理、病患医疗信息管理等。

包括英特尔® 至强® 可扩展处理器、英特尔® MKL-DNN 等在内的一系列英特尔先进产品和技术，在上述的应用场景中也被证明可以助力多种深度学习模型高效释放潜能。通过英特尔® MKL-DNN 进行优化的 TensorFlow 框架，可以使 LeNet-5 卷积神经网络发挥更优效能，并有效提升它们的训练和推理效率，使 AI 应用的处理效率获得巨大提升。

在解放军总医院的方案中，采用了基于英特尔® 至强® 可扩展处理器、DRAM 内存以及传统英特尔® NAND 固态盘的硬件平台。在未来，用户可选择性能更强，且在 AI 领域有着更多优化举措的第二代英特尔® 至强® 可扩展处理器，以及由英特尔® 傲腾™ 数据中心级持久内存提供的大内存方案，和它们与高性能英特尔® 傲腾™固态盘的组合，来构建性能更卓越的解决方案。

英特尔也将继续推动各类 AI 识别技术在医疗行业中的应用，为智能化医疗体系的构建献技献策。



技术篇

第二代英特尔® 至强® 可扩展处理器

第二代英特尔 ® 至强 ® 可扩展处理器专为数据中心现代化变革而设计，提供比前代产品高出 25%-35% 的性能 [1]，且具备多项新特性，提升了灵活性与安全性，增强了内存性能，能够帮助用户提高各种基础设施、企业应用及技术计算的运行效率，打造性能更强的敏捷服务和更具价值的功能，进而改善总体拥有成本 (Total Cost of Ownership, TCO)，提升生产力。

英特尔® 至强 ® 金牌处理器 6200 系列，特别是主流的英特尔 ® 至强 ® 金牌 6248 处理器、英特尔 ® 至强 ® 金牌 6240 处理器、英特尔 ® 至强 ® 金牌 6230 处理器，作为英特尔 ® 至强 ® 可扩展处理器平台的中流砥柱，能够支持更高的内存速度、增强的内存容量和四路可扩展性，并在性能、高级可靠性和硬件增强型安全技术方面取得了显著改进，且针对要求苛刻的主流数据中心、多云、网络和存储等工作负载进行了优化，能够适应更复杂、更多样化的应用场景。此外，英特尔 ® 至强 ® 金牌 6200 系列首次支持双 FMA 通道，意味着 FMA 性能提升了 2 倍。

此外，第二代英特尔 ® 至强 ® 可扩展处理器集成深度学习加速技术 (矢量神经网络指令 VNNI)，扩展了英特尔 ® AVX-512，赋予平台更多、更强的 AI 能力，可加速人工智能和深度学习推理，并针对工作负载进行了优化。这使其拥有了集成 AI 加速能力的 CPU 架构。基于这一架构，大多数推理工作被集成在工作负载或应用程序中，让用户可以获得加速带来的性能和更高的灵活性等优势，在以数据为中心的时代，帮助在多云与智能边缘之间高效进行无障碍性能切换，以及 AI 开发与应用。

作为新一代至强可扩展平台的 “核心”，第二代英特尔 ® 至强 ® 可扩展处理器支持英特尔 ® 傲腾™ 数据中心级持久内存这一全新产品类别。而英特尔 ® 傲腾™ 数据中心级持久内存通过与第二代英特尔 ® 至强 ® 金牌以及铂金处理器搭配，可以作为 DRAM 内存的有力补充，来显著提高系统性能，加速工作负载处理和服务交付 (具体请参考《英特尔 ® 傲腾™ 数据中心级持久内存》部分)。

功能特性：

- 更高的每核性能：多达 56 核 (9200 系列)和多达 28 核 (8200 系列)，在计算、存储和网络应用中，为计算密集型工作负载提供更高的性能和可扩展性。
- 基于 VNNI 的英特尔 ® 深度学习加速 (英特尔 ® DL Boost) 技术：提高了在 CPU 上运行人工智能推理的表现，与上一代产品相比，性能提升高达 30 倍，有助于从数据中心到边缘，充分支持 AI 部署和应用。
- 业界领先的内存和存储支持，更大的内存带宽 / 容量：支持英特尔 ® 傲腾™ 数据中心级持久内存，与传统 DRAM 结合使用可支持高达 36TB 的系统级内存容量；内存带宽和容量提高 50% [2]，每路支持 6 个内存通道和多达 4TB DDR4 内存，速度高达 2933 MT/s (1 DPC)。还支持英特尔 ® 傲腾™ 数据中心级固态盘和英特尔 ® QLC 3D NAND 固态盘，对于数据密集型的工作负载，突破性的内存和存储内存创新可以显著提高其效率和性能。
- 英特尔 ® Infrastructure Management 技术 (英特尔 ® IMT)：该资源管理框架能够将英特尔的多种能力结合起来，有效支持平台级检测、报告和配置。
- 面向数据中心的英特尔 ® Security Libraries (英特尔 ® SecL-DC)：该软件库和组件实现了基于英特尔硬件的安全功能。

