解决方案简介

英特尔® 至强® 可扩展处理器
英特尔® 高级矩阵扩展

东软医保 OCR 票据识别解决方案 采用第四代英特尔® 至强® 可扩展处理器 加速 AI 推理

Neusoft

“医疗票据录入是医保机构在办理费用报销业务时经常面临的场景之一，通过应用智能 OCR 产品，东软能够将手动录入流程转化为自动流程，从而提高医保报销效率，实现医保业务的智能化精细化管理。第四代英特尔® 至强® 可扩展处理器的应用让我们能够显著提升智能 OCR 的推理效率，帮助客户打造更加现代化的智能医保平台，助力客户践行服务型政府和数字化政府的建设目标。”

— 刘兵
东软集团医疗保障事业部总经理

挑战

为了支撑医保票据的高效识别与录入，东软希望能够构建高效的算力系统，化解智能 OCR 推理在推理性能、成本、精度等方面的挑战：

- **性能：** 为了提升效率，医保票据 OCR 识别通常需要在较短的时间内完成。由于票据识别规模较低，智能 OCR 推理服务器需要具备较高的推理性能；
- **成本：** 虽然独立 GPU 等专用的加速器具备较高的 AI 推理性能，但是通常有着较高的应用和开发成本。东软希望能够尽可能利用现有的 CPU 服务器资源，以帮助客户降低基础设施层面的支出；
- **精度：** 医保票据录入关系到病患的费用报销，信息识别的准确性至关重要。东软希望在提升性能的同时，使模型推理精度能够满足应用所需。

解决方案概述

医疗保障（医保）是为了补偿劳动者因疾病风险造成的经济损失而建立的一项社会保险制度，在医疗系统整体运行中扮演着重要的角色。在医保业务中，由于系统故障、异地就医等原因，部分场景下的医保业务难以实现联网结算，通常需要医保机构人工录入、审核单据，这一过程耗时耗力，影响医保业务的高效运转。

为了帮助医保机构提升纸质单据的处理效率，释放人力资源，同时降低人工录入存在的信息疏漏等风险，东软推出了医保光学字符识别（OCR）票据识别解决方案。该方案能够通过由人工智能（AI）赋能的 OCR 应用，将相当一部分的医保票据识别转为自动化的工作流程，可将流程处理的时间缩短三分之二[1]。为了解决智能 OCR 票据识别在算力资源、总体拥有成本（TCO）等方面的挑战，东软采用了基于第四代英特尔® 至强® 可扩展处理器的服务器作为基础算力设备，并通过 OpenVINO™ 工具套件进行优化，实现了高性能、高性价比的 AI 推理。

[1] 数据援引自东软内部测试结果，通过对比传统手工报销流程（30 分钟）和新模式下报销流程（10 分钟）计算得出。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。

东软医保 OCR 票据识别方案

医保单据识别是医保业务中的一个重要场景。在无法联网结算时，医院需要将所有相关的住院、用药、就诊信息打印为纸质单据，并将纸质单据提交给医保结算柜台。医保机构随后会录入这些纸质单据中的信息并进行处理。在传统模式上，这一流程需要通过手动录入，不仅耗时耗力，而且还可能因为人为疏忽导致错录、漏录等问题。

为了帮助医保部门提高医保结算效率、响应服务型政府号召，使医保经办人员摆脱重复性、事务性工作，实现精细化管理，东软提供了医保 OCR 票据识别方案。该方案能够通过纸质单据电子化、OCR 文字识别、人工辅助校改、目录智能比对等流程，最终形成符合业务系统报销要求的医保电子结构化数据，降低人工成本、优化医保经办工作流程，保障医保基金安全。

