# 英特尔中国
# 教育行业
# AI实战手册

intel ai

# Contents

## 目 录

### 趋势篇



### 实战篇



### 技术篇




**主编：** 赵朝卿，秦莉

**作者**（排名不分先后，按姓氏首字母排序）

崔爽，赖美璇，陆礼明，邱亮，温炜，吴缘，夏磊，徐焰庆，颜彦，伊红卫，于超，俞巍，臧战，赵玉萍，赵桢



此外，本手册的编撰工作也得到了合作伙伴及诸多英特尔同事们给予的大力支持帮助，在此表示感谢。


# 趋势篇

# 助力教育智能化转型
# 服务教育现代化进程

百年大计，教育为本。教育是民族振兴、社会进步的基石，强国必先强教。多年来，中国以教育信息化支撑和引领教育现代化，并将应用现代技术，加快教育信息化基础设施建设，加强网络教学资源体系建设，以及构建适合信息时代的教与学模式，作为推进教育创新发展以及推动人才培养模式改革的重要举措。

为抓住新一轮科技革命带来的新机遇，加速教育现代化，“十三五”期间，教育部结合国家“互联网 +”、大数据、新一代人工智能等重大战略任务安排，提出实施《教育信息化 2.0 行动计划[1]》，将教育信息化作为教育系统性变革的内生变量，提出了为国际教育信息化发展提供中国智慧和中国方案的目标，在新时代赋予了教育信息化新的使命，以真正走出一条中国特色的教育信息化发展之路。步入“十四五”，中国开启了建设教育强国的新征程，利用新一代信息技术，推动信息化时代教育创新，实施教育新基建工程，大力开发优质数字教育资源，被确定为推动教育改革创新与建设高质量教育体系的根本动力[2]。

在国家政策引领推动、新一代信息技术不断迭代更新，以及教育治理能力日益优化的教育发展新格局下，“信息技术与教育教学的深度融合”已经成为共识。随着“三通两平台”建设取得巨大成就，国家数字教育公共服务体系建设日趋完善，教育信息化基础设施建设已全面覆盖，数字化教育资源得到极大丰富，师生网络学习空间已达 6,300 多万个[3]，中小学网络接入率达 99.7%，拥有多媒体教室的中小学校比例达 95.2%，已接受过不同程度信息技术应用能力培训的教师人数也超过 1,000 万[4]，基于网络开展教与学的大环境已经基本形成。

在信息技术应用规模快速扩展的同时，K12 和高等教育领域的信息化基础环境和教师信息化素养也已得到全面提升。新冠疫情下的“停课不停教，停课不停学”大规模在线教学实践，进一步推动了信息技术与教育教学深度融合与应用，展现了信息技术与教学融合创新带来的强大合力，加速了中国教育信息化进程由 1.0 迈进 2.0 新时代。

以人工智能为代表的新一代信息技术在为各行各业跨越式发展带来广阔前景的同时，也不断融入教育领域，使得“人工智能 + 教育”成为教育行业创新发展的绝对热点，得到高度关注。2017 年国务院印发的《新一代人工智能发展规划的通知》中指出，要推动人工智能在教学、管理、资源建设等方面的全流程应用。2018 年教育部印发的《教育信息化 2.0 行动计划》进一步明确提出，要利用智能技术加快推动人才培养模式、教学方法改革，探索泛在、灵活、智能的教育教学新环境建设与应用模式。随着教育信息化 2.0 新征程步伐的加速，智能化正在政策驱动和产学研用等教育生态共同推进下，与数字化、网络化、泛在化同行并领跑，以数据驱动为核心的智慧教育逐渐深入人心。

与此同时，业界对“人工智能 + 教育”这一教育信息化新方向也期许热烈。有数据显示，目前整个“人工智能 + 教育”相关市场规模已达数千亿[5]。近年来，相关市场融资总量亦达数百亿元，涵盖了 K12 阶段教育、职业培训、学前教育等不同的教育细分领域[6]。


图 1-1-1 未来教育信息化正向智能化方向切换

[1]《教育信息化 2.0 行动计划》http://www.moe.gov.cn/srcsite/A16/s3342/201804/t20180425_334188.html

[2] 坚持以建设高质量教育体系为统领 谋划推动“十四五”时期教育发展，http://www.scio.gov.cn/xwfbh/xwbfbh/wqfbh/44687/45183/zy45187/Document/1701372/1701372.htm

[3] 教育信息化从 1.0 到 2.0——走具有中国特色的发展之路，https://www.ict.edu.cn/news/jrgz/xxhdt/n20200509_67683.shtml

[4] 中央电教馆与英特尔联合推出的《2020 教育信息化年度蓝皮书》

[5] 千亿级 AI+ 教育市场规模，会是下一个风口吗？ https://www.sohu.com/a/435798627_120411615

[6] 2020 年中国 AI+ 教育行业投融资热度不减，http://market.chinabaogao.com/wenti/03204TH52020.html

大规模的投入加快了智能技术在教育行业的创新应用，语音识别、自然语言理解、深度学习、AR/VR 等前沿技术和相关教育教学解决方案正在更多场景落地，包括自适应学习系统、智能导师系统、智能测评系统、基于虚拟现实 / 增强现实的场景式教学都正在成为现实。而这又推动了以智能技术为代表的新一代信息技术，更全面深入地渗透到教育多环节以及教与学模式的融合创新过程之中，为面向未来打造智能型泛在学习环境，构建智能化、网络化、个性化、终身化的现代化教育体系添加了新动力。

## 人工智能在教育行业各场景中的应用探索与实践

目前，在教育行业各场景中，如图 1-1-2 所示，由数据驱动，基于各类机器学习 / 深度学习方法构建，涵盖计算机视觉（CV）、自然语言处理（NLP）以及自动语音识别（ASR）等技术领域的人工智能应用探索和方案部署，可以分为教学环节、练习测评和教学管理三个核心场景：

- **教学环节场景：**与传统教学模式相比，智慧教育引入更多人工智能应用来提升教学环节中的互动性和精准性。一方面，通过交互式电子白板、双师课堂、AR/VR 教学等应用，不仅可通过增强师生间的交互来提升线上线下课堂的教学效果，也能有效缩小区域、城乡、校际间的教育质量差距，实现教育的优质、均衡发展；另一方面，借助课堂行为分析等应用，能够通过动态、实时的教学数据分析，帮助教师和教学管理人员获得智能化的课堂观察和分析能力，让“教”与“学”实现精准化。

- **练习测评场景：**人工智能技术的引入大大丰富了练习测评的形式。一方面，师生可将教学练习与测评的场景延展至课堂之外，并获得来自智能系统和教师的双重反馈；另一方面，在 NLP、ASR 等人工智能技术的帮助下，课程评测的内容得到大幅扩展，评测结果也能更加精准和科学，并帮助学生依据评测结果制定和优化后续学习过程。目前，诸如口语测评等测评类人工智能应用已在各级教育机构获得广泛欢迎。

- **教学管理场景：**数据驱动的智慧教育场景更注重数据的快速处理、分析与反馈。得益于 5G、边缘计算等技术的成熟，教育机构得以在校园部署更多的人工智能应用，进而能够通过更为快捷的数据交互和处理方式来对教学过程实施灵活的调度和管理。例如，通过智能教学辅助能力，学校管理人员可以快速远程巡课、智能排课、获取教学大数据，教师可以进行智能备课，以及基于知识点地图和资源库开展教学等。


图 1-1-2 人工智能在教育行业各场景的应用

## “人工智能 + 教育”转型带来的挑战与机遇

由新技术、新模式驱动的行业转型，在提升效率与优化产品或服务质量的同时，势必也会对行业既有基础设施能力带来挑战，在教育领域也同样如此。随着更多教育机构变道切向新的智能化方向，其 IT 基础设施也面临着严峻挑战。如图 1-1-3 所示，这些挑战包括：

- **人工智能应用对算力的高要求：**与学校已有的校园网、电子白板、平板电脑等信息化设备相比，人工智能应用因其数据量大、推理要求高、计算负载密集等特点，要求学校信息化系统具有更高的算力。

- **人工智能应用如何与教学环节无缝对接：**与教学环节的紧密结合，使人工智能应用有别于传统多媒体课件等校园信息化应用，尤其是在应用时效性上，要求能够与教师授课讲解、学生实操练习等环节无缝对接，并给与实时反馈；反之，明显的时延会带来使用体验的大幅下降。

- **对校园既有 IT 设备兼容性需求：**与互联网、通信、软件等始终站在新技术潮头的企业不同，教育机构的信息化建设通常是基于实际需求而不断增补，同时校园环境的复杂性也使其软硬件基础设备纷繁复杂，因此对新的人工智能方案与既有系统的兼容性都颇具考验。

- **对校园既有 IT 系统架构的要求：**由数据驱动的智慧教育系统也给传统校园 IT 架构带来挑战。部署在教室等系统末端的设备通常处理能力较弱，而全部通过远端数据中心或云端进行处理，又容易受到网络因素的影响。为应对这些问题，智慧教育系统的系统架构正由简单走向复杂。

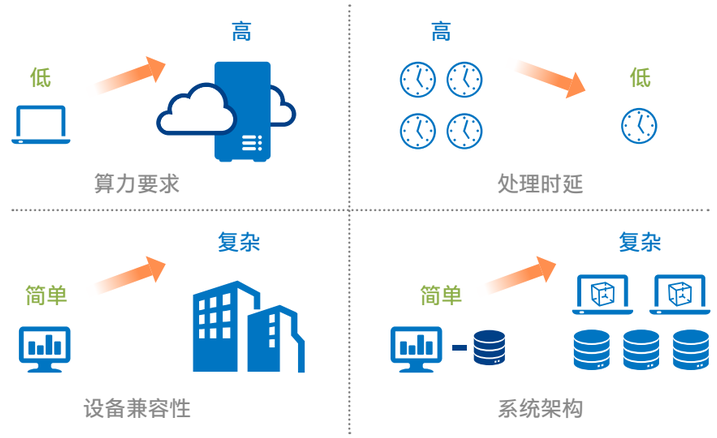



图 1-1-3 “人工智能 + 教育”转型带来诸多挑战

挑战往往伴随着机遇。面对以上挑战，各教育行业解决方案厂商在为各级教育机构打造新一代“人工智能 + 教育”系统时，选择与英特尔携手，大胆引入更多、性能更强劲的软硬件设备和更全面、更前沿的系统架构设计。事实上，采用一系列英特尔先进产品与技术的解决方案已在诸多教育场景中获得了验证，并取得了良好的应用反馈，这些产品与技术包括：

- 多样化的硬件产品矩阵，包括英特尔® 至强® 可扩展处理器、英特尔® 酷睿™ 处理器、英特尔凌动 ® 处理器、英特尔® Movidius™ Myriad™ X 视觉处理单元等。


图 1-1-4 多样化的英特尔硬件产品矩阵

- 全方位的软件产品助力，包括 OpenVINO™ 工具套件、英特尔® 深度学习加速技术、英特尔® AVX-512、面向英特尔 ® 架构优化的深度学习框架等。

- 英特尔在 5G、边缘计算（MEC）平台以及“云 - 边 - 端”架构上的一系列成功部署和实战经验。

## 英特尔携手合作伙伴推动 “人工智能 + 教育” 实践

为了让人工智能技术在落地过程中更好地与教育教学融合创新，从而在教育行业智能化转型过程中发挥其强大的驱动作用，英特尔积极发挥性能领导者、软硬件创新引领者以及在人工智能领域领先的全栈解决方案提供商的优势，释放多年通过系统化的投入和生态力量，推动教育变革的经验，运用创新技术，与众多合作伙伴一起拥抱智能化教育潮流，针对不同的教育场景，打造多个适应普遍性需求，且经过实践部署验证的成功案例，为学生与老师提供更加简洁高效的教育环境，推动教育智能化变革。

如图 1-1-5 所示，这些案例围绕着教师与学生、教学过程与效果评估这两组核心关键词展开，并由此形成人才培养、教育教学、教学管理和考核测评等四种典型的场景象限。在本手册中，将为每个象限选取一个专门的人工智能应用方向与读者探讨，同时在每个方向中也选取了数个实际应用案例。

- **人才培养：**本手册在人才培养象限将以人工智能教学与实训作为方向。探讨在 K12 与高等教育阶段的人工智能教学与实训过程中，如何在既有校园环境下，通过英特尔高性能软硬件产品与创新 “云 - 边 - 端” 架构的引入，对人工智能教学各环节提供支撑，全面解决人工智能人才培养中的数据、算力、算法三大需求。

- **教育教学：**本手册在教育教学象限将以课堂行为分析作为方向。探讨如何以深度学习、计算机视觉等人工智能方法，在英特尔软硬件产品和技术的支持及优化下，对教学过程进行实时性分析评估，全面实现师生教与学模式、情绪状态等信息的量化统计和可视化呈现，在有效降低教师工作负担、优化教学过程的同时，大幅提高教育机构的管理能力。此方法也适用于对在线教学（一对一或网络课堂）的分析。

- **教学管理：**本手册在教学管理象限将以智能教学辅助作为方向。探讨如何将基于语音识别的智能辅助能力引入教育机构的日常管理及师生的教学环节，以提高教学质量和效率。后文将以语音识别技术在智能会议、智慧课堂等场景中的应用为例，阐述师生如何从基于英特尔产品与技术的各项语音识别智能应用中获益。

- **考核测评：**本手册在考核测评象限将以口语测评作为方向。探讨如何有效利用英特尔先进软硬件产品带来的高性能算力和深度学习加速能力，用深度学习方法为教育行业构建涵盖发音准确度、流畅度、自然度、完整度维度和多项指标的综合测评智能系统，从而高效、快速和准确地帮助教育机构和学生，对各类语言的口语学习成果进行智能化测评。

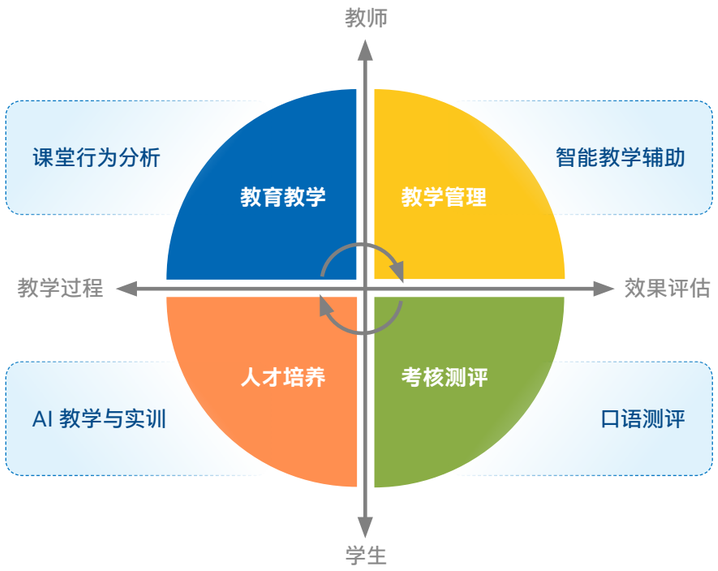



图 1-1-5 面向 “人工智能 + 教育” 实践的四种典型场景

推动教育与科技深度融合是顺应智能环境下教育发展的必然选择，其核心是充分激发信息技术的革命性影响，解决数字教育资源开发与服务能力不强、信息化学习环境建设与应用水平不高、教师信息技术应用能力基本具备但信息化教学创新能力尚显不足，以及高端研究和实践人才依然短缺等问题，推动教育观念更新、模式变革和体系重构。以上由英特尔与其众多合作伙伴共同开展的一系列 “人工智能 + 教育” 探索和实践，就是要以可落地的方案，帮助各级各类教育机构颠覆传统教与学模式，打造适应智能时代的现代化教育教学体系。

在下一篇章中，本文将就以上四个教育领域场景象限内的案例，围绕人工智能技术在其中的部署情况，共同探讨英特尔相关技术与产品在这些真实场景中的应用和优化方案。

# 实战篇

# 打造高效人工智能教学与实训解决方案

## 英特尔携手合作伙伴持续探索人工智能教学场景建设

### 人工智能教育市场现状与趋势

随着人工智能逐渐进入技术成熟度曲线（The Hype Cycle）中的生产成熟期（Plateau of Productivity），人工智能已在各领域得到广泛的应用，对行业的发展速度、内涵及质量产生了深刻的影响，并成为行业实施数字化、智能化转型的基石。与此同时，技术的飞速发展也带来了巨大的人才缺口。有统计数据表明，目前我国人工智能人才需求缺口达 500 万人[7]，在人才需求结构上，基础层人才需求尤为迫切。

为此，无论是教育行政主管部门，还是科研院校、中小学以及培训学校等各级各类教育机构，都把打造高质量的人工智能教育体系做为面向未来人才培养的重要方向和目标之一。国务院在 2017 年 7 月印发的《新一代人工智能发展规划》中也明确指出，人工智能成为国际竞争的新焦点，要完善人工智能教育体系，加强人才储备和梯队建设，逐步开展全民智能教育项目。

在政策激励和需求驱动下，众多教育机构积极落实行动。在 K12（即学前教育至高中教育阶段，一般代指“基础教育”）阶段，许多省市已经启动了人工智能教育的实验性建设。在高等教育阶段，已有数百所本科院校和高职院校开设了人工智能相关专业，并建立起一大批人工智能实验室供师生开展人工智能学习和实训。


图 2-1-1　K12 阶段的人工智能教育目标

在人工智能教育的不同阶段，对教学目标和教学环节的设置也有所不同。在K12阶段，如图2-1-1所示，人工智能教学主要是通过一系列基于场景实验的感性引导和动手实验，让中小学生通过感知、认知、创造和应用的初体验，了解典型的人工智能实现过程，提升AI素养，并在此基础上推动创新思维的发展。


图 2-1-2　高等教育阶段人工智能教育与人才培养体系

在高等教育阶段，如图 2-1-2 所示，人工智能教育则更多是注重针对人才需求实施分级培养。围绕着人工智能人才需求的三个层次：算法科学家、技术专家和应用工程师，在设计和开展教学时，通常遵循产业分析、课程建设、综合实训和就业认证四个主要环节展开。其中，产业分析是通过对企业场景进行细化和分析，确定热门领域以及亟需的人工智能岗位；课程建设是根据岗位需要，有针对地进行课程设置，确定一般专业课和核心专业课，打造完备的线上 / 线下培训体系；综合实训是设置贴近不同行业生产运行环境的真实案例应用，进行项目制的综合实训；就业认证是指导学生获取由人工智能龙头企业，如英特尔等提供的相关技术能力认证，并提供生态体系内的岗位就业指导。

### 人工智能教育面临的挑战及对策

人工智能从诞生伊始，就是一门需要将理论与实践充分融合，并在实际应用场景中开展实训论证的学科。比如，人工智能的算法演进通常都是为了解决某一场景中的具体需求，应用场景的变化以及对更高训练、推理效率和精度的要求，使更多新模型、新算法被提出。而新算法在提出后，也需要在实际场景中不断进行实践应用，才能积累更多的结果数据，进而对算法实施反向迭代优化。

所以人工智能教育的本质，也是通过合理的课程设置和实训环境，科学引导学生完成从理论到实践，再到创新应用的过程。

[7] 数据援引自工业与信息化部人才交流中心发布的《人工智能产业人才发展报告（2019-2020 年版）》一文

如图 2-1-3 所示，这个过程可以分解为：

- **从感知到认知：**通过场景实验帮助学生感性地了解人工智能在生活生产中的应用，触发学生的思考，揭开人工智能的神秘面纱，了解人工智能背后的基础理论，进而引导学生用多元的视角感知人工智能世界。
- **从认知到应用：**通过动手实验体验人工智能实现过程，让学生习得的知识得到实践，从而体会人工智能技术带来的成就感，激发其探究和应用技术的热情。
- **从应用到创新：**通过理论联系实际的教学活动，帮助学生进一步掌握知识和工具的实际应用方法，采用项目式教学等方式为学生搭建贴近实际场景的人工智能应用，在此基础上进一步开拓创新。


图 2-1-3 人工智能教育的不同阶段

无论是在 K12 阶段，还是在高等教育阶段的人工智能教学，面向不同应用场景开展人工智能课程设计和实训都是确保教学质量的关键环节。但在当前各类教育机构中，这两个环节还处于相对薄弱的状况。在课程设计上，由于人工智能是新兴的学科，因此教育机构往往欠缺完善的课程体系，课程特色不足，同时理论知识设置分散，针对不同层次人才的专指性课程欠缺；在师资层面上，由于人工智能技术不断推陈出新，新的算法、模型、技术和产品出现后，教师也需要大量时间重新进行体系建设，加之授课教师水平参差不齐，使人工智能专业师资队伍的建设也成为其推进过程中的一大难点。为改变这一现状，许多教育机构正与英特尔等人工智能行业先行者携手合作，利用他们在人工智能领域丰富的经验，实现知识点与技能点之间的完整链接，并构建起从理论到硬件，再到软件的综合性课程体系。

如图 2-1-4 所示，在某教育机构的人工智能课程设计中，不仅纳入了机器学习、深度学习、时间序列分析等人工智能领域常见技术方向，也在英特尔提供的软硬件基础之上，开设了使用英特尔® FPGA 进行深度学习推理，使用英特尔® 至强® 可扩展平台进行人工智能设计，以及利用 OpenVINO™ 工具套件开设人工智能推理加速等实用性非常强的课程，使现有学习内容与未来工作实操对接，让人工智能教育更具实践意义。

在人工智能实训阶段，从零起步的中小学、高校等教育机构往往缺乏适用于人工智能教育的实训环境，相关的实训实验室通常是在原有的电教室、微机室基础上改造而成，在应对大规模


图 2-1-4 典型的人工智能课程设计

学生进行人工智能实操时，往往存在以下问题：

- 缺乏规模化人工智能训练、推理所需的算力储备，传统 PC 在执行人工智能训练、推理时效率低下，而要大规模采购专用设备又必然使教育机构面临巨大成本开支。
- 缺乏面向不同应用场景、不同软件框架的软硬件优化方案，同时异构设备之间也难以实施有效协同。
- 可选实验场景、器材和软件套件混乱，无法满足真实场景的学习需要，更无法贴近行业实践要求。

为应对以上挑战，英特尔与众多人工智能教育技术合作伙伴和教育机构一起，引入一系列英特尔先进产品与技术，并通过针对 K12 阶段和高等教育阶段的不同教学目标进行专门优化，帮助教育机构打造基于“云 - 边 - 端”架构的、高效、灵活和可扩展的人工智能教学与实训环境。

## 基于英特尔产品与技术，打造“云 - 边 - 端”架构人工智能教育实训环境

### “云 - 边 - 端”架构构建高效人工智能教育实训环境

#### ■ 人工智能教育实训需要高效的“云 - 边 - 端”协同

传统的教育机构信息化系统，如校园网、课程管理系统、多媒体教室等通常采用烟囱式、岛屿式的系统架构，在面向人工智能教学需求时显现明显短板。一方面，校园内各 IT 系统数据彼此割裂，无法打通人工智能教学所需的从理论到实践的闭环；另一方面，人工智能教学管理者也缺乏有效的手段，针对不同实践场景进行课程编排、开发 / 运行环境部署和配置，以及对训练 / 推理任务调度实施管理。

同时，教育机构部署在教室等处的 IT 设备在以往大多用于支持课件演示、课程管理等应用，在算力输出上很难应对规模化大并发的人工智能实训所需，而依赖云端数据中心又很容易受到网络因素的影响，造成性能不稳定。

为助力打造更高效的人工智能教育实训环境，英特尔凭借其不断创新的产品与技术体系，以及在“云边协同”上积累的丰富实战经验，与人工智能教育解决方案厂商一起为教育机构打造“云 - 边 - 端”架构的端到端人工智能实训解决方案，为师生提供高性能、高可用和灵活可扩展的人工智能动手实践平台。

#### ■ 典型的“云 - 边 - 端”架构人工智能教育实训环境

典型的“云 - 边 - 端”人工智能教学与实训环境如图 2-1-5 所示，在校园环境内，可以由数据中心 / 私有云与边缘平台一起，通过校园网络构建 BS ( Brower-Server ) 或 CS( Client-Server )

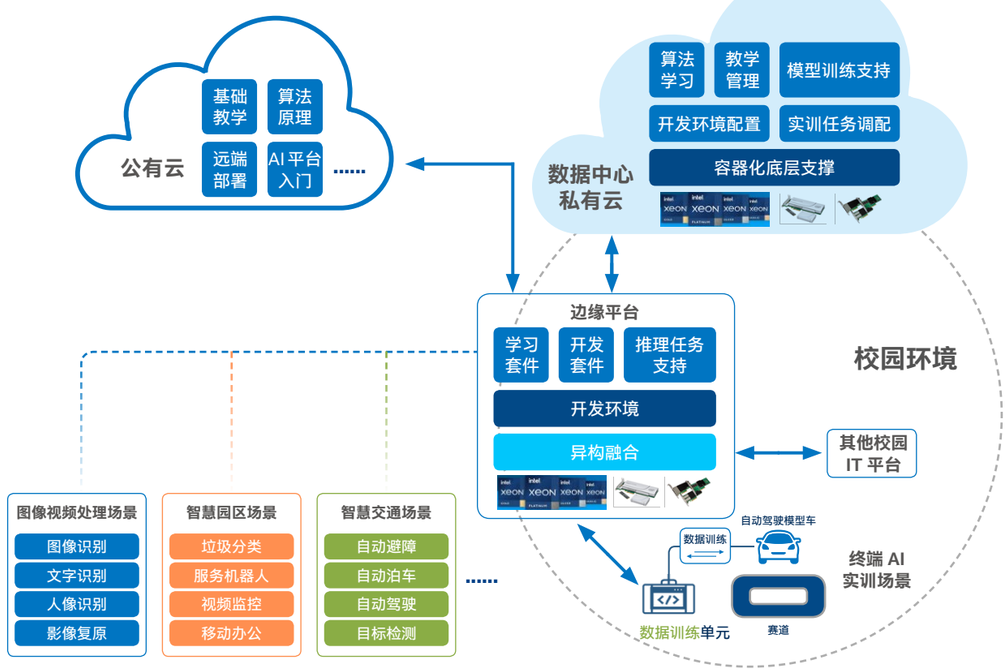



图 2-1-5 典型的“云 - 边 - 端”人工智能教育实训环境

架构的云边协同环境。其中，数据中心 / 私有云主要承载实训任务调配、模型训练支持，以及算法学习、教学管理等功能，并可通过容器等虚拟化方式，为师生提供相互独立的并行实训环境；而在边缘平台则可以就近部署人工智能实训所需的学习套件、开发套件以及推理任务等，对师生的实训过程直接提供算力支持。此外，将实训任务直接部署到边缘平台还有另一个优势，就是可针对不同的硬件基础设施，如通用处理器（CPU）、视觉处理器（VPU）、FPGA 等，在边缘平台上实施异构融合，从而打造更为高效、灵活和可扩展的平台支持能力。

基于数据中心 / 私有云与边缘平台提供的云边协同能力，教育机构可以在其上挂载各个人工智能实训任务终端，例如智能交通场景中的自动驾驶小车、智慧园区场景中的服务机器人等。同时，边缘平台还可借助互联网与公有云实施协同，向学生提供算法原理讲解、人工智能平台入门等远程教学能力，并供系统管理者进行远程部署和维护。

### ■ “云 - 边 - 端” 架构提供层层堆叠、灵活扩展优势

在实践中，云边协同架构能够带来的突出优势之一在于其能够层层堆叠、灵活扩展，让教育机构可根据人工智能教育实训课程的实际需求，以成本可控的方式，灵活开展方案的部署、优化和升级。

如图 2-1-6 所示，在实训课程开设之初，教育机构可在实训教室搭建基本的边缘平台并部署少量实训终端单元。随着课程规模和影响力的扩大，教育机构可以灵活地去扩展平台和训练单元的规模和数量，并逐步与校园数据中心以及其他 IT 平台打通数据通道，形成完备的人工智能教学实训方法体系。

而当校园内部课程、资源已无法满足学生下一阶段人工智能学习所需时，利用这一灵活可扩展的架构，教育机构可进一步打通与公有云教学资源的连接，利用互联网浩瀚的资源优势，如接入一些先进的开源深度学习平台项目等，来有效提升人工智能教学效率。

此外，基于这一架构部署的人工智能教学实训方案，也能帮助落后、边远地区的学生有机会开展人工智能教学。通过与公有云的连接，边远地区的学生可以远端接入本地实训系统，同本地学生一起接受高质量的人工智能教学，从而缓解不同地区间的教育不均衡现象。

## 借力英特尔® 至强® 处理器平台，实现低成本 AI 实训 “云”

在人工智能教育实训阶段，通常会出现数十名学生同时进行模型训练或推理的场景，需要云端算力中心或边缘计算平台提供稳定高效的并行多路算力支持。而这首先需要方案中的算力设备有能力支持足够多的虚拟机或容器，其次需要部署足够多的内存以及性能可靠的网络设备，来保证每个学生都能获得所需的实训算力。