作为至强 ® 平台的创新之作，第二代英特尔 ® 至强 ® 可扩展处理器，基于突破的设计，从平台层面融合计算、内存、存储、网络以及安全等功能，并将它们提升到了新的高度。

[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

适用于人工智能应用的第二代英特尔® 至强® 可扩展处理器

* 仅在特定处理器上受支持。						
	英特尔® 至强® 金牌处理器（6200 系列）	英特尔® 至强®金牌处理器（6200 系列）			英特尔® 至强® 金牌处理器（8200 系列）	英特尔® 至强® 金牌处理器（9200 系列）
		英特尔® 至强® 金牌 6230 处理器	英特尔® 至强® 金牌 6240 处理器	英特尔® 至强® 金牌 6248 处理器		
普适的性能和安全性						
支持的最大内核数	24 核	20 核	18 核	20 核	28 核	56 核
支持的最高频率	4.4 GHz	3.90 GHz	3.90 GHz		4.0 GHz	3.8 GHz
支持的 CPU 路数	多达 4 个				多达 8 个	多达 2 个
英特尔® 超级通道互连（UPI）	3	3	3	3	3	4
英特尔® UPI Speed	10.4 GT/s				10.4 GT/s	10.4 GT/s
英特尔® 高级矢量扩展512（英特尔® AVX-512）	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA	2 FMA
支持的最高内存速度（DDR4）	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s	2933 MT/s
每路支持的最高内存容量*	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 支持	●	●	●	●	●	●
采用矢量神经网络指令（VNNI）的英特尔® 深度学习加速（英特尔® DL Boost）	●	●	●	●	●	●
英特尔® 傲腾™ 数据中心级持久内存模块支持*	●	●	●	●	●	
英特尔® Omni-Path 架构（独立式PCIe* 卡）	●	●	●	●	●	●
英特尔® QuickAssist技术（集成在芯片组中）	●	●	●	●	●	
英特尔® QuickAssist技术（独立式PCIe* 卡）	●	●	●	●	●	●
英特尔® 傲腾™ 数据中心级固态盘	●	●	●	●	●	●
英特尔® 固态盘数据中心家族（3D NAND）	●	●	●	●	●	●
PCIe 3.0	●	●	●	●	●	●
英特尔® QuickData技术（CBDMA）	●	●	●	●	●	●
非透明桥（NTB）	●	●	●	●	●	●
英特尔® 睿频加速技术 2.0	●	●	●	●	●	●
英特尔® 超线程技术（英特尔® HT 技术）	●	●	●	●	●	●
节点控制器支持	●	●	●	●	●	●
* 仅在特定处理器上受支持。						

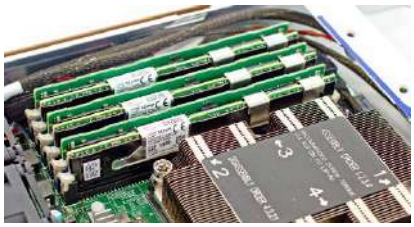
了解更多第二代英特尔 ® 至强 ® 可扩展处理器信息，请访问：

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

英特尔® 傲腾™ 数据中心级持久内存

英特尔® 傲腾™ 数据中心级持久内存是英特尔的革命性产品。其通过创建新的存储层来填补内存 - 存储之间的性能差距，从而颠覆传统的内存 - 存储架构，以合理的价格提供大型持久性内存层级，可为内存密集型工作负载、虚拟机密度和快速存储带来更优的性能，以及更加出色的效率和经济性。进而帮助用户通过比以往更快的分析、云服务、人工智能训练与推理以及下一代通信服务，来加速 IT 转型，满足数据时代的需求。

英特尔® 傲腾™ 数据中心级持久内存兼顾了成本、容量、非易失性和性能等特性，且单一模块可提供 128/256/512 GiB 三种选择，并兼容 DDR4 插槽。在与传统 DDR4 DRAM 内存一同应用在基于第二代英特尔® 至强® 可扩展处理器的平台上时，可以更低的成本在八路系统上实现高达 24TiB 的容量（每路最高可配置容量达 3TiB 的傲腾数据中心级持久内存），进而帮助用户在靠近处理器的内存系统上加载规模远超以往的数据集，满足包括内存数据库、更大规模数据集分析，以及 AI 推理与迭代等对大内存有苛刻要求的应用负载需求，让更多数据的处理和分析走向实时化。



