案例研究

深度学习
安全计算

百度飞桨与第三代英特尔® 至强® 可扩展处理器为深度学习产业落地打造速度与安全新支点

本文介绍百度深度学习平台飞桨如何与第三代英特尔® 至强® VNNI 可扩展指令及软件防护扩展技术 SGX 结合，通过模型量化及加速帮助开发者更快更轻松地部署深度学习模型，同时通过机密计算能力帮助更多企业以安全可信的方式为深度学习模型提供更多源的数据。

引言

深度学习技术是推动人工智能时代发展的强大引擎，通过对文字、图形、声音和多媒体等数据的分析和推理推动产业向智能化迈进。然而，因为数据量庞大、技术门槛高，对很多企业而言，深度学习技术似乎高不可攀。开发者迫切希望找到深度学习模型在产业中落地的支点。

比如，大多数深度学习模型使用 32 位浮点精度 (FP32) 构建，复杂度高，模型参数量大，限制了其在一些场景和设备上进行部署，特别是在移动嵌入式设备的部署。如果能对模型进行“瘦身”，减少存储空间的同时加快预测速度，即找到一个速度新支点，将能极大推动深度学习的落地可行性。

再者，数据是模型学习的基础，如果能突破企业自有数据瓶颈，聚合更多来源、更多维度和更高质量的数据，同时以机密计算确保数据的安全性，即找到一个安全新支点，将能释放出超乎想象的数据潜力。

针对这两点需求，百度开源深度学习平台百度飞桨结合第三代英特尔® 至强® 可扩展处理器给出了令产业开发者满意的解决方案，为深度学习技术在实际场景用落地提供了有力支撑。

关于百度飞桨

百度飞桨以百度多年的深度学习技术研究和业务应用为基础，是中国首个开源开放、技术先进、功能完备的产业级深度学习平台，集深度学习核心训练和推理框架、基础模型库、端到端开发套件和丰富的工具组件于一体。目前，飞桨已凝聚超 265 万开发者，服务企业 10 万家，基于飞桨开源深度学习平台产生了 34 万个模型。[1] 飞桨能助力开发者快速实现 AI 想法，快速上线 AI 业务，从而帮助越来越多的行业完成 AI 赋能，实现产业智能化升级。

2021 年 1 月，飞桨正式进入 2.0 时代，通过全新升级的 API 体系，让深度学习技术的创新和应用更简单；以成熟完备的动态图模式带来最佳编码体验；以更强大的分布式训练能力提升易用性与灵活性；并构建出更繁荣的硬件生态，实现软硬一体的深度优化。

飞桨模型落地的速度新支点：量化及加速

飞桨模型量化及加速的整体方案包括两个环节：以PaddleSlim产出量化的飞桨模型，然后通过Paddle Inference在英特尔® CPU上部署和加速量化模型，如图一所示。

