案例研究

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
深度学习

导入英特尔® 深度学习加速技术，百度飞桨 INT8 方案显著提升深度学习推理速度

作为一家持续创新，致力于“用科技让复杂世界更简单”的高科技公司，百度* 正不遗余力地推进人工智能（Artificial Intelligence，AI）最前沿技术的发展与探索，并为各行业与 AI 的无缝对接提供基础支撑。其中，飞桨*（PaddlePaddle*）作百度推出的开源深度学习平台，就能帮助企业与开发者快速、便捷地创建自己的深度学习应用。

随着飞桨在 AI 应用领域的不断拓展，百度发现，AI 训练与推理所需的巨大计算量和部署复杂度，给其在企业市场上的商用部署造成了较高的门槛，对可用性造成了影响。同时，它也看到，在用于图像识别、图像分类等场景的深度学习模型中，INT8 等低精度的定点计算在推理准确度上与传统的 32 位浮点计算相差无几，但在计算速度、功率消耗上却有着更大的优势。

通过与英特尔开展紧密合作，百度在其飞桨深度学习平台中发布了 INT8 离线量化方案。实际验证表明，基于第二代英特尔® 至强® 可扩展处理器平台，并利用其集成的、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，加速到使用 FP32 时的 2-3 倍之多[1]，大大提升了用户深度学习应用的工作效能。

在图像分类等深度学习应用中，使用 INT8 替代 FP32 来提升推理效率、降低功耗和部署复杂度，是目前 AI 技术发展的重要方向。飞桨基于第二代英特尔® 至强® 可扩展处理器的高性能算力以及英特尔® 深度学习加速(VNNI 指令集)的技术，对应的 INT8 方案在不影响推理准确度的情况下，推理速度实现显著的提升。

高铁柱
高级经理
百度深度学习平台部

百度飞桨开源深度学习平台实现的解决方案优势：

- 在图像分类等场景所用的深度学习模型中，采用 INT8 等低精度定点计算方式，可以更高效地利用高速缓存，减少带宽瓶颈，并更大限度地利用计算资源，降低功率消耗；
- 在 ResNet-50* 和 MobileNet-V1* 等多个深度学习模型上的实践表明，基于第二代英特尔® 至强® 可扩展处理器，特别是它所集成的英特尔® 深度学习技术的支持，INT8 可以实现与 FP32 相近的深度学习模型推理准确度，两者差值在 1% 以内[2]；
- 在这些深度学习模型上的实践，同时还表明，基于第二代英特尔® 至强® 可扩展处理器及英特尔® 深度学习技术的支持，INT8 可实现更快的深度学习模型推理速度，其推理速度约为 FP32 的 2.2 倍~2.79 倍[3]。

飞桨深度学习框架

作为一个易学易用、安全高效的分布式深度学习平台，百度飞桨具备优异的训练和推理性能。它在核心框架基础上，提供了 VisualDL*、PARL*、AutoDL*、EasyDL* 以及 AI Studio* 等全流程深度学习工具组件和服务平台，能够满足不同开发者和企业学习、开发及部署 AI 应用与服务的需求。目前，该框架已在智能推荐、机器视觉、自然语言处理等一系列 AI 应用场景的探索中，获取了丰硕的实践成果，得到了众多 AI 开发者的青睐。

为了进一步提升飞桨在商业部署中的可用性，百度团队从多个维度着手，进一步探索提升其工作效率、降低其部署门槛和复杂度的“捷径”。传统上，深度学习的训练和推理过程都采用精度较高的浮点数值，例如 FP32。高精度数值意味着 AI 系统会承载更大的计算、存储压力，有更高的功率消耗以及更复杂的系统设计。而随着 AI 应用部署的“战场”越来越广，网络边缘节点乃至移动终端都开始接入 AI 应用，这对 AI 系统的计算、存储能力和功耗提出了更为苛刻的要求。为此，飞桨也在寻求更多样、更适合的方式来提升自身的可用性和资源利用效率。

在图像识别、图像分类等场景的深度学习场景中，采用 INT8 等较低精度的数值替代 FP32 是一种可行的方案。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，从而能够更为充分地利用计算和存储资源，并降低系统功率。这意味着，在同样资源的支持下，INT8 可为深度学习的推理带来更多的每秒操作数（Operations Per Second，OPS）。