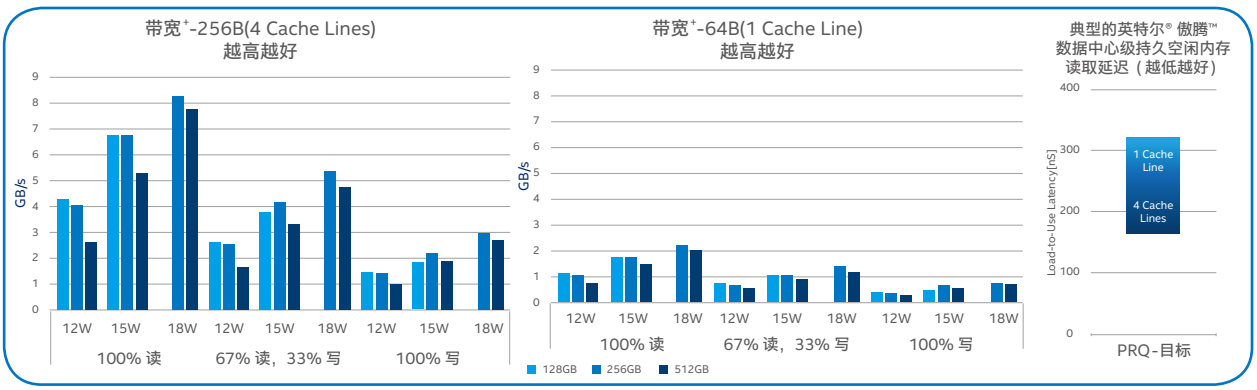
基于将内存和存储特性的融合，傲腾™ 数据中心级持久内存具有两种不同的工作模式：内存模式和 App Direct 模式。通过这两种截然不同的工作模式， 客户可以灵活地跨越多个工作负载，显著提升系统性能。

内存模式。内存模式非常适合提供大容量内存，且不需要对应用进行更改。在内存模式下，CPU 内存控制器会将英特尔® 傲腾™ 数据中心级持久内存用作可寻址的主内存，来扩展操作系统支持的可用易失性内存容量，而将 DRAM 用作其缓存。这可让虚拟化及容器技术直接受益，因为它可以借此以更低的成本在单个物理服务器上提升虚拟机或容器的密度，或为每个虚拟机及容器提供更大的内存容量，且无需重新编写软件。

App Direct 模式。在这一模式下，操作系统会将 DRAM 内存和英特尔® 傲腾™ 数据中心级持久内存视为两个独立的内存池，使得英特尔® 傲腾™ 数据中心级持久内存可以像内存一样寻址，并像存储设备一样具备数据持久性。这一数据持久特性，使得系统即使在维护或重启期间，持久内存也能保留数据，从而能够增加系统的业务弹性，缩短重启时间，并减少由此带来的成本。

双重模式。属于 App Direct 的子集，可通过预配置，让英特尔® 傲腾™ 数据中心级持久内存部分处于内存模式，其余部分则处于 App Direct 模式，进而来满足具备双重需求的工作负载或应用环境需求。

英特尔® 傲腾™ 数据中心级持久内存随其他英特尔® 至强® 可扩展处理器平台产品一并面市，并针对第二代英特尔® 至强® 可扩展处理器做了优化。其独到的应用优势在其上市伊始就吸引了很多 ISV 合作伙伴着手对旗下相关软件进行调优。同时，也有云基础设施及数据分析用户，也导入了英特尔® 傲腾™ 数据中心级持久内存。这些合作伙伴和用户在相对应的应用调优和应用实践中，已初步见证了英特尔® 傲腾™ 数据中心级持久内存在多重模式下的应用价值。



产品家族	英特尔® 傲腾™ 数据中心级持久内存					
外形规格	持久内存模块 (PMM)					
DCPMM SKU[1]	128 GiB		256 GiB		512 GiB	
用户容量	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
产品代码	NMA1XXD128GPSU4	NMA1XXD128GPSUF	NMA1XXD256GPSU4	NMA1XXD256GPSUF	NMA1XXD512GPSU4	NMA1XXD512GPSUF
模型字符串	NMA1XXD128GPS		NMA1XXD256GPS		NMA1XXD512GPS	
技术	英特尔® 傲腾™ 技术					
有限保修	5 年					
平均年故障率 (AFR)	⩽ 0.44					
耐用性 100% 写入 15W 256B	292 PBW		363 PBW		300 PBW	
耐用性 100% 写入 15W 256B	91 PBW		91 PBW		75 PBW	
耐用性 100% 读取 15W 64B	6.8 GB/s		6.8 GB/s		5.3 GB/s	
耐用性 100% 写入 15W 256B	1.85 GB/s		2.3 GB/s		1.89 GB/s	
耐用性 100% 读取 15W 64B	1.7 GB/s		1.75 GB/s		1.4 GB/s	
耐用性 100% 写入 15W 64B	0.45 GB/s		0.58 GB/s		0.47 GB/s	
DDR 频率	2666、2400、2133、1866 MT/s					
最大热设计功耗 (TDP)	15W		18W			
温度 (TJMAX)	⩽ 84℃ (85℃ 关闭, 83℃ 默认) 介质温度					
温度 (环境温度)	10W: 54℃ @ 2.4m/s					
温度 (环境温度)	12W: 49℃ @ 2.4m/s					
温度 (环境温度)	15W: 44℃ @ 2.7m/s					
温度 (环境温度)	N/A		18W: 40℃ @ 3.7m/s			