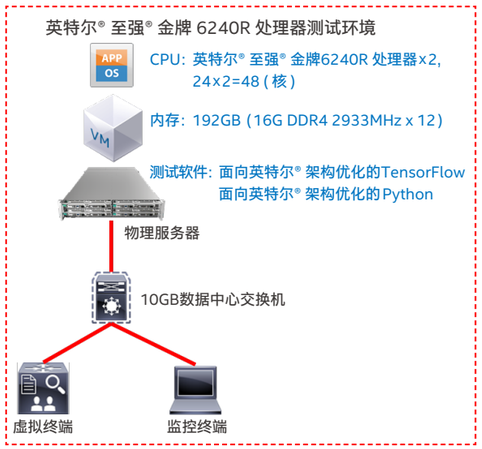



图 2-1-7 基于英特尔® 至强® 可扩展处理器的实训环境构建方案

在实践中，可采用图 2-1-7 中所示架构来构建人工智能教育实训环境。该架构以基于双路英特尔® 至强® 金牌 6240R 处理器，内置 192GB DRAM 内存的物理服务器为核心，通过 10GB 数据中心交换机将算力资源推送到学生使用的各个虚拟终端中。

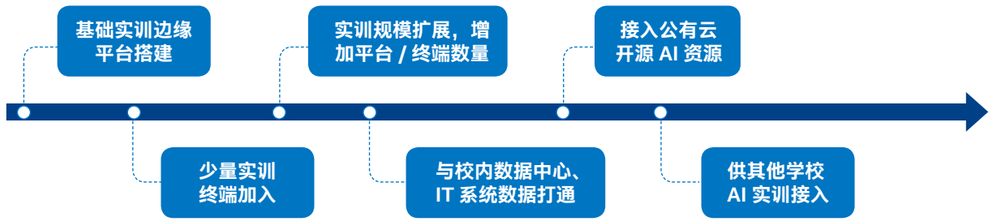



图 2-1-6 不断优化完善的 “云 - 边 - 端” 人工智能教育实训环境

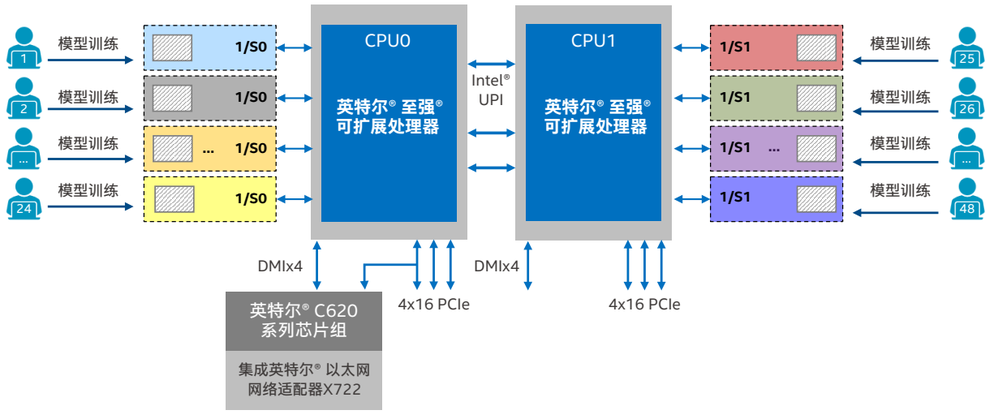



图 2-1-8 英特尔® 至强® 可扩展处理器提供多路并行算力输出

一直以来，为满足人工智能教学实训环境中，多人同时开展训练或推理的需求，实训方案中的算力设备需在提供充足算力总量的同时，也需能将算力、内存等资源均匀地并行分配给每个虚拟终端使用。一些算力设备虽然同样具有大规模算力输出的能力，但却缺乏多路并行输出的能力。教育机构要实现数十名学生同时进行实训操作，就必须采购大量的算力设备，也因此抬升了实训环境的建设成本。

而英特尔® 至强® 可扩展处理器能以其优异的微架构设计，以及更多的核心、线程和高速缓存数量，一方面保证边缘实训平台获得足够的算力总量，支撑整个实训课程过程中的全体师生所需；另一方面，如图 2-1-8 所示，处理器平台也能通过核心绑定（Core Binding）技术以及非一致存储访问（Non-Uniform Memory Access，NUMA）架构等技术，实现先进的多路并行算力输出能力，使每位实训学生都可获得所需的算力。

多核处理器平台通常会根据工作负载的变化不断对工作核心实施调整，以期获得最优性能输出，但这也带来多路并行算力输出的不稳定。而核心绑定技术是让系统进程不管如何调整，分配到每个虚拟终端的工作负载都始终在同一个核心上执行，这就保证了多路并行算力输出的稳定性。

而英特尔® 至强® 可扩展处理器所采用的 NUMA 架构通过处理器和内存之间的相对位置，将内存分为近端内存和远端内存，每个处理器节点优先访问对应的近端内存，从而有效解决多核处理器访问内存时的性能问题，获得更快的访问和处理速度，在提升云端算力中心或边缘计算平台并行任务处理能力的同时，也可让学生获得一致的人工智能实训体验。

以利用双路英特尔® 至强® 金牌 6240R 处理器（每处理器包含 24 个物理核，系统共包含 48 个物理核）系统实现 48 个学生课堂实训为例，通过 NUMA 架构分配计算资源如下表 2-1-1 所示：

| 分配给学生实训使用的系统核编号 | 学生实训座位号 |
|---|---|
| #0,48 | #1 |
| #1,49 | #2 |
| … | … |
| #47,95 | #48 |

表 2-1-1 课堂实训中通过 NUMA 架构分配计算资源方案

利用 numactl 这样的系统指令可以对处理器节点以及访问本地内存进程予以精确控制，从而获得更优的计算性能。例如通过 numactl –H 可以获知当前服务器的处理器节点以及对应本地内存分布情况：

```
# numactl -H
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66
67 68 69 70 71
node 0 size: 191925 MB
node 0 free: 188113 MB
node 1 cpus: 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
42 43 44 45 46 47 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87
88 89 90 91 92 93 94 95
node 1 size: 193511 MB
node 1 free: 190707 MB
node distances:
node   0   1
  0:  10  21
  1:  21  10
```

假设现在执行 mnist.py ( 所执行模型文件使用 Python3 编写 ) 时，只使用了 #0 处理器中的 0-23 核心和 48-71 核心，且只访问 #0 处理器对应的近端内存，可以使用如下命令：

```
numactl -C 0,48 -m 0 python3 mnist.py
numactl -C 1,49 -m 0 python3 mnist.py
…
numactl -C 23,71 -m 0 python3 mnist.py
```

在一些教学实训的实践中，基于图 2-1-7 ( 第 18 页 ) 所示方案搭建的实训环境，已被证明能在语音识别训练等实训任务中，同时满足 48 人 ( 一个实训教室 ) 的同时使用需要，很好匹配了教学实训的场景需求。虽然在针对其它模型训练的实训任务中，方案所支持的并发数可能有所改变，但总体而言，来自教育机构的反馈已经证明，采用英特尔® 至强® 可扩展处理器搭建的实训环境，可以大幅提升实训教室的使用效率，让人工智能学习者获得更多动手实践的机会，提高人工智能学习的效率。

按照以上资源分配方法，如表 2-1-2 所示选择不同的英特尔® 至强® 可扩展处理器金牌系列的型号，可灵活满足不同规模的实训要求：

| 实训并发规模 | 英特尔® 至强® 可扩展处理器 | 双路服务器的物理核数 |
|---|---|---|
| 多至32学生席位 | 英特尔® 至强® 金牌6226R处理器 | 32 |
| 多至40学生席位 | 英特尔® 至强® 金牌6242R处理器 | 40 |
| 多至48学生席位 | 英特尔® 至强® 金牌6240R处理器 | 48 |
| 多至52学生席位 | 英特尔® 至强® 金牌6230R处理器 | 52 |
| 多至56学生席位 | 英特尔® 至强® 金牌6238R处理器 | 56 |

表 2-1-2 不同英特尔® 至强® 可扩展处理器金牌系列型号可灵活满足不同规模实训需求

## 基于英特尔® NUC 平台，打造可灵活扩展的人工智能边缘终端

人工智能教育的重要目标之一，是让学生通过熟悉不同硬件环境的搭建、软件框架的配置，进而快速了解不同类型人工智能应用场景的设计、搭建和实验方法。因此，在人工智能教育实训环节中，教育机构希望引入对老师和学生友好、开发周期短、学习曲线低、易兼容各种系统的产品套件，来打造通用性更强、且能快速上线落地的实训环境。

根据学习阶段的不同，教育机构通常希望部署以下三种不同应用场景实训产品套件：

- **学习套件：** 主要是满足学生对人工智能推理实践的学习场景，其可以安装各种开发框架和推理开发套件，帮助学生基于该环境进行算法模型的推理实践开发；
- **开发套件：** 主要是满足算法模型开发的场景，其通常配置了强劲的算力，使模型不必跑在大型服务器上。而小型化的设计，也令算法模型的开发和学习场所不再拘泥于实验室内；
- **部署套件：** 主要是用于部署实践，为人工智能学习和实践者提供小身材、大能量的边缘部署终端。

同时，对快速响应性能、灵活可扩展能力等方面的高标准要求，使人工智能教育实训环境天然成为 “人工智能 + 边缘计算” 的落地方向之一。为此，基于英特尔® NUC 平台所开发的一系列人工智能边缘终端，满足人工智能学习者在学习、开发和部署的不同阶段对硬件基础设施的需求。如表 2-1-3 所示，基于以上三种不同应用场景，英特尔® NUC 平台能够以英特尔® 酷睿™ 处理器平台、英特尔® 至强® 处理器平台以及英特尔® VPU 产品为算力核心，通过不同组合给出相应的推荐配置。

| 实现场景 | NUC规格 | CPU | VPU | FP16算力 | 内存 |
|---|---|---|---|---|---|
| 学习套件  | • 2.5G以太网口；<br>• 内置Wi-Fi6和蓝牙5.0；<br>• 支持4屏显示输出；<br>• 4个USB3.1接口； | 第十一代智能英特尔® 酷睿™ i5处理器及以上； | 1颗MA2485 | 最高可达 4.94T FLOPS | 16G DDR4 |
| 开发套件  | • 双千兆以太网卡；<br>• 6个USB 3.1；<br>• 2个雷电3；<br>• 2个USB2.0内置接口； | 第十一代智能英特尔® 酷睿™ i7处理器或 英特尔® 至强® E-2286M 及以上； | 1颗MA2485 | 最高可达 10.3T FLOPS | 32G DDR4 |
| 部署套件  | • 双千兆以太网卡；<br>• 双HDMI；<br>• 3个USB 3.0及3个USB2.0接口；<br>• 内置串口；<br>• 12-24V宽压供电； | 第十一代智能英特尔® 酷睿™ i5处理器及以上； | 2颗MA2485 | 最高可达 5.94T FLOPS | 8G DDR4 |

表 2-1-3 基于不同应用场景的英特尔® NUC 平台产品套件

### ■ 基于英特尔® NUC 平台产品套件的自动驾驶场景实验 / 实训环境

作为一种典型的人工智能应用方向，自动驾驶所运用的人工智能核心概念包括大数据、机器学习、深度学习、计算机视觉，人机交互、AI 推理、智能控制等。其目的是实现能够感知、推理、行动和调整的程序，达到机器模仿人类心智，比如 “学习” 和 “解决问题” 的 “认知功能” 来实现自动驾驶。

为此，英特尔与联合伟世一起，基于英特尔® NUC 平台产品套件与训练系统搭配，共同打造完整的、面向 K12、高职高教等不同教育阶段的人工智能自动驾驶场景实验 / 实训解决方案。

如图 2-1-9 所示，产品方案包含自动驾驶模型车与专业赛道，配套了完善的实训课程和实验，内容涵盖了深度学习、监督式学习、非监督式学习和强化学习等人工智能知识点，实现自动驾驶中自动避障、红绿灯识别、标志识别和行人识别等应用场景。

方案中，基于英特尔® NUC 平台构建的上位机系统以英特尔® 酷睿™ 处理器为算力枢纽，主要承担数据融合处理、人机交互等功能，是实现自动驾驶场景实验 / 实训的核心大脑。同时，平台所集成的英特尔® VPU 产品以 2 颗前置摄像头作为视觉主传感器，能从海量数据中提取特征并使用多层神经网络，以良好深度学习推理加速能力为实验小车提供行进过程中的多视觉识别与决策。

基于以上能力，实验小车在行进过程中无需联网后端交互，先进的边缘终端架构设计让实验小车可自行完成行进过程中的推理任务，面对红绿灯、停车指示牌、行人标识可以直接做出推理判断。

通过这一实训环节和实训环境，可以让学生直观地感受数据采集、数据清洗、数据标注、模型搭建、模型训练、模型调优和模型推理等一系列复杂的人工智能全流程。从而由浅入深地了解人工智能的核心问题、技术方法及实际应用。

## OpenVINO™ 工具套件提供推理加速和异构融合方案

如前文所述，对于大多数高校、中小学等教育机构而言，人工智能教育的目标更多是激发学生对人工智能知识的兴趣并提升学生的动手能力，而非承担具体的人工智能应用开发。因此，其人工智能实训环境在性能上往往并不宽裕，并可能根据不同的应用场景基于多样化的硬件，如通用处理器、内置图像处理器或 FPGA 等，构建了人工智能平台。

但是，为了有效支持教学活动，教育机构在构建实训平台方案时，其实还需要充分考虑学生进行的推理等人工智能任务是否可在课堂时间内完成。同时， 出于校园实训环境建设方案的差异，同一项人工智能任务可能需要在不同硬件平台，例如 CPU、FPGA 等上执行，如何避免或减少因硬件平台差异带来的影响，也需要在方案设计时予以考虑。

由英特尔开源的 OpenVINO™ 工具套件能够为以上两种挑战提供应对方案。这一工具套件产品通过模型优化器 ( Model Optimizer ) 和推理引擎 ( Inference Engine ) 这两个核心组件，为人工智能教育实训提供了良好的推理加速和异构融合能力。

其中，模型优化器具有的量化功能能基于不同人工智能框架，如 TensorFlow、PyTorch、MXNet 等训练后的 FP32 数据格式模型，在精度损失可接受的前提下转化为 INT8 等低精度模型。然后通过英特尔® 至强® 可扩展处理器所集成的英特尔® 深度学习加速 ( Intel® Deep Learning Boost, 英特尔® DL Boost ) 技术为后续推理提供大幅加速。


图 2-1-9 自动驾驶场景实验 / 实训涉及的人工智能环节

而 OpenVINO™ 工具套件的推理引擎组件，则能基于通用 API（基于 C/C++ 或 Python 开发）为人工智能实训场景提供强大的异构融合能力，让学生的人工智能应用在一次编写之后，就可以更智能、也更有针对性地选择英特尔的 CPU、VPU、GPU、FPGA 等平台，来实现异构部署及更优的加速能力。

一般地，师生在实操时，可以使用以下命令进行模型转换：

```
python mo_tf.py --input_model /usr/models/model.pb
```

或

```
python mo.py --framework tf --input_model /usr/models/model.pb
```

以图像分类场景为例，采用经典的 ResNet50 v1.5 模型[8]。在使用英特尔® 架构处理器平台时，学生首先可以使用以下命令进行模型转换：

```
python mo_tf.py --input_model ./model/ResNet50_v1.pb --input_shape
    [1,224,224,3] --mean_values [123.68,116.78,103.94] --output softmax
    _tensor --output ./model --model_name ResNet50_v1.5_fp32
```

下一步再使用以下命令完成模型的推理验证：

```
python classification_sample_async.py -m ./model/resnet50_v1.5_fp32
    .xml -i ./data/test.jpg -d CPU --labels ./data/imagenet1001_labels.txt
```

## 英特尔面向不同人工智能框架提供性能优化

为更大程度地发挥处理器硬件带来的性能优势，英特尔为云端算力中心或边缘计算平台中的人工智能系统提供了一系列人工智能框架，这些框架能够通过所集成的英特尔® MKL、英特尔® oneDNN 等函数库，面向基于英特尔® 架构的硬件基础设施提供多种针对机器学习 / 深度学习的训练和推理优化能力。

以面向基于英特尔® 架构优化的 TensorFlow 为例，其既是流行的深度学习框架，也是人工智能教学与实训阶段最常用于实操的框架之一。为此，英特尔基于硬件架构特性，为其提供了一系列优化，包括调整处理器设置、引入对 NUMA 架构的支持，以及引入英特尔® oneDNN 函数库等。使用面向英特尔® 架构优化的 TensorFlow 框架，可以采用以下方法。

### ■ 环境变量设置

1) 清空系统的缓存（cache），以 CentOS 为例，将处理器设置为性能优先的模式：

```
echo 0 > /proc/sys/vm/compact_memory
echo 0 > /proc/sys/vm/drop_caches
echo 0 > /sys/devices/system/cpu/intel_pstate/min_perf_pct
echo 0 > /sys/devices/system/cpu/intel_pstate/no_turbo
echo 0 > /proc/sys/kernel/numa_balancing
cpupower frequency -set -g performance
```

2) 将线程在执行完当前任务并进入休眠之前需要等待的时间设置为 1 毫秒：

```
export KMP_BLOCKTIME=1
```

3) 线程绑定设置为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，可以充分利用多级缓存带来的优势：

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

4) 将并行执行线程的数量设置为处理器核心数（本样例中设为 2）：

```
export OMP_NUM_THREADS=20
```

### ■ 添加线程控制

进行 tf.ConfigProto() 初始化，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。其中，intra_op_parallelism_threads 用于控制单一运算符 op，如矩阵乘法、reduce_sum 等内部的并行操作，通常设置为处理器物理核心数目；inter_op_parallelism_threads 控制多个运算符 op 之间的并行计算，通常设置为 1 或 2：

```
config = tf.ConfigProto()
config.allow_soft_placement = True
config.intra_op_parallelism_threads = FLAGS.num_intra_threads
config.inter_op_parallelism_threads = FLAGS.num_inter_threads
```

除 TensorFlow 框架之外，英特尔也对人工智能教育实训环节中其他常见的人工智能软件框架进行了大量有针对性的优化，

更多详情，请访问英特尔相关页面：https://software.intel.com/content/www/cn/zh/develop/tools.html

## 基于英特尔® 架构平台推动迁移学习在人工智能教学中的应用

众所周知，课堂教学是有很强时间限制的场景，而传统全量人工智能训练过程往往无法在几十分钟时间内完成。因此，迁移学习（Transfer Learning）因其效率优势，不仅广泛用于商业化人工智能场景，也日益受到人工智能教学场景的关注。

许多人工智能任务都存在相关性，将已训练好的模型参数迁移到新的模型中，可以有效地避免重复“造轮子”的过程。人工智能教学通常是面向成熟的数据集进行课程设计，一方面，已有预训练模型的数据集与新数据集有很多相似性；另一方面，人工智能教学实训的许多环节，并不需要学生对全面训练过程进行跟踪，例如仅仅需要通过特征向量去训练模型，此时选择迁移学习的方式就特别合适，其不仅能提升实训环节模型训练任务的执行效率，也可使实训平台的算力需求以及相关成本大幅下降。

根据训练模式和范围的不同，迁移学习可以分为几种模式。例如根据预训练模型的特征向量去训练新的模型，即特征提取（Extract Feature Vector）模式；或者是冻结预训练模型的部分卷积层，训练剩下的卷积层和全连接层，即参数微调（Fine-Tune）模式等。

通常地，根据新数据集的大小以及其与原数据集的相似度，教学平台可以引导学生以不同的迁移学习模式组合成相应的策略，如图 2-1-10 所示，

- 当新数据集比较大且和原数据集相似度较低时，可以对整个网络进行重新训练（图中策略 A）；
- 当新数据集比较大且和原数据集相似度较高时，可以采用参数微调模式微调整个网络（图中策略 B）；
- 当新数据集比较小且和原数据集相似度较低时，可以使用前面的特征来训练分类器（图中策略 C）；
- 当新数据集比较小且和原数据集相似度较高时，可以使用预训练网络当作特征提取器，用提取的特征训练线性分类器（图中策略 D）。

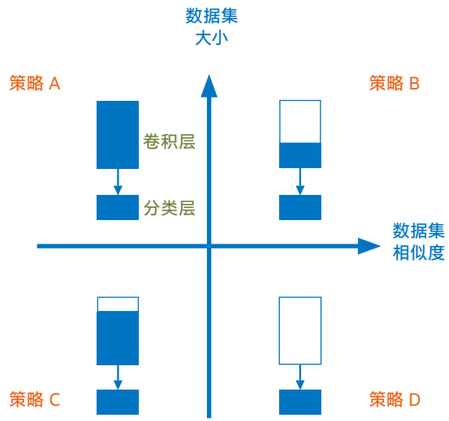



图 2-1-10 基于不同数据集特性的迁移学习策略

迁移学习大幅降低了进行完整深度神经网络训练时所需的巨大算力资源，使用户可以灵活便捷地调度基于英特尔® 架构处理器的算力资源来完成训练任务，这对于教育机构无疑是重大利好。例如，一方面学校可以通过人工智能任务调度系统，利用学校 IT 系统的空余算力完成训练；另一方面，在实训环节中，也可以将处理器算力尽可能细分，以保证更多学生得到实训机会。

以 Cifar100 图像分类为例，其采用 TensorFlow 框架实现迁移学习的代码示例如下：

```
import tensorflow as tf
from tensorflow.keras import Keras

# prepare model and freeze layers
base_model=Keras.applications.VGG16(include_top=False, weights
    ='imagenet', input_shape=(32,32,3))
base_model.trainable=False
model=Keras.Sequential(name='vgg16-finetune')
model.add(base_model)
model.add(Keras.layers.Flatten())
model.add(Keras.layers.Dense(1024, activation=tf.nn.relu))
model.add(Keras.layers.Dense(512, activation=tf.nn.relu))
model.add(Keras.layers.Dense(100, activation=tf.nn.relu))
model.summary()
# start fine-tune
optimizer=tf.keras.optimizers.Adam(learning_rate=0.001)
eval_func=tf.keras.metric.CategoricalAccuracy()
model.compile(optimizer=optimizer, loss='categorical_crossentropy',
    metrics=['acc'])
model.fit(x_train, y_train, batch_size=64, epochs=20, validation_
    data=(x_val, y_val))
```

一项面向图像识别场景的验证测试，充分展现了迁移学习方法为用户带来的效率优势。如图 2-1-11 所示，在使用深度学习方法进行训练时，耗时约 20 小时，而采用迁移学习方法，训练时间可缩短至 15 分钟，即使得训练效率提升可达 80 倍之多。

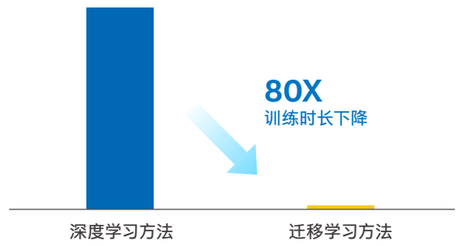



图 2-1-11 迁移学习带来训练时长大幅降低

更多测试详情，请参阅：https://software.intel.com/content/www/us/en/develop/articles/use-transfer-learning-for-efficient-deep-learning-training-on-intel-xeon-processors.html

[8] 模型可参考：https://zenodo.org/record/2535873/files/resnet50_v1.pb

# 基于英特尔优化方案的应用案例

## 联合伟世：“云 - 边 - 端” 协同，采用先进硬件与创新理念打造高效人工智能教学实训平台

### 引言

*“基于英特尔产品与技术打造的人工智能教学组件，结合‘云 - 边 - 端’分层部署方式，让我们的新方案可以高效、灵活可扩展地为不同学习阶段的学生，提供高品质的人工智能教学与实训体验，大幅提升人工智能教学效率。”*

——联合伟世

### 背景与挑战

在 K12 和高等教育阶段开展不同层级的人工智能教学，是弥补我国现有人工智能人才缺口，提升人工智能行业全球竞争力的有效手段之一。作为一门实践性较强的学科，人工智能教育在理论基础、算法框架和平台入门教学之外，面向不同应用场景开展有针对性的人工智能实训，也是快速提升学生人工智能水平的良好手段。

由于教育机构的传统 IT 设施并非专门为人工智能教学而设置，因此，在学生进行规模化并行人工智能实训时，往往囿于系统性能和处理时延而造成效果不佳，同时学校也缺乏对人工智能教学进行统一调度、管理和优化的手段。

为解决这些问题，英特尔与深度合作伙伴联合伟世协作，基于英特尔硬件强大的计算能力和数据训练方法，并结合学科教学、师资建设和学生发展的需求，向各级教育机构提供一套完整适用的人工智能教学端到端全栈实验室解决方案，为人工智能教学以及相关的实训环节提供平台，并取得了良好的实践效果。

### 解决方案

#### ■ 方案解析

如图 2-1-12 所示，联合伟世与英特尔共同打造的人工智能教学实训系统在架构上采用了 “云 - 边 - 端” 的分层部署方式，其核心组件包括（可根据实际需要增加或删减）：

- **人工智能 LAB:** 部署于数据中心 / 私有云，用于人工智能实训管理，对实验数据进行存储、交互和管理，并可扩展调用公有云算法模型；
- **AI NUC:** 部署于边缘平台，内置一系列计算学习和开发套件，支持人工智能推理、训练任务以及异构融合，并可用于嵌入式场景创新开发；
- **实训项目:** 用于开展前端数据采集并反馈到边缘平台，以自动驾驶为例，自动驾驶实验小车带有数个摄像头，其可以通过机器视觉的方式抓取图像数据。

在具体实践上，由实验小车自动采集完的数据放到边缘平台的训练机中，通过相应算法对所有数据进行加工，包括数据预处理、打标签以及后续的建模、训练等，最后形成一个用于自动驾驶的模型。以上完整的、与行业特点深度结合的、从数据到实验结果的完整流程，能有效帮助学生培养对于 “人工智能 + 行业” 的思维模式和兴趣，了解人工智能基本概念，理解人工智能核心知识点及实现方法，从而发挥自身创新力，逐步实现从学习到认证、再到实习就业的完整教育全周期。

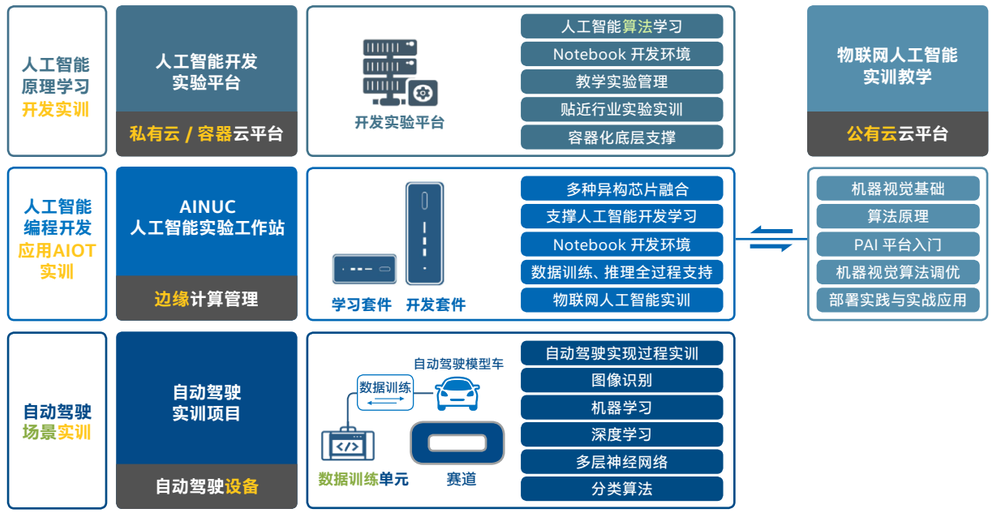



图 2-1-12 联合伟世人工智能教学实训系统架构

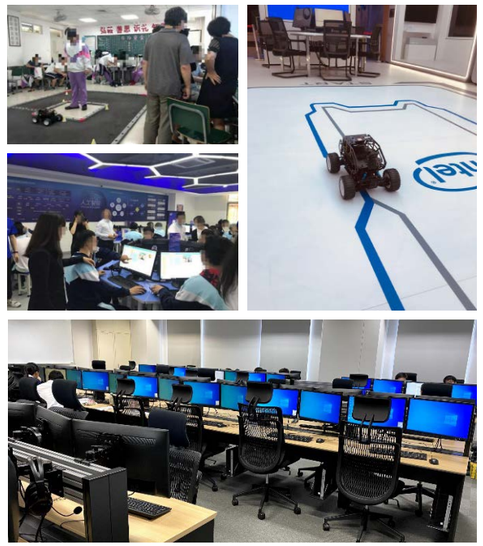


图 2-1-13 联合伟世联合教育机构开展人工智能教学实训现场

#### ■ 方案亮点

**在架构设计以及针对人工智能教学的功能设置上：**

- 新方案构建了 “云 - 边 - 端” 的系统架构，可以按需搭建、灵活扩展。保证了人工智能教学既可以在纯终端实现，也可以通过云端加边缘的远程交互式环境来实现；
- 新方案采取了分层递进的系统实训，从基于实验平台传感器进行的数据采集，到开展自动驾驶的人工智能实训，再到通过可视化编程，为学生提供分层向上、逐步递进的学习过程。

