白皮书

第四代英特尔® 至强® 可扩展处理器
英特尔® 高级矩阵扩展（英特尔® AMX）
自然语言处理（NLP）

验证至强内置全新 AI 加速技术：AMX 助百度 ERNIE-Tiny 性能提升达 2.66 倍[1]

目录

前言概述

得益于近十年来在自有“飞桨”人工智能（Artificial Intelligence, AI）框架上的前瞻布局和大力投入，以及在语言与知识技术上积累的丰厚经验与成果，百度已在自然语言处理（Natural Language Processing，以下简称 NLP）领域构建起完整的产品体系与技术组合。ERNIE 3.0 作为其飞桨文心 · NLP 大模型的重要组成部分，也在各种 NLP 应用场景，尤其是中文自然语言理解和生成任务中展现出卓越的性能。

随着 NLP 逐渐进入技术和产业结合的快车道，并在更多行业中实现商业化落地，用户对 ERNIE 3.0 也提出了更多细分需求，例如更高的处理效率和更广泛的部署场景等。为此，百度不仅借助其创新技术优势，推出了轻量版 ERNIE-Tiny，也与合作伙伴英特尔携手，提前引入即将发布的全新第四代英特尔® 至强® 可扩展处理器，作为 ERNIE-Tiny 未来的硬件承载基座。

为了让 ERNIE-Tiny 在第四代英特尔® 至强® 可扩展处理器及其内置的全新英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extension，英特尔® AMX）技术的加速支持下实现更优推理性能，双方也携手开展了多项优化工作。而来自对比测试的数据表明，相比通过英特尔® AVX-512_VNNI 技术来实现 AI 加速的、面向单路和双路的第三代英特尔® 至强® 可扩展处理器，ERNIE-Tiny 在升级使用内置英特尔® AMX 技术的第四代英特尔® 至强® 可扩展处理器后，其整体性能提升高达 2.66 倍[2]，取得了令人满意的效果。

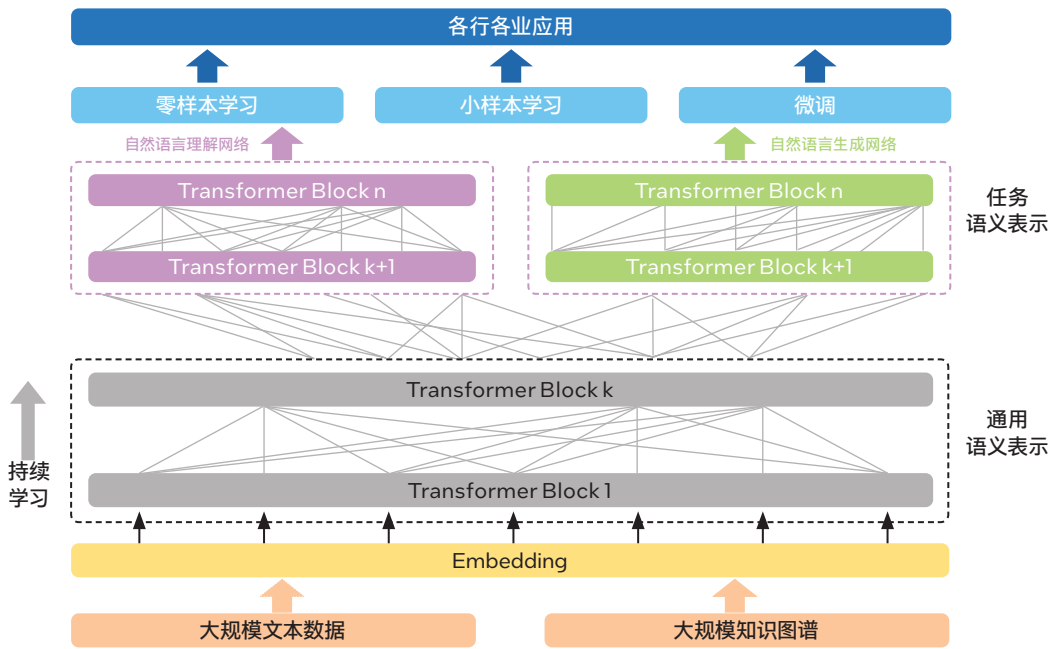
“作为百度面向 NLP 领域的重要技术方案，基于轻量化技术的 ERNIE 3.0 轻量版可在搜索排序、推荐、信息抽取、地图检索、自然语言推断等应用场景中为用户提供响应迅速、质量可靠的能力输出。全新第四代英特尔® 至强® 可扩展处理器及英特尔® AMX 技术的引入，使得轻量版 ERNIE 3.0 在通用 CPU 平台上也能获得令人满意的推理效能，从而能帮助更多用户在其既有 IT 设施中更为方便地部署 ERNIE 3.0，从而进一步普及其应用范围。”