注： [1]GiB = 2^{30}；GB = 10^9

更多信息请参阅：

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

英特尔® 傲腾™ 固态盘与基于英特尔® QLC 3D NAND 技术的英特尔® 固态盘

英特尔® 傲腾™ 固态盘和采用英特尔® QLC 3D NAND 技术的英特尔® 固态盘以创新的存储架构，助力数据中心面向未来，加速变革与跨越。

作为英特尔在固态盘产品线上的高端成员，英特尔® 傲腾™ 固态盘采用创新的 3D XPoint™ 存储介质，并结合了一系列的先进系统内存控制器、接口硬件和软件技术，实现了在低延迟、高稳定等多方面均有上佳表现，可帮助消除数据中心存储瓶颈，并允许使用更大型、更经济实惠的数据集，进而加快应用程序速度、降低延迟敏感型工作负载的事务处理成本，并改善数据中心的 TCO。英特尔® 傲腾™ 固态盘这一更全面、更优秀，也更为均衡的 IT 基础设施能力，无疑能够为数据密集型的 AI 模型训练和推理带来更高的效率。以英特尔® 傲腾™ 固态盘 DC P4800X 为例，其具有高达 55 万 IOPS 的随机读写能力，低至 10 微秒的读写延迟，可更好地应对多用户、高并发场景，帮助应用方案获得更强的性能表现。同时，其写入寿命（Drive Writes Per Day，DWPD）高达 60，远超 NAND 固态盘，能够赋予存储系统更长的生命周期，也带来了更佳的经济性。

英特尔® 固态盘采用具有突破意义、可信的 3D NAND 技术来提升存储经济性，进而为替代传统硬盘（HDD）提供了性价比更高的选择，能够帮助用户改善体验、提升应用与服务的性能，并降低 TCO。作为固态盘中的“新兴生力军”，英特尔® 固态盘 D5-P4320 系列依托英特尔领先的 64 层 3D NAND 技术，可使得 QLC 固态盘单盘容量达到 7.68TB（TeraByte，万亿字节），从而有效应对数据中心等基础设施用户对“大容量”存储的需求。同时，其随机读取的 IOPS 高达 42.7 万，通过与第二代英特尔® 至强® 可扩展处理器搭配，尤其适用于 AI 训练等应用场景中对于“一写多读”的性能需求，为支持复杂的多样化工作负载提供了高效性、高稳定性和低能耗的存储框架。

了解更多信息，请访问：

- https://www.intel.cn/content/www/cn/zh/products/memory-storage/optane-memory/optane-memory-h10-solid-state-storage.html
- https://www.intel.cn/content/www/cn/zh/products/memory-storage/solid-state-drives/data-center-ssds.html

面向深度神经网络的英特尔® 数学核心函数库

面向深度神经网络的英特尔 ® 数学核心函数库 (Intel® Math Kernel Library, 英特尔 ® MKL-DNN) ，是一款面向深度学习应用的开源性能增强库，也是英特尔为了帮助开发人员充分利用英特尔 ® 架构，推进深度学习的研究和应用而创建的基础库。(源代码地址: https://github.com/intel/mkl-dnn)

英特尔 ® MKL-DNN 作为专为在英特尔 ® 架构上加快深度学习框架的运行速度而设计的一个性能增强库，包含了高度矢量化和线程化的构建模块，支持利用 C 和 C++ 接口实施深度神经网络，具备广泛的深度学习研究、开发和应用生态系统，适用于：Caffe、TensorFlow、PyTorch Apache、 Mxnet、BigDL、CNTK、OpenVINO ™ 工具包等丰富的深度学习软件产品。

为了有效提升深度学习模型在英特尔 ® 架构基础设施上的运行速度，以及提升各类神经网络中其他性能敏感型应用的效率，英特尔 ® MKL-DNN 提供了众多优化的深度学习基元，可应用于不同的深度学习框架，以确保通用构建模块的高效实施。这些模块除了矩阵乘法和卷积以外，还包括：

- 直接批量卷积；
- 内积；
- 池化：最大、最小、平均；
- 标准化：跨通道局部响应归一化 (LRN)，批量归一化；
- 激活：修正线性单元 (ReLU)；
- 数据操作：多维转置 (转换) 、拆分、合并、求和和缩放。

这些高效的函数模块可以应用于广泛的深度学习模型，如：

应用类型	拓扑结构
图像识别	AlexNet, VGG, GoogleNet, ResNet, MobileNet
图像分割	FCN, SegNet, MaskRCNN, U-Net
体积分割	3D-Unet
目标检测	SSD, Faster R-CNN, Yolo
机器翻译	GNMT
语音文字识别	DeepSpeech, WaveNet
对抗网络	DCGAN, 3DGAN
强化学习	A3C

为大幅提升了深度学习在 CPU 上的性能，英特尔还和众多开源社区合作，把英特尔 ® MKL-DNN 集成进多种深度学习框架。如早在 2016 年，经过英特尔 ® MKL-DNN 优化的 Caffe，利用英特尔 ® 至强 ® 处理器 E5-2697 v3，相对于原始的 Caffe 性能获得高达 10 的提升[1]。经过优化后的 ResNet-50 也在英特尔 ® 至强 ® 铂金 9282 处理器上实现了每秒 7736 张图像的领先性能[2]。