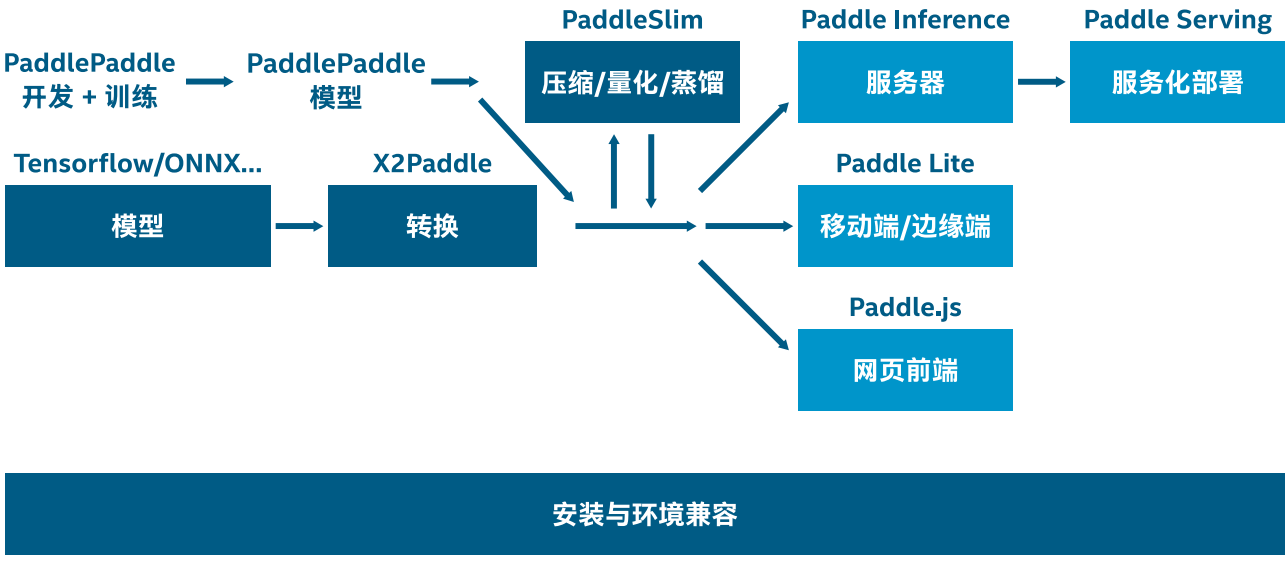

图1. 产出量化模型并在英特尔® CPU上部署

产出量化模型

为帮助用户快速满足模型“瘦身”的需求，百度飞桨提供了PaddleSlim深度学习模型压缩工具库，包含模型剪裁、定点量化、知识蒸馏、超参搜索和神经网络结构搜索(NAS)等一系列模型压缩策略。最新升级的PaddleSlim 2.0在静态图的基础上，还推出了动态图支持功能。

在众多瘦身方法中，量化是指用低比特数值替换FP32数值进行存储和计算，重点是INT8量化。模型量化的优点包括：减小存储空间、加快预测速度、降低能耗。PaddleSlim支持量化训练和静态离线量化方法，可以覆盖计算机视觉(CV)和自然语言处理(NLP)模型，能产出较为准确的scale并统一采用对称量化方式。

通过英特尔® 技术部署和加速量化模型

在英特尔® CPU上部署和加速百度飞桨量化模型时，关键步骤是与第三代英特尔® 至强® 可拓展处理器的内置AI加速技术及英特尔® oneAPI工具包的结合。

VNNI可扩展指令集

第三代英特尔® 至强® 可扩展处理器支持英特尔® 深度学习加速(英特尔® Deep Learning Boost，包括更新的VNNI)技术，扩展了英特尔® 高级矢量扩展指令集512(英特尔® AVX-512)。这一新的嵌入式加速器可将算力增加4倍，将内存要求降至¼(图二)。内存的减少和频率的提高加快了低数值精度运算的速度，最终加速AI和深度学习推理，适合图像分类、语音识别、语音翻译、对象检测等众多方面。

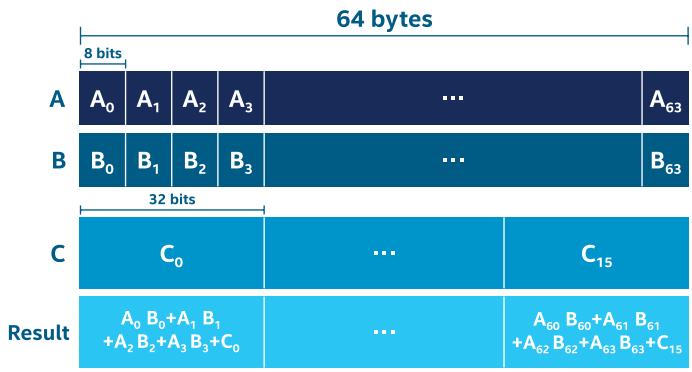

图2. 英特尔® DL Boost AVX512_VNNI VPDPBUSD指令可通过1条u8×s8→s32指令执行8位乘法和32位累加，INT8 OP理论算力峰值增益为FP32 OP的4倍

oneAPI

为激活VNNI加速功能，百度飞桨压缩量化方案广泛使用英特尔® oneAPI工具包。英特尔® oneAPI是一个统一的、简化的编程模型，集成了多平台下向量计算算子的JIT (just-in-time)代码库，使开发者在不同的架构上(CPU、GPU和FPGA)都可以方便地调用oneAPI算子的即时代码通用接口，而无需担心平台不兼容问题。

与百度飞桨融合及量化

借助VNNI扩展指令和统一编程接口oneAPI的支持，PaddleSlim产出的模拟INT8模型可以转化成真实的INT8模型，并部署到第三代英特尔® 至强® 可扩展处理器上。转化部署过程中的核心步骤包括：

1. 收集微调模拟模型所得的scale等数据。
2. 对模型算子进行融合，如conv+relu算子融合并精简图。
3. 根据所支持的oneDNN INT8算子列表，结合模拟训练所得的scale数据，做量化/反量化算子的插入操作。
4. 最后，平台支持量化后的INT8模型保存，以便进行后续推理部署。