**在基础设施构建和人工智能性能加速上：**

- 新方案采用一系列基于英特尔® 架构的产品作为算力基础，包括英特尔® 至强® 可扩展处理器在内的先进产品，以更优异的微架构设计、更多的核心和更大的内存支持，结合 NUMA 架构带来的优势，使众多学生在人工智能实训环节中可以体验到一致的训练推理效果；
- 新方案中引入了 OpenVINO ™ 工具套件、面向英特尔® 架构优化的 TensorFlow 深度学习框架等，不仅有效提升了人工智能训练、推理的效率，也进一步优化了最终结果；
- 新方案中，引入英特尔® NUC 平台，融合通用处理器、内置 GPU 和 VPU 等多种异构算力平台构建了 AI NUC 产品，提供了大量预训练好的模型库以及丰富的示例代码，为学生进行人工智能实训提供了良好平台。

### 实战成效

为了验证 NUMA 架构下，英特尔® 至强® 可扩展处理器可为人工智能实训环节中的各个学生提供独立稳定的性能输出，英特尔与联合伟世一起面向 K12 阶段的人工智能教学实训场景进行了测试。测试服务器配置了双路英特尔® 至强® 金牌 6240R 处理器，该款处理器具有 24 个核心、48 个线程。

面向 K12 阶段的人工智能教学实训的主要目的，是让学生对人工智能的训练等过程形成基本认识并产生学习兴趣。因此，实训课程的设计以短小精炼为特色，训练时间一般控制在 300 秒左右，且需要保证每个学生（单个班级一般为 48 位学生）都能分配到计算资源，参与到训练过程中去。

在测试中，系统会将处理器的每个物理核心以及对应的虚拟核心分配给一位学生，使他（她）们进行语音识别场景与另一个分类场景中的人工智能训练任务。在测试开始前，双方工程师首先进行了环境配置，包括：

1) 基于面向英特尔® 架构优化的 TensorFlow 框架进行环境配置：

环境配置命令请参考第 22 页 "英特尔面向不同人工智能框架提供性能优化“ 部分。

2) 基于 oneMKL 对 Kaldi 库进行重编译：

```
Modify compile options in kaldi.mk:
CXXFLAGS = -std=c++11 -I.. -isystem $(OPENFSTINC) -O3 $(EXTRA_
CXXFLAGS) \
            -Wall -Wno-sign-compare -Wno-unused-local-typedefs \
            -Wno-deprecated-declarations -Winit-self \
            -DKALDI_DOUBLEPRECISION=${DOUBLE_PRECISION) \
            -DHAVE_EXECINFO_H=1 -DHAVE_CXXABI_H -DHAVE_MKL \
            -I${MKLROOT)/include -m64 -msse -msse2 -pthread  -mavx2 \
            -mfma  -mavx512f -mavx512vl -mavx512bw -mavx512dq \
            -mavx512cd -mavx5124vnniw
Compile Kaldi library
]# make -j10
```

测试结果如图 2-1-14 所示，在两种不同的训练任务中，每个参与人工智能实训的学生均取得了令人满意结果，不仅训练时长均在预期之内（面向 K12 阶段的实训课程中，一般预留训练时长为 300 秒），训练效果（ACC 值）也分别达到了 100% 和 88%。

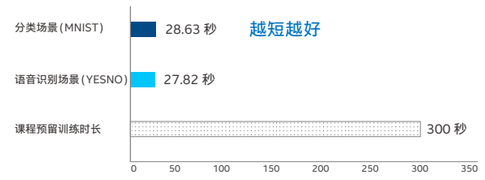



图 2-1-14 联合伟世方案中多人同时开展人工智能训练效果

该项测试结果及新方案在其他教育机构的实践反馈均表明，基于英特尔® 架构基础设施的新方案能够很好地满足各层级教育机构开展人工智能教学和实训的需要，从而有效地对人工智能教学所需的学科体系建设提供支撑。

## 客户证言

*“在人工智能发展进入新阶段的时代背景下，华南农业大学顺应大湾区现代农业发展的趋势，充分发挥我校在人工智能方向上学科发展和人才培养的优势，积极开展了人工智能专业建设工作。*

*人工智能专业的课程体系尤其是实践、实训课程的人才培养路径，仍处于探索研究阶段。联合伟世基于英特尔技术的开发实验平台集成了较为丰富的课程资源，也蕴涵了大量实战型的实训案例。我校电子工程学院（人工智能学院）人工智能系即采用了联合伟世基于英特尔技术的人工智能开发平台，以支撑人工智能专业实践、实训课程的顺利开展，从而有力保障了基础和专业课程实践教学环节的实施。”*

陆健强

华南农业大学

## 解决方案中的软硬件配置建议

### 硬件配置

| 名称 | 规格 |
|---|---|
| 处理器 | 双路英特尔® 至强® 金牌 6240R 处理器 |
| 基础频率 | 2.40GHz |
| 核心/线程 | 24/48 |
| HT | On |
| Turbo | On |
| 内存 | 192G (16G DDR4 2933MHz x 12) |
| 硬盘 | 英特尔® 固态盘 DC P4610 系列及以上 |
| 网卡 | 英特尔® 以太网网络适配器 X722 |

### 软件配置

| 名称 | 规格 |
|---|---|
| 操作系统 | CentOS 7.8.2003 |
| 内核版本 | 3.10.0-1127.18.2.el7.x86_64 |
| 编译器 | GCC 8.0 |
| 框架 | Kaldi-7a50987/ 面向英特尔® 架构优化的 TensorFlow2.4.0 及以上 |
| 库 | 英特尔® oneAPI Math Kernel Library 2021.2.0 |

# 五舟科技：高性能硬件助力打造高校人工智能教学平台

## 引言

*“我们的教学平台由英特尔® 至强® 可扩展处理器等产品提供强劲算力，并由英特尔® 深度学习加速提供人工智能加速力，为高校进行贴近行业场景的人工智能教学与实训课程设计提供了关键支撑。”*

——五舟科技

## 背景与挑战

作为高新技术人才重要的输出地，高等院校正不遗余力地通过人工智能学科建设，推动人工智能领域的产学研融合。在这一进程中，高校将人工智能的教学与科研创新及行业需求相结合，以促进前沿人工智能技术落地应用作为重要方向，有的放矢地进行课程设置和实训。例如，面向医疗与金融领域，课程强调与大数据的结合；在语音图像领域，则更加重视深度学习方法的设计；而在智能制造领域，时序性预测正获得重点关注。这一紧贴行业需求的教学模式，注重培养学生们的实践动手能力，显然需要高效和易于管理的人工智能教学实训平台予以支持。

为此，英特尔与深度合作伙伴广州五舟科技股份有限公司（以下简称“五舟科技”）携手，构建了深度学习与人工智能科研环境管理平台解决方案。新的平台方案以一系列英特尔先进产品与技术为抓手，借助英特尔® 至强® 可扩展处理器等产品提供的算力，有效应对高校开展人工智能教学和实训时遇到的挑战。

## 解决方案

### ■ 方案解析

如图 2-1-15 所示，五舟科技深度学习与人工智能科研环境管理平台采用“中心 - 边缘”多层 BS 架构的模块化设计。针对数据中心 / 私有云环境，该方案可根据不同行业场景人工智能教学实训的需要，构建一系列基于英特尔® 架构产品的服务器集群，如人工智能计算资源集群、大数据融合集群以及项目工程案例集群，并采用以容器为主的虚拟化技术，来组成面向数据存储运算、人工智能模型训练及人工智能模型部署应用所需的资源池，通过万兆数据交换矩阵实现互联互通，充分满足学生实训所需。


图 2-1-15　五舟科技深度学习与人工智能科研环境管理平台架构

采用容器化的虚拟方式，可以有效降低学校数据中心的部署压力。平台通过容器镜像方式，供师生进行不同人工智能实训场的创建和使用。此外，平台对流行深度学习框架，包括 TensorFlow、PyTorch、Caffe /Caffe2 等，都有着充分支持，并预装了主流深度学习工具和驱动程序。无论是高校科研人员还是学生，都无需过多关注环境部署而将更多精力投入到人工智能任务本身。

在边缘侧，方案支持基于通用处理器、FPGA、英特尔® Movidius™ VPU 等设备构建的边缘实训环境，能够为师生提供异构融合、训练、迁移学习等能力，并有效连接实训项目中的边缘计算智能设备。在平台的使用过程中，师生、项目实训团队可以基于平台提供的不同登录门户，执行统一的硬件资源编排管理、人工智能工作流调度、性能监控以及系统调配、功能升级等任务。得益于基于英特尔® 架构的处理器强大的多核性能，平台可以划分出大量并行任务处理进程供不同人工智能任务调用，确保了各项科研项目的高效进行。

### ■ 方案亮点

作为一个高效的人工智能科研与教学平台，方案经由五舟科技与英特尔紧密合作，基于创新的软硬件，具备一系列领先功能与特性：

**一键部署算力资源**

平台提供了一站式人工智能计算资源配置模板，将过去繁杂的容器设置参数集成到统一的配置页面中。使用者可根据需求，灵活自主地配置容器系统镜像、算力设备、内存以及存储空间；平台可自动下发到底层容器进行资源调度，快速启动人工智能实训进程。

**强劲的算力支撑与人工智能加速**

在算力部署上，平台服务器集群均采用了基于英特尔® 架构的处理器产品，其在核心数量、内存容量支持等方面的优势，可令平台快速启动及部署分布式人工智能训练和推理任务，令运行效能大幅提升。

## 实战成效

为验证基于英特尔® 架构的产品构建的新平台在进行一系列特定优化后的性能表现，英特尔与五舟科技一起，以面向高等教育人工智能专业实训场景为例，进行了相关的验证性测试。测试服务器配置了双路英特尔® 至强® 金牌 6230R 处理器，该款处理器具有 26 个核心、52 个线程。

与面向 K12 的方案不同的是，高等教育人工智能实训在难度、深度以及复杂度上有着较大提升，实训时间也会增长到 40 分钟左右。因此在测试中，系统需要将处理器的 2 个或 6 个核心分配给一个实训组开展分类场景（使用 IMDB 数据集）、图像识别场景（使用 Cifar100 数据集）两项人工智能训练任务。在测试开始前，双方工程师首先进行了环境配置，包括：

1) 基于面向英特尔® 架构优化的 TensorFlow 框架进行环境配置：

环境配置命令请参考第 22 页 "英特尔面向不同人工智能框架提供性能优化" 部分。

测试结果如图 2-1-16 所示，在两种不同的训练任务中，各个参与人工智能实训的实训组均取得了令人满意的结果，与以往单项实训任务 40 分钟左右的设计时长相比，新平台完成两项训练任务分别只需要 459 秒和 928 秒，大幅提升了实训效率；同时，在仅训练 10 个 Epochs 的情况下，训练效果（ACC 值）就达到了 80% 以上。

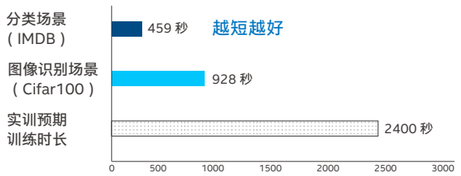



图 2-1-16 五舟科技新平台多人同时开展人工智能训练效果

测试结果表明，基于英特尔® 架构产品构建的深度学习与人工智能科研环境管理平台性能卓越，能够有效助力高等院校在培养学生深度学习、机器学习等人工智能基础技能的基础上，面向不同应用场景开展丰富的人工智能教学实训。

## 解决方案中的软硬件配置建议

### 硬件配置

| 名称 | 规格 |
|---|---|
| 处理器 | 双路英特尔® 至强® 金牌 6230R 处理器 |
| 基础频率 | 2.10GHz |
| 核心/线程 | 26/52 |
| HT | On |
| Turbo | On |
| 内存 | 384G (32G DDR4 2933MHz x 12) |
| 硬盘 | 英特尔® 固态盘 DC P4610 系列及以上 |
| 网卡 | 英特尔® 以太网网络适配器 X722 |

### 软件配置

| 名称 | 规格 |
|---|---|
| 操作系统 | CentOS 7.8.2003 |
| 内核版本 | 3.10.0-1127.18.2.el7.x86_64 |
| 编译器 | GCC 8.0 |
| 框架 | 面向英特尔® 架构优化的 TensorFlow2.4.0 及以上 |

## 小结

作为一门由行业需求驱动的新兴学科，人工智能从课程设计之初就非常注重动手实践。而各级教育机构在开展人工智能教育时遇到的一大挑战，就是如何基于已有的校园 IT 系统为人工智能教学各个环节提供支撑。要应对这一挑战，不仅需要新的支撑方案能对人工智能算法等理论教学提供助力，更重要的是构建起高效、灵活且可按需扩展的人工智能教学实训平台。

基于英特尔® 架构的处理器平台、英特尔® 深度学习加速、英特尔® 人工智能框架以及由英特尔开源的 OpenVINO™ 工具套件等一系列产品和技术，凭借其创新架构和强劲性能，正帮助人工智能教育系统厂商和各级教育机构借助云计算等 IT 新技术，以 "云 - 边 - 端" 等灵活的系统架构设计，打造适于 K12 及高等教育等不同阶段师生使用的人工智能教学和实训平台。来自基于英特尔 ® 架构的处理器平台的多核心优势，通过对 NUMA 架构的良好支持，使开展规模化人工智能实训任务的学生都能分配到足用的算力和内存等资源；同时，源于 OpenVINO ™ 工具套件等提供的人工智能性能加速，也使人工智能教学获得更高效率。

# 优化方案设计、提升推理性能，助力智能课堂行为分析

## 英特尔与合作伙伴共同探索课堂行为分析在智慧教育场景中的应用

### 人工智能行为分析解决方案开发及挑战

在现代教育理念中，教师对学生的关注以及对教学反馈信号（如学生的行为状态、表情变化等）的科学解读会对教学过程带来巨大影响。上述内容也构成了教师有的放矢调整教学内容、方法与策略的依据，进而提升教学质量和教学效果。然而，随着教学方式和内容变得愈加丰富，例如更多的电子课件、电子设备等被引入课堂，课堂已变得更加复杂。教师在基于这些新方式、新设备进行授课时，由于自身精力和课堂环境的限制，很难关注到每一位学生的表现和状态。在线教育等远程授课模式的出现，更加剧了这一现象。受制于教室光线、摄像头角度等因素，教师在教学过程中更难以全面掌握学生的面部表情、身体语言反馈等教与学的信号，也就无法依此调整教学节奏，从而阻碍了教学效率的提升。

计算机视觉（Computer Vision，CV）与人工智能技术的发展，尤其是在图像、音视频分析等领域有着巨大优势的深度学习方法，在教育机构中正获得越来越多的应用，并有望帮助解决上述问题。例如，有研究表明，人工智能科学家可以借助循环神经网络（Recurrent Neural Networks，RNN）模型来提升对交互式行为和手势的识别结果[9]，从而预测目标人群的下一步行动走向；而在另一项研究中，基于卷积神经网络（Convolutional Neural Network, CNN）模型的方法也被用于区分目标人群的不同运动状态，并有着很好的识别准确率[10]。这些基于人工智能的行为识别研究成果也正逐步运用于智慧教育领域，例如利用深度学习方法对线上线下课堂中的学生面部表情变化、身体语言反馈等进行分析。

一种典型的、适用于课堂的行为分析解决方案如图 2-2-1 所示，方案中可通过所部署的高清摄像头，在经过人员识别后，对师生教学过程中的行为和表情，如举手、起立等动作，以及思考、皱眉等不同表情予以捕获。

这些捕获的数据会被送至边缘或云端的人工智能服务器进行基于深度学习方法的分析预测，所得结果将被进一步传送至专门的数据分析应用中进行汇总分析，进而在教师或教育管理机构的终端上产生统计、分析或提示信息。


图 2-2-1　基于深度学习方法的课堂行为分析解决方案

但类似解决方案在教育机构的部署、实践过程中也遇到诸多的挑战，包括：

- **行为分析准确度有限：**传统面向视频分析、图像分割的深度学习方法往往针对的是静止的，或有着固定位置的目标物，因此系统在工作时，很少受到拍摄角度不佳、光线不足等因素的影响。但是，在学习过程中，学生的行为模式是无序和不确定的，一些经典的深度学习算法模型和方案设计在此种场景中可能面临分析结果准确度有限的问题。
- **实时推理效率不足：**教学过程通常是一个动态且变化的过程，教师对学生行为如果不能做出快速的反馈，就会失去很多互动交流和调整教学的机会，尤其在远程教学模式中，如果教师不能马上根据学生的行为和表现做出反馈，就更加削弱了教师的临场感，加剧了学生的孤立感，带来“老师无法看到我、理解我”的学习体验。因此，课堂行为分析往往对实时性有较高要求，而深度学习方法的高效推理过程有赖于强劲的算力以及优化的框架对其提供支撑，这就对教育机构的既有 IT 设施能力提出新挑战。
- **方案部署成本高昂：**一般地，应对基于深度学习方法的课堂行为分析解决方案对算力、数据等的需求，需对既有 IT 基础设施实施改造。而对采用非通用处理器作为算力的方案进行改造，显然会给教育机构带来高昂的再采购和再部署成本。
- **硬件兼容性造成部署困难：**随着在课堂环境中部署的电子设备和终端越来越多，教育机构往往消耗大量的精力与运维成本来解决不同设备间的兼容性以及数据有效传输问题，这也是造成人工智能课堂行为分析解决方案无法得到快速推广的重要原因之一。

为帮助教育机构更有效地部署基于深度学习方法的智能课堂行为分析解决方案，英特尔正与众多合作伙伴一起，基于课堂教学的特点，一方面提出更有针对性的算法模型和方案设计，另一方面也帮助厂商将更多先进英特尔产品与技术，如融合了

[9] 具体可参见 Hongsong Wang and Liang Wang. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[10] 具体可参见 Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2017.

英特尔® 深度学习加速的英特尔® 至强® 可扩展处理器、英特尔® 酷睿™ 处理器、英特尔® VPU产品以及OpenVINO™ 工具套件等引入方案之中，并通过引入遵循英特尔® OPS/英特尔® OPS-C（英特尔® 开放式可插拔规范/英特尔® 开放式可插拔规范加强版)的设备来降低兼容性问题，从而帮助更多教育机构，基于既有英特尔® 架构的IT基础设施，打造卓有成效的课堂行为分析解决方案，为教育信息化2.0新阶段提升教与学效率提供直接助益，同时服务于构建“互联网+”条件下的人才培养新模式。

## 面向教育场景的行为分析方案设计

### 教学场景中行为分析方法的构建

计算机视觉技术和深度学习方法的发展，正让更多面向视频分析的应用成为人工智能技术落地的重要方向，在智慧教育领域，也涌现出大量基于视频的行为分析的智能应用。

在面向教育场景的行为分析上，如图 2-2-2 所示，典型的方案架构可以归纳为视频采集、人物检测和行为分类三个核心步骤。


图 2-2-2　面向教育场景的行为分析方案架构

**视频采集：**由摄像头、录像机等前端采集设备捕获课堂视频，在进行去噪声、归一化等预处理后传送至部署于边缘或数据中心的行为分析平台。

**人物检测：**行为分析平台首先进行人物检测，其目的是找出视频帧中所有的学生信息，包括位置和大小，并以矩形框表示。人物检测是一种典型的目标检测问题，目前常用的基于深度学习的算法模型包括 Faster R-CNN、YOLO（You Only Look Once）、RetinaNet、SSD（Single Shot MultiBox Detector）等。

以 YOLO 算法为例，如图 2-2-3 所示，算法将检测任务转换成一个回归问题，其将每个输入图像都划分成 S × S（例如取为 7 × 7）个网格，每个网格都要预测两个 bounding box 的坐标（x, y, w, h）、box 内包含检测目标的置信度（confidence），以及检测目标属于预设类别中每一类的概率。在后续的筛选层中，则会选出合适的 bounding box 作为结果。

YOLO 算法的优势在于更快的检测速度，原始 YOLO 算法的检测速度就可达 45 帧 / 秒，可满足课堂行为分析所需的实时性要求，而优化后的 Fast YOLO 等新版本，更可将速度提升至 155 帧 / 秒。[11]

更多 YOLO 算法详情，请参阅 You Only Look Once: Unified, Real-Time Object Detection，Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi

https://arxiv.org/pdf/1506.02640.pdf

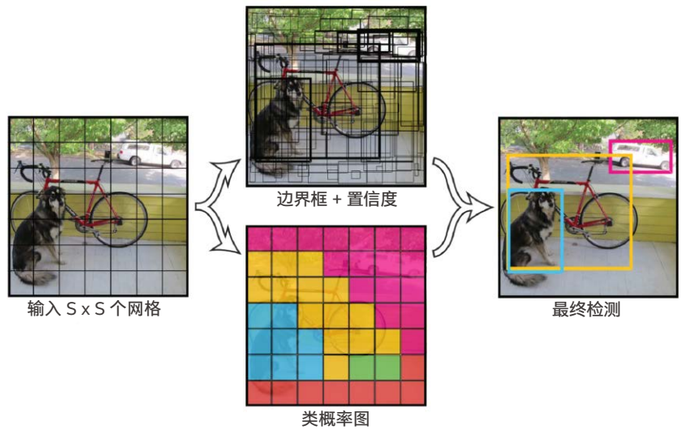



图 2-2-3　YOLO 算法模型

[11] 数据援引自 You Only Look Once: Unified, Real-Time Object Detection，Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi

**行为分类：**在区分出待检测的各个学生后，就需要对学生当前行为进行识别与分类。通常而言，最简单的行为识别就是以单一视频帧为对象，然后利用 CNN 对其进行识别，但这种方法在画面比较复杂时，如人物有遮挡、重叠时会受到较大干扰，准确性不高。为提升准确率，许多基于 CNN 网络扩展的优化方案应需而生，例如 Two-Stream（双流）类方法、C3D 方法以及 CNN-LSTM 方法等。

以目前常见的 CNN-LSTM 方法为例，其网络架构如图 2-2-4 所示。首先，算法在输入侧通过连续多帧计算光流图，获取相应运动信息；然后，在进行特征池化后，使用 LSTM（Long Short-Term Memory, 长短期记忆）网络来得到视频级描述，其中每帧特征池化后将连接五个 LSTM 层，后一帧的 LSTM 输出将输入至下一帧的 LSTM，网络最后可采用 softmax 分类器来进行分类。

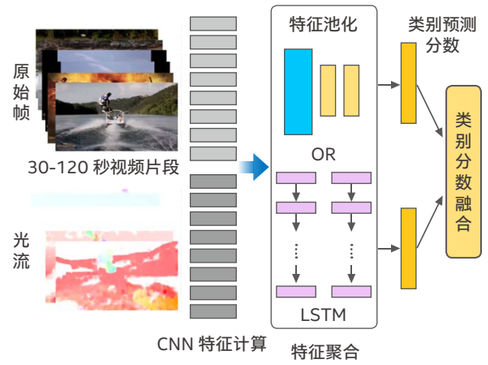



图 2-2-4　采用 LSTM 模型的行为识别网络架构

LSTM 网络的特性使其能够对 CNN 特征进行更长时间的融合，因此能对更长视频进行表达。同时由 LSTM 网络引入的记忆特性，能使方案有效地表达帧的先后顺序，从而对行为做出更有效的分类。

更多用于行为识别和分类的 LSTM 算法详情，请参阅 Beyond Short Snippets: Deep Networks for Video Classification，Joe Yue-Hei Ng，Matthew Hausknecht，Sudheendra Vijayanarasimhan，Oriol Vinyals，Rajat Monga，George Toderici

https://arxiv.org/pdf/1503.08909.pdf

### 传统方案面临的挑战以及基于“事件统计”的行为分析方案设计

从上节可以看到，在传统行为分析方案中，行为分析通常是采用“帧统计”的方法。但在教育领域的智能行为分析场景中，这些方法也面临一些新的技术挑战。尤其在课堂环境中，无论是学生还是教师的交互行为往往无序和不确定，且存在大量的中间状态。

如图 2-2-5 所示，当方案以帧为单位进行统计时，以“坐下”和“起立”为例，各状态之间都存在数帧图像处于中间状态，这些中间状态具有很强的不确定性，传统的分析算法很难对每一帧视频进行准确预测。而当用户希望用基于帧的分析来推断最终行为状态时，可能会导致准确率的下降。

为应对这一问题，英特尔根据课堂等教学场景的实际状况，以及教学过程中的普遍性行为模式，提出了一种创新的、基于“事件统计”的行为分析方案设计。这一方案的基础架构如

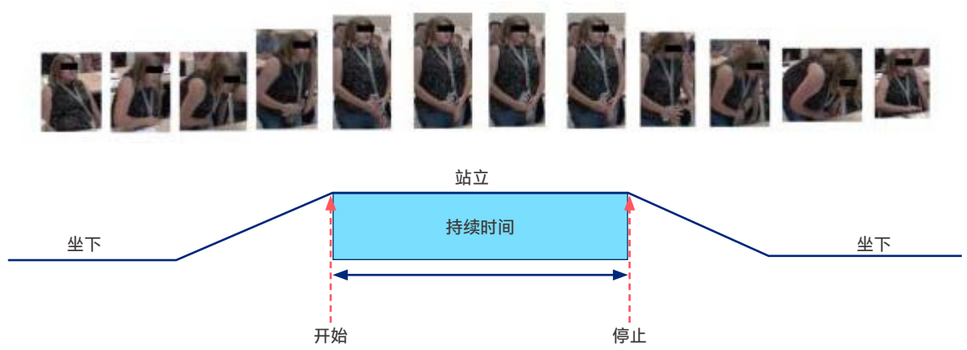



图 2-2-5　按帧统计时的行为状态分析

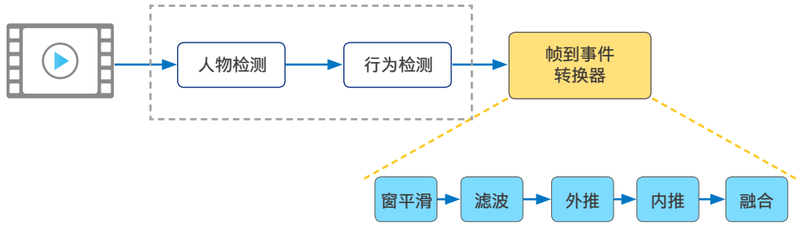



图 2-2-6　基于 “事件统计” 的行为分析方案架构

图 2-2-6 所示，当完成 “人物检测” 和 “行为分类” 后，新方案会加入一个新的后处理框架，其作用是通过对基于 “帧统计” 的结果进行重建，来帮助系统获得基于 “事件统计” 的、更为精准的结果。

如图 2-2-6 所示，这一框架的工作步骤包括：滑动窗口（Smooth）、短事件滤波器（Short Event Filter）、外推（Extrapolation）、插值（Interpolation）以及合并（Merge）等。

- **滑动窗口：** 如图 2-2-7 所示，滑动窗口过程是通过加入一个滑动窗口的方法，来降低由于闭塞和拥挤的环境带来的假阳性（FP）和假阴性（FN）判定；
- **短事件滤波器：** 该过程的目的是为了减少假阳行为导致的不相关的短间隔动作，可以简单地忽略或过滤短的间隔；
- **外推：** 这一过程用于将预测结果推断到事件的初始时间范围，从而校正滑动窗口滤波引起的偏移误差；
- **插值：** 该模块可以通过观察较长周期的相邻事件，通过差值的方式填补缺失部分，以提高预测结果的采样率；
- **合并：** 其作用是将同一个动作的连续事件合并在一起，输出最终事件的分析结果。

可以将上述工作步骤以伪代码形式表述如下：

- **输入：** 所有目标的行为检测结果，起始帧，终止帧，窗大小，事件最短时长限制，默认行为
- **输出：** 事件统计结果 EventMap
- **过程：**

```
for 每个目标 do
  # 将目标的第一个行为作为事件 (Event)加入到事件列表 event_list
  event_list.insert(Event(第一个行为))
  for 当前目标的每一个行为 do
    cur_event ← event_list.back()
    if cur_event 和当前行为在同一个窗口内且行为类别相同 then
      cur_event.insert(当前行为)
    else
      if cur_event 小于最短时长限制 then
        丢弃 cur_event

      event_list.insert(Event(当前行为))

  if 最后一个事件小于最短时长限制 then
    丢弃该事件

  if event_list = ∅ then
    event_list.insert(Event(默认行为))
  else
    event_list.front().begin_frame ← 起始帧
    event_list.back().end_frame ← 终止帧

  # 处理相邻事件{event1, event2}
  for {event1, event2} in event_list do
    middle_frame ← (event1.begin_frame + event2.end_frame) / 2
    event1.end_frame ← middle_frame
    event2.begin_frame ← middle_frame

  # 将事件列表的第一个事件加入当前目标的最终事件列表
  final_event_list ← event_list.front()
  for event in event_list do
    last_event ← final_event_list.back()
    if last_event.action == event.action then
      last_event.end_frame ← event.begin_frame
    else
      final_event_list.insert(event)

  EventMap[当前目标] ← final_event_list
```

框架相关示例代码可参考：https://github.com/opencv/open_model_zoo/tree/master/demos/smart_classroom_demo