英特尔 ® MKL-DNN 目前已成为众多深度学习框架在 CPU 上运行时的基本配置。开发者可在深度学习框架的安装和应用中，直接获得英特尔 ® MKL-DNN 带来的性能提升。

了解更多信息，请访问：

- https://software.intel.com/zh-cn/articles/intel-mkl-dnn-part-1-library-overview-and-installation
- https://software.intel.com/zh-cn/articles/introducing-dnn-primitives-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

面向英特尔® 架构优化的Caffe

面向英特尔 ® 架构优化的 Caffe，从最初版本即集成了当时最新版的英特尔 ® MKL，专门面向当时至强 ® 处理器产品已经集成的英特尔 ® AVX 2 和 英特尔 ® AVX-512 做了优化。它完全兼容 BVLC Caffe，并且包含 BVLC Caffe 的所有优势，且具备处理器优化功能，在英特尔 ® 架构处理器上展现了非常出色的性能，并支持多节点分布程序训练。

面向英特尔 ® 架构优化的 Caffe 支持完备的 Post-training 量化方案，并在大量 CNN 模型中得到实践。特别是在基于第二代英特尔 ® 至强 ® 可扩展处理器的平台（参考处理器介绍章节），利用其集成的、对 INT8 有优化支持的英特尔 ® 深度学习加速技术（VNNI 指令集），可在不影响预测准确度的情况下，使多个深度学习模型在使用 INT8 时的推理速度能够加速到使用 FP32 时的 2-4 倍之多（见下图）[1]，从而大大提升了用户深度学习应用的工作效能。

了解更多信息，请访问：

- https://software.intel.com/en-us/articles/caffe-optimized-for-intel-architecture-applying-modern-code-techniques
- https://software.intel.com/zh-cn/videos/what-is-intel-optimization-for-caffe

优化的深度学习框架和工具包

采用英特尔® 深度学习加速技术的ResNet-50增益

第二代英特尔® 至强® 铂金8280处理器 vs 英特尔® 至强® 铂金8180处理器

英特尔®至强®可扩展处理器	第二代英特尔® 至强®可扩展处理器	mxnet	PYTORCH	TensorFlow	Caffe	OpenVINO
FP32	INT8 W/ Intel® DI Boost	3.0x	3.7x	3.9x	4.0x	3.9x
INT8	INT8 W/ Intel® DI Boost	1.8x	2.1x	1.8x	2.3x	1.9x

[1] 配置信息，请参见：https://www.intel.cn/content/www/cn/zh/benchmarks/server/xeon-scalable/xeon-scalable-artificial-intelligence.html

面向英特尔® 架构优化的TensorFlow

面向英特尔 ® 架构优化的 TensorFlow，是英特尔为了应对在 CPU 上运行深度学习模型面临的性能挑战而推出的优化版，能够确保深度学习类工作负载在各种情况下都可利用英特尔 ® MKL-DNN 基本运算单元高效运行。

为了显著提升性能，英特尔还在持续采用其他举措对 TensorFlow 进行了优化。

计算图优化

英特尔推出了大量计算图优化通道，以便在 CPU 上运行时，将默认的 TensorFlow 操作替换为英特尔优化版本，确保用户能运行现有的 Python 程序，并在不改变神经网络模型的情况下提升性能；同时，消除不必要且昂贵的数据布局转换，以及通过将多个运算融合在一起，确保在 CPU 上高效地重复使用高速缓存，并处理支持快速向后传播的中间状态。

这些计算图优化措施进一步提升了性能，且没有为 TensorFlow 编程人员带来任何额外负担。此外，数据布局优化是一项关键的性能优化措施。此前，将来自 TensorFlow 本地格式的数据布局转换运算插入内部格式，在 CPU 上执行运算，并将运算输出转换回 TensorFlow 格式时，会因为转换造成性能开销的问题。而今利用英特尔 ® MKL 优化运算完全执行的子图，可消除子图运算中的转换。自动插入的转换节点能在子图边界执行数据布局转换，并通过另一个关键优化，即融合通道，将多个运算自动融合为高效运行的单个英特尔 ® MKL 运算。