应用示例

目前，百度飞桨模型量化及加速方案已广泛应用在百度多个服务中，如百度文字识别 (OCR) 服务。百度商业化OCR覆盖多场景、多语种、高精度的文字检测与识别服务，适用于远程身份认证、财税报销、文档电子化等场景，旨在为企业降本增效。百度OCR提供稳定易用的在线API、离线SDK、软件部署包等多种服务。在模型量化及加速方面，百度打造了一套以量化为核心的Slim OCR工具，取得了显著的性能提升。鉴于百度多个OCR模型均以图像分类模型ResNet50为基础，此处以ResNet50为例展示性能收益。

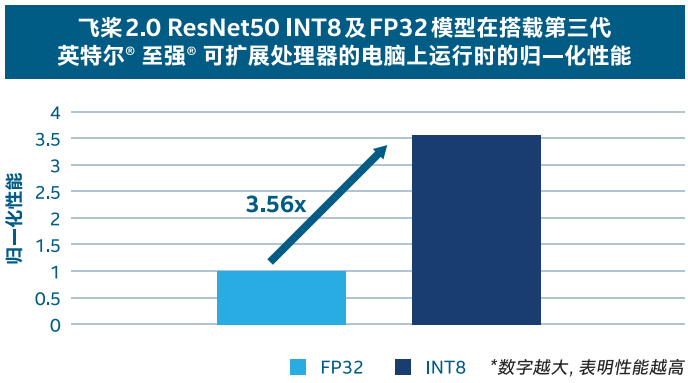

图3. 图像分类模型INT8量化前后在英特尔® 至强® Platinum 8358 CPU单核上性能[2]

测试结果表明，在Full ImageNet Val完整验证数据集上，ResNet50 INT8推理吞吐量是FP32的3.56倍 (图三)。可见，在第三代英特尔® 至强® 可扩展处理器上使用百度飞桨量化策略可让多种深度学习模型的推理速度显著提升，有效提升用户深度学习应用的工作效能。

飞桨模型落地的安全新支点: 机密计算

通过英特尔® SGX在飞桨中引入机密计算能力

在深度学习应用中聚合多源数据是众多企业面临的一大安全挑战，特别是在涉及敏感数据时。为此，百度飞桨实现了与百度安全计算平台MesaTEE的对接。MesaTEE即内存安全可信计算环境(Memory-safe Trusted Execution Environments)，基于两大技术：内存安全的Rust语言及英特尔®软件防护扩展(英特尔® Software Guard Extensions，简称英特尔® SGX)。英特尔® SGX是基于硬件的安全解决方案，绕过操作系统和虚拟机软件层，帮助将敏感的程序代码和数据加载到指定的受CPU保护的内存分区“飞地”(enclave)里，提供更强的保护以防止其被泄露或更改。借助英特尔® SGX，MesaTEE得以提供完善的机密深度计算能力，从而保护敏感数据。商业版MesaTEE基于Apache Teaclave(以Rust语言开发的通用安全计算平台，目前正处在孵化阶段)开发，为厂商客户提供深度定制的商用解决方案，引入了独创的协同机密计算引擎，实现了分布式的TEE集群计算，让大规模的隐私数据分析及训练成为可能。

在与百度飞桨对接时，MesaTEE的角色是协作平台和任务调度者，以Executor插件的形式将飞桨作为其一种任务执行环境，将特定的深度学习任务投递到飞桨的TEE运行环境中执行。此外，MesaTEE还支持多种TEE厂商产品，适配其远程证明(Remote Attestation)流程的正确性，并且整合至通信协议上，确保其运行环境的安全性是可度量的。在文件访问方面，支持S3等远程文件存储协议，文件内容全部使用密文存储。飞桨在不改变使用方式的前提下，在训练过程中在TEE中访问的所有数据均为密文。在TEE外无论从磁盘、内存或网络通信上都不存在途径可以窥查其计算内容，从而可以抵抗恶意攻击，确保整体安全性。图四显示了百度飞桨如何与MesaTEE联动。