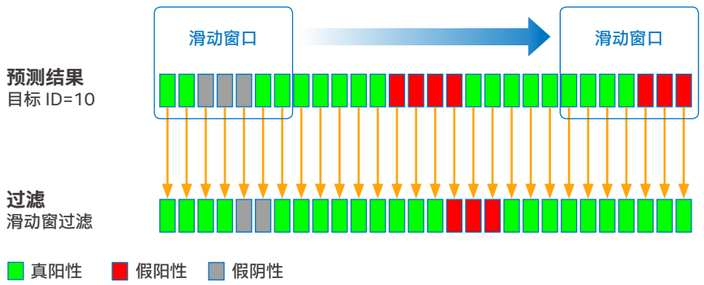



图 2-2-7　滑动窗过滤假阳和假阴判定

## 针对行为分析的英特尔产品优化方案

### 英特尔® 深度学习加速技术提升训练与推理效率

传统用于视频分析、图像分割等场景的深度学习应用，在训练和推理工作负载中往往采用 32 位浮点精度（FP32）的数据格式，这虽然可带来更精确的结果，但会使方案在执行训练和推理时，陷入内存瓶颈而降低计算效率。而一些研究表明，以较低精度的数据格式进行深度学习训练和推理，特别是在音视频、图像相关的应用场景中，并不会对结果准确性带来太多影响。因此，借助 OpenVINO™ 工具套件将 FP32 数据格式的模型量化为 INT8 等低精度数据格式，用户可以通过英特尔® 深度学习加速技术对低精度数据格式的良好支持实施人工智能训练与推理加速。

集成在英特尔® 至强® 可扩展处理器平台（第二代 / 第三代英特尔® 至强® 可扩展处理器）中的英特尔® 深度学习加速，分别以英特尔® AVX-512_VNNI（矢量神经网络指令）和 AVX-512_BF16（bfloat16），对 INT8 数据格式和 BF16 数据格式提供了良好支持，并能够与英特尔® oneDNN 库等相结合，广泛地为商用化人工智能应用的训练和推理过程提供加速。其中，AVX-512_VNNI 理论上可使计算效率提升 4 倍 [12]，而 AVX-512_BF16 则能帮助训练性能提升达 1.93 倍 [13]。

如欲了解英特尔® 深度学习加速技术更多技术细节，请访问：https://www.intel.cn/content/www/cn/zh/now/your-data-on-intel/deep-learning-boost-video.html

### 面向边缘复杂人工智能需求的基于英特尔® 架构的软硬件产品组合

随着人工智能应用在教育场景中获得愈来愈广泛地运用，英特尔也正为用户提供产品形态更多样、适用场景更广泛，且预先经过优化、集成和认证的人工智能软硬件产品与解决方案以供选择。

以面向教育场景的课堂行为分析方案为例，如图 2-2-8 所示，针对方案中对处理性能、兼容性、可扩展性以及部署便捷性等方面的要求，英特尔不仅在硬件层面上提供了多种遵循英特尔® OPS 规范的硬件产品与英特尔® VPU 产品的组合，也在软件层面上提供了 OpenVINO™ 工具套件，以软硬件结合来协同提升方案的人工智能处理能力和兼容性。

#### ■ 遵循英特尔® OPS 规范的硬件产品

英特尔® OPS 及其扩展版本英特尔® OPS-C 是英特尔针对各类数字显示系统，例如智慧教室中常见的电子白板、嵌入式一体机等设备推出的开放式可插拔规范，其不仅在可用性、可维护性、功耗等方面具有优势，更在显示输出和连接性能等方面提供了良好的兼容性和易管理性，为教育场景中课堂行为分析方案的设计和部署提供有效助力。

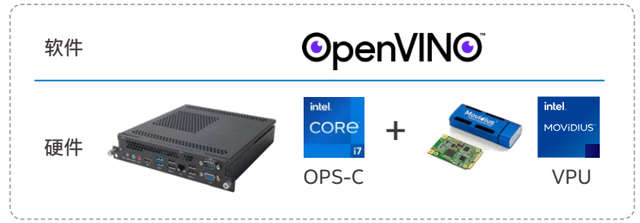



图 2-2-8　软硬结合，以灵活的基础设施应对复杂人工智能需求

[12] 数据援引自：https://www.intel.com/content/www/us/en/artificial-intelligence/posts/lowering-numerical-precision-increase-deep-learning-performance.html。实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex；

[13] 测试配置如下：测试组配置：单节点，4 个安装在英特尔参考平台 (Cooper City) 的第三代智能英特尔 ® 至强 ® 可扩展 8380H 处理器（预生产 28C，250W），总内存 384 GB（24 个插槽 /16 GB/3200），ucode 0x700001b，超线程开启，睿频开启，带有 Ubuntu* 20.04 LTS，Linux* 5.4.0-26,28,29-generic，英特尔 800 GB 固态盘 OS 驱动器，ResNet-50 v 1.5 吞吐量，https://github.com/Intel-tensorflow/tensorflow -b bf16/base，commit #8287386427603 58b388d8f615ded0c213f10c9 9a，Modelzoo：https://github.com/IntelAI/models/ -b v1.6.1，Imagenet 数据集，oneAPI 深度神经网络库 (oneDNN) 1.4，BF16，BS=512，英特尔于 2020 年 5 月 18 日进行测试。

基准组配置：英特尔参考平台 (Lightning Ridge) 上的 1 个节点、4 个英特尔® 至强® Platinum 8280 处理器，总内存 768 GB（24 个插槽 /32 GB/2933），ucode 0x4002f00，超线程开启，睿频开启，Ubuntu* 20.04 LTS，Linux* 5.4.0-26,28,29-generic，英特尔 800 GB 固态盘 OS 驱动器，ResNet-50 v 1.5 吞吐量，https://github.com/Intel-tensorflow/tensorflow -b bf16/base，commit #8287386427603 58b388d8f615ded0c213f10c99a，Modelzoo：https://github.com/IntelAI/models/ -b v1.6.1，Imagenet 数据集，oneAPI 深度神经网络资料库 (oneDNN) 1.4，FP32，BS=512，英特尔于 2020 年 5 月 18 日进行测试。

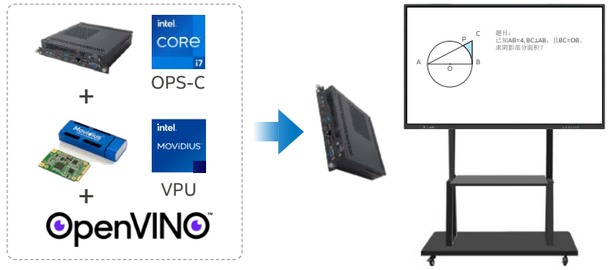



图 2-2-9　符合英特尔® OPS-C 规范的课堂嵌入式一体机设备

以遵循英特尔® OPS-C 规范的嵌入式教学一体机设备为例，如图 2-2-9 所示，其一方面可以选配英特尔® 酷睿™ 处理器等算力平台来形成强大的计算处理能力、优异的图形处理能力以及强有力的视频编解码性能，并通过与英特尔® VPU 产品的结合，借助 OpenVINO™ 工具套件提供的人工智能性能加速，使教育机构在狭小的嵌入式设备中也能部署高效的边缘人工智能处理能力。

另一方面，借助英特尔® OPS-C 提供的 80PIN 的输出接口，能为嵌入式教学一体机设备提供高达 2 路的 4K 显示输出，这使得课堂行为分析方案所需的高清码流视频信号传输变得更为顺畅便捷。

在人工智能加速和图形处理能力之外，英特尔® OPS-C 规范也能帮助嵌入式教学一体机设备降低与其它课堂设备相互协同连接时的复杂度，将电子白板等教学设备有机融合在一起，减少传统设备维护对教学的影响，提供一种性能稳定且兼容性强的智慧教室设备解决方案。

高度的整体性让符合英特尔® OPS-C 规范的课堂嵌入式一体机设备在部署后，一方面有力提升了部署在教室边缘设备中的人工智能课堂行为分析方案的工作效率；另一方面也帮助教育机构大幅减轻系统运维压力，降低了方案的部署门槛。

### ■ 英特尔® VPU 产品

受限于场地、环境以及部署成本等因素，一些教育机构无法在面向教学过程的边缘侧部署算力强劲的推理服务器等设备。同时，在线上教学模式中，为了尽可能地降低网络传输带来的延迟，也需要尽可能地将处理能力进行前置。为此，英特尔通过在方案中加入敏捷高效的 VPU 英特尔® Movidius ™ Myriad ™ X，来帮助用户在边缘部署人工智能推理能力。

英特尔® Movidius ™ Myriad ™ X 视觉处理器架构引入了全新深度神经网络处理单元：神经计算引擎。其专门的设计，可以高速、低功耗地运行深度神经网络。结合 16 个 SHAVE 内核，神经计算引擎在执行神经网络推理时能实现 716G FLOPS 的计算性能。

为使 VPU 产品在实战中为智能课堂行为分析方案提供更强助力，英特尔为其提供了两种不同的工作模式：Squeeze Scheduler 模式和 ByPass Scheduler 模式，来应对不同工作场景中的需求。

**Squeeze Scheduler 工作方式特点如下：**

- 在该方式下，每个网络模型会重复加载到所有 VPU，例如有两个 VPU，则网络模型会同时加载到第一个 VPU 和第二个 VPU 上；
- 加载到 VPU 上的网络模型总大小需小于 512MB；
- 每个 VPU 将获得一个默认的 cache graph 配置，默认大小为 4，当加载到 VPU 上的网络模型数量大于 cache graph 数量时，会导致系统性能变慢。cache graph 数量可以在配置文件中配置，但是最大只能配置成 10，超过 10 会报错。配置命令为：

```
"max_cached_graph_number": "10"
```

**ByPass Scheduler 工作方式特点如下：**

- 在该方式下，可将特定的网络模型装载到特定的 VPU 上，而非装载在所有 VPU 上；
- 在性能不受影响的情况下，ByPass Scheduler 方式支持的最大网络模型数为 10 x ( VPU 的个数 )；
- 可将某个使用较频繁的网络模型同时加载到多个 VPU 上，并指定优先级，这样该网络模型优先运行在优先级高的 VPU 上，当优先级高的 VPU 繁忙而优先级低的 VPU 空闲时，优先级低的 VPU 能够继续运行该网络模型，从而保证 VPU 资源充分利用。

与 CPU 相比，VPU 在并行数据处理能力上的优势，使其更适于模型推理等工作负载。在实际的方案设计中，通常可以采用 VPU 与 CPU 混合配置的模式。以其在阅面科技人脸识别与课堂行为分析方案中的应用为例，在模型推理的加载和运行阶段，混合工作模式如图 2-2-10 所示：

- **在模型加载阶段，**采用 ByPass Scheduler 的工作方式可以让不同工作负载分别加载到指定的 VPU 与通用处理器中，其中人头检测加载到 VPU 1 上，跟踪网络 1 和跟踪网络 2 加载到 CPU 上。而行为检测和行为验证则采用了多 VPU 优先级加载模式，其可以同时被指向加载 VPU 2 和 VPU 1，且 VPU 2 优先级高于 VPU 1。
- **在推理运行阶段，**系统可以通过分线程处理的方式来提升效率。其中线程 1 用于视频的解码且缓存到 Buffer 1 中。线程 2 则从 Buffer 1 中取数据，使用 VPU 1 进行人头检测，返回检测结果给 CPU，然后基于 CPU 执行跟踪网络 1 和跟踪网络 2 两个模型，并将结果缓存到 Buffer 2。线程 3 则从 Buffer 2 中取数据，使用 VPU 2 和 VPU 1 执行行为检测和行为验证工作负载，一般情况下工作负载优先在 VPU 2 上执行，当 VPU 2 繁忙且 VPU 1 空闲时，系统会自动将其调度到较为空闲的 VPU 1，从而保证负载稳定运行，得到最终的举手行为的检测结果。

如欲了解英特尔® Movidius ™ Myriad ™ X 视觉处理单元更多技术细节，请访问：https://www.intel.cn/content/www/cn/zh/products/processors/movidius-vpu/movidius-myriad-x.html

### ■ 利用 OpenVINO™ 工具套件提升推理性能

由英特尔开源的 OpenVINO™ 工具套件能够为智慧教育场景中的各类视频应用和推理负载提供强劲的“加速引擎”。

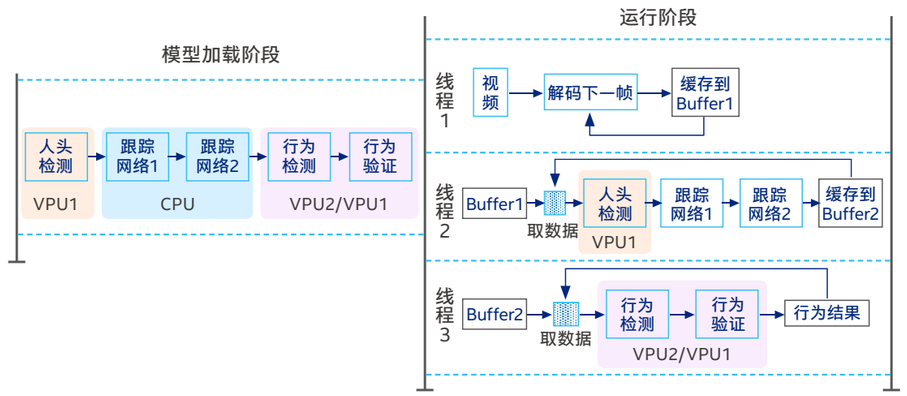



图 2-2-10　面向举手场景的 CPU 与 VPU 异构模式配置

OpenVINO™ 工具套件一方面对传统的 OpenCV、OpenVX 等图像处理库进行了大量指令集优化，实现了性能与速度的显著提升；另一方面，其通过内置的模型量化工具，帮助推理引擎可以轻松地运行在 INT8 数据格式上，进一步实现推理速度的有效提升。

同时，OpenVINO™ 工具套件也能帮助用户以通用 API 接口来实现跨环境的异构执行，即人工智能方案在开发完成后，利用 OpenVINO™ 工具套件可以在通用处理器、VPU、FPGA 等不同硬件设施环境中部署使用，进而简化智慧教育厂商的开发流程，加速方案落地。

为便于使用者调用，OpenVINO™ 工具套件提供了简单清晰的 API 接口。以 CPU 为例，可以通过如下代码进行网络初始化：

```
Core ie;
CNNNetwork net = ie.ReadNetwork("model.xml");
ExecutableNetwork executable_network = ie.LoadNetwork(network=net, device_name="CPU");
```

其中 device 可设置成 CPU、GPU 以及 HDDL 等等，config 为可选项，可以添加附加设置，例如绑定、优先级设置等。

同时，该工具套件还提供了网络同步运行与异步运行两种推理方式，同步运行的接口为 Infer()，异步运行的接口为 StartAsync()。前者属于阻塞模式，执行时需等待推理结果的返回才能继续往下执行；后者属于非阻塞模式，执行时无需等待结果返回，直接往下执行程序。针对最大化推理吞吐率的场景，可以采用异步的方式来最大限度的利用资源。与

VPU 产品相结合，使用 FP16 数据格式的异步推理相关代码可参考如下：

```
#define MAX_ACTION_BATCH_SIZE 16
std::string input_name = net.getInputInfo().begin()->first;
InputInfo::Ptr input_info = net.getInputsInfo().begin()->second;
# 修改输入配置
input_info->setLayout(Layout::NHWC);
input_info->setPrecision(Precision::U8);
# 准备输入
InferRequest inferRequest = executable_network.CreateInferRequest();
Blob::Ptr inputBlob = inferRequest.GetBlob(input_name);
SizeVector dims = inputBlob->getTensorDesc().getDims();
size_t num_channels = dims[1];
size_t image_size = dims[3] * dims[2];
MemoryBlob::Ptr minput = as<MemoryBlob>(inputBlob);
auto minputHolder = minput->wmap();
auto data = minputHolder.as<PrecisionTrait<Precision::U8>::value_type*>();
# TODO: 复制图像数据到 data 缓存
# 异步推理
size_t numIterations = action_request_num;
size_t curIteration = 0;
std::condition_variable condVar;
inferRequest.SetCompletionCallback([&] {
    curIteration++;
    if (curIteration < numIterations) {
        inferRequest.StartAsync();
    } else {
        condVar.notify_one();
    }
});
# 开始第一次异步请求推理
inferRequest.StartAsync();
# 等待所有异步请求
std::mutex mutex;
std::unique_lock<std::mutex> lock(mutex);
condVar.wait(lock, [&] {
    return curIteration == numIterations;
});
```

如果选择第三代英特尔® 至强® 可扩展处理器（Cooper Lake）平台来部署推理服务，用户可以采用 BF16（Brain Floating Point 16）这一 16 位的数据格式来提升性能。如图 2-2-11 所示，BF16 是单精度浮点 FP32 数据格式的截断 16 位版本，与 FP32 有相同的动态范围。

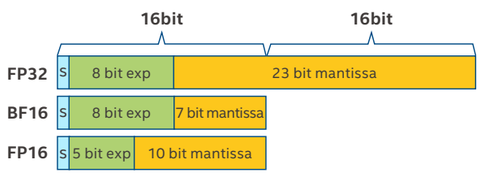



图 2-2-11　三种浮点数据格式

基于以下几个维度的优势，选择 BF16 做推理可以最大化利用第三代英特尔® 至强® 可扩展处理器平台的算力来实现最大性能，同时将计算精度保持在可接受的范围内，帮助用户提高方案整体性能：

- BF16 数据的尾数更短，因而 2 个 BF16 数相乘更快；
- 特定性能优化方案使其不需要支持异常处理；
- 可实现与 FP32 相互快速转换；
- 减少内存占用，使内存可以容纳更大的模型；
- 减少数据量传输，降低数据转换时间。

通常用户可以通过以下 2 种方式检查处理器是否支持 BF16：

1. 执行 lscpu | grep avx512_bf16 或者 cat /proc/cpuinfo | grep avx512_bf16

2. 通过 OpenVINO™ 工具套件提供的 API 来查询：

```
InferenceEngine::Core core;
auto cpuOptimizationCapabilities = core.GetMetric("CPU", METRIC
_KEY(OPTIMIZATION_CAPABILITIES)).as<std::vector<std::string>>();
```

OpenVINO ™ 工具套件默认支持 BF16，这种情况 KEY_ENFORCE_BF16 默认设置为 YES，下面代码示例演示了如何检查是否设置支持 BF16：

```
InferenceEngine::Core core;
auto network = core.ReadNetwork("sample.xml");
auto exeNetwork = core.LoadNetwork(network, "CPU");
auto enforceBF16 = exeNetwork.GetConfig(PluginConfigParams::
KEY_ENFORCE_BF16).as<std::string>();
```

如果要禁用 BF16 的内部转换，将 KEY_ENFORCE_BF16 设置为 NO 即可：

```
InferenceEngine::Core core;
core.SetConfig({ { CONFIG_KEY(ENFORCE_BF16), CONFIG_VALUE(NO)
} }, "CPU");
```

如欲了解 OpenVINO ™ 工具套件更多技术细节，请访问：
https://docs.openvinotoolkit.org/
https://docs.openvinotoolkit.org/latest/openvino_docs_IE_DG_Bfloat16Inference.html

# 基于英特尔优化方案的应用案例

## 阅面科技：借力人脸识别与课堂行为分析提升教学互动效果

### 引言

*“得益于英特尔® 架构处理器和 VPU 提供的强劲算力，以及由 OpenVINO ™ 工具套件带来的推理加速，新的人脸识别与课堂行为分析方案能帮助教师提升教学的互动性与有效性。”*

——阅面科技

### 背景与挑战

通过摄像机来分析拥挤的教室或线上不同授课环境中每个学生的行为，已被证实可以有效帮助教育机构、教师和家长获得教学效果、学生学习状态等有效信息，并据此帮助教师有的放矢调整教学方式。同时，行为分析还能帮助教师在课堂中开展抢答、分组竞赛等饶有趣味的授课模式，提升教学质量与效率。

但在线上线下教学过程中实施行为分析同样都面临着巨大的挑战。由于学生行为通常具有无序的特性，传统人脸识别等人工智能能力虽然可以有效地分辨学生身份，但对学生的不同行为，例如起立、举手、聊天等实施分析在实时性和准确性方面仍显不足。为解决这些问题，上海阅面网络科技有限公司（以下简称“阅面科技”）与英特尔一起，通过更具针对性的算法和方案设计，以及对模型推理过程的有效优化，来为教育机构提供高可用性的课堂行为分析解决方案。

### 解决方案

#### ■ 方案解析

阅面科技与英特尔携手采用人工智能技术，构建的面向课堂场景的人脸识别与课堂行为分析解决方案架构如图 2-2-12 所示，由人脸注册、人脸识别和行为分析三个主要模块组成。其中，人脸注册模块通过将学生的既有人脸图像导入后台，进行特征提取后作为系统特征库供后续流程调用；人脸识别模块由前端高清摄像头捕获学生视频后，通过推理服务器进行人脸检测，并输出识别结果。

而核心的行为分析模块则基于高清摄像头捕获的视频，在进行头像检测跟踪后，利用经英特尔优化的行为分析算法，对学生行为如举手等进行判断识别，并输出反馈结果。教师可以根据系统给出的结果，进行下一步的教学流程。

在人脸识别与课堂行为分析方案的设计和迭代过程中，阅面科技与英特尔合作，对基于“帧统计”和基于“事件统计”的新方案设计进行了详实的预研和验证，并最终选择了基于“事件统计”的方案设计。由于教育机构采用该方案时，部署行为分析模块的主要目的是为了统计学生上课期间的教学活动参与度，例如针对教师提出的某个话题，学生是否有参与到这个话题中的行为发生，比如举手抢答、站立回答等。如采用“帧统计”的方式，则不利于后续进行数据分析。

更为重要的是，基于“事件统计”的方案有助于提升方案的行为分析准确率，降低由统计标准不一致而造成的分歧（详细技术分析请参见第 33 页“传统方案面临的挑战以及基于**事件统**

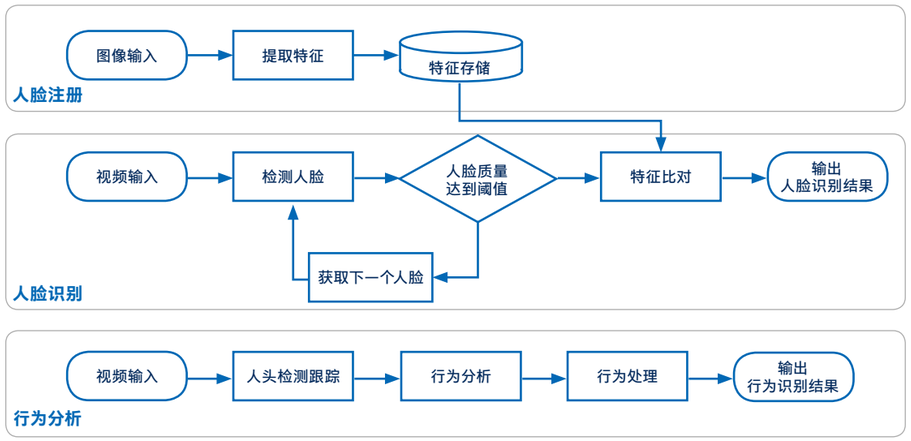



图 2-2-12　阅面科技人脸识别与课堂行为分析方案架构

计的行为分析方案设计”部分）。如图 2-2-13 所示，典型课堂动作行为的发生往往没有很清晰的分割边界，图中许多学生的手都处于一种似举非举的模糊状态。面对这一场景，如果采用“帧统计”方式，结果准确度可能比较差。而采用按事件统计的方式，则可以明确表示一次举手行为的发生，结果完全不受动作模糊状态的影响。

图 2-2-13　阅面科技人脸识别与课堂行为分析方案面临的典型场景

### ■ 英特尔产品和技术发挥的作用及效果：

**OpenVINO™ 工具套件**

方案通过引入 OpenVINO™ 工具套件来加速模型推理速度，从而提高整个方案的可用性（OpenVINO™ 工具套件优化方法请参见第 37 页“利用 OpenVINO™ 工具套件提升推理性能”部分）。为验证该工具套件为方案带来的收益，阅面科技将相同硬件配置下的优化方案（使用 OpenVINO™ 工具套件）与未优化方案（使用开源深度学习框架 Caffe）进行了性能比对，结果如图 2-2-14 所示，在人脸检测和行为检测两个场景下，通过 OpenVINO™ 工具套件优化的方案在吞吐率性能上较优化前分别提升约 31% 和 23%[14]。

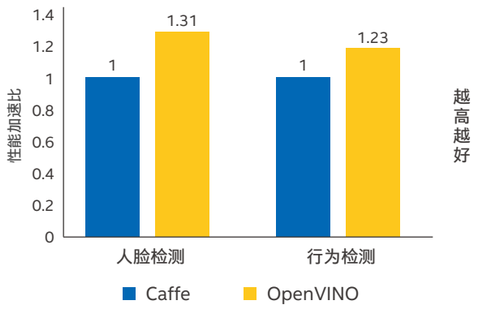



图 2-2-14　OpenVINO™ 工具套件优化方案带来的吞吐率性能提升

**英特尔 ® Movidius Myriad X 处理器**

方案采用了通用处理器和 VPU 的混合配置模式，来应对部分教学场景中无法全量部署通用处理器进行模型推理的问题。例如在一些实践部署中，方案采用 “2+3” 模式（即 2 个模型是基于英特尔® 架构处理器、3 个模型是基于 Movidius™ Myriad ™ X VPU 的并行负载配置），借助 OpenVINO™ 工具套件，方案可以将不同网络模型分别加载到英特尔® 架构处理器或 Movidius™ Myriad™ X VPU 上。这一模式已在行为分析模块的推理工作中获得了良好效果，在推理效率下降不大的情况下，可令处理器负载维持在 30% 以下，从而保证了方案中其他工作负载不会因算力不足而受到堵塞。

## 方案收益

阅面科技新的人脸识别与课堂行为分析解决方案目前已在诸多教育机构获得了部署与实践，来自教育机构、教师、学生以及学生家长的反馈表明，新方案可以带来以下显著效果：

- 对各级教育机构而言，新方案的部署，使教育方法的优化和教育理论的演进有了更为翔实的数据支撑，教育主管部门和教育机构可以据此制定进一步的教学质量提升计划，加强教学过程的数字化、精细化管理，提升决策效率。
- 对教师而言，方案能够准确识别课堂中常见行为的发生，统计各类行为的比例，并快速直观以可视化方式提供给教师。同时方案还能有效减轻教师日常上课过程中点名等繁琐任务，提升课堂时间的利用率；另外，通过与其他智慧教育应用相结合，方案也能帮助教师积极合理地安排教学内容，提升教学的互动性与有效性。
- 对学生和家长而言，基于新方案的统计分析数据，可以对学习过程、学习方法进行快速有效的复盘，找出自身薄弱点并制定学习计划，使信息科技对学习质量的助益落到实处。

## 解决方案中的软硬件配置建议

| 名称 | 规格 |
|---|---|
| Socket | 1 |
| 处理器 | 第十代英特尔 ® 酷睿™ i7- 10510U |
| 基础频率 | 1.8GHz |
| 核心/线程 | 4/8 |
| HT | On |
| Turbo | On |
| 内存 | 16G（8 G DDR4 2666MHz x 2） |
| 加速处理器 | 英特尔® Movidius™ Myriad™ X 视觉处理器 x 2 |
| 操作系统 | Microsoft Windows* 10 64-bit |
| 编译器 | MSVC++ 15.0 |
| 框架 | OpenVINO ™ 工具套件 2021.2 及以上 |

[14] 测试配置：英特尔® 酷睿™ i5-8300H @2.30GHz；内存 8GB DDR4 2666GHz,；2 * 英特尔® Movidius Myriad X 处理器；操作系统：Windows10 64bit；OpenVINO ™ 工具套件版本 2019.3, Caffe 版本 1.1.6

# 百家云：基于课堂行为分析实现双师课堂教学效果评估

## 引言

*“得益于第三代英特尔® 至强® 可扩展处理器提供的 AI 加速力，让基于人工智能的课堂行为分析解决方案在推理效率上获得了更大的优势，进而令双师课堂的教学评估更具实时性且更为精准。”*

——百家云

## 背景与挑战

利用互联网、人工智能等新技术构建兼具质量和体验的教学环境已成为各级教育机构的共识。这其中，整合线上线下教学需求，兼顾教学资源和学生体验的双师课堂教学模式，正获得教学效益和市场效益的双重肯定。

双师课堂是指基于网络互动视频直播技术开展的 “名师直播教学 + 线下辅导老师服务” 的模式。如图 2-2-15 所示，这一模式通常由两名老师远程配合共同完成教学，左侧主讲老师通过大屏幕远程直播授课，而右侧辅导老师在课堂内负责课堂管理、答疑等。

图 2-2-15　百家云双师课堂教学模式

这一模式能够带来的优势包括：

- 对 K12 等常规教育而言，双师课堂模式能有效解决优秀教师复制困难的问题，提升优秀教师的产能，缓减目前优质教学资源不足、不均衡问题。同时，通过在边远农村地区的部署，这一模式还能肩负扶弱助学的重任，推动乡村教育振兴，让教育资源更普惠。
- 在职业培训、素质教育等场景中，双师课堂模式能简单接入第三方双师直播内容，快速拓展教学科目，从而有效帮助教育机构平衡好学习效果和教师人效，降低异地扩张成本，提升利润率。