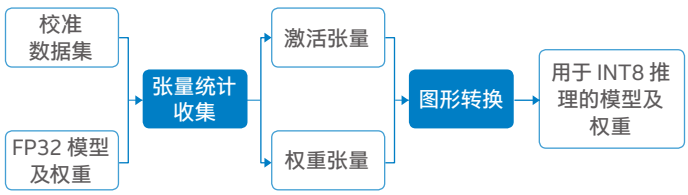
为帮助用户便捷地采用 INT8 构建其深度学习应用，并取得与 FP32 同样优异的推理性能，百度与英特尔开展了深入的技术合作，利用第二代英特尔® 至强® 可扩展处理器集成的英特尔® 深度学习加速技术对 INT8 更优的支持，在最新版本的飞桨中，为用户提供了离线 INT8 深度学习推理能力，并在随后进行的一系列验证测试中，印证了这一方案的应用优势。

更具效能的 INT8 推理方案

与 FP32 相比，INT8 有着更小的数值精度和动态范围，因此在深度学习中采用 INT8 推理方式，需要着重解决计算执行时的信息损失问题。在飞桨新版中，离线 INT8 推理功能通过校准的方式来形成待推理的 INT8 模型，实现了在将 FP32 转换为INT8 时尽可能减少其信息损失的目标。

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尽可能减少信息损失的关键。在 FP32 向 INT8 映射的过程中，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。

基于以上概念，如图一所示，飞桨将校准过程分成了张量统计收集和图形转换两个步骤。前者会输入预设的校准数据集（可根据用户需求来确定校准数据集的 Batch Size 大小和迭代次数）和经优化的 FP32 模型及权重，平台会在校准数据集上执行 FP32 推理，并获得激活张量（Activation Tensor）和权重张量（Weight Tensor）两个参数。



图一 飞桨离线 INT8 推理校准流程

在随后的图形转换过程中，平台首先会通过用户所指定的量化算法，并根据上述两个参数来确定量化比例。其次，在对需要计算的图形进行分析后，平台会根据所支持的 INT8 操作列表（飞桨定义了两个新的 Kernel: Conv2D* 和 Pool2D*）进行量化/反量化的

插入操作，从而将量化比例添加到 Conv 作为 OP 属性。最后，平台将量化后的 INT8 模型保存，以便进行后续推理部署。

除了量化和反量化操作，飞桨还定义了 INT8 Kernel Conv 和 INT8 Kernel Pool 两种操作，前者输入 S8/U8 数值，采用 S8 的权重；输出为 S8/U8/FP32 三类数值，用于支持 INT8 Conv2D 计算；后者输入输出均为 S8/U8 数值，用于支持 INT8 Pool2D 计算。

来自英特尔新一代至强的技术助力

在新版飞桨中，百度导入了英特尔® 深度学习加速技术——这一集成在第二代英特尔® 至强® 可扩展处理器中的全新 AI 技术的支持。英特尔这款全新处理器不仅以优化的微架构、更多及更快的内核和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硬件加速能力。

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宽融合乘加 (FMA) 内核指令，可用于支持 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%[4]，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。

利用第二代英特尔® 至强® 可扩展处理器提供的强劲算力，以及英特尔® 深度学习技术提供的加速能力，百度团队已在 ResNet-50 和 MobileNet-V1 两种图像分类应用常用的网络模型上，对离线 INT8推理进行了缜密的测试验证。在推理准确度验证中，平台采用了英特尔® 至强® 金牌 6271 处理器，测试在拥有 50,000[5] 张图像的 Full ImageNet Val* 完整验证数据集上完成。如表 1 所示，从测试结果来看，在 ResNet-50 和 MobileNet-V1 两种模型的 TOP-1 准确率 (预测出最大概率的分类是正确的概率) 上，INT8 分别只有 0.4% 和 0.31% 的准确度损失[6]，基本可视为没有准确度损失。

模型	数据集	FP32 准确率	INT8 准确率	准确率差值
ResNet-50	Full ImageNet Val	76.63%	76.23%	0.40%
MobileNet-V1	Full ImageNet Val	70.78%	70.47%	0.31%

表一 FP32 和 INT8 推理准确度结果比较[6]

在接下来的推理吞吐量 (速度) 性能测试中，平台采用英特尔® 至强® 金牌 6271 处理器单核部署，并根据百度的业务部署要求，Batch Size 配置为 1 来衡量吞吐量。如表 2 所示，测试结果表明，在 Full ImageNet Val 完整验证数据集上，ResNet-50 和 MobileNet-V1 两种模型的 INT8 推理吞吐量是 FP32 的 2.2 倍到 2.79 倍[7]。由此可见，在不影响预测准确度的情况下，使用 INT8，可让多种深度学习模型的推理速度得到显著提升，这就意味着能有效提升用户深度学习应用的工作效能。

模型	数据集	FP32 准确率	INT8 准确率	INT8/FP32 吞吐量比比率
ResNet-50	Full ImageNet Val	11.54 images/s	32.2 images/s	2.79
MobileNet-V1	Full ImageNet Val	49.21 images/s	108.37% images/s	2.2