孙宇
百度杰出架构师
百度自然语言处理部

方案背景：ERNIE 3.0 走向轻量化 为更多行业用户提供 NLP 应用助力

作为 AI 领域的重要分支，NLP 正获得前所未有的市场关注与技术追踪。有预测数据表明，到 2024 年全球 NLP 市场规模将达 264 亿美元[3]，并将大幅缓解金融、医疗、法律等行业中人力密集型工作环节带来的效率和成本压力。例如在医疗健康领域，利用 NLP 应用对医疗记录开展筛查有助于发现药物的长期不良反应；在法律领域，NLP 应用也在帮助人们从浩如烟海的记录中梳理出案件的来龙去脉。

作为拥有强大互联网基础的领先 AI 公司，百度凭借其旗下飞桨文心 · NLP 大模型所具备的创新技术优势，在语言理解、语言生成等 NLP 场景中已获取了明显的市场优势，并在中国 AI 公有云 NLP 领域连续多年取得市场份额第一[4]。如图一所示，在大模型中，基于知识增强的多范式统一预训练框架 ERNIE 3.0 融合了自回归网络与自编码网络，并加入了大数据预训练与多源丰富知识相结合、持续学习等特性，在理解任务、生成任务、零样本学习任务和常识推理任务上均有着优秀的表现，在 14 种类型/45 个 NLP 数据集取得了 SOTA（State Of The Art Model，当前最佳模型）的结果。在中文领域，ERNIE 3.0 的表现则更为突出，不仅刷新了 54 个中文 NLP 任务基准，并登顶 SuperGLUE 全球榜首，同时也因具备非常出色的语言理解能力，还可以完成写小说、歌词、诗歌等的文学创作任务[5]。



图一 ERNIE 3.0 框架基本架构

在 ERNIE 3.0 的实际落地应用过程中，许多细分领域也根据自身业务特点，对它提出了特定化需求。众所周知，模型中更多的层数、参数意味着更大的模型体积、更强的计算资源需求以及更长的推理耗时，对于业务响应速度和构建成本敏感的用户而言，无疑提高了引入和使用门槛。

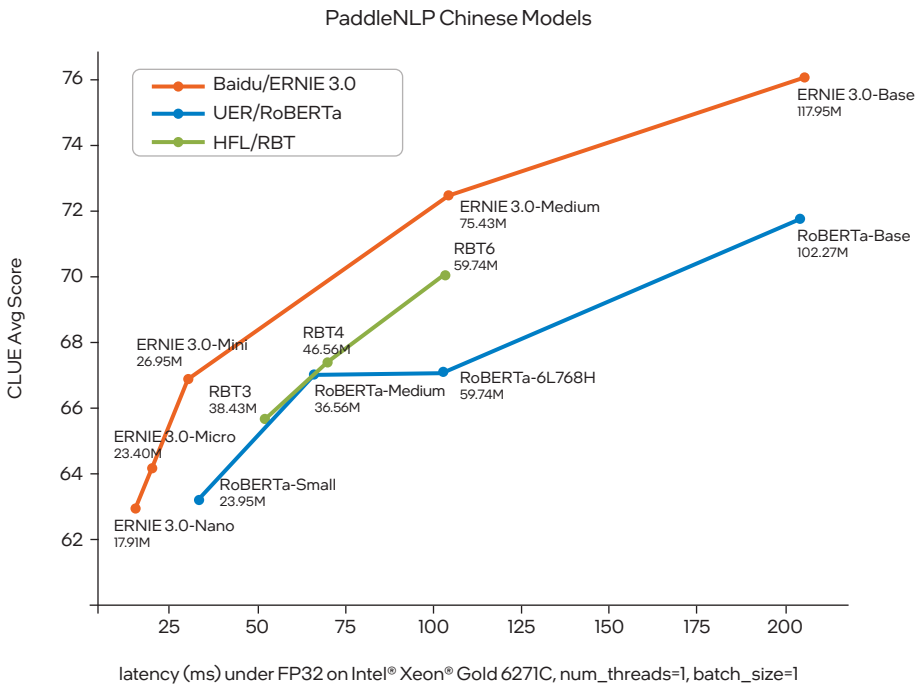
借助模型轻量化技术对 ERNIE 3.0 大模型进行蒸馏压缩，是助力 ERNIE 3.0 推广到更多行业与场景的有效方法。为此，百度基于其“在线蒸馏”等创新技术方案，推出多个 ERNIE 3.0 轻量化版本 ERNIE-Tiny，在保持模型平均精度的前提下实现了更短的运算时间以及更少的算力需求。同时，ERNIE-Tiny 在推理时，也无需再为之配备昂贵的专用 AI 算力设备，在通用平台，如 CPU 平台上即可高效率完成推理作业。这无疑能让用户在既有公有云或数据中心的 IT 配置上即可使用该模型，而无需增添额外硬件或服务。