作为视频领域专业性和影响力都备受认可的服务平台，百家云基于多年服务教育行业的探索与实践，以一站式双师课堂解决方案为不同类型的教育机构提供高效、灵活和弹性可扩展的产品及服务，目前已在全国部署 5,000 余间双师教室，涵盖领域包括 K12 教学、职业培训、素质教育、国际培训等。

为使双师课堂获得更优的授课效果，百家云正与英特尔一起，利用基于人工智能的课堂行为分析技术对学生的专注度、情绪、上课坐姿等行为进行检测和分析，并同时对教师授课过程进行监测，最后生成课堂报告发送给教育机构和学生家长，家长可以随时了解孩子的学习情况，真实感知孩子在线教学体验，及时干预、监督和辅导；教育机构可以做多维度、可视化教学全过程数据分析，帮助教师 KPI 考核，从而大大降低机构日常管理成本，实现利用数据分析驱动精细化管理和决策。

## 解决方案

### ■ 方案解析

用于百家云双师课堂的课堂行为分析解决方案的基本工作流程如图 2-2-16 所示，系统通过客户端分别截取教师授课和学生学习的视频帧，经过数据清洗和预处理后，上传到部署在边缘或云端的人工智能服务器。在这里，百家云引入了多种基于深度学习方法的图像分类、目标检测识别、图像分割算法，通过良好的识别准确率对师生双方的视频数据进行分析，生成课堂报告并发送给教育机构和学生家长。

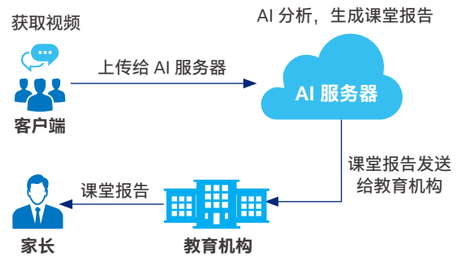



图 2-2-16　百家云智能学情评估解决方案流程

为更有效地对双师课堂中师生的行为进行分析，提升方案在实际应用中的运行成效，百家云在方案中融入了自研的学生学习类分析算法和教师授课监测算法：

- **学生学习类分析算法**包含学生专注度算法、情绪识别和人体骨骼检测。专注度算法输出学生专注度时序、眼神时间、空间分布、脸部朝向、人员干扰情况；而情绪识别可以甄别学生在课堂上的喜、怒、哀、乐；人体骨骼检测用来识别学生在上课过程中的坐姿，多方面反馈学生在本次课堂上的学习情况。

- **教师授课监测算法**则包含了教师激励手势识别算法、教学动作识别算法。鼓励手势主要是 OK 手势、鼓掌、点赞手势；教学动作包括数数、开口、观看、思考、聆听、荧屏等动作，用于考核教师和学生在课堂上的互动情况。

基于上述算法以及方案整体的架构设计，百家云在双师课堂方案中实现了较为完善的师生课堂行为分析能力，并以此构建多维度的课堂报告。例如方案可以根据不同的教师引导方式下所得到的学生反应进行打分，然后将分值形成数据矩阵，再利用 K-means 聚类算法等进行分析，从而针对不同学生学习进度、交流风格，在课堂报告中给出具有数据支撑的翔实说明。

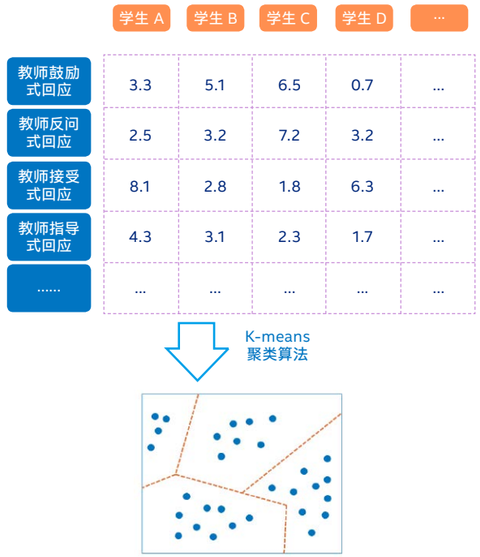



图 2-2-17　基于师生交互情况开展数据分析 [15]

基于更为详实的可视化报告，教师可以针对学生学习情况制定更为细化的教学策略，学生家长也能根据双师课堂中教师与学生的交互情况予以更好的家庭教育配合，形成 “学校 - 家庭” 的双向教育促进。而教育机构也能据此制定更实用的教学计划，实现数据分析驱动下的精细化管理和决策。

### ■ 英特尔产品和技术发挥的作用及效果：

**第三代英特尔® 至强® 可扩展处理器**

如图 2-2-18 所示，百家云双师课堂教学环境在系统构建上，一般由主讲教室和一系列远程教室构成，并接入部署在校园网 / 互联网上的云服务，在一些场景中还会加入专门的远程授课节点。

[15] 测图中打分数据，K-means 算法过程均为模拟数值，仅作为方案描述。


图 2-2-18　百家云双师课堂教学系统架构

而百家云双师课堂教学方案的最核心优势是能够形成主讲教室与远程教室之间的远程实时互动，并可根据需要灵活、快速地扩展。要充分发挥这些优势，首先，需要硬件基础设施对于灵活和高扩展性的云服务有着更好的支撑能力，其次，能对课堂行为分析中所需的人工智能应用提供有效加速。

为此，百家云正在方案中引入全新的第三代英特尔® 至强® 可扩展处理器来保持和扩大上述核心优势。一方面，得益于更多的内核、更大的内存和 I/O 带宽、创新的架构设计以及全面优化的软硬件，新处理器平台不仅为双师课堂教学方案提供了性能卓越，且具有广泛生态系统支持的优化平台，同时还能根据方案对云服务资源需求，轻松进行横向和纵向扩展，并依托英特尔提供的多元化行业知识和协作能力，帮助教育机构将经过实践检验有效的课堂行为分析方案快速部署到更多双师课堂中，精简成本并加强数据管理。

另一方面，第三代英特尔® 至强® 可扩展处理器集成的英特尔® 深度学习加速，支持 INNT8 和 BF16 两种低精度数据格式，能够让方案在不影响整体推理精度的情况下，有效加速训练和推理效率。如图 2-2-19 所示，经在基于第三代英特尔® 至强® 可扩展处理器平台上的验证测试表明，结合英特尔® 深度学习加速可为方案带来 1.77 倍的吞吐率性能提升（与第二代英特尔® 至强® 可扩展处理器平台，且未使用英特尔® DL Boost 技术相比较）。

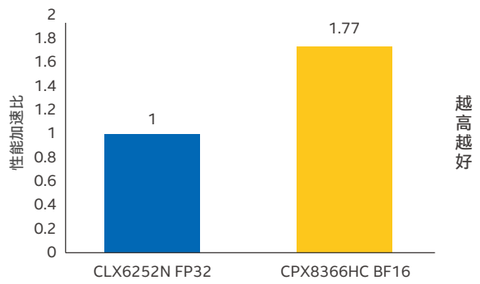



图 2-2-19　第三代英特尔® 至强® 可扩展处理器表情识别吞吐率性能优化效果

## 方案收益

通过在多个头部教育机构中的部署实践，面向百家云双师课堂场景的人工智能课堂行为分析解决方案已被证明可以让课堂学情、教情报告更加科学和便捷，不仅能够真实反映学生学习情况，受到家长的一致认可，还能评估教师在课堂上的表现，成为教师教学能力和效果评估的重要参考，提高了教育机构的管理和运营效率。其分别可为教育机构、教师、学生和家长带来以下的显著收益：

- **教育机构：**通过多维度、可视化教学全过程数据分析，教育机构能够帮助教师制定更为细化的 KPI 考核计划，并大幅降低机构日常管理成本，实现利用数据分析驱动精细化管理和决策。
- **教师：**结合课堂教学过程中不同行为分析得到的数据报告，能让教师对教学互动过程形成正向反馈，不断优化教学计划，提升教学质量与效果。
- **学生：**基于课堂行为分析的人工智能可视化学情分析，能帮助学生直观感知、了解自身学习情况，及时纠正错误的学习习惯，提升学习成效。
- **家长：**可以轻松获取孩子课堂全过程可视化学情分析，通过数字化复现课程学习行为，让家长了解学生学习过程表现，真实感知孩子在线教学体验，及时干预、监督和辅导，减轻家长焦虑感。

## 解决方案中的软硬件配置建议

### 硬件配置

| 名称 | 规格 |
|---|---|
| 实例规格 | ecs.hfc7.12xlarge（阿里云 ECS 实例） |
| Socket | 1 |
| 处理器 | 英特尔® 至强® 铂金 8369 HB 处理器 |
| 基础频率 | 3.3GHz |
| 核心/线程 | 24/48 |
| HT | On |
| Turbo | On |
| 内存 | 6 x DDR4-3200 16GB |

### 软件配置

| 名称 | 规格 |
|---|---|
| 操作系统 | Centos Linux release 8.2.2004 |
| 内核版本 | 4.18.0 |
| 编译器 | GCC 8.0 |
| 框架 | OpenVINO ™ 工具套件 2021.2 及以上 |

## 小结

利用人工智能方法，尤其是深度学习的方法来对教学过程进行实时性的评估和反馈，可以有效降低教师授课过程中的工作负担，提升师生之间的互动效率，并有效提高教育机构的管理能力。同时，对学习行为的监控、分析和反馈，也可以有效应对目前远程教育模式中，师生无法面对面交流带来的交互困难等问题。适应这一新需求，英特尔携手合作伙伴所推出的一系列智能课堂行为分析解决方案和产品，已经成为教育机构加快教育信息化 2.0 进程，推进信息技术与教育深度融合的一大抓手，并在实践中获得了教师、学生以及学生家长的好评。

伴随这一加速教学智能化转型的进程，包括英特尔® 架构处理器平台、英特尔® VPU 产品、英特尔® 深度学习加速技术以及由英特尔开源的 OpenVINO™ 工具套件等一系列先进产品和技术，正在为众多智慧教育应用场景中多个新方案大幅提升效率提供动力。例如，由英特尔® 深度学习加速技术和 OpenVINO™ 工具套件提供的对低精度数据格式的良好支持，能有效提升模型推理速度，配合英特尔提出的优化方案设计，使面向教育场景的行为分析方案在落地实战中获得成功，无论是在结果准确率，还是在系统实时性上，都很好满足了教育行业用户的需求。

# 以先进人工智能技术赋能语言教学，打造更优口语测评方法


## 英特尔与合作伙伴共同探索基于人工智能的智能口语测评方法

### 基于人工智能的智能口语测评

#### 行业趋势

口语环节正在语言类教育课程中获得更多重视，无论是 K12、高校阶段的外语教学，还是普通话培训等职业能力教学，口语测试都变得必不可少。因此教育机构正致力于通过更专业的口语测评、口语纠正等方式来提升学生的口语能力。这在提升学生学习兴趣，提高口语教学质量的同时，也对相应的教学和测评方法提出挑战。

在传统口语教学场景中，口语水平的评判往往需要师生间以“一对一”的方式进行，老师有限的时间和精力与学生大量的口语评判需求之间形成了尖锐的矛盾。因此，教育机构希望部署更为智能化的口语测评系统，来满足口语学习的旺盛需求。

得益于语音识别、语义识别以及语音测评技术的突飞猛进，以及由先进硬件设备提供的强劲算力支持，使基于人工智能技术的口语测评方案构建成为可能，由此推动大量智能化口语测评系统应运而生。这些智能测评系统能借助计算机辅助语言学习（Computer Assisted Language Learning，CALL）领域的技术方法，将口语发音所形成的音频数据进行特征提取后输入到声学模型中。随后系统将声学模型与语言模型进行融合计算，针对学生口语发音水平和差错，进行自动评价、检错，并提供相应的指导和纠正建议。

伴随人工智能技术，尤其是深度学习方法的高速发展，这类平台产品目前已经能面向普通话、外语等不同语言环境，对包括发音准确度、流畅度、自然度、完整度等维度在内的各项指标进行综合评估，从而能高效、快速和便捷地帮助教育机构和学生对口语学习成果进行智能化检测。

现在，基于人工智能的智能口语测评方案已在商用落地上获得巨大成功，并在许多中、英文发音标准程度、口语表达能力等测评任务中被证明可超越人类口语测评专家的水平，已经广泛使用在一些口语测评和定级中。如图 2-3-1 所示，一些数据表明，到 2020 年末，口语测评市场规模已达 10 亿元左右[16]，成为“人工智能 + 教育”领域不可忽视的细分领域。

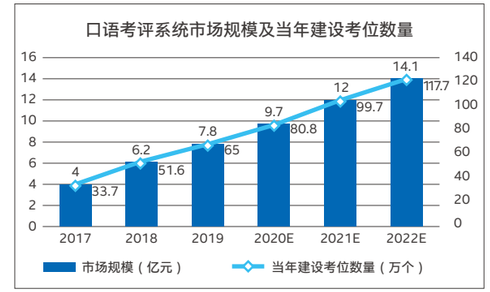



图 2-3-1 人工智能口语测评市场规模趋势

#### 应用挑战

语言是人类智慧的重要体现，同样一句话，在不同的语境、语调、断句等情形下，可能表达出截然不同的含义。因此，基于人工智能的口语测评解决方案所面临的困难之一，是语言学习教学和数字技术的支持中间没有明显的匹配。如图 2-3-2 所示，人们评价一套人工智能口语测评系统的有效性，通常基于三个核心维度，即专业性、实时性和稳定性。

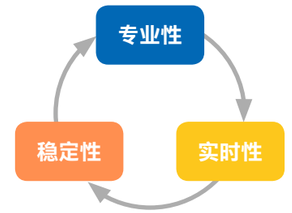



图 2-3-2 人工智能口语测评系统核心评价指标

- **专业性：**口语测评系统的专业性指标是指与人类专家测评之间的差异度。业内通常采用皮尔森相关系数（Pearson Correlation Coefficient，PCC）来衡量两者之间的向量相似度，经过标准化处理之后的 PCC 系数范围在 -1 到 1 之间，数值越大表示相关度越高。由于英语不是中国学生的母语，因此人工智能口语测评系统的设计也需要根据非母语使用者的特点，以丰富的语料库（基于不同的应用场景、语言环境、学生发音情况等）以及老师的经验来不断迭代优化；
- **实时性：**人工智能口语测评通常采用线上模式提供服务，巨大的服务规模往往会给后端处理引擎带来巨大系统压力。例如在中小学教育场景，每天的 19:00-21:00 是学生在线进行口语学习与测评的高峰期，服务瞬时并发量极高，一旦后端人工智能基础设施平台无法承载如此巨大的服务并发量，

[16] 数据援引自公开媒体报道：https://www.chyxx.com/industry/202003/843106.html

就可能导致服务速度变慢。此外，用户对于人工智能口语测评的性能正在提出越来越高的要求，高实时性与低延迟正成为用户体验的重要组成部分，因此，人工智能口语测评系统在方案设计时需要着重考虑实时率指标，以提升用户体验；

- **稳定性：**由于在语言教学和测评场景中，学生的使用状况各不相同，因此系统在设计时需要针对不同特定场景，如背景噪音、无关语音等，以及不同的学生情况，如口音、漏读、错读等有着良好的容错性和鲁棒性。

而在以上三个核心指标之外，总拥有成本（Total Cost of Ownership，TCO）也是方案设计时重要的考量因素。基于非通用处理器的人工智能基础架构虽然能在性能上满足声学训练与推理的需求，但是采购、部署与运维成本较高，同时也难以实现快速的系统扩展。而基于通用处理器的人工智能基础架构则在以上方面具有优势，其能够高效利用教育机构既有的硬件基础设施，带来更好的成本优势。

## 面向英特尔® 架构优化的人工智能口语测评解决方案

### 人工智能口语测评解决方案

人工智能口语测评解决方案是利用事先已知的文本信息，计算输入语音对应于已知文字信息的相似性。相似性越高，说明发音越标准。因此，典型的方案架构可如图 2-3-3 所示构建。

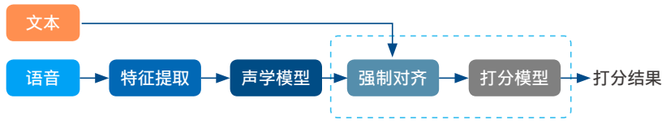



图 2-3-3　典型的人工智能口语测评解决方案架构

用户通过声音采集设备输入待测语音后，系统首先会基于自动语音识别（Automatic Speech Recognition，ASR）技术，开展特征提取和声学模型构建流程，然后将其与已知文本进行强制对齐（Force Alignment），强制对齐后再通过打分模型进行打分，以得到的分值做为发音好坏的评价。各主要工作流程内容如下：

#### ■ 特征提取

口语测评提取特征的过程是：先提取 MFCC（Mel-Frequency Cepstral Coefficients）特征，然后进行均值方差归一化（Cepstral Mean and Variance Normalization，CMVN）。首先，MFCC 特征具有更好的判别度，做 CMVN 的目的是减小麦克风和音频通道的影响，使模型的输入特征趋近于正态分布，模型描述性更好，更准确，可以增加模型的鲁棒性。但是，如果数据量够大，CMVN 处理可以省去。提取特征时，窗函数采用 25ms 的汉明窗，帧移 10ms，得到的是 39 维的 MFCC 特征。


图 2-3-4　特征提取流程

提取 MFCC 特征的 Python 脚本示例如下：

```
import librosa
import numpy as np

wav_data, sample_rate=librosa.load('test.wav', sr=None)
# 提取 MFCC 特征
feat=librosa.feature.mfcc(y=wav_data, sr=sample_rate, n_mfcc=39)
# 倒谱均值归一化
feat=(feat-feat.mean(axis=1)[:,np.newaxis])/(feat.std(axis=1)+1e-16)[:,np.newaxis]
```

使用 Kaldi 来提取 MFCC 特征和做 CMVN 处理的方法如下：

```
$ compute-mfcc-feats --use-energy=false scp:wav.scp ark:feats.ark
$ copy-feats ark:feats.ark ark,scp:feat.ark,feat.scp
$ compute-cmvn-stats scp:feat.scp ark,scp:cmvn.ark,cmvn.scp
```

#### ■ 声学模型

目前主流的口语测评方法，是先通过大量母语发音人以及适量非母语发音人的录音来训练一个语音识别系统，然后将其中的声学模型部分拿出来，作为口语测评的声学模型模块，用来计算音素对应的声学状态概率。随着深度学习的发展，基于深层网络 - 隐马尔可夫模型（DNN-HMM）的方法相比传统的基于混合高斯 - 隐马尔可夫模型（GMM-HMM）的语音识别系统有更好的性能表现。

声学模型的任务是根据输入特征帧来计算对应的音素的后验状态概率，一般会使用前后的一些帧，一般取 5 到 11 之间，上下文的窗口越大，就能提供更多的上下文信息。使用 nnet3 中的 nnet3-compute 来计算对数似然概率矩阵，核心计算过程示例如下：

```
NnetSimpleComputationOptions opts;
opts.acoustic_scale = 1.0;
Nnet raw_nnet;
ReadKaldiObject("model.mdl", &raw_nnet);
Nnet &nnet = raw_nnet;
CachingOptimizingCompiler compiler(nnet, opts.optimize_config);
SequentialBaseFloatMatrixReader feature_reader(feats_file);
for (; !feature_reader.Done(); feature_reader.Next()) {
  const Matrix<BaseFloat> &features(feature_reader.Value());
  DecodableNnetSimple nnet_computer(opts, nnet, \
                                    false, features, \
                                    &compiler, NULL, NULL, 0);
  Matrix<BaseFloat> matrix(nnet_computer.NumFrames(), \
                           nnet_computer.OutputDim());
  for (int32 t = 0; t < nnet_computer.NumFrames(); t++) {
    SubVector<BaseFloat> row(matrix, t);
    nnet_computer.GetOutputForFrame(t, &row);
  }
}
```

其对应的 Kaldi 的调用脚本如下：

```
$ nnet3-compute [options] <model-in> <feats-in> <output_wspecifier>
```

#### ■ 强制对齐

区别于语音识别任务，口语测评是事先知道语音的参考文本的，可以理解为在有限范围内语音识别，其搜索空间局限于跟读文本的范围内，允许多读、回读和跳读等。识别出文本之后，可以拿到对应的 HMM 序列，根据这些信息可以确定文本的每个音素、每个单词的时间轴信息，包括开始时间、结束时间、时长等，这个过程就叫强制对齐。实现中最常用的算法是加权有限状态转化器（WFST）和维特比（Viterbi）解码。

用 Kaldi 实现对齐过程，是先用跟读的文本构建解码图 HCLG，Kaldi 中使用 compile-train-graph 来加载声学模型，用 faster-decoder 解码器进行维特比解码，解码器的输出结果会包含识别文本序列、pdf-id 序列。其中，pdf-id 序列可以计算出音素的起始帧和帧数，根据音素的时间轴信息可以还原识别文本的单词、句子的时间轴信息。根据这些时间轴信息在声学模型中提取分数特征用来结算每个颗粒度的发音分数。对应的 Kaldi 的调用脚本如下：

```
$ compile-train-graphs [options] <tree-in> <model-in> <lexicon-fst-in> <transcriptions-rspecifier> <graphs-wspecifier>
$ align-compiled-mapped [options] <model-in> <graphs-rspecifier> <feature-rspecifier> <alignments-wspecifier>
```

#### ■ 打分模型

在口语测评中，通常每个颗粒度、音素、单词、句子都需要一个分数模型，这其中，音素的分数模型是最基础且最核心的部分。在 DNN-HMM 系统中，往往采用 GOP（Goodness of Pronunciation）及其衍生方法对音素来进行打分，其核心思想，是强制对齐后得到该音素在该音频片段中的似然分数，与该音频片段中识别的最佳音素的似然分数比较，以这个似然比（Likelihood Ratio）做为发音质量的评价。其基本计算公式（基于 DNN-HMM 模型）如下所示：

首先，定义音素的似然分数

$$LPP(p) = \log(p|o;t_s,t_e)$$

然后，该因素的 gop 分数可以定义为：

$$GOP(p) = \log(\frac{LPP(p)}{\max_{q\in Q} LPP(q)})$$

$LPP(p)$ 可以根据如下的公式计算

$$LPP(p) \approx \frac{1}{t_e-t_s+1}\sum_{t=t_s}^{t=t_e}\log p(p|O_t)$$

$$p(p|O_t) = \sum_{s\in P} P(s|O_t)$$

其中，p 为参考文本中当前要打分的音素，o 是强制对齐后 p 对应的语音段，$t_e$-$t_s$+1 是这段语音的帧数。从公式可以看到，GOP 打分表达的是一个条件概率，其衡量了在观测到测评者语音 o 的情况下，这段语音对应音素 p 的概率。概率值越高，发音越准确，反之则越差。具体计算过程可以参考 Kaldi 的 compute-gop.cc 脚本。

GOP 及其衍生方法在面向母语口语练习测评时有着较好的效果，但在一些外语口语测评场景下，如测评者的口音比较重、发音很不标准时，就可能由于似然比区分度过低而带来较大偏差。为此，近年一些厂商在方案设计时，会采用融合 GOP 分数和一些分类器如支持向量机（Support Vector Machine，SVM）、逻辑回归（Logistic Regression，LR）或决策树（Decision Tree）等进行监督性的训练，来进行更贴合打分标准的拟合。

评分决策树本质上是通过一系列规则对数据进行评分的监督学习过程，其具有的优势是评分更为细化。方案在构建之前，需要搜集大量的口语数据集，并进行逐一标注。然后，再结合

使用者的实际学习目标、学习内容以及学习深度等，制定一系列评分标准，例如长元音 / 短元音的发音标准、音素位置等等，而上述的 GOP 分值也会融合成为评分决策树中的一个维度。

## 基于英特尔® 架构，开展多云部署方案

一方面，为保证人工智能口语测评系统中，使用者的语音等相对敏感的数据能获得更有效的信息安全和隐私保护，并降低因网络不稳带来的处理延迟，形成更好的系统稳定性，教育机构一般采用私有云方式部署服务的核心部分。但另一方面，由于稳定性和实时性也是口语测评系统的重要评价指标，而对稳定性和实时性能影响较大的因素之一是口语测评使用时的“潮汐效应”，因此在方案构建时也需要加入相应的优化措施。

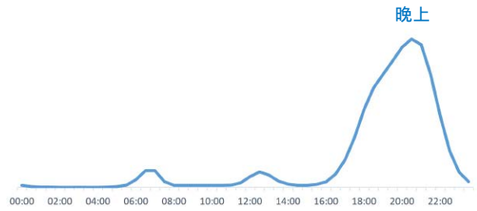



图 2-3-5　口语学习的潮汐效应

如图 2-3-5 所示，从一些统计分析得到的“口语训练”实时热点变化可以看到，就全天而言，在线口语学习系统的使用率有着明显的波动，峰值往往出现在放学后的 18 点至 20 点，而当观察范围扩大到周、月乃至年，可以发现潮汐效应在更大时间范围也同样存在，例如每年 9 月初开学后，老师对学生进行暑假学习成果考察，会令线上口语测评系统迎来一个小高峰，而每年 12 月和 5 月的期末复习季也是线上口语测评系统的热门使用时段。

这种潮汐效应下，要保证口语测评系统在高峰时段，大并发量接入时始终保持良好的稳定性和实时性，就需要系统按照热度峰值进行资源配置，但在平时和低峰时段，资源无疑会被闲置，这显然会提升成本且造成巨大浪费。现在，借助公有云具有的敏捷、弹性和易扩展的优势，口语测评解决方案厂商正基于英特尔® 架构，打造多云部署解决方案来应对这一挑战。

多云部署是目前云服务领域的热点之一，根据业务需求，将负载部署到本地私有云以及不同公有云服务上。多云部署能为在线口语测评系统带来多种助益，例如，更好的弹性伸缩能力可以帮助用户实现 IT 资源按需使用，并覆盖更多服务区域，而将敏感数据置于私有云区域，则可更好地保障数据安全。同时，这一混合部署方式也能有效降低用户的部署和业务维护成本。

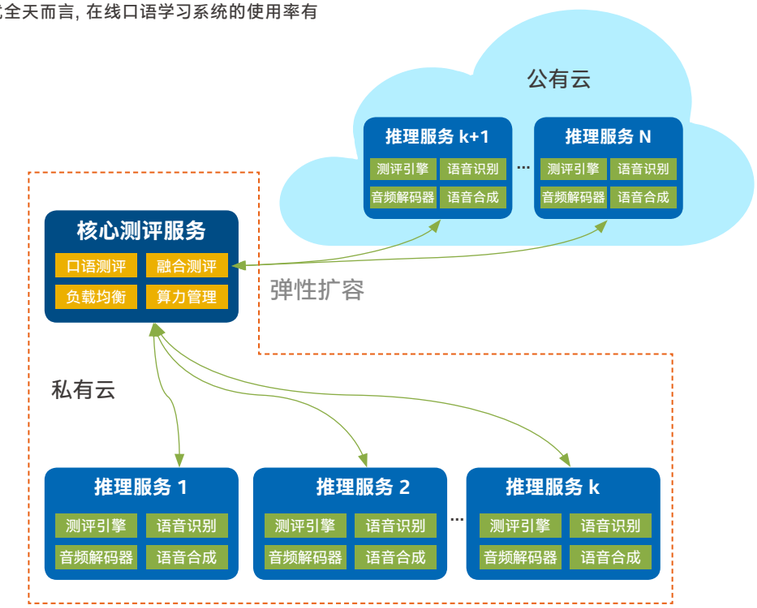



图 2-3-6　基于英特尔® 架构开展多云部署

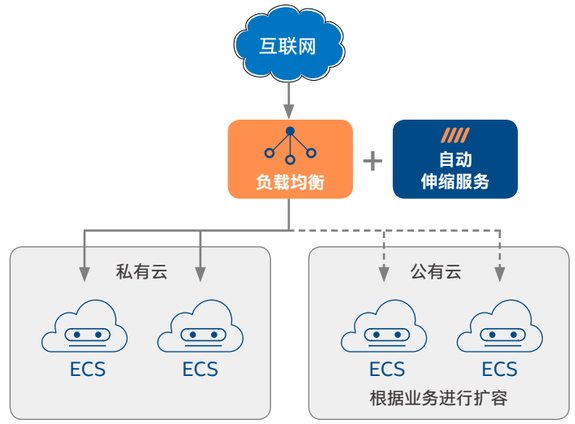



图 2-3-7　公有云扩充流程

实现高效多云部署的关键，是在不同公有云和私有云之间形成平滑、高效的容器和虚拟机迁移机制。出于服务连续性考虑，在线口语测评系统一般要求方案采取热迁移方式，这需要在进行系统设计和基础硬件设施选择时，就对性能做出充分考虑。

公有云扩充流程如图 2-3-7 所示，通过自动伸缩服务，用户可根据业务任意扩展 ECS（Elastic Compute Service）云服务。

通过集成开源的英特尔® oneDNN，英特尔优化了大量的、公开可用的深度学习框架（如 TensorFlow、PyTorch 等）。安装这些优化的开源框架，能充分挖掘英特尔® 至强® 可扩展处理器的算力，图 2-3-8 中给出了一个融合英特尔® 架构软硬件产品与其它各种资源的云服务系统架构图。