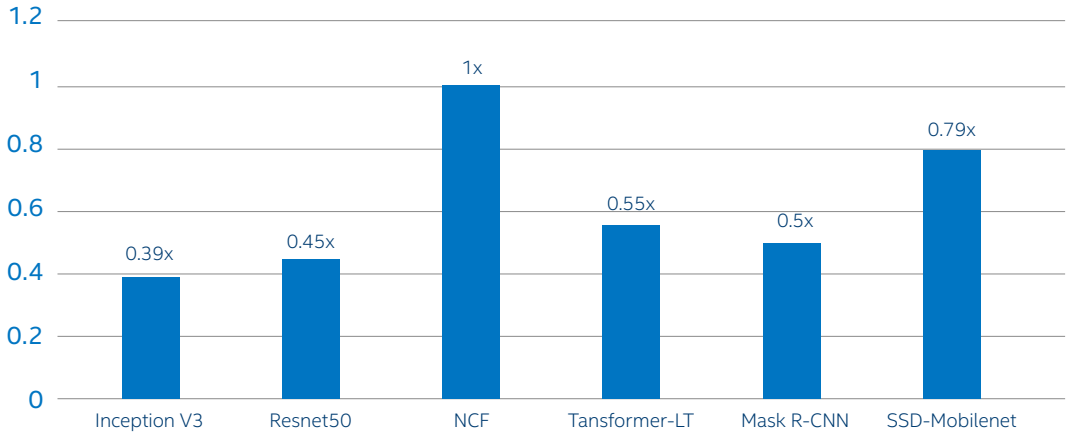
其他优化

为确保在多种深度学习模型上实现最高的 CPU 性能，英特尔有针对性地调整了众多 TensorFlow 框架组件。比如，使用 TensorFlow 中现成的内存池分配器开发了一款自定义内存池分配器，确保其与英特尔 ® MKL 共享相同的内存池（使用英特尔 ® MKL imalloc 功能），从而不必过早地将内存返回至操作系统，避免了昂贵的页面缺失和页面清除。此外，对多个线程库（TensorFlow 使用的 pthread 和英特尔 ® MKL 使用的 OpenMP）页进行了深度优化，以使其能共存，避免了互相争抢 CPU 资源的情况，提升了资源的综合利用率。

了解更多信息，请访问：

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

延时的结果（英特尔® MKL-DNN/ Eigen库）——越低越好



基于英特尔® MKL-DNN 的TensorFlow延时性能[1]

[1] 系统配置：英特尔 ® 至强 ® 铂金 8180 处理器 @2.50GHz; OS: CentOS Linux 7 (Core); TensorFlow 源代码 : https://github.com/tensorflow/tensorflow; TensorFlow 版本号 : 355cc566efd2d86fe71fa9d755ceabe546d577a

面向英特尔® 架构优化的Python分发包

面向英特尔® 架构优化的 Python 分发包，是一个由英特尔主导开发，功能强大的软件开发工具套件，提供了编写 Python 原生扩展所需的一切，包括 C 和 Fortran 编译器、数学库和分析器，并且集成了多个高性能数据分析和数学库，如 NumPy、SciPy、Scikit-learn、Pandas、Jupyter、matplotlib、mpi4py。

英特尔® Python 分发包是英特尔® Parallel Studio XE 的重要工具集之一，具备多项特性与高效率：

- 通过提供开箱即用的 uMath、NumPy、SciPy 和 Scikit-learn 等工具，加速包括数字、科学、数据分析、机器学习等在内的计算密集型应用。
- 集成英特尔® 性能库（如英特尔® MKL），内置最新的矢量化指令，如英特尔® AVX-512 和多线程指令、Numba 和 Cython，并应用对多线程构建模块库英特尔® TBB 的可组合并行，解锁 Python 基于多核处理器的并行应用功能，进而在基于英特尔® 架构的平台上提升 Python 程序运行性能，保证良好的系统兼容性，且不需要对代码进行任何更改。
- 支持 Python2.7、Python3.6 和最新一代英特尔® 架构处理器，提供优化的 TensorFlow、Caffe 等深度学习库和机器学习库，如支持向量机 (SVM) 和 K-means 预测、随机森林和 XGBoost 算法等，便于面向科学计算和机器学习等工作负载构建和扩展生产就绪算法。

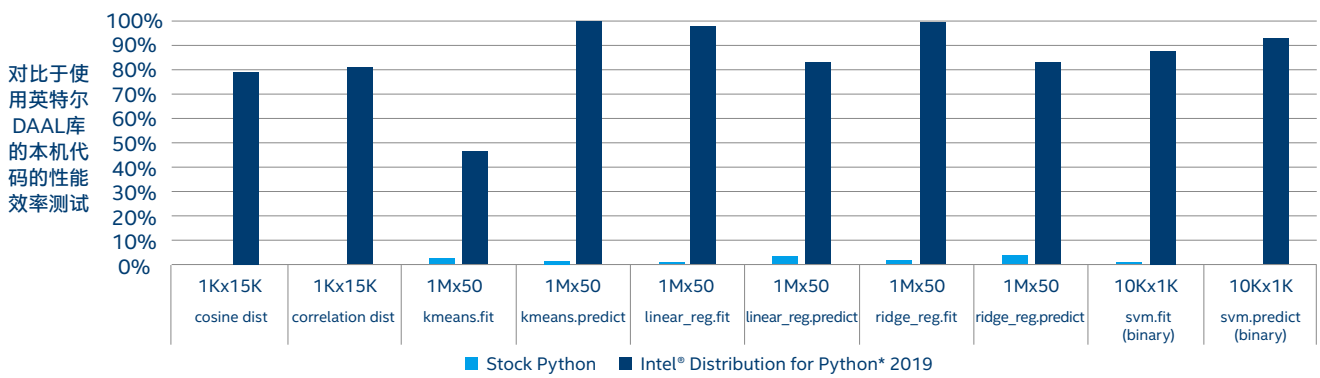
经过基准测试，对比英特尔® Python 分发包与其它开放源码 Python 中 Scikit-learn 工具包（一个广泛用于数值计算、科学计算和机器学习的工具包）的效率，显示面向英特尔® 架构优化的 Python 指标有显著提升（见下图，效率越高，函数速度越快，与本机 C 语言速度越接近），如英特尔® Python 分发包 K-means 聚类、线性回归等算法的效率可以达到英特尔® 数据分析加速库（Intel® Data Analytics Acceleration Library，英特尔® DAAL）中 C 语言效率的 90%。