与此同时，引入更强的通用计算平台和优化方案，也是助力 ERNIE-Tiny 获得更优效率的另一项重要手段。百度为此与英特尔开展深度技术合作：一方面将全新第四代英特尔® 至强® 可扩展处理器引入 ERNIE-Tiny 的推理计算过程；另一方面，也推进了多项优化措施，例如通过英特尔® oneAPI 深度神经网络库 (Intel® oneAPI Deep Neural Network Library，英特尔® oneDNN) 来调用英特尔® AMX 指令等，以确保 ERNIE-Tiny 可以更为充分地利用这款处理器内置的全新 AI 加速技术带来的性能加速“红利”。

解决方案：三项优化方案 助 ERNIE-Tiny 尽释新平台 AI 加速潜能

■ 新一代英特尔® 至强® 可扩展处理器为 ERNIE 3.0 带来更强 AI 加速引擎

百度与英特尔携手开展的优化方案，首先围绕 ERNIE-Tiny 系列中 Medium 版展开，这一轻量版本与基础版 ERNIE 3.0 相比，其网络层数从 12 层减少到了 6 层，以此可有效降低所需的算力资源并缩短推理时长。



图二 ERNIE-Tiny 模型精度-推理时延对比度[6]

优化方案中，英特尔为 ERNIE-Tiny Medium 版本（以下简称 ERNIE-Tiny）提供了第四代英特尔® 至强® 可扩展处理器作为推理工作的算力输出引擎。这一采用 Intel 7 制程工艺的新一代至强® 可扩展处理器，可凭借全新的性能核微架构设计来提升处理速度，并在低时延和单线程性能上实现突破。

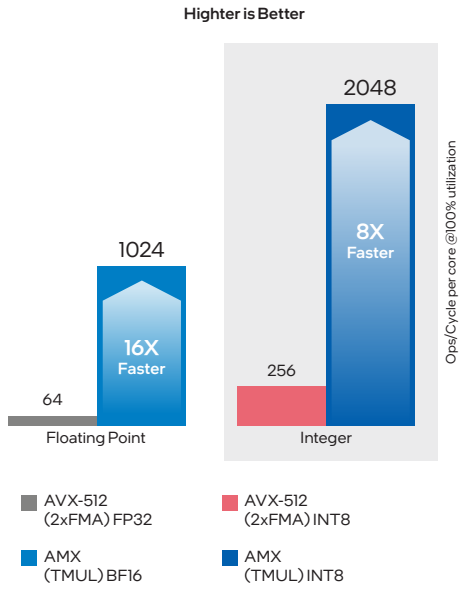
在整个芯片的架构层面，第四代英特尔® 至强® 可扩展处理器通过使用嵌入式多芯片互连桥接（Embedded Multi-die Interconnect Bridge，EMIB）技术，可在保持既有单核优势的同时，大幅提升可扩展性。同时，新处理器也提供了对先进内存和下一代 I/O 技术，包括 DDR5、PCIe 5.0、CXL 1.1 以及高带宽内存（High Bandwidth Memory，HBM）技术的支持，这些技术特性能为 ERNIE 3.0 这样的 AI 模型所需的高强度工作负载提供更可靠的全局性能加速。

更重要的是，第四代英特尔® 至强® 可扩展处理器还增添了多种新的内置加速引擎来助力用户在不同应用场景中提升性能，如英特尔® 加速器接口架构指令集（英特尔® AIA）、英特尔® 数据流加速器（英特尔® DSA）和英特尔® 高级矩阵扩展（英特尔® AMX）。其中，AMX 对于 AI 工作负载的加速尤为显著。