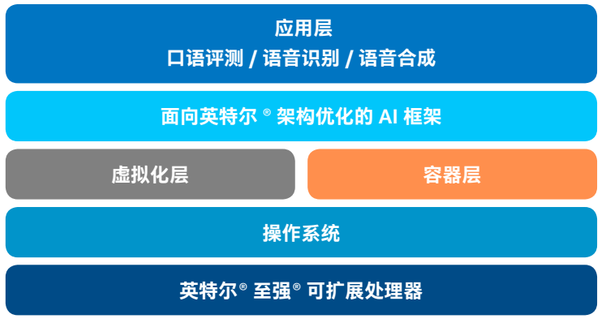



图 2-3-8　融合基于英特尔® 架构的软硬件产品的云服务系统架构图

## 英特尔® C++ 编译器助力提升方案整体效率

目前很多口语测评系统都基于 Kaldi 开发，还有部分合作伙伴为了进一步提升性能，用 C++ 重写了框架，不管是哪一种方式实现，脚本都是 C++ 语言。利用英特尔® C++ 编译器，可以快速优化口语测评系统的性能。英特尔® C++ 编译器对基于英特尔® 架构的处理器做了深度的优化，能更好地匹配和支持硬件的特性，比如英特尔® AVX-512 和英特尔® DL Boost（VNNI）等。在优化 Kaldi 框架时，将 Kaldi 依赖的加速库换成 oneMKL 库，然后用英特尔® C++ 编译器进行重新编译，安装好 oneMKL 库后，执行下面脚本，生成新的 MakeFile:

```
1.  $ export CC=icc
2.  $ export CXX=icpc
3.  $ ./configure -shared --mkl-root=/opt/intel/mkl --use-cuda=no
```

然后，将 kaldi.mk 中的 -g 去除，-O1 换成 -O3，如果处理器支持英特尔® AVX-512 和 VNNI 指令集，增加如下编译选项，然后编译即可。

```
-xHost -fp-model=consistent -mfma -mavx2 -mavx512f -mavx512vl
-mavx512bw -mavx512dq -mavx512cd -mavx5124vnniw
```

如果是使用 C++ 重构测评系统，可借助英特尔® VTune ™ Profiler 可视化性能分析器来分析和优化应用。作为英特尔面向软硬件性能瓶颈推出的分析工具，英特尔® VTune ™ Profiler 能够帮助用户迅速确定应用中的热点（Hotspot），并通过性能数据收集、数据间关系分析与展示以及分析潜在性能问题这一流程，对口语测评等人工智能应用进行优化。

在某应用场景的实战中，用户首先通过英特尔® VTune ™ Profiler 锁定热点函数 quantized_vector_product，然后利用 VNNI 指令集来实施优化。在使用英特尔® C++ 编译器进行编译时，需要选择添加选项 -D USE_VNNI，并另外添加英特尔® 编译器优化选项：

```
-xCORE-AVX512，-ipo，-no-inline-max-size 和 -no-inline-max-total
-size
```

代码脚本示例如下：

```
float quantized_vector_product(const size_t vectorSize, const unsigned
char *quantizedInput, const char *weights) {
    __m512i sum = mm512_setzero_si512();
    for (size_t j = 0; j < vectorSize; j += sizeof(__m512i)) {
        const __m512i input = __mm512_load_si512(reinterpret_cast<const
__m512*>(&quantizedInput[j]));
        const __m512i weight = __mm512_load_si512(reinterpret_cast<const
__m512*>(&weights[j]));
#define USE_VNNI
        asm("vpdpbusds %2,%1,%0":"+x"(sum):"x"(input),"mx"(weight));
#else
        const __m512i c = _mm512_maddubs_epi16(input, weight);
        const __m512i lo = _mm512_cvtepi16_epi32(_mm512_constsi512
_si256(c));
        const __m512i hi=_mm512_cvtepi16_epi32(_mm512_extracti64x4
_epi64(c,1));
        sum = _mm512_add_epi32(sum, _mm512_add_epi32(lo,hi));
#endif
    }
    return _mm512_reduce_add_epi32(sum);
}
```

# 基于英特尔优化方案的应用案例

## 一起教育科技：基于英特尔的产品与技术，打造先进人工智能口语测评平台

### 引言

*“来自英特尔的先进产品与技术，帮助人工智能口语测评平台的性能得到持续优化，从而为老师们提供了更加精准的口语评判体系，为学子们提供了更加有效的学情反馈。”*

——一起教育科技

### 背景与挑战

随着口语教学在英语教学中扮演越来越重要的角色，基于人工智能技术的口语测评应用也开始逐渐在教育领域获得青睐。高效、智能和使用便捷的人工智能口语测评平台，不仅能显著减轻学校与老师的教学负担，提高日常口语教学质量，还能通过信息技术手段，让标准化、普适化的服务帮助偏远地区的学生享受到优质的教育资源，有利于弥补区域之间的教育质量差，推动教育公平。

人工智能口语测评平台能够提供专业、实时服务的背后，有赖于优秀的测评解决方案、引擎设计丰富的口语语料库积累、以及强劲的人工智能算力提供的支撑。为帮助师生获得更精准的口语测评体系和学情反馈，长期致力于智慧教育方案设计研发的一起教育科技与英特尔协作，基于英特尔先进产品与技术，打造了全新人工智能口语测评平台。

### 解决方案

#### ■ 方案解析

为满足广大师生对智能化口语测评的需求，一起教育科技在其旗下的免费学习工具 “一起作业” 中推出了基于人工智能技术构建的口语测评平台。新平台通过课本点读、同步练习、口语交际、趣味配音以及绘本阅读等功能模块，有针对性地对学生英语口语水平进行准确测评，能够为校内老师们提供更加精准的口语评判体系，并为学生们提供更加有效的学情反馈。

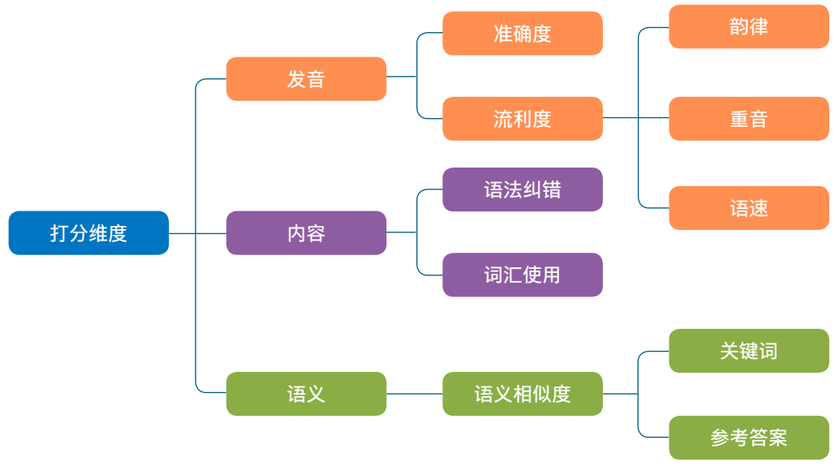



图 2-3-9　更多测评维度的一起教育科技口语测评引擎

以上平台能力的背后，是一起教育科技所打造的创新口语测评引擎。如图 2-3-9 所示，为了让口语测评更符合校内教学需求，且更贴近学生使用习惯，引擎在设计时一方面基于教学大纲进行考察，充分涵盖音素准确率、内容完整度、流利度、重音、韵律（语调，停顿）等维度；另一方面，也通过引入自研的评分决策树打分方法，使之更符合校内口语测评场景所需。

在使用中，学生只需选择一定的练习模式，通过麦克风等音频采集设备将口语音频上传至平台，平台就能基于口语测评引擎，以严谨的口语水平评分算法对学生的作答进行自动评判，并运用技术手段和大批量数据统计的方式进行结果复核，最终为学生反馈具体的评测结果以及相应的提示信息，以此帮助学生准确评估当前口语练习水平。

现在，基于口语测评引擎的口语测评、语音识别以及语音合成等人工智能应用，已经成为一起教育科技产品体系中的有机组成部分，并在学校中获得了广泛的部署实践。

#### ■ 方案亮点：

一起教育科技与英特尔协作，从算法设计与基础设施两个方面对口语测评解决方案进行了有针对性的设计与优化。从算法设计来看，方案从课堂教学、统一考试、日常练习等不同场景的实际要求出发，通过认真分析全国近三十个省市的英语口语考试测评标准和《中学英语课程标准》，通过反复迭代优化，为校内师生提供了一套符合课标、地方考试统一要求、以及满足日常练习、课后作业、形成性和过程性评价的口语水平评分算法。同时，一起教育科技还使用全国各地的题型对算法进行了多维度的检验，以证明其对于各式题型、各地区评分要求的适应性。

同时，在基础设施层面，一起教育科技与英特尔携手，基于英特尔的先进产品与技术构建全新的智能口语测评引擎。得益于算力的提升和算法的优化，新平台在测评准确性、测评效率等方面都实现了突破，能够真实地反映出学生当前口语的实际水平，并给出纠错建议。

为了让平台发挥更优效能，英特尔也为之提供了英特尔® 至强® 金牌 6230 处理器来作为智能口语测评服务的核心算力引擎。这一处理器属于第二代英特尔® 至强® 可扩展处理器产品家族，能够为计算密集型工作负载提供高性能和良好的可扩展性。同时，该处理器所集成的英特尔® 超级通道互联（英特尔® UPI）、英特尔® AVX-512等技术特性，也可满足平台中所需的密集 I/O 工作负载。

此外，处理器中内置的采用矢量神经网络指令（VNNI）的英特尔® DL Boost 技术也能显著提高平台中人工智能推理的工作效率，在保证测评结果准确性的同时，对平台的实时性性能提供了有效保障。

### ▪ 英特尔产品和技术发挥的作用及效果：

为验证在引入一系列英特尔产品与技术后，围绕口语测评引擎提供的能力以及学校日常开展口语学习和练习的需要，一起教育技术与英特尔一起，着重对口语测评、语音识别和语音合成进行了专门的优化和测试验证。

在实时性提升指标上，英特尔的优化方案包括，在方案中部署英特尔® 至强® 金牌 6230 处理器，并调用其 VNNI 指令集优化热点函数，然后将GCC 更换成英特尔® C++ 编译器，进行重新编译（编译方法详见第49页 “英特尔® C++ 编译器助力提升方案整体效率” 部分）。

在完成优化设置后，在一项 40 路并发多线程推理的过程中[17]，围绕语音练习的各项任务上都获得了提升，其中，口语测评工作流的实时率（越低越好）下降了 12.7%，优化后和优化前的性能对比如图 2-3-10 所示：

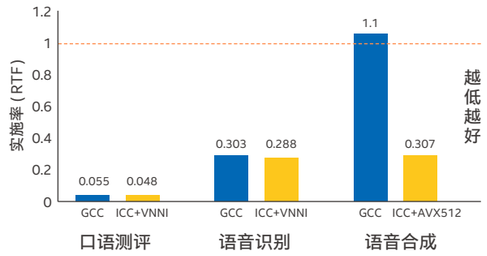



图 2-3-10　一起教育科技实战效果

在语音识别功能上，在预先对方案进行了重构和深度优化的情况下，来自VNNI 指令集和英特尔® C++ 编译器的优化（同样采用40 路并发多线程推理），也仍然能进一步帮助英特尔® 至强® 可扩展处理器发挥出更多算力潜能，如图2-3-11所示，优化方案使语音识别功能的实时率（越低越好）比优化前下降了5%。而在语音合成功能上，其声学器的 C 语言版本在不改变代码的情况下，使用英特尔® C++ 编译器的高级选项来进行优化，使整个工作流的实时率比优化前降低了 72.1%。

### 实战成效：

随着一起教育科技与英特尔合作推动的人工智能口语测评平台在更多学校获得部署和实践，其也为广大师生带来以下显著收益：

- **口语测评结果 “秒速” 提供：** 由于口语测评实时率的降低，学生在进行口语练习并提交平台测评之后，平台能够接近实时地将测评结果反馈给学生，以便其进行有针对性的纠正与学习。而语音识别与语音合成性能的提升，则为用户的口语学习提供了更高效的跟读训练、智能对话等服务。
- **口语测评服务始终稳定如一：** 得益于第二代英特尔® 至强® 可扩展处理器的高性能，以及一起教育科技在混合云架构与服务方面的创新，一起教育科技可以确保在高峰期也能为用户提供稳定高效的测评服务，避免响应缓慢等问题。
- **TCO 得到显著控制：** 相较于其他计算硬件，基于第二代英特尔® 至强® 可扩展处理器的人工智能服务器有着更为显著的 TCO 优势，能够帮助一起教育科技为校内师生带来更具价值的教学服务，更好地实现 “减负增效”。

### 解决方案中的软硬件配置建议

**硬件配置**

| 名称 | 规格 |
|---|---|
| 处理器 | 双路英特尔® 至强® 金牌 6230 处理器或更高 |
| 基础频率 | 2.10GHz |
| 核心/线程 | 20/40 |
| HT | On |
| Turbo | On |
| 内存 | 192GB (6x32G DDR4 2933MHz) |
| 硬盘 | 英特尔® 固态盘 D5 P4610 系列及以上 |

**软件配置**

| 名称 | 规格 |
|---|---|
| 操作系统 | CentOS Linux 8 (Core) |
| 内核版本 | 4.18.0-193.19.1.el8_2.x86_64 |
| 编译器 | ICC 19.1/GCC 7.3 |

[17] 测试配置：英特尔® 至强® 金牌 6230 处理器 @ 2.10GHz，192G DDR4 2933*6，CentOS Linux 8（Core），Kernel 4.18.0-193.19.1.el8_2.x86_64，英特尔® C++ 编译器 v19.1，GUN 编译器套件 v7.3，Python v3.6。

## 小结

得益于数据的不断积累、算力的不断突破以及算法的持续创新，基于人工智能技术构建的口语测评等语音交互、自然语言理解应用正在教育行业获得更为广泛的发展和落地，并在实际使用中为教学工作提供出色支持，帮助显著减轻老师的教学负担，提升了英语等语言类学科教学的精准性、有效性、针对性。

这一过程中，英特尔也以基于英特尔® 架构的处理器平台、英特尔® 深度学习加速技术等先进产品和技术，帮助合作伙伴和教育机构不断优化与革新各类口语测评平台的实战能力，使其更有效地完成全开放的口语测评环境建设、多维度的口语评价标准，以及地区化的差异测评方案适配等新的方案目标。

面向未来，英特尔将与众多合作伙伴一起，围绕创新硬件选型、人工智能 性能优化等方面进行更加深入的合作，发挥英特尔在端到端人工智能产品与技术上的优势，为人工智能教育应用提供强大的算力支持，实现跨架构的算法移植与优化，进而服务智慧教育，提供高效、公平、个性化的教育服务。

# 借力人工智能语音识别，打造高效教学辅助能力

## 英特尔携手合作伙伴探索基于语音识别的智能教学辅助能力

### 语音识别等人工智能技术在智慧教育场景中的应用

包括人物识别、目标检测、行为分析在内的一系列智能化应用已逐渐成为教育行业实施数字化、智能化转型的重要辅助手段。一方面，得以充分整合的人工智能应用方案和技术，正帮助教育机构大幅扩展教学场景和环节的广度和深度，让师生之间形成 360 度教学体验；另一方面，在教育机构的日常管理工作中，语音识别等人工智能技术也可从不同维度为广大师生提供各类服务。

这其中，基于人工智能的语音识别技术，也称为自动语音识别（Automatic Speech Recognition，ASR），是常见的教学辅助应用之一，目标是将自然语音中的内容准确、快捷地转换为文本并与其它人工智能、大数据和互联网技术一起，构建多样化的智能应用供师生使用。在本篇中，我们将聚焦语音识别在智能会议构建及提升日常教学效率两个场景中的应用。

**■ 基于语音识别打造智能会议流程**

交流协作的重要性正越来越为教育行业所认同。虽然邮件、电话、办公自动化（OA）系统等方式也能实现大量的信息交流，但面对面的会议沟通仍是各教育机构最主要的交流协作方式。在会议中，学校管理者与教师之间，可以进行教学任务的上传下达；教师与教师之间，可以进行教学经验的交流；而教师与家长 / 学生之间，可以对教学问题进行深度复盘。

面对面的语言交流虽然有着交流效率高，沟通速度快等优势，但会议中产生的高价值信息同样也有着生命周期短、辐射范围窄等短板。随着越来越多的人工智能应用被引入会议流程，许多教育机构也通过智能会议系统的构建，来推动会议价值的提升。如图 2-4-2 所示，这类系统首先可通过语音识别能力，将非结构化的音视频信息文本化，然后通过信息提取和信息利用过程打造各类教学辅助能力。

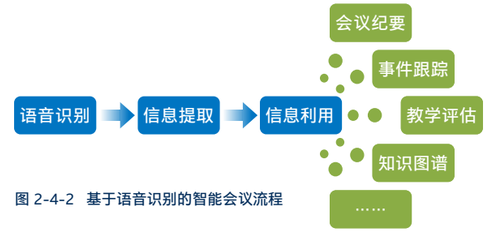



图 2-4-2　基于语音识别的智能会议流程

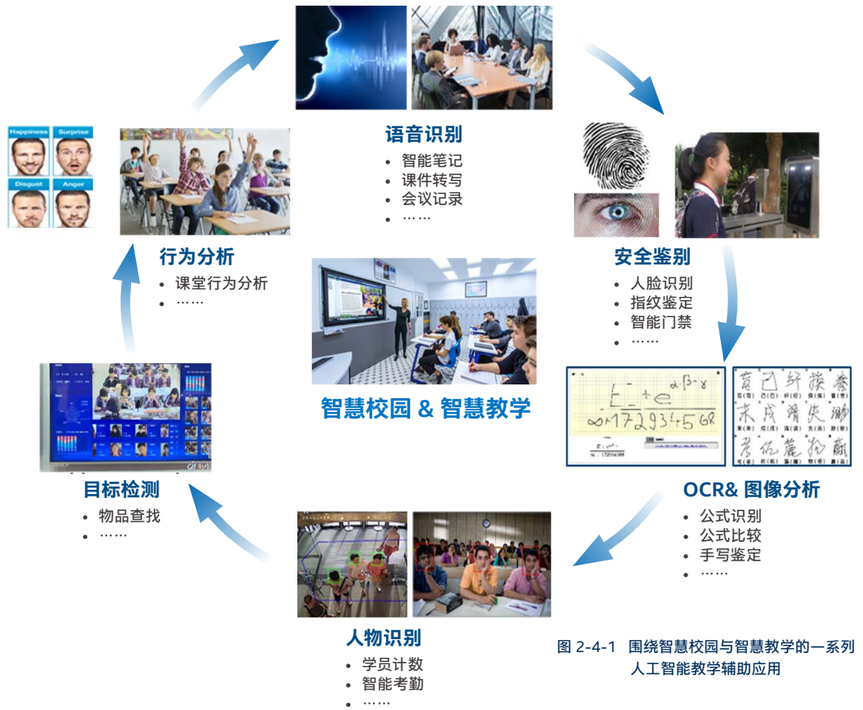



图 2-4-1　围绕智慧校园与智慧教学的一系列人工智能教学辅助应用

例如智能化的会议纪要，系统可实时将其中的关键字形成可检索的标签，并与音轨一一对应，方便后续资料整理、回溯时通过音频检索技术迅速定位。同时，借助一些新兴技术，还可以将会议纪要进一步解析，成为知识图谱等人工智能应用的数据基础。而其它一些人工智能应用，如学科知识自动归纳整理、教学质量评估、校园事件跟踪等，也能帮助教育机构、教师、学生以及家长之间进一步提升交流与沟通效率。

**■ 利用语音识别提升日常教学效率**

在日常教学过程中，课前、课中和课后不同阶段的教学需求，语音识别功能可有效地提高教学效率。如图 2-4-3 所示，在课前，学生可利用语音识别产品在人机互动上的优势，开展线上课程预习，并使用口语测评等应用提升语文、英语等语言类学科的学习效率。在课中，基于实时语音识别技术构建的智能笔记类产品，可以帮助学生自动记录教师的授课内容，使学生能更加专注于学科知识的理解，而不因教师的口音、外界的噪声等因素分心。同时这类产品还可以通过引入知识库等人工智能能力，帮助学生更好梳理各学科的知识脉络。此外，智能翻译等应用，则能够帮助留学生等人群更方便地适应非母语教学环境。


图 2-4-3　可用于智慧课堂全流程的语音识别功能

在课后，利用语音识别技术，无论是教师还是学生都可以方便地进行课程回顾，复盘教与学过程中的得失。同时，得益于文本表述在阅读速度和表述准确性上的优势，通过语音识别转化文本的方式，也能提升师生在线上答疑和课程咨询时的效率。

本篇章将以上述两个场景的应用为例，介绍如何打造在校园环境部署的语音识别方案边缘端。首先，下文将介绍方案在教育机构部署时面临的挑战，然后给出一个旨在提供高效边缘人工智能算力的英特尔的优化方案。

## 基于语音识别能力构建教学辅助能力

### 语音识别在教学辅助中应用时面临的挑战

语音识别解决方案的评估指标一般为字错误率（Word Error Rate,WER）和实时率（Real Time Factor, RTF），面向教育机构的应用场景还有数据隐私和安全性的要求。当方案厂商与教育机构一起优化基于英特尔® 架构的 AI 负载的过程中，往往会遇到以下挑战：

**■ 边缘算力挑战：**

诸如智能笔记、互动提示等人工智能应用在为智慧课堂提供新颖实用的教学辅助功能时，对底层硬件设备提出了更高的算力要求。传统教室课堂中部署的教学 PC 等设备往往性能有限，无法承受高并发、低时延的处理需求。

采用云部署方案虽然可以部分缓解使用端算力不足的难题，但囿于网络带宽限制、网络干扰以及传输时延等因素，完全依赖云端解码的语音识别系统依然面临实时率不佳的问题，在交互性很强的课堂使用环境下，这一问题无疑将大幅降低师生使用体验。

同时，未经优化的处理器等硬件设备在执行训练、推理等人工智能任务时，往往无法将潜在的算力优势发挥到最大，要么不能实时处理，要么不能支持更多路并发，导致 TCO 过高。因此，在选择这些参考硬件配置时，同样也需要根据应用场景、所使用的软件框架等开展相应的优化。

**■ 硬件兼容性带来的挑战：**

教育机构在其 IT 建设过程中，往往会根据不同的需要，采购多样化的硬件配置，例如 CPU, GPU, FPGA, VPU 等，导致承载人工智能应用的硬件平台变得多种多样。如果面向每一种硬件平台都需要进行模型集成开发等工作，无疑会加重教育机构的负担。因此，各类人工智能应用方案在部署时，需要充分考虑硬件的兼容性，尽可能做到一套人工智能模型和部署工具兼容各种硬件平台。

**■ 数据隐私安全性挑战：**

在智能会议或日常教学中使用语音识别功能时，采集到的语音或转换的文本往往包含大量敏感信息，例如学校的教学情况、人员动向、学生的个人信息、考试成绩等。而语音识别中的语音转文本流程，以及后续的数据智能化应用所涉及的各种操作，例如训练、推理、存储、备份等，如全部采用公有云部署方案，无疑会产生较多的信息暴露面和攻击点，增加泄密风险。因此，用户在进行方案设计时，需要同时考虑更安全的部署方案来保护数据隐私。

**■ 复杂环境对准确度的挑战：**

与商务会议等场景不同，部署在课堂等处的人工智能应用通常会面对更为嘈杂的使用环境，因此方案设计时需要考虑更多的噪音环境和人为干扰因素。同时，一些教育管理类应用，例如智能化的会议信息同步等，在提升行政和沟通效率的同时，也对系统的准确率和效率提出了更高的要求。

而从实战来看，要进一步提升语音识别相关教学辅助应用在实际场景中的使用表现，解决方案厂商可以采取以下措施：

- 针对教育环境不同场景的部署提供最优的参考硬件配置，引入更高效、更具性价比的基础算力；
- 针对不同类型教育机构的需求，提供面向不同场景的人工智能优化方案。
- 针对教育环境既有的复杂算力设备提供有效的兼容性方案；
- 根据教育机构实际的应用场景，IT 环境以及对数据隐私安全的要求等级，采用更安全的私有化部署方案。

### 基于英特尔® 架构的产品在边缘侧构建实时智能会议解决方案

基于以上挑战，英特尔正与众多智慧教育解决方案合作伙伴一起，根据语音识别应用在智能会议、智慧课堂等场景中的部署要求，提出基于英特尔® 架构产品的边缘侧实时智能会议解决方案。

由于传统教室 PC、笔记本电脑等一线 IT 设施很难承载高负荷的信号处理和模型推理任务，因此，传统语音识别解决方案往往需要云端或远端数据中心提供支持，方案中核心的特征提取、模型训练和解码等都部署在远端，语音被采集后通过网络上传。这种模式下，当网络带宽或处理能力受限时，语音数据和识别后的文本就无法实现实时同步，造成使用体验的下降。特别是在智能会议字幕等实时在线语音识别应用中，过长的时延会带来字幕卡顿、不同步等现象，严重影响与会者体验。

采用边缘部署的方式可以有效解决这一问题。边缘方案能在贴近师生一线的边缘侧构建高效的计算处理和模型推理能力，打破传统方案中的网络和处理性能瓶颈。

一种典型的边缘部署方案如图 2-4-4 所示，这是一个基于语义理解的智能会议纪要解决方案，能够在基础的语音识别能力之上，一方面通过声纹识别，使会议中各发言人的发言内容能够被轻松回溯，另一方面通过智能标点和语义理解模块，大幅提升会议纪要的准确性。

在方案中，由麦克风采集的音频数据经模 / 数（A/D）转换后，通过微控制单元（MCU）传送到遵循英特尔® OPS 规范的终端中（例如嵌入式一体机）。在这里，借助 OPS 终端内置英特尔® 酷睿™ 处理器在处理性能以及多媒体处理方面的优势，能够高效完成音频编解码以及回声消除、去混响和波束形成等信号处理流程。

经处理后的数据流就近传输至部署在校园、教室周边的边缘节点中，借助英特尔® 至强® 可扩展处理器、OpenVINO™ 工具套件等软、硬件产品，通过语音识别、声纹识别、智能标点和语义理解等人工智能应用模块，完成整个基于语义理解的智能会议纪要流程。

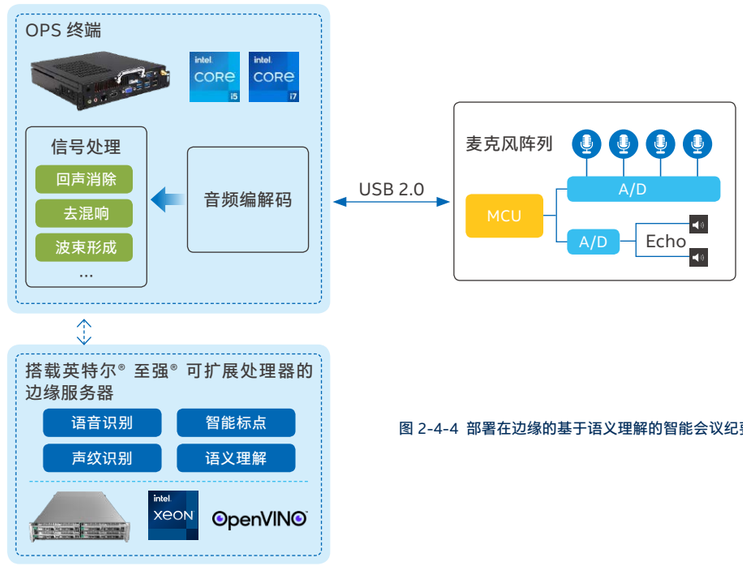



图 2-4-4　部署在边缘的基于语义理解的智能会议纪要解决方案

## 扩展 OpenVINO™ 工具套件自定义层，提升语音识别推理效率

由英特尔开源的英特尔® 发行版 OpenVINO™ 工具套件（以下简称为 “OpenVINO™ 工具套件”），能够在基于人工智能的语音识别解决方案中支持基于多种深度学习框架训练的模型，如 TensorFlow、Caffe、 MXNet、Kaldi 以及 ONNX 等，帮助方案有效加速推理速度，提升方案整体性能。由于不同框架支持的层不一样，用户在具体使用和部署时可参考具体框架的支持列表 [18]。

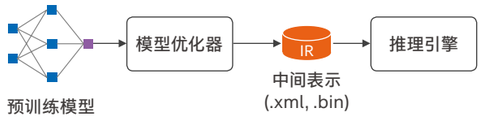



图 2-4-5　OpenVINO™ 工具套件正常处理流程

OpenVINO™ 工具套件正常的处理流程如图 2-4-5 所示，先将预训练模型通过模型优化器转换为 IR 格式文件，然后调用推理引擎 API 执行后序推理计算。

虽然 OpenVINO™ 工具套件所支持的深度学习框架提供了各式各样的层，但在语音识别等人工智能应用场景中，有时候用户依然需要使用自定义层来扩展现有框架功能，此时，用户在使用 OpenVINO™ 工具套件进行方案优化时就可能遇到异常，例如用模型优化器转换包含自定义层的模型时，会出现 “存在不支持的层” 的错误提示，导致无法进行下一步 IR 文件转换。这就需要通过拓展 OpenVINO™ 工具套件的自定义层，对模型优化器实施优化来实现模型优化转换，并拓展推理引擎来实现推理引擎的封装。