表二 FP32 和 INT8 推理吞吐量结果比较[7]

总结与展望

在过去数年中，百度与英特尔就深度学习等先进技术与平台的发展，一直在开展紧密的技术协作。在飞桨平台开发过程中，英特尔工程师以丰富的硬件调优经验，助力解决了许多关键问题，例如并行性能问题、处理器缓存竞争带来的性能问题等，从而在框架层次上整体提升了飞桨的预测性能。同时，在英特尔的支持下，百度也基于英特尔® 数学核心函数库 (Intel® Math Kernel Library，英特尔® MKL) 和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 Network，英特尔® MKL-DNN) 提供的特性，对不同类型的深度学习模型进行了特定的性能优化，并已在自然语言处理、图像处理等模型上获得了显著的进展。

而在基于英特尔® 架构处理器的推理效率提升方面，双方的技术协作更是结出了丰硕的成果。目前，百度内部业务和开源平台上的一些深度学习模型已实现了英特尔® 架构处理器资源利用率与性能的提升。同时，针对特定型号的服务器经优化后，已在平台中实现了现有最佳（State Of The Art, SOTA）效率并逐步上线。同时，在最新的飞桨中，还集成了面向各种设备和框架的英特尔® nGraph 深度学习模型编译器，大大提升了用户部署深度学习环境的效率。

未来，百度与英特尔在AI方面的合作，还将围绕四个方向持续展开。首先，百度计划将英特尔的量化和压缩方面的技术，集成到飞桨的整体工具链中；其次，英特尔® MKL-DNN 作为英特尔面向深度学习的核心数学函数库，也将进一步提供低精度数值函数支持，可作为处理器上的高性能 Kernel，成为更多飞桨 OP 的首选；第三，英特尔® nGraph 编译器也将灵活地作为子图，参与飞桨在英特尔® 架构平台上的高性能执行；最后，双方团队还计划在百度内部和外部完成更多飞桨部署实例的优化，携手为用户带去更多、更优异的深度学习平台能力。

[1,2,3,7] 数据源自 2019 年 4 月 4 日，百度基于第二代英特尔® 至强® 可扩展处理器开展的测试。
测试平台配置：双路英特尔® 至强® 金牌 6271 处理器，主频为 2.6GHz，24 核心/48 线程，启用睿频，启用超线程，搭载 512 GB 内存（16 slots/ 32GB/ 2666 MHz）。
操作系统为 CentOS 6.3，5.16 (ucode 0x4000013)，测试基于深度学习框架飞桨 1.4.0 版本，Kernel 版本为 3.10.0，编译器版本为 GCC4.8.2，使用英特尔® MKL-DNN 库 0.18 版本;工作负载：在 Full ImageNet Val 完整验证数据集上使用 ResNet50 and MobileNet-V1, Batch Size=1, Datatype: INT8
基准配置：双路英特尔® 至强® 金牌 6271 处理器，主频为 2.6GHz，24 核心/48 线程，启用睿频，启用超线程，搭载 512 GB 内存（16 slots/ 32GB/ 2666 MHz）。
操作系统为 CentOS 6.3，5.16 (ucode 0x4000013)，测试基于深度学习框 飞桨 1.4.0 版本，Kernel 版本为 3.10.0，编译器版本为 GCC4.8.2，使用英特尔® MKL-DNN 库 0.18 版本;工作负载：在 Full ImageNet Val 完整验证数据集上使用 ResNet50 and MobileNet-V1, Batch Size=1, Datatype: FP32

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

[5] http://www.image-net.org/challenges/LSVRC/2012/nnoupb/ILSVRC2012_img_val.tar

[6] 数据援引自“以数据为中心 加速数字经济落地——2019 英特尔® 创新产品发布会”上的百度分享环节

以上性能测试结果可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。

没有任何产品或组件是绝对安全的。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

性能测试中使用的软件和工作负荷可能仅在英特尔微处理器上进行了性能优化。

诸如SYSmark和MobileMark等测试均系基于特定计算机系统、硬件、软件、操作系统及功能。上述任何要素的变动都有可能导致测试结果的变化。请参考其他信息及性能测试（包括结合其他产品使用时的运行性能）以对目标产品进行全面评估。更多信息，详见 www.intel.com/benchmarks。

英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务才能激活。更多信息，请见 Intel.com，或从原始设备制造商或零售商处获得更多信息。描述的成本降低情景均旨在在特定情况和配置中举例说明特定英特尔产品如何影响未来成本并提供成本节约。情况均不同。英特尔不保证任何成本或成本降低。