■ 基于全新处理器和内置 AI 加速能力，双方携手为 ERNIE-Tiny 加入三项优化方案

1、全新 AI 加速引擎优化提升处理效率

与英特尔在此前的至强® 可扩展处理器中提供的两种 AI 加速能力，即英特尔® AVX-512_VNNI_INT8 与英特尔® AVX-512_VNNI_BF16 不同，英特尔® AMX 采用了全新的指令集与电路设计。在实际的工作负载中，英特尔® AMX 能同时支持 BF16 和 INT8 数据类型，其每个物理核在每个时钟周期可实现 2,048 次 INT8 运算和 1,024 次 BF16 运算[7]，与上一代 AI 加速引擎相比，大幅提升了 AI 工作负载的效率，这显然有助于提升 ERNIE-Tiny 在推理环节的性能表现。



图三 与英特尔® AVX-512 相比，英特尔® AMX 可带来 8 倍以上的效率提升[8]

2、利用英特尔® oneDNN 实现对英特尔® AMX 指令的调用

为了让英特尔® AMX 的加速能力能直接作用于 ERNIE-Tiny，百度与英特尔一同借助英特尔® oneDNN 来实现英特尔® AMX 指令的调用。作为开源的、跨平台的性能库，英特尔® oneDNN 可有效助力用户提升其 AI 应用与框架在英特尔® 架构平台上的性能，而且它也已加入了对英特尔® AMX 的支持。

在本次合作中，双方除携手完成了英特尔® oneDNN 与飞桨开源深度学习平台（*PaddlePaddle，https://www.paddlepaddle.org.cn/*）的集成外，也根据 ERNIE-Tiny 的实际运行需求开展了一系列增量工作，包括将 Linux 操作系统的内核更新为支持英特尔® AMX 的版本等。

3、内存性能优化

借助第四代英特尔® 至强® 可扩展处理器与英特尔® AMX 获得计算性能的大幅提升之后，内存性能的优化自然也不可或缺，为此百度与英特尔也制定了针对性的优化方案。双方通过分析发现，ERNIE-Tiny 在推理过程中有许多串行操作，即每次运算都会先读数据再写数据，然后下一次运算也是如此，这会消耗大

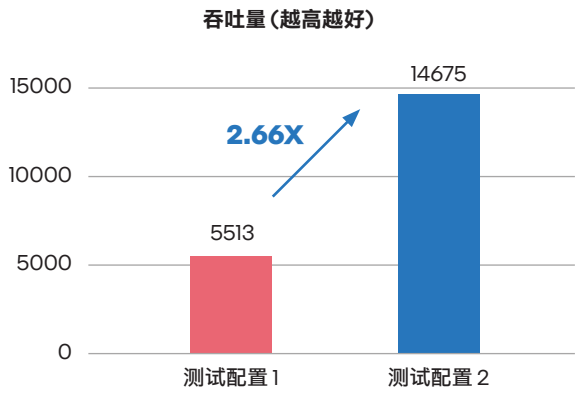
量操作时间。优化方案则是将矩阵乘法与元素的运算及激活融合在一起，即把连续的操作合并为一个大操作，可使内存的运行效率显著提升。

同时，方案中针对多线程的优化也被证明可助力 ERNIE 3.0 提升推理计算性能，与上一版本相比，方案进一步优化了多线程的效率，并提升了多核的扩展性。

优化方案验证：第四代英特尔® 至强® 可扩展处理器可大幅提升 ERNIE-Tiny 性能表现

为了验证第四代英特尔® 至强® 可扩展处理器与上述多项优化方案对 ERNIE-Tiny 性能提升的实际作用，英特尔协助百度推进了性能对比测试。测试在第四代英特尔® 至强® 可扩展平台与第三代英特尔® 至强® 可扩展平台之间展开。后者使用英特尔® AVX-512_VNNI 对模型进行了 INT8 量化提速，而前者则启用英特尔® AMX 技术进行加速。

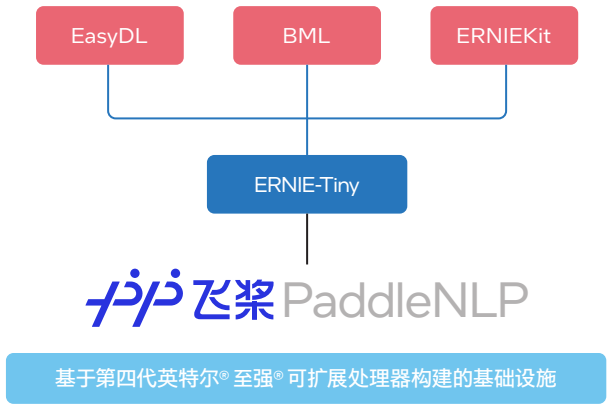
测试结果如图四所示，ERNIE-Tiny 的性能（测试采用吞吐量（Throughput）作为测评指标）获得了显著的提升，对比上一代英特尔® 至强® 可扩展平台，其吞吐量提升到了它的 2.66 倍[9]。