下面将介绍这一方法的具体流程。

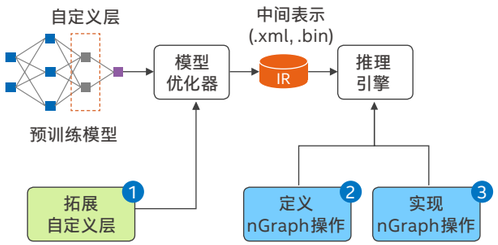



图 2-4-6　OpenVINO™ 工具套件处理自定义层的流程

如图 2-4-6 所示，实现自定义层拓展通常可分为三个步骤：

1. 在模型优化器中添加对自定义层的支持，以便模型优化器用来生成 IR 文件；
2. 创建一个操作集（Operation Set），并且实现自定义 nGraph 操作；
3. 根据推理部署的平台插件（plugin）为推理引擎实现 nGraph 操作；

以语音识别方案中常见的快速傅立叶变换（Fast Fourier Transform，FFT）操作在通用处理器（CPU）平台的优化为例，其实现自定义层的操作步骤如图 2-4-7 所示。首先，用户需要为模型优化器实现模型的自定义层拓展，创建一个目录，如 mo_extension，用于定义上述自定义层所在的位置。

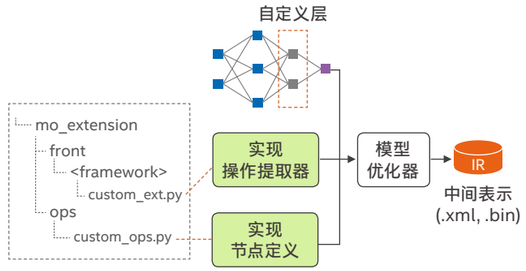



图 2-4-7　模型优化器的自定义层实现

然后，在该 mo_extension 目录下创建 2 个子目录：front/<framework> 和 ops，其中 <framework> 用于指明需要为何种模型框架创建该自定义层，而 ops 则用于定义节点。

在接下来实现自定义节点和抽象提取器步骤中，用户需要在目录 ops 下创建 Python 文件 FFT.py，该文件用来定义节点，示例代码如下：

```
from mo.front.common.partial_infer.elemental import copy_shape_infer
from mo.graph.graph import Graph
from mo.ops.op import op
class FFT(Op):
    op='FFT'
    enable=True

    def __init__(self, graph: Graph, attrs: dict):
    #  infer 属性需给一个计算输出张量 shape 的函数，
    #  这里 copy_shape_infer 表示该操作输入和输出的 shape 相同
        super().__init__(graph, {
            'type': __class__.op,
            'op': __class__.op,
            'in_ports_count': 1,
            'out_ports': 1,
            'infer': copy_shape_infer}, attrs)
    def supported_attrs(self):
        return ['inverse']
```

然后在 front/<framework> 目录下创建 Python 文件 fft_ext.py，该文件是用于在模型优化器的拓展接口上实现 FFT 操作的抽象提取器，示例代码如下：

```
from mo.front.extractor import FrontExtractorOp
from ../ops.FFT import FFT

class FFTFrontExtractor(FrontExtactorOp):
    # 定义算子在模型框架中的操作类型
    op='FFT'
    enable=True

    @classmethod
    def extract(cls, node):
        data = {'inverse': 0}
        # 更新操作节点属性
        FFT.update_node_stat(node, data)
        return cls.enabled
```

在实现自定义节点和抽象提取器后，模型优化器就可以正确生成包含 CustomOps 节点的 IR 模型文件。在进行模型转换时需要添加 --extension 选项，用于指定 mo_extensions 目录，即上述模型优化器拓展接口实现的自定义层所在的目录。

```
python3 /opt/intel/openvino/deployment_tools/model_optimizer/mo_onnx.py \
    --input_model model.onnx \
    --extensions mo_extensions
```

以上是一个用来说明自定义层拓展流程的简单示例，如需了解更详细的信息，请参见 OpenVINO™ 工具套件说明文档 [19]。

在完成模型优化器的扩展后，用户也需要通过实现一个扩展插件，来对 OpenVINO™ 工具套件中的推理引擎进行扩展。作为一个动态库，如图 2-4-8 所示，这一扩展插件通常通过实现以下三个步骤来实现扩展操作所需要的接口，完成推理引擎的扩展：

- 推理引擎自身提供了 InferenceEngine::IExtension 接口，因此在进行推理引擎扩展时，扩展插件首先要实现这个接口，该接口包括了扩展插件所支持的操作等 API；
- 推理引擎中同样也需要实现一个新的 nGraph 操作，使其在读入 IR 文件时，能够正确解析自定义的网络层；
- 推理引擎的扩展插件的同时还需要实现新操作的计算核（kernel）函数。Kernel 函数的实现，与目标平台是相关的。


图 2-4-8　推理引擎的扩展

下面仍然以 FFT 操作在 CPU 平台的优化为例，介绍上述推理引擎扩展的三个实现步骤：

### ■ 实现扩展库（Extension Library）接口

推理引擎库是一个动态执行库，需要实现 InferenceEngine::IExtension 接口，InferenceEngine::IExtension 为推理引擎库的一个基类，需要继承该基类定义一个新的扩展类，示例代码如下：

```
#include <ie_iextension.h>
#include <ngraph/ngraph.hpp>

class NewExtension: public Inference::IExtension {
public:
    NewExtension() = default; // 缺省构造函数
    void GetVersion(const InferenceEngine::Version *& versionInfo) const noexcept override;
    void Unload() noexcept override{} // 空函数，unload 时不做其他工作
    void Release() noexcept override { delete this; }
    std::map<std::string, ngraph::OpSet()> getOpSets() override;
    std::vector<std::string> getImplTypes(const std::shared_ptr<ngraph::Node& node) override;
    InferenceEngine::ILayerImpl::Ptr getImplementation(const std::shared_ptr<ngraph::Node>& node, const std::string& implType) override;
};
```

其中，getOpSets 函数是将 nGraph 的扩展定义加入到操作集合 (opset) 里， getImplType 函数中设置使用哪种平台，getImplementation 函数根据节点类型返回对应核函数。

此外还要加入一个全局注册函数 InferenceEngine::CreateExtension()，推理引擎加载动态库时会自动调用该函数，得到创建的 NewExtension 对象，示例代码如下：

```
INFERENCE_ENGINE_API(InferenceEngine::StatusCode)InferenceEngine::CreateExtension(InferenceEngine::IExtension *&ext, InferenceEngine::ResponseDesc * resp) noexcept {
    try {
        ext = new NewExtension();
        return OK;
    } catch (std::exception &ex) {
        if (resp) {
            std::string err = ((std::string) "Error for creating extension: ");
            err.copy(resp->msg, 255);
        }
        return InferenceEngine::GENERAL_ERROR;
    }
}
```

[18] https://docs.openvinotoolkit.org/latest/openvino_docs_MO_DG_prepare_model_Supported_Frameworks_Layers.html

[19] https://docs.openvinotoolkit.org/latest/openvino_docs_HOWTO_Custom_Layers_Guide.html

### ■ 自定义 nGraph 操作

推理网络隔层操作在推理引擎内部通过 nGraph 的图来表示，因此需要扩展 nGraph 的实现来使推理引擎能够解析自定义层。在定义新的操作时，需要通过继承 ngraph::op::Op 基类实现，示例代码如下：

```
class FFTOp : public ngraph::op::Op {
public:
    static constexpr ngraph::NodeTypeInfo type_info {“FFT”, 0};
    const ngraph::NodeTypeInfo& get_type_info() const override { return type_info; }
    FFTOp() = default;
    FFTOp(const ngraph::Output<ngraph::Node>& inp, bool inverse);
    void validate_and_infer_type() override;
    std::shared_ptr<ngraph::Node> clone_with_new_inputs(const ngraph::OutputVector& inputs) const override;
    bool visit_attributes(ngraph::AttributeVisitor& visitor) override;
    bool inverse;
};
```

其中，validate_and_infer_types 函数的主要功能是根据输入数据的形状和类型，设置自定义节点的输出数据的形状和类型，clone_with_new_inputs 函数根据输入节点构造一个新的操作节点，visit_attributes 函数返回操作支持的属性。在自定义新的 nGraph 操作后，就可以通过 getOpSets 函数添加到 opset 里面。

### ■ 实现 CPU 的执行 kernel

实现执行 kernel 时，需要从基类 InferenceEngine::ILayerExecImpl 继承一个新类，定义示例代码如下：

```
class FFTImpl :: public InferenceEngine::ILayerExecImpl {
public:
    explicit FFTImpl(const std::shared_ptr<ngraph::Node>& node);
    InferenceEngine::StatusCode getSupportedConfiguration(std::vector<InferenceEngine::LayerConfig> &conf, InferenceEngine::ResponseDesc *resp) noexcept override;
    InferenceEngine::StatusCode init(InferenceEngine::LayerConfig &config, InferenceEngine::ResponseDesc *resp) noexcept override;
    Inference::StatusCode execute(std::vector<InferenceEngine::Blob::Ptr> &inputs, std::vector<InferenceEngine::Blob::Ptr> &outputs, InferenceEngine::ResponseDesc *resp) noexcept override;
private:
    ngraph::Shape inpShape;
    ngraph::Shape outShape;
    bool inverse;
    std::string error;
};
```

其中，getSupportedConfiguration 函数返回 kernel 支持的输入与输出数据格式，init 函数主要是用来检查推理引擎选择的 config 的合法性，execute 函数是 kernel 的计算的实现函数，里面可以使用各种加速指令集编写执行代码。定义了该类，就可以通过 getImplimentatino 进行注册。

在应用程序中，可通过类似于下面的示例代码，将扩展动态库提供推理引擎：

```
InferenceEngine::Core core;
//动态库名字为 custom_fft.so
auto ext = make_so_pointer<InferenceEngine::IExtension>("custom_fft.so");
core.AddExtension(ext);
core.ReadNetwork("example_model.xml", "example_model.bin");
……
exe_network = core.LoadNetwork(network, dev);
……
infer_request = exe_network.CreateInferRequest();
infer_request.infer();
```

上述流程与正常的 API 推理代码相比，其不同点在于需另外调用 InferenceEngine::AddExtension() 接口来将扩展动态库加载到推理引擎中。

通过以上步骤，用户就完成了语音识别等人工智能模型在 CPU 平台上进行扩展所需的所有工作。简单来说，这一过程需要同时实现一个扩展库的接口，定义新的 nGraph 的操作和实现一个执行计算的 Kernel 函数。更多信息可参考 OpenVINO™ 工具套件相关文档 [20]。

## 解决方案中的软硬件配置建议

### 硬件配置

| 名称 | 规格 |
|---|---|
| 处理器 | 双路英特尔® 至强® 金牌 6240R 处理器 |
| 基础频率 | 2.40GHz |
| 核心/线程 | 24/48 |
| HT | On |
| Turbo | On |
| 内存 | 192G (16G DDR4 2933MHz x 12) |
| 硬盘 | 英特尔® 固态盘 DC P4610 系列及以上 |

### 软件配置

| 名称 | 规格 |
|---|---|
| 操作系统 | CentOS 7.8.2003 |
| 核心 | 3.10.0-1127.18.2.el7.x86_64 |
| 编译器 | GCC 8.0 |
| 框架 | OpenVINO ™ 工具套件 2021.2 及以上 |

[20] https://docs.openvinotoolkit.org/latest/openvino_docs_IE_DG_Extensibility_DG_Intro.html

# 基于英特尔优化方案的应用案例

## 思必驰：与英特尔携手打造精准、高效的语音识别应用，加速智慧教育前行步伐

### 引言

*“基于英特尔® 架构的处理器在并行计算处理方面的优势，以及 OpenVINO™ 工具套件在模型推理上的加速能力，让我们在模型和解码器上的优化方案更具效能，能有效提升语音识别系统的识别准确率、并发量和实时率。”*

——思必驰

### 背景与挑战

作为领先的对话式人工智能平台公司，思必驰科技股份有限公司（以下简称“思必驰”）通过全链路的智能语音、语言技术，致力于为众多行业场景和合作伙伴提供自然语言交互解决方案。为帮助教育行业加速“人工智能 + 教育”转型，思必驰与英特尔一起，使用基于英特尔® 架构的处理器、OpenVINO™ 工具套件等软硬件产品，构建了高效边缘 MEC 平台方案，并利用英特尔产品在并行计算处理、模型推理效率等方面的优势，对语音识别系统中核心的声学、语言模型以及解码器组件等进行了重构优化，以全新方案设计为教育领域的语音识别智能应用提供动力引擎。

### 解决方案

#### ■ 方案解析

思必驰语音识别系统架构可采用图 2-4-9 的方式构建。其可分为语音识别和模型训练两个主要阶段。在语音识别阶段，当需要识别的语音语料到达系统时，系统首先会对语音进行声学特征提取，将数字音频信号转换为语音识别系统能够处理的输入形式。再根据模型库提供的声学模型、语音模型和发音字典，对声学特征进行解码搜索，获得对应的文本结果。

基于以上流程，思必驰方案能根据教育机构需求，既用于实时在线识别，也用于非实时离线识别。其中，实时在线识别一般是在会议、教学等场景的进行过程中，对语音进行随时采集、随时识别，典型应用如会议实时字幕系统，就对系统时延等性能有着较高要求。而非实时离线识别一般是对较长的录音文件进行识别，典型应用如会议录音文本化存档，对处理并发量以及识别准确率等有着较高要求。

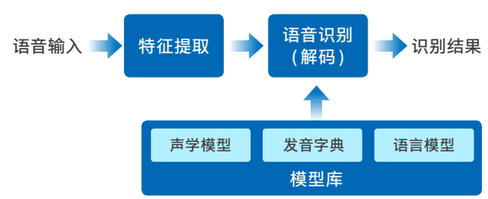



图 2-4-9　思必驰语音识别系统架构

**实时在线语音识别**

在实时在线语音识别过程中，方案首先需要将不定长的语音信号分成适于处理的语音小段，一般可采用语音活动检测（Voice Activity Detection，VAD）方法进行处理，VAD 主要作用是在音频会话期间去除非语音片段，减少网络数据传输。传统的串行 VAD 方式，音频流先经过 VAD 判断是否静音，如果是静音就不继续后续流程，如果是语音就将音频数据送到后续的 ASR 模块进行识别，这样 ASR 模块处理数据严格受 VAD 控制，因此 ASR 模块和实时语音数据之间会有一个延迟（约 20~30ms）。并行 VAD 方式是将音频流数据同时传输给 VAD 模块和 ASR 模块，两个模块同时处理，ASR 模块不需要等待 VAD 处理完毕的结果（除语音开头）。当并行处理时，VAD 模块发现此时出现静音段，需要终止识别时，VAD 模块会发送一个终止信号给 ASR 模块，ASR 模块接受信号，终止识别。

为提升识别质量并降低识别时延，在实操中，实时在线语音识别通常可采用边缘侧部署的方式来加快响应速度。被麦克风等设备采集的语音信号首先需要进行去噪音、回声消除以及去混响等处理，然后如图 2-4-10 所示，可以采用并行化 VAD 与 ASR 的设计来加快识别效率，降低时延。

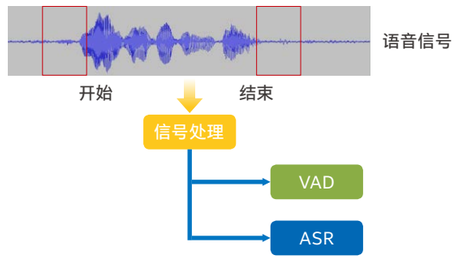



图 2-4-10　VAD 与 ASR 并行化设计的实时在线语音识别架构

**非实时离线语音识别**

面向大音频文件的离线语音识别一般对识别实时率不敏感，但对识别准确率要求较高。因此在系统设计时通常采用在云端 / 数据中心进行识别。

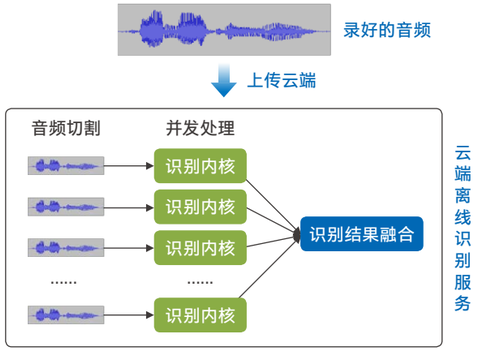



图 2-4-11 多路并发的非实时离线 语音识别系统

如图 2-4-11 所示，用户音频被上传到云端后，可分割为多个音频段落进行并发处理。这一设计的优势在于能使不同说话人的语音分离，独立进行识别过程，使识别结果不受口音、方言等因素影响。更重要的是，并发处理的方式能充分发挥基于英特尔® 架构的处理器带来的多内核优势，加速识别进程。

各个识别内核分别完成各自的识别任务后，再由系统内置的，基于人工智能模型（例如 BERT）的融合策略对各个识别结果进行排错、去重、断句以及字词句融合，形成完整的文本信息。

## ■ 方案亮点

**创新 VDCNN-CTC 声学模型**

为避免传统 HMM 模型需要的帧级别对齐问题，并充分利用 OpenVINO™ 工具套件在深度学习模型推理上的优势，思必驰在新的语音识别解决方案中加入其自研的创新 VDCNN-CTC 声学模型，在提升识别准确率的同时，加速模型解码效率。

作为一种深度卷积神经网络，VDCNN（Very Deep Convolutional Neural Network）可以融合深度卷积网络，并加入了残差连接、池化降采样、低帧率加速等技术，避免传统 DNN 网络中训练困难、推理速度慢的问题。同时，新模型还通过加入行业领先的字级别建模单元，并结合 OpenVINO™ 工具套件提供的模型优化器和推理引擎，使模型在保证高识别准确率的同时，也能大幅提升解码速度。

**语言模型优化方案**

针对教育环境中复杂的语言环境，思必驰对传统 Ngram 统计模型进行了两方面的优化。首先是将 Ngram 模型分为两级打分模式，低阶模型用于一阶段快速打分，高阶模型在一阶段得分基础上进行重打分。这一优化方案虽然对硬件基础设施提出了更多算力要求，但可以有效提升教室等背景噪声复杂的环境中语音识别的准确率。而方案中基于英特尔® 架构的处理器的加入，更是为新增算力需求提供了保障。

此外，思必驰也基于大规模神经网络语言模型技术，为方案额外增加了 NN 语言模型，在有效提升识别准确率的同时，还结合方案中的“多路解码融合”技术，以帮助用户实现语音识别的场景化快速定制。

**智能标点**

断句并加上准确的标点符号，一直是人工智能语音识别技术领域的技术难点，在语言博大精深且语境复杂多变的中文环境中更为如此。而与此同时，智能标点在会议记录等场景中又有着巨大的需求。为此，思必驰在方案中对模型采用基于 BERT（Bidirectional Encoder Representations from Transformers）大规模语料预训练，并结合多任务学习（Multi-Task Learning）和知识蒸馏（Knowledge Distillation）方法来提升智能标点等语音识别后处理技术的效率和准确度，获得了良好的实践成果。

**基于英特尔® 架构处理器的多路融合解码器设计**

得益于基于英特尔® 架构的处理器提供的英特尔® AVX 2、英特尔® AVX-512 等指令集带来的并行计算处理增强能力，思必驰在方案中设计了新的 “多路解码融合” 技术。如图 2-4-12 所示，其可以通过采用多个代表场景不同的语言模型的结合来保障通用场景的高识别率，同时结合快速的场景化模型定制，尤其是场景化语言模型定制能力，保证在一些特殊场景下的识别率。通过采用该架构，一方面提高了语音识别在通用场景下“开箱即用”的能力，另一方面也让使用者可以进行快速的大规模自定制。

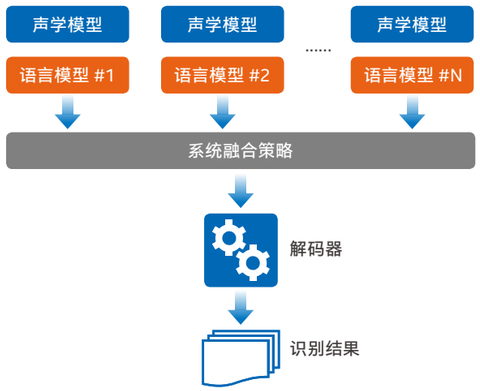



图 2-4-12 支持多路融合的解码器

## ■ 英特尔产品和技术发挥的作用及效果

为验证在引入一系列英特尔产品与技术，尤其是基于英特尔® 发行版 OpenVINO™ 工具套件开展优化后，人工智能语音识别应用在不同使用场景中的性能表现，思必驰与英特尔一起，着重对在线和离线模式下的语音识别，以及智能标点符号预测三个模型进行了优化和测试验证[21]。优化过程中，三个模型都包含了一个 OwnQuantMul 的自定义操作，需要利用 OpenVINO™ 工具套件的自定义层扩展来优化模型。

在实时在线语音识别场景中，经过 OpenVINO™ 工具套件优化后，方案实时率（越低越好，在大于 1 的情况下可视为失去实时性）由 0.0509 下降至 0.0289，下降了约 43.2%，优化后和优化前的性能对比如图 2-4-13 所示：

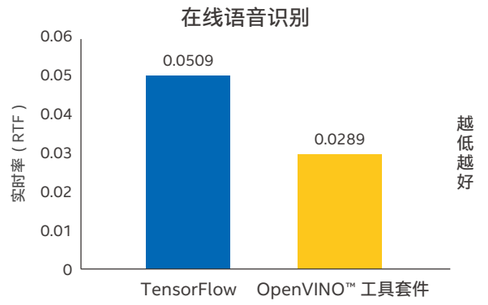



图 2-4-13 OpenVINO™ 工具套件优化为在线语音识别带来的提升

在非实时离线语音识别场景中，经过 OpenVINO™ 工具套件优化后，方案实时率由 0.0223 下降至 0.0161，下降了约 27.8%，优化后和优化前的性能对比如图 2-4-14 所示：

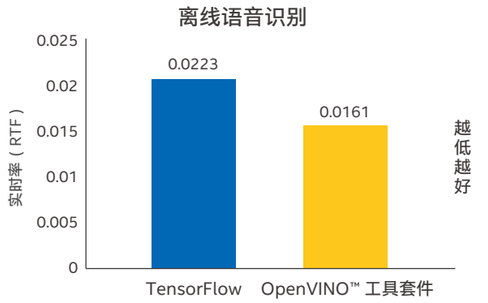



图 2-4-14 OpenVINO™ 工具套件优化为离线语音识别带来的提升

而在语音识别的重要后处理技术，智能标点符号预测场景中，经过 OpenVINO™ 工具套件优化后，预测时延由 57.74 毫秒下降至 12.52 毫秒，下降了约 78.3%，同时这一数值也已远低于流畅人机交互所需时长，让用户切实感受到了流畅度极强的标点符号标注。优化后和优化前的性能对比如图 2-4-15 所示：

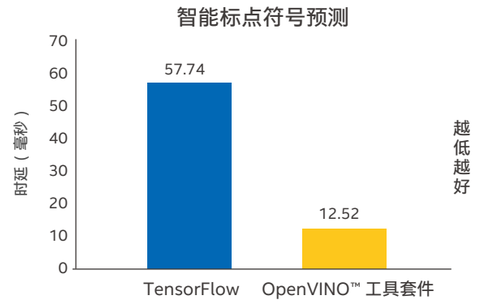



图 2-4-15 OpenVINO™ 工具套件优化为智能标点符号预测带来的提升

## 小结

诸如智能会议、智能笔记等基于语音识别技术的各类教学辅助能力，在引入教育机构办公管理以及日常教学过程后，一方面能帮助教育机构利用语音识别技术将会议记录等音频信息转化为文本，不仅能够方便信息的管理与留存，也有助于管理者和教师对教学状况、教学方式等进行复盘和整理，开展管理模式的革新，提升整体效率。另一方面，基于语音识别的诸多智慧课堂应用也能有效帮助学生更专注于教学过程中的听讲与互动，避免因漏听、误听带来的知识点混淆。

包括基于英特尔® 架构的处理器平台、英特尔® 发行版 OpenVINO™ 工具套件等在内的一系列先进产品和技术，正通过在边缘侧的部署，为基于语音识别的教学辅助人工智能应用与方案提供高效算力支持和推理性能优化，使面向教育场景的教学辅助人工智能应用方案在实战中获得更多成功。

[21] 三项测试均采用测试配置均为：处理器：双路英特尔® 至强® 金牌 6240R 处理器 @2.4GHz；内存：192GB（12 * 16GB DDR4 2933）；操作系统：CentOS Linux 7.9（Core）；操作系统内核版本：3.10.0-1160.31.1.el7.x86_64；Python 版本：v3.6；TenforFlow 版本：v1.5.1；OpenVINO 版本：v2020.4；

# 技术篇

# 第二代英特尔® 至强® 可扩展处理器

随着数字世界的不断演进，包括人工智能在内的新兴应用正使从云端、数据中心到边缘的各类数据节点面临更复杂、且更多样化的工作负载。因此这些数据节点需要配备更具性能优势，且能根据各类计算处理需求快速应变的处理器平台。

一直在数据中心担纲主力的第二代英特尔® 至强® 可扩展处理器具备多项创新特性，在提升灵活性与安全性的同时，也通过增强的内存性能，内置的 AI 加速能力以及更低 TCO，帮助用户在不同工作负载中都能获得更优的计算处理效率，打造性能更强的敏捷服务和更具价值的功能，为教育等场景中的人工智能应用提供了更具数据洞察的生产力平台。

## 性能驱动突破性功能输出

与上一代产品相比，第二代英特尔® 至强® 可扩展处理器着力为用户提供性能更强、服务更敏捷的算力平台，来应对用户在不同工作负载上的性能需求。除了由更高每核性能、更大内存带宽/容量、扩展的 I/O 以及英特尔® UPI 等技术带来的基础性能大幅增强外，如表4-1所示，为满足人工智能、云服务等不同工作负载的需求，第二代英特尔® 至强® 可扩展处理器也分别给出了突破性的功能输出：

| | 目标 | 突破性功能 |
|---|---|---|
| 人工智能 | • 提供基于 CPU 的强劲算力平台 | • 集成全新的英特尔® 深度学习加速技术，加速人工智能推理效率 |
| 云服务 | • 通过兼容的虚拟化基础设施显著降低复杂性 | • 与其他英特尔® 架构产品协同，构建可快速部署英特尔® 虚拟机，提升云服务扩展能力 |
| 高性能计算 | • 大幅提升矢量浮点性能和效率 | • 以更少的服务器实现高性能计算能力输出 |
| 存储 | • 确保确定性的存储响应 | • 得益于核心、缓存、内存、IO 等基础性能的提升，提供更佳响应确定性 |

表4-1 第二代英特尔® 至强® 可扩展处理器
为各种工作负载提供突破性功能

这其中，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（矢量神经网络指令VNNI）及英特尔® AVX-512，可通过对人工智能工作负载的优化，和对深度学习推理的有效加速，赋予平台更强的人工智能性能表现。基于这一架构，大多数深度学习推理工作能够被集成在基于英特尔® 架构的处理器平台构建的工作负载或应用程序中，这帮助用户的人工智能应用可在多云环境与智能边缘之间高效地进行无障碍性能切换。

## 实现存储突破性创新

第二代英特尔® 至强® 可扩展处理器的另一项功能突破是对英特尔® 傲腾™ 持久内存等存储产品提供了良好支持。其中，英特尔® 傲腾™ 持久内存作为新型内存和存储创新产品，能为人工智能等场景中的工作负载提供高密度的内存容量，这帮助用户在一些深度学习任务中，能够启动更多的任务进程，加速工作负载处理和服务交付。而这一内存产品所具有的非易失性特性，也能有效提升用户系统的可用性和稳定性。

同时，第二代英特尔® 至强® 可扩展处理器也对英特尔® 傲腾™ 固态盘和英特尔® QLC 3D NAND 固态盘有着良好的支持，能帮助用户系统将出色的高吞吐量、低延迟、高服务质量（QoS）和超强耐用性集于一体，突破数据访问瓶颈。

为用户所广泛选择的第二代英特尔® 至强® 可扩展处理器，包括英特尔® 至强® 金牌处理器 6200 系列、英特尔® 至强® 铂金处理器 8200 系列等。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和更多的内存容量，在性能、可靠性和硬件增强型安全技术方面具有优势，且针对要求苛刻的人工智能、多云环境等工作负载进行了优化。其中，英特尔® 至强® 金牌 6240R 处理器、英特尔® 至强® 金牌 6248R 处理器等作为更新版本，在保持成本优势的同时，为用户提供了更多的内核数量、缓存以及更大的睿频，帮助用户适应更复杂、更多样化的应用场景。