部署简单、易于使用也是英特尔® Python 分发包的一大特性——使用 Conda 软件包管理器和 Anaconda 云，用户只需执行一个命令，既可安装核心 Python 环境：

- conda install intelpython3 -c intel

此外，英特尔® Python 分发包是预先构建的二进制文件，也可以通过 pip、Docker images、YUM 和 APT repos 等各种渠道获得。

了解更多信息，请访问：
https://software.intel.com/en-us/distribution-for-python



英特尔的优化提升了scikit-learn程序包的效率，使其更加接近于本机代码在英特尔® 至强® 处理器上的运行速度

OpenVINO™ 工具套件

OpenVINO™ 工具套件是英特尔推出的一款加速深度学习推理及部署的软件工具套件，用以加快高性能计算机视觉处理和应用。该工具允许异构执行，支持 Windows 与 Linux 系统，以及 Python/C++ 语言，能够有效推进计算机视觉技术在从智能摄像头、视频监控、机器人，到智能交通、智能医疗等领域的深入应用。

本工具包套件提高了计算机视觉解决方案的性能，缩短了开发时间，简化了从英特尔提供的丰富硬件选项中获得效益的途径，而这些选项可以提高性能、降低功耗并最大化硬件利用率——让用户可以低资源获得高收益，并为新的产品设计提供个性化空间。

通过基于深度卷积神经网络（CNN），扩展英特尔® 硬件（包括加速器）的工作负载，使得 OpenVINO ™ 工具套件可依托英特尔® 架构处理器集成的显卡（Integrated GPU）、FPGA、英特尔® Movidius™ VPU 等芯片，来增强视觉系统的功能和性能。最新发布的 OpenVINO ™ 版本已能支持第二代英特尔® 至强® 可扩展处理器，并通过英特尔® AVX-512 以及采用 VNNI 的英特尔® 深度学习加速（英特尔® DL Boost）技术来提升推理性能，可帮助客户在不改变软件的基础上，快速完成硬件产品升级和算法移植，从而助其在边缘侧快速实现高性能计算机视觉与深度学习应用的开发：

- 在英特尔® 架构平台上提升计算机视觉相关深度学习性能达 19 倍以上[1]；
- 释放 CNN-based 的网络在边缘设备的性能瓶颈；
- 对 OpenCV、OpenXV 视觉库的传统 API 实现加速与优化；
- 基于通用 API 接口在 CPU、GPU、FPGA 等设备上运行；
- 基于英特尔® 平台优化的 OpenVINO™ 工具套件主要包括深度学习部署工具包和传统的计算机视觉工具包两大部分。深度学习部署工具包括模型优化器 (Model Optimizer) 和推理引擎 (Inference Engine) 两个核心组件；
- 模型优化器（Model Optimizer）：将给定的模型转化为标准的中间表示 (Intermediate Representation，IR)，并对模型优化，支持 ONNX、TensorFlow、Caffe、MXNet、Kaldi* 等深度学习框架；
- 推理引擎（Inference Engine）：支持硬件指令集层面的深度学习模型加速运行，支持的硬件设备：CPU、GPU、FPGA、VPU。

提供跨平台工具，支持计算机视觉和深度学习推理加速

同时，对传统的 OpenCV 图像处理库也进行了指令集优化，实现了性能与速度的显著提升。计算机视觉工具包括：

- OpenCV（3.3 版本）：预编译 OpenCV 和全新英特尔 CPU 优化视觉库（Intel Photography Vision Library），具有人脸检测 / 识别、眨眼检测、微笑检测等功能；
- OpenVX：基于图形的 OpenVX 实现，支持传统的 CV 操作和 CNN 原语。 支持 Khronos OpenVX 神经网络扩展 1.2；
- 杂项：包括 OpenCL™ 驱动程序和运行库，以及媒体驱动程序，以简化与英特尔® Media SDK 和英特尔® SDK OpenCL™ 应用程序一起工作的计算机视觉 SDK，英特尔® MKL-DNN、CLDNN 都包含在其中，不需要单独下载。