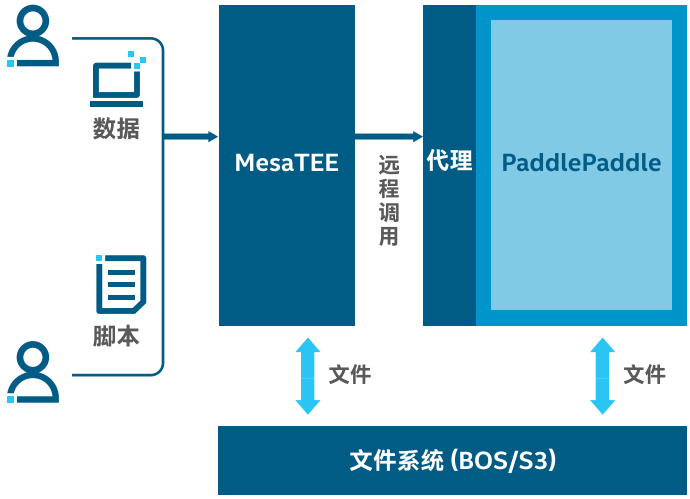

图4. PaddlePaddle与MesaTEE的联动

应用前景及案例

百度飞桨将深度学习平台与机密计算能力有机结合，响应了业界趋势，能够在众多领域发挥重要作用。以金融行业为例。传统以来，金融机构以简单的二维数字型数据来勾勒借贷人画像，并使用机器学习技术展开建模和分析。深度学习技术的日趋成熟为信贷风险评估开拓了全新的可能。深度学习模型能突破传统机器学习技术的局限，高效利用语音识别、图像识别等方式，将多媒体数据和流数据纳入模型，比如支付数据(第三方支付)、消费数据(零售商及电商平台)、信贷数据(第三方信贷机构)，甚至是社交人脉(社交平台)、工作履历(招聘平台)、旅游出行信息(出行平台)等。

然而，从数据提供方角度而言，这些用户隐私数据敏感性很高，并且可能受到法律保护不允许流通。这导致现实场景中，金融行业数据孤岛情况非常普遍。数据的提供方、建模方、评估方和使用方之间彼此割裂。通过在TEE中运行的百度飞桨深度学习平台，如百度度小满一样的金融机构将得以安全地与更多数据源合作，获得更多维的用户画像数据，打造出更健壮的信贷风险评估模式。

总结

本文介绍的英特尔® 深度学习加速VNNI指令集及英特尔® 软件防护扩展SGX技术是全新推出的第三代英特尔® 至强® 可扩展处理器的核心特色，能够推动安全高性能的AI应用。第三代英特尔® 至强® 可扩展处理器采用具有内置加速和高级安全功能的平衡架构，通过数十年的创新设计，可满足苛刻的工作负载要求。该处理器针对云、企业应用、AI、高性能计算、网络、安全性和物联网工作负载进行了优化，具有8至40个功能强大的内核以及广泛的功能。

此外，第三代英特尔® 至强® 可扩展处理器还是唯一具有内置AI加速功能、端到端数据科学工具和智能解决方案生态系统的数据中心CPU。这一强大的能力组合可在从边缘到云的每个应用程序中解锁更多数据价值。

未来，凭借上述量化技术、部署工具及与英特尔® 硬件的深度整合优化，飞桨主平台中丰富的模型资源及越来越成熟的多种应用开发套件（如PaddleOCR、PaddleDetection等），都将无缝地在英特尔® 平台上线，为用户提供最优的模型 + 硬件加速体验。

此外，百度飞桨的部署工具Paddle Inference和Paddle Lite等也将分别针对OpenVINO™ 工具套件等英特尔® 推理加速库进行原生整合。种种持续努力将帮助深度学习框架的用户以更低门槛体验在英特尔® 硬件上的加速效果。

了解更多

英特尔及百度飞桨团队欢迎更多用户在GitHub上关注飞桨项目并给予宝贵意见。

https://github.com/PaddlePaddle/Paddle

https://github.com/PaddlePaddle/Paddle-Lite

[1] 数据来自百度。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。

[2] 百度：Platinum 8350C：由英特尔于2021年3月19日进行测试。2节点，2x英特尔® 至强® Platinum 8350C CPU，2.60GHz，32核，HT开启，Turbo关闭。总内存256 GB（4插槽/64GB/2933 MHz），BIOS：WLYDCRB1.SYS.0020.P92.2103170501(ucode：0x8d000270)，Ubuntu 18.04.3 LTS，4.15.0-55-generic，gcc 9.3.0编译器，ResNet50模型，深度学习框架：飞桨2.0，下载：https://paddle-inference-lib.bj.bcebos.com/2.0.0-cpu-avx-mkl/paddle_inference.tgz BS=1，ImageNet Val，1个实例，数据类型：FP32/INT8 。

性能因用途、配置和其他因素而异。请访问www.Intel.com/PerformanceIndex了解更多信息。

性能结果基于截至配置中显示的日期的测试，可能无法反映所有公开可用的更新。有关配置的详细信息，请参见备份。没有任何产品或组件能够做到绝对安全。

成本及结果均不同。

英特尔技术可能需要支持的硬件、软件或服务得以激活。