# 第三代英特尔® 至强® 可扩展处理器


得益于英特尔数十年来针对不同工作负载需求进行的创新，并与全球软件领导者和解决方案提供商进行紧密合作与深度集成，英特尔对面向四路和八路的第三代英特尔® 至强® 可扩展处理器（Cooper Lake）和面向单路和双路的第三代英特尔® 至强® 可扩展处理器（Ice Lake）在多样化的工作负载类型和性能需求方面进行了优化，并通过平衡的架构以及多种集成加速和先进的安全功能，帮助用户将迫切的工作负载安全地放置在从边缘到云的最佳性能位置上。

## 基础性能大幅提升：

- 第三代英特尔® 至强® 可扩展处理器基于平衡、高效的架构而构建，该架构可提升内核性能、内存和 I/O 带宽，为处理从数据中心到边缘的各种工作负载提速。在单路和双路配置中，支持每处理器多达 40 个内核，在四路和八路配置中则支持每处理器达 28 个内核，在八路配置下，单平台支持多达 224 个内核。
- 单个处理器支持8 条 DDR4 内存通道（Cooper Lake）或 6 条 DDR4 内存通道（Cooper Lake），最高速率为 3200 MT/s。同时每路多达 64 条 PCI Express Gen4 通道，实现更高的每核 I/O带宽；
- 多达六条英特尔® 超级通道互联（英特尔® UPI）通道有效提高了平台可扩展性以及I/O 密集型工作负载的 CPU 间带宽，从而在提高吞吐量和能效之间达成平衡；
- 对全新英特尔® 傲腾™ 持久内存 200 系列提供支持，在双路平台上，每路最高可支持 4TB 的英特尔® 傲腾™ 持久内存容量，以大规模内存池加速不同工作负载中数据获取与洞察能力。

## 增强的AI加速与安全能力

- 第三代英特尔® 至强® 可扩展处理器加入了增强版英特尔® 深度学习加速技术，同时支持 16 位 Brain Floating Point（BF16）和矢量神经网络指令（VNNI），有效加速人工智能推理和训练性能。其中BF16适用于特定型号的第三代英特尔® 至强® 可扩展处理器，其在视觉、自然语言处理（NLP）和强化学习（RL）等需要兼顾吞吐量和准确率的AI应用场景可以提供更有效的训练与推理加速能力。而矢量神经网络指令（VNNI）能够充分提高计算资源和缓存的利用率、减少潜在的带宽瓶颈，以此增强推理工作负载。
- 单路和双路配置的第三代英特尔® 至强® 可扩展处理器对英特尔® SGX 提供支持，帮助用户无论是从边缘到数据中心还是到多租户公有云，都可以在确保数据和应用代码安全的前提下，采用联邦学习等方法，以多源数据加强 AI 应用的应用效能。

## 自定义性能助推各种工作负载：

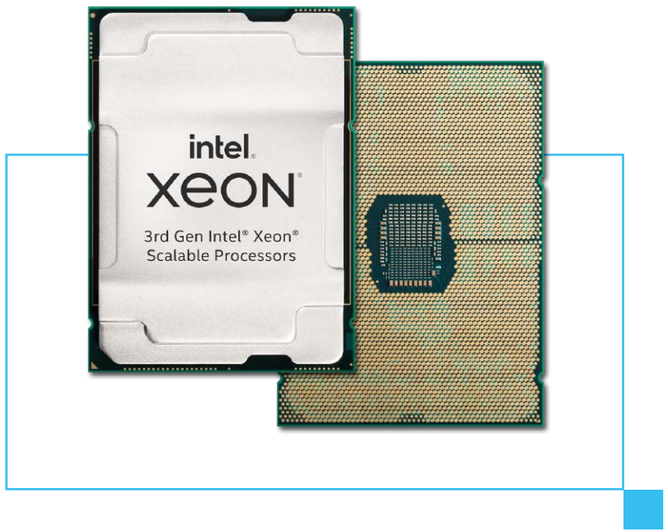
- 英特尔® 高级矢量扩展 512（英特尔® AVX-512) 可为数据分析、机器学习、可视化等应用中高密度的计算任务提高性能和吞吐量。在第三代英特尔® 至强® 可扩展处理器的铂金、金牌和银牌处理器中，融合乘加（FMA）单元增加了 2 倍。与上一代英特尔® AVX 2 技术相比，英特尔® AVX-512 带来

了更为出色的性能。

- 第三代英特尔® 至强® 可扩展处理器增强了英特尔® SST（英特尔® Speed Select）功能，其可以对处理器性能实施精细控制，有助于优化TCO。大部分第三代英特尔®至强®铂金和金牌处理器都支持英特尔® SST– BF、英特尔® SST–CP和英特尔® SST–TF等不同模式，而第三代英特尔® 至强® 可扩展处理器 Y SKU 支持新的英特尔® SST-PP模式，可以为用户提供更多内核、频率、外形尺寸和功率配置选择。

## 适用于不同工作负载的第三代英特尔® 至强® 可扩展处理器:

- 英特尔® 至强® 铂金 8300 处理器是打造可靠、敏捷的混合云数据中心的基石。处理器具备增强型硬件安全功能以及出色的多路处理性能，适用于关键业务的实时分析、机器学习、人工智能、高性能计算和多云工作负载。借助可靠的硬件增强型数据服务交付，该处理器在 I/O、内存、存储和网络技术方面实现了诸多改进，因此能有效利用从这个日益由数据推动的世界获得的可行洞察。英特尔® 至强® 铂金 8300 处理器专为高级分析、人工智能以及高密度的基础设施而设计，提供出色的性能、平台功能和工作负载加速。
- 英特尔® 至强® 金牌 6300 和 5300 处理器支持更高的内存速度、更大的内存容量以及多达四路的可扩展性，带来更出色的性能和内存功能、硬件增强型安全和工作负载加速。这些处理器已针对要求苛刻的主流数据中心、多云计算、网络和存储工作负载进行了优化。其高达四路的可扩展性，非常适合范围更广的工作负载。
- 英特尔® 至强® 银牌 4300 处理器提供基本性能、更快的内存速度以及更高的能效，为入门级数据中心计算、网络和存储带来所需的硬件增强性能。




# 英特尔® 傲腾™ 持久内存100系列、200系列

英特尔® 傲腾™ 持久内存采用创新的内存技术，将高性价比的大容量内存与对数据持久性的支持巧妙结合，具备高耐用性、一致性和低时延等高性能特性。已推出100系列、200系列产品，且均具备128 GB、256 GB和512 GB模组，是从云到数据库，再到内存分析、虚拟化基础设施、内容分发网络等数据密集型和计算密集型工作负载所需大规模持久内存的理想选择。

## 多项优势:

- 将更多数据保存在更靠近处理器的位置，帮助加速大内存计算，在不同带宽压力下，延时均可小于1微妙。
- 可按字节访问，精确找到读和写的数据位置，消除读写放大问题
- 5 年全盘写质保可以达到 410 PBW 的写入寿命。
- 将数据长久保存在内存，无需从存储设备重新加载，可加快重启时间并减少 I/O。
- 降低大内存节点的功耗。

英特尔® 傲腾™ 持久内存可通过灵活配置，为用户提供内存模式和 App Direct 模式2种不同模式以应对不同场景需求:

- **内存模式：**无需更改应用即可提供大内存容量，且性能接近DRAM（具体视工作负载而定）。
- **App Direct 模式：**能够实现大内存容量和数据持久性，且在该模式下，傲腾™ 持久内存支持行业标准持久内存编程模型的软件和应用直接与持久内存通信，降低了堆栈的复杂性。

作为具有突破意义的内存技术创新产品，英特尔® 傲腾™ 持久内存通过与第三代英特尔® 至强® 可扩展处理器组合，可创建两层内存和存储分层架构，打造出色的低延迟、高带宽、服务质量（QoS）和耐用性，尤其是新推出的 200 系列较 100系列带宽平均提升 32%，且总内存每路高达 6 TB，可进一步优化工作负载的性能与成本[22]。

## 出色性能:

- 200 系列通过扩大内存容量，较 100 系列可使得数据库性能提升达2.5倍;
- 200 系列较 100 系列图形分析计算速度提升高达2倍;
- 在保持原有性能的前提下，200 系列较 100系列可使每台虚拟机的成本降低 25%;
- 200 系列较 100 系列通过提升内容分发网络吞吐量，使视频直播分辨率提高 63%;
- 200 系列较 100系列功耗降低幅达 16 %。

基于优良特性和高性能，以及合作伙伴、OEM、OSV、CSP 和 ISV 等庞大生态系统的支持，英特尔® 傲腾™ 持久内存能够帮助应对当今数据中心面临的大内存节点 DRAM 成本高、断电和维护期间的数据保护，以及利用分层架构支持新兴工作负载等挑战，进而创造新的计算可能。

[22] https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/achieve-insight-data-intel-optane.html

# 英特尔® 傲腾™ 固态盘 P5800X/P5801X

英特尔® 傲腾™ 固态盘 P5800X 系列以填补关键的存储性能缺口为突破点，是热数据快速缓存或分层的理想选择，能够提供数据密集型工作负载所需的性能。

作为高性能数据中心级固态盘，英特尔® 傲腾™ 固态盘 P5800X 系列集出色的低时延、高质量（QoS）、快速吞吐和高耐用性于一身，尤其同时支持读取和写入而不会降低性能的特色。此外，P5800X 系列固态盘产品符合 PCIe 4.0 规范，在低队列深度下可实现高达 160 万次 IOPS 的随机读取/写入性能，能够帮助企业和云服务商更快访问大型复杂的数据集，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，并改善数据中心的整体 TCO，与上一代英特尔® 傲腾™ 固态盘 DC P4800X 相比，性能大幅提升 [23]：

- 与上一代相比，4 KB 随机读取平均时延降低 40%。
- 与上一代相比，稳定性提升达 50%，且服务质量水平达 99.999%。
- 与上一代相比，耐用性提高 67%。
- 能够在 70/30 混合负载下，实现超越技术参数的性能（IOPS高达200万次）。
- 具备专为元数据用例设计的 512 B（小于 4 K 的数据块）读取能力，在混合工作负载下 IOPS 能够达到 500 万次。

英特尔® 傲腾™ 固态盘 P5800X 系列为混合工作负载环境而设计，克服了传统存储产品处理日益密集的工作负载时的性能短板，为构建超融合基础设施、人工智能、高性能计算，以及高可靠性数据库等提供了高效而稳健的存储架构。


[23] https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/welcome-storage-media-revolution.html

# 英特尔® Movidius™ 视觉处理器 (VPU)

英特尔® Movidius™ Myriad™ X 视觉处理器是英特尔第一个包含神经计算引擎（用于深度神经网络推理的专用硬件加速器）的视觉处理器，旨在显著提高深度神经网络的性能，且不会影响其低功耗特性。

作为专用的硬件加速器，英特尔® Movidius™ 视觉处理器基于将数据移动最小化的独特架构，通过全新的神经计算引擎，强大的 SHAVE 内核和高吞吐量智能内存结构，将高度并行的可编程计算与面向特定工作负载的硬件加速相结合，在电源效率和计算性能之间实现了平衡，能够高效运行要求苛刻的计算机视觉和边缘人工智能工作负载，成为在设备端运行深度神经网络和计算机视觉应用程序的理想选择。

通过以下方面的优势加成，英特尔® Movidius™ Myriad™ X 视觉处理器可在视觉零售、安全、工业自动化等多领域，为智能摄像头、边缘服务器和人工智能等应用提供支持，帮助用户有效应对成本和功耗限制，提升各领域边缘AI能力。

## 超低功耗，卓越性能

英特尔® Movidius™ Myriad™ X 视觉处理器为计算机视觉和深度神经网络推理应用提供出色性能。作为以超低功耗著称的 Movidius 视觉处理器家族的一员，英特尔® Movidius™ X 视觉处理器正以一系列全新性能增强技术来为高能效解决方案提供支持，可将先进的视觉和人工智能应用引入无人机、智能相机、智能家居、安全、VR/AR 可穿戴设备和 360 相机等设备。

## 新一代深度神经网络性能

英特尔在英特尔® Movidius™ Myriad™ X 视觉处理器架构中引入了全新深度神经网络处理单元：神经计算引擎。神经计算引擎经过专门设计，可以高速、低功耗地运行深度神经网络，使英特尔® Movidius™ Myriad™ X 视觉处理器能够在神经网络推理上实现 716G FLOPS 的计算性能。神经计算引擎被集成作为高能效的 Movidius 视觉处理器架构的一部分，通过减少片上数据移动将功耗降至最低。

## 可定制的成像和视觉管道

Movidius VPU 家族始终提供独特、灵活的图像处理、计算机视觉和深度神经网络架构。该架构提供了一种用于配置成像和视觉工作负载的模块化方法，因为它结合了一组成像和视觉硬件加速器（例如立体深度或神经计算引擎）以及一系列 C-可编程 VLIW 矢量处理器，所有这些处理器均可访问片上内存。除了交叉存取的计算机视觉和深度神经网络推理应用程序管道外，这种方法还可以实现出色的图像信号处理（ISP），而无需进行内存访问以实现最佳的电源效率，所有这些都源自一种通过使数据移动最小化来降低功耗的数据流方法 。Movidius VPU 可在低功耗情况下在可编程性和性能之间实现最佳平衡。

## 支持 8 个高清传感器和 4K 编码

英特尔® Movidius™ Myriad™ X 视觉处理器拥有16 条 MIPI Rx 通道，支持直连多达16 个Rx 通道和8个MIPI DPHY Tx 通道，支持1/2/4通道 MIPI sensor 接入。高吞吐量内联 ISP 可确保高速处理数据流，而新的硬件编码器则以 30 Hz（H.264/H.265）和 60 Hz（M/JPEG）帧率支持 4K 分辨率。

英特尔® Movidius™ Myriad™ X 视觉处理器接口包括：

- MIPI DPHY 1.2（2.5 Gbps per lane）16 Rx Lances;
- 8x Tx Lanes;
- USB 3.1;
- Ethernet 10G LLI;
- PCI-E Gen 3;
- 1 Lane 5x I2C;
- 4x UART;
- 3x SPI;
- 1x Quad-SPI;
- 4x I2S;
- PDM;
- 2x SD 3.0;
- eMMC 5.1

# OpenVINO™ 工具套件

为各类AI解决方案提供模型优化、推理加速提及异构平台部署方案的OpenVINO™ 工具套件，是由英特尔开源的，面向深度学习网络和计算机视觉领域推出的性能加速工具包。无论是在云端还是边缘环境，OpenVINO™ 工具套件都能以其丰富的功能组件，为各类AI应用，包括计算机视觉、语音识别、自然语言处理以及推荐系统等所需的高性能深度学习推理提供有效加速。

如图4-1所示，在最基本的工作流程中，OpenVINO™ 工具套件提供了模型优化器（Model Optimizer）和推理引擎（Inference Engine）两个核心组件。其中，模型优化器是一种跨平台命令行工具，它可将训练后的网络模型从其源框架转换为开源、且与nGraph 兼容的IR（中间表示）文件以用于推理操作，IR文件是由bin（经训练的数据文件）和xml（描述网络拓扑的文件）两种格式文件组成。

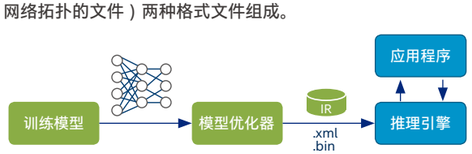



图4-1 OpenVINO™ 工具套件基本工作流程

一方面，模型优化器可以在导入由Caffe、TensorFlow、MXNet、PyTorch、Karas 以及 ONNX 等流行框架训练好的模型后，执行相应优化，包括去除多余的层，并在可能的情况下将操作分组为更简单、更快的图（Graph）等。

另一方面，OpenVINO™ 工具套件也可以进行模型量化过程。PyTorch等流行框架中训练的模型通常为FP32精度数据格式，而英特尔® 至强® 可扩展处理器等计算平台已经支持INT8等低精度数据格式下的模型推理，其可在损失很小精度的前提下实现更高的推理效率。量化过程是将基于高精度数据格式的模型（由IR文件表示）转为低精度，并加入校准过程以保证精度不受损失。

推理引擎是用于管理和优化深度学习模型的加载和编译，对输入数据运行推理操作并输出结果。推理引擎可以同步或异步执行。如下图所示，利用统一的API，推理引擎能让深度学习模型在不同的硬件平台上进行高性能推理过程，包括基于英特尔® 架构的处理器、英特尔® 集成显卡、英特尔® 神经计算棒二代、采用英特尔® Movidius™ 视觉处理单元（VPU）的英特尔® 视觉加速器设计等。同时也可基于Windows、Linux、macOS甚至树莓派等平台部署。实现用户一次开发，即可面向不同平台部署并获得一致的性能表现，有效提升AI开发与部署效率。

此外，开发者还可借助OpenVINO™ 工具套件中Open Model Zoo 组件所提供的强大AI应用模板，来快速实现特定的深度学习场景。其提供的40多个经过预先训练，且优化过的模型和代码示例（涵盖物体检测、物体识别、行为检测、图像处理等一系列常见且实用的AI模型），能帮助开发者有效加速AI应用过程。同时，工具套件中的 DL Workbench 组件能让开发者在基于英特尔® 架构配置，以可视化方式调用各类组件去实施算法调优、模型校准等工作。

而在面向计算机视觉的AI方案中，OpenVINO™ 工具套件也已对OpenCV、OpenCL™ 等标准计算机视觉库开展了优化，并集成英特尔® Media SDK（仅在面向 Linux 的英特尔® 分发版 OpenVINO™ 工具包）来提供媒体性能增强。

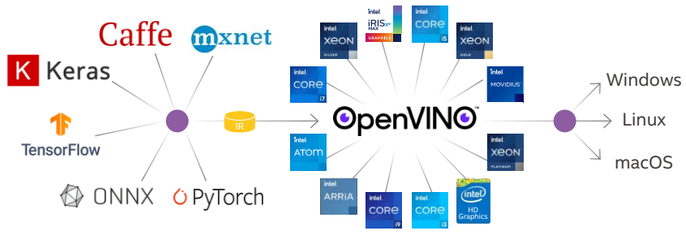



图4-2 OpenVINO™ 工具套件为深度学习提供多平台推理、部署能力

# 面向英特尔® 架构优化的 TensorFlow

基于英特尔® 架构优化的 TensorFlow，是英特尔为了在处理器上更高效运行深度学习模型而推出的优化版，为了显著提升性能，英特尔持续采用多种措施对 TensorFlow 进行优化，包括：

- **计算图优化：**英特尔推出了大量计算图优化通道，以便在处理器上运行时，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，确保在处理器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 开发者带来任何额外负担。

- **低精度优化：**配合全新第三代英特尔® 至强® 可扩展处理器，TensorFlow 开发者可以通过低精度数据类型 Bfloat16 为训练与推理模型加速，获得至多2倍的性能提升且保持模型精度不变。基于英特尔® 架构优化的 TensorFlow 提供了包括手动、自动等多种 Bfloat16 模型转换解决方案，用户可以根据实际情况选择不同的方案来满足对于性能或使用体验的要求。

更多详情可以参考：https://blog.tensorflow.org/2020/06/accelerating-ai-performance-on-3rd-gen-processors-with-tensorflow-bfloat16.html

- **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），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，并通过正确配置线程池设置，避免了互相争夺处理器，提升了资源的综合利用率。

# 面向英特尔® 架构优化的 PyTorch 扩展包

面向英特尔® 架构优化的PyTorch扩展包（Intel® Extensions for PyTorch，IPEX）在原生PyTorch的基础上，添加面向深度学习优化的扩展能力yin，同时让用户在英特尔® 架构上使用PyTorch 时，获得“开箱即用”式的优化体验。面向英特尔® 架构优化的 PyTorch 扩展包会根据基于英特尔® 架构的硬件特性的变化而不断升级，随着全新第三代英特尔® 至强® 可扩展处理器的全面铺开，新版扩展包也针对新处理器平台的各项特性进行了专门优化。

以IPEX v1.2.0为例，其安装首先需要获取 PyTorch v1.7.0 源码：

```
git clone --recursive https://github.com/pytorch/pytorch
cd pytorch # checkout 源代码到指定版本
git checkout v1.7.0 #更新指定 PyTorch 版本的子模块
git submodule sync
git submodule update --init --recursive
```

获取 IPEX v1.2.0 的源代码：

```
git clone --recursive https://github.com/intel/intel-extension-for-pytorch
cd intel-extension-for-pytorch
git submodule sync
git submodule update --init --recursive
```

在 PyTorch 中增加对IPEX v1.2.0 的支持

```
#将 git 补丁应用到 pytorch 代码
cd ${pytorch_directory}
git apply ${intel_extension_for_pytorch_directory} /torch_patches /xpu-1.7.patch
```

编译和安装 PyTorch

```
cd  ${pytorch_directory}
python setup.py install
```

在安装完毕后，用户只需要将模型（model）和输入张量（tensor）转换到扩展设备（‘xpu’），然后IPEX将自动启用。可以参考以下代码用例：

```
import torch
import torch.nn as nn

# Import Extension
import intel_pytorch_extension as ipex

class Model(nn.Module):
    def __init__(self):
        super(Model, self).__init__()
        self.linear = nn.Linear(4, 5)

    def forward(self, input):
        return self.linear(input)

# Convert the input tensor to the Extension device(‘xpu’)
input = torch.randn(2, 4).to(xpu)
# Convert the model to the Extension device
model = Model().to(ipex.DEVICE)

res = model(input)
```

此外，面向基于英特尔® 架构优化的 PyTorch 扩展包也能够对混合精度予以支持。这意味着模型中的某些运算可以运行在 Float32 数据格式下，而其他一些运算可以使用 BFloat16 或 INT8 数据格式。传统上，如果要运行低精度类型的模型，则需要手动将参数和输入张量转换为低精度类型。如果模型包含一些不支持低精度格式的运算或者低精度运算会带来较大精度损失，则需要将该运算转换回 Float32。而 IPEX 可以有效帮助用户简化这一过程，

# 面向英特尔® 架构优化的 Python

面向英特尔® 架构优化的 Python 是一个由英特尔主导开发、功能强大的 Python 分发包，提供了编写 Python 原生扩展所需的一切：编译器、语言以及支持模块，并且集成了多个高性能数据分析和数学库（包括 Numba, CPython, NumPy、SciPy、scikit-learn、pandas、Jupyter、matplotlib、mpi4py 等）。

面向英特尔® 架构优化的 Python 是英特尔® oneAPI AI Analytics Toolkit 的重要工具集之一，具备多项先进特性，且有着可与 C 媲美的高效率：

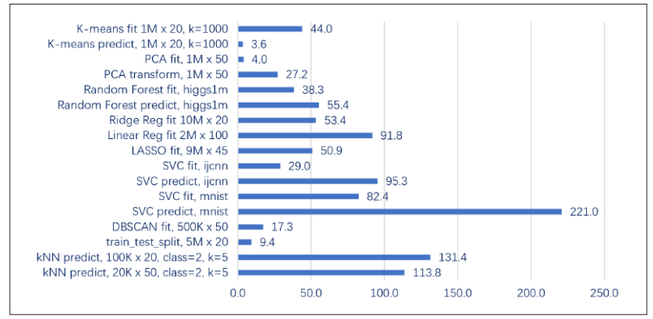
- 通过提供开箱即用的 uMath、NumPy、SciPy 和 scikit-learn 等工具，加速包括数字、科学、数据分析、机器学习等在内的计算密集型应用。
- 集成英特尔® 高性能库（如 Intel® oneMKL），内置最新的矢量化指令，并应用对多线程构建块库 Intel® oneTBB 的可组合并行，解锁 Python 基于多核处理器的并行应用功能，进而在英特尔® 平台上提升 Python 程序运行性能，且不需要对代码进行任何更改。
- 支持最新一代处理器，提供英特尔® 优化的数据分析加速库（Intel® oneDAL）用于加速数据处理、深度学习和机器学习的速度，如支持向量机（SVM）、K-Means、随机森林等，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。

部署简单、易于使用也是面向英特尔® 架构优化的 Python 的一大特性，且其可通过 pip、Docker images、YUM 和 APT repos 等各种渠道获得，2021 年新版本又提供了多项新特性和新功能：

- 数据并行 Python（DPPy），这是 Python 开发在跨平台程序上所必需的一个工具组包。并且包括帮助协调主机和设备间的数据并行执行控制和管理的工具 dpCtl。
- 优化的 XGBoost、Intel Extension for Scikit-learn 和高级数据处理方法 Modin（包括使用的多个设备），实现更快的机器学习。
- 用于图像扭曲、图像过滤和形态学操作的 Scikit ipp，包括对使用 OpenMP 的过滤器的转换函数多线程和部分多线程的支持。
- 更新了 NumPy 版本；用于 SYCL 设备上阵列计算的新 dpnp 包。
- 改进 Numba 编译器，加速针对 CPU 和 GPU 执行的自定义 Python 代码。

了解更多信息，请访问：
https://software.intel.com/content/www/us/en/develop/tools/oneapi/components/distribution-for-python.html

硬件配置：双路英特尔® 至强® 铂金 8276L 处理器 @ 2.20GHz，每路 28 个核心。更多详情请参阅：
https://medium.com/intel-analytics-software/accelerate-your-scikit-learn-applications-a06cacf44912




由英特尔® oneDAL 技术加强的 Scikit-Learn，相比开源 Scikit-Learnsk 带来的性能提升

## 本手册涉及的专业词汇表

| 英文全称 | 英文缩写 | 中文全称 |
|---|---|---|
| Application Programming Interface | API | 应用程序编程接口 |
| Artificial Intelligence | AI | 人工智能 |
| Augmented Reality | AR | 增强现实 |
| Automatic Speech Recognition | ASR | 自动语音识别 |
| Cepstral Mean and Variance Normalization | CVMN | 均值方差归一化 |
| Computer Assisted Language Learning | CALL | 计算机辅助语言学习 |
| Computer Version | CV | 计算机视觉 |
| Convolutional Neural Networks, | CNN | 卷积神经网络 |
| Deep Learning | DL | 深度学习 |
| Deep Natural Networks | DNN | 深度神经网络 |
| Field-Programmable Gate Array | FPGA | 现场可编程门阵列 |
| Gaussian Mixed Model | GMM | 高斯混合模型 |
| Goodness of Pronunciation | GOP | |
| Hidden Markov Model | HMM | 隐马尔可夫模型 |
| Intel C++ Compiler | ICC | 英特尔 C++ 编译器 |
| Intel® Advanced Vector Extensions 512 | Intel® AVX -512 | 英特尔 ® 高级矢量扩展 512 |
| Intel® oneAPI Math Kernel Library | Intel® oneMKL | 英特尔 ® 数学核心函数库 |
| Intermediate Representation | IR | 中间表示 |
| Knowledge Distillation | | 知识蒸馏 |
| Logistic Regression | LR | 逻辑回归 |
| Long Short-Term Memory | LSTM | 长短期记忆 |
| Machine Learning | ML | 机器学习 |
| Mel-Frequency Cepstral Coefficients | MFCC | 梅尔频率倒谱系数 |
| Microcontroler Units | MCU | 微控制单元 |
| Mobile Edge Computing | MEC | 移动边缘计算 |
| Multi-Task Learning | | 多任务学习 |
| Natural Language Processing | NLP | 自然语言处理 |
| Open Pluggable Specification | OPS | 开放式可插拔规范 |
| Open Pluggable Specification-China | OPS-C | 中国开放式可插拔规范 |
| Pearson Correlation Coefficient | PCC | 皮尔森相关系数 |

| 英文全称 | 英文缩写 | 中文全称 |
|---|---|---|
| Real-time Factor | RTF | 实时率 |
| Recurrent Neural Network | RNN | 循环神经网络 |
| Reinforcement Learning | RL | 强化学习 |
| Residual Net | ResNet | 残差网络 |
| Single Shot MultiBox Detector | SSD | |
| Support Vector Machine | SVM | 支持向量机 |
| Text to Speech | TTS | 语音合成 |
| Total Cost of Ownership | TCO | 总拥有成本 |
| Transfer Learning | | 迁移学习 |
| Vector Neural Network Instructions | VNNI | 矢量神经网络 |
| Very Deep Convolutional Neural Network | VDCNN | |
| Virtual Reality | VR | 虚拟现实 |
| Visual Processing Unit | VPU | 视觉处理器 |
| Voice Activity Detection | VAD | 语音活动检测 |
| Weighted Finite-State Transducers | WFST | 加权有限状态转化器 |
| Word Error Rate | WER | 字错误率 |
| You Only Look Once | YOLO | |

法律声明:

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex。

性能测试结果基于配置信息中显示的日期进行测试，且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。

在此提供的信息可随时更改，恕不另行通知。英特尔可以随时在不发通知的情况下修改产品生命周期、规格和产品说明。以上信息是按“原样”提供，英特尔对该信息的准确性、产品的特性、可用性、功能或列出产品的兼容性不做任何形式的声明或担保。请联系系统厂商，了解关于上述特定产品或系统的更多信息。

描述的产品可能包含可能导致产品与公布的技术规格有所偏差的、被称为非重要错误的设计瑕疵或错误。一经要求，我们将提供当前描述的非重要错误。

本文件不授予任何关于知识产权的许可（包括通过明示、暗示、以禁反言方式或其他方式的许可）。唯一的例外是，本文件中包含的代码是通过零条款 BSD 开源许可（0BSD）进行授权的，https://opensource.org/licenses/0BSD。