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

另外，作为旨在快速、有效地在多个应用中实施计算机视觉和深度学习而推出的工具包，基于英特尔® 平台优化的 OpenVINO ™ 工具套件目前提供预先转换的 Caffe、TensorFlow、Mxnet 模型的 MO 文件，如 VGG-16, VGG-19、Squeezenet、Resnet、Inception、CaffeNet、SSD、Faster-RCNN、FCN8 等，还具备超过 20 个预先训练的模型。软件开发人员和数据科学家等可以利用这些工具，快速实现个性化的深度学习应用，且可以使用 OpenCV、OpenVX 的基础库，去创建特定的算法，进行定制化和创新型应用的开发。

OpenVINO™ 工具套件在英特尔平台上让视觉成为现实，已帮助众多用户轻松开发和快速部署计算机视觉应用程序，在多种深度学习应用场景展示了人工智能解决方案所蕴藏的巨大潜力。

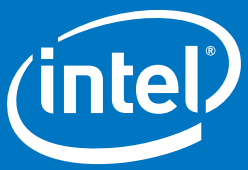
了解更多信息，请访问：

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

本手册涉及的专业词汇表

英文全称	英文缩写	中文全称
Automatically Tuned Linear Algebra Software	ATLAS	自动调优线性代数系统
Basic Linear Algebra Subroutine	BLAS	基本线性代数子程序
Batch Size		批处理参数
Cardiac Magnetic Resonance	CMR	磁共振成像
Clinical Information System	CIS	临床信息系统
Computed Tomography	CT	电子计算机断层扫描
Concat Ops		连接操作
constant folding		常量折叠
Convolution Ops		卷积操作
Convolutional Layer		卷积层
Convolutional Neural Network	CNN	卷积神经网络
Cosine Similarity		余弦相似度
Deep Supervision		深度监督
Deeping Learning	DL	深度学习
Double Data Rate	DDR	双倍速率
Dynamic Random-Access Memory	DRAM	动态随机存取存储器
Electronic Medical Record	EMR	电子病历
Euclidean Distance		欧氏距离
Feature extraction		特征抽取
Feature map		特征地图
Fully Connected Layer		全连接层
Fully Convolutional Network	FCN	全卷积网络
Fused Multiply Add	FMA	宽融合乘加
General matrix multiply	GEMM	通用矩阵乘法
Geometric pattern		几何特征
GNU Compiler Collection	GCC	GNU 编译器套件
High Content screening	HCS	高内涵筛选
Hospital Information System	HIS	医院信息管理系统
Intel® d Vector Extensions	Intel® AVX	英特尔 ® 扩展指令集
Intel® Deep Learning Deployment Toolkit	Intel® DLDT	英特尔 ® DLDT
Intel® Streaming SIMD Extensions	Intel® SSE	英特尔 ® 流式单指令
Intel® Ultra Path Interconnect	Intel® UPI	英特尔 ® 超级通道互联
Last Level Cache	LLC	末级高速缓存
Layer Fusion		层融合
Layer-wise Relevance Propagation	LRP	层级相关性传播
Learning Rate	LR	学习率
Liquid-Based Cytologic Preparation	LBP	宫颈液基细胞学制片
Logistics Regression	LR	逻辑回归

英文全称	英文缩写	中文全称
Magnetic Resonance Imaging	MRI	核磁共振成像
Mahalanobis distance		马氏距离
Math Kernel Library for Deep Neural Networks	MKL-DNN	数学核心函数库
Matrix multiplication		矩阵乘法
Multi-scale Convolutional Neural Networks	M-CNN	多尺度卷积神经网络
Multi-Scale Prediction		多尺度预测
Non- Uniform Memory Access Architecture	NUMA	非统一内存访问架构
Open Message Passing Interface	OpenMPI	开放消息传递接口
Open Multi-Processing	OpenMP	
Open Neural Network Exchange	ONNX	开放神经网络交换
Operations Per Second	OPS	每秒操作数
Optical Character Recognition	OCR	光学字符识别
Picture Archiving and Communication Systems	PACS	影像归档和通信系统
Picture Archiving and Communication Systems	PACS	影像归档和通信系统
Pixel Intensity		像素亮度
Platform as a Service	PaaS	平台即服务
Pooling Layer		池化层
Position-Sensitive RoI Pooling		敏感的 ROI 池化操作
position-sensitive score map		位置敏感得分映射
Positron Emission Tomography CT	PET-CT	正电子发射计算机断层显像
Primitive		多种原语
Region of Interest	ROI	兴趣区域
Region Proposal Network	RPN	区域生成网络
Reorder Ops		重排序操作
Resample Ops		重采样
Residual Net	ResNet	残差网络
Single Instruction Multiple Data (SIMD)	SIMD	单指令多数据流
Sliding Window Algorithm		滑动窗口算法
Software as a service	SaaS	软件即服务
Standard Uptake Value	SUV	标准化摄取值
Standardized Euclidean distance		标准化欧氏距离
Support Vector Machines	SVM	支持向量机
Total Cost of Ownership (TCO)	TCO	总拥有成本



加速AI实践，请访问：

官网
Intel.cn/ai

微博
@ Intel Business

微信
英特尔商用频道