图四 ERNIE-Tiny 在不同处理器平台上的性能对比[10]

如图五所示，目前，各个 ERNIE-Tiny 不仅已部署在零门槛 AI 开发平台 EasyDL、全功能 AI 开发平台 BML 和 ERNIEKit（旗舰版）产品中，它们也将与平台和产品的其它能力一起协同，在基于第四代英特尔® 至强® 可扩展处理器的基础设施上，为使用者提供文本分类、关系抽取、文本生成以及问答等能力。同时，它们也将作为百度飞桨 PaddleNLP 自然语言处理开发库的核心模型，搭配训练-压缩-推理端到端全流程应用与丰富的产业实践范例，全力加速 NLP 技术产业落地。

（如欲了解更多详情，请访问：*https://github.com/PaddlePaddle/PaddleNLP/tree/develop/model_zoo/ernie-3.0*）



图五 ERNIE-Tiny 对外能力输出

未来展望

百度与英特尔本次协作优化的成功，再一次证明各个行业用户在通用的 CPU 平台上也能同样方便地部署高效能的 ERNIE-Tiny，用以应对越来越多的 NLP 应用需求。使用这一方案，用户不必额外采购昂贵的专用 AI 算力设备，这将大幅降低企业借助 NLP 能力提升业务效率的门槛，并加速更多 NLP 技术与应用的商业落地过程。

面向未来，英特尔还将与百度一起在 NLP 领域开展更多也更为深入的技术合作，让新一代至强® 可扩展处理器及其内置的更强 AI 加速技术为更多 AI 应用的落地与实践提供更多助力。

[1, 2, 9, 10] 测试配置如下：测试配置 1：处理器：双路英特尔® 至强® 铂金 8358P 处理器，主频 2.6GHz，32 核心 64 线程，睿频开启；内存：512GB DRAM（16x32GB DDR4 3200 MT/s [2933 MT/s]）；存储：INTEL_SSDSC2KG960G8,INTEL_SSDSCKKB480G8；网络适配器：2x Ethernet Controller 10G X550T；BIOS 版本：06.00.01；操作系统版本：Ubuntu 20.04.4 LTS（Kernel：5.8.0-43-generic）；GCC 版本：8.4；英特尔® oneDNN 版本：2.6；工作负载：PaddlePaddle 2.3，Ernie-3.0 optimization for INT8；
测试配置 2：处理器：双路第四代英特尔® 至强® 可扩展处理器，铂金型号，主频 2.6GHz，48 核心 96 线程，睿频开启；内存：512GB DRAM（16x32GB <OUT OF SPEC> 4800 MT/s [4800 MT/s]）；存储：INTEL SSDPEKNW020T8，INTEL SSDPEDMD800G4；网络适配器：1x Ethernet Controller I225-LM；BIOS 版本：EGSDCRB1.SYS.0090.D03.2210040200；操作系统版本：CentOS Stream 8（Kernel：5.15.0-spr.bkc.pc.8.8.5.x86_64）；GCC 版本：8.5；英特尔® oneDNN 版本：2.6；工作负载：PaddlePaddle 2.3，Ernie-3.0 optimization for INT8。
该数据由百度提供，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。

[3] 数据援引自公开媒体报道《The State of Enterprise NLP in 2020》：https://opendatascience.com/the-state-of-enterprise-nlp-in-2020

[4] 数据援引自公开媒体报道：https://baijiahao.baidu.com/s?id=1736317499268483396；https://news.fx678.com/202112241641002280.shtml

[5] 数据援引自百度 ERNIE 3.0 官网介绍：https://wenxin.baidu.com/wenxin/nlp

[6] 数据援引自百度内部测试，如欲了解更多详情，请访问：https://github.com/PaddlePaddle/PaddleNLP/tree/develop/model_zoo/ernie-3.0，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。

[7, 8] 数据援引自英特尔官网，如欲了解更多详情请访问：https://edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/