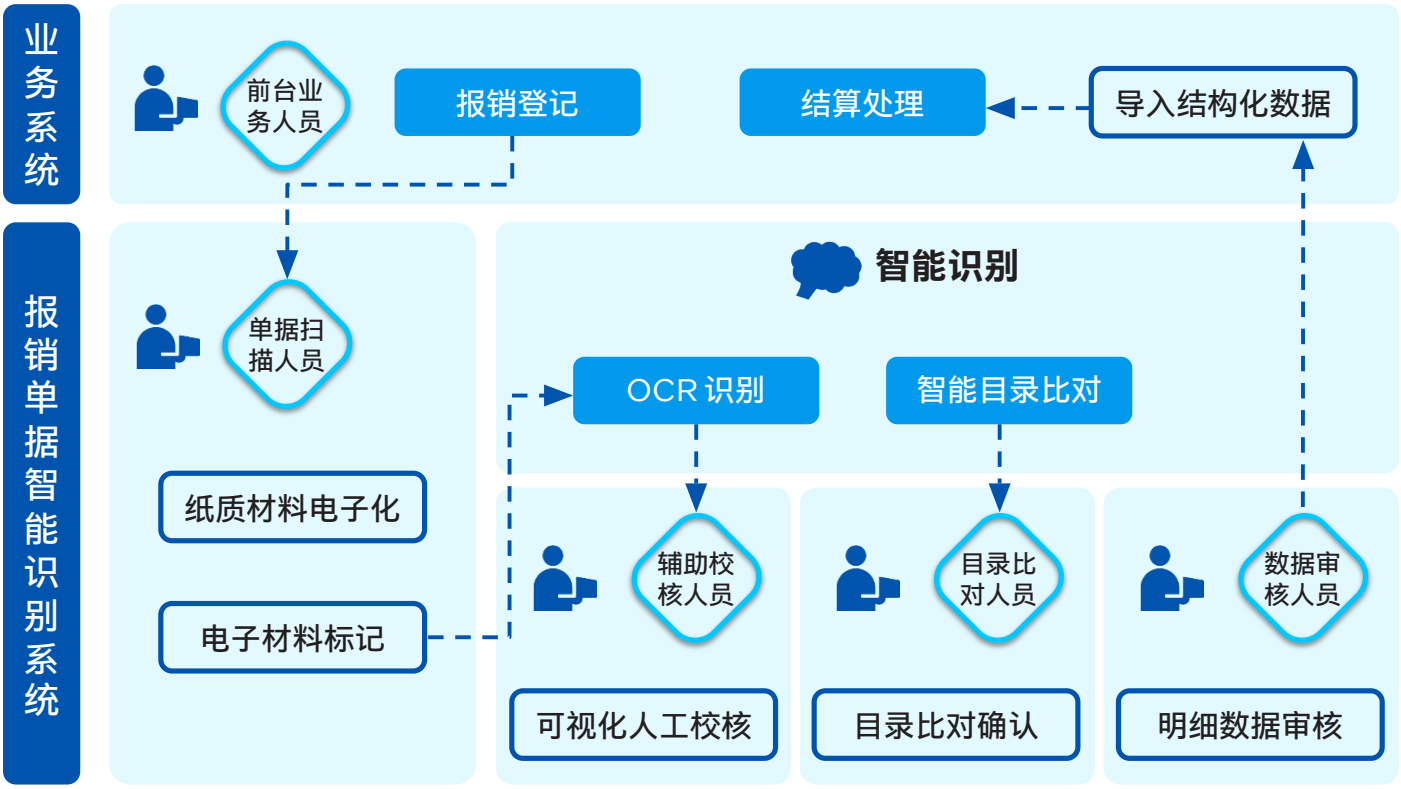

图 1. 东软医保 OCR 票据识别方案应用流程

智能 OCR 是该方案的关键技术，为识别不同医院打印出的处方、明细、项目名称、数量和单价等信息，东软自研智能 OCR 算法，能够准确地在复杂背景下，识别出不同医院出具的不同格式单据，实现了较高的识别准确率。该方案在通过 OCR 将纸质单据转换为电子数据后，还会对数据进行智能化的匹配，以便于后续的数据处理。

采用第四代英特尔® 至强® 可扩展处理器提升 OCR 推理性能

为了实现高性能、低成本的 OCR 推理，东软采用了基于第四代英特尔® 至强® 可扩展处理器的服务器，并进行了性能验证。

第四代英特尔® 至强® 可扩展处理器通过创新架构增加了每个时钟周期的指令，每个插槽多达 60 个核心，支持 8 通道 DDR5 内存，有效提升了内存带宽与速度，并通过 PCIe 5.0（80 个通道）实现了更高的 PCIe 带宽提升。第四代英特尔® 至强® 可扩展处理器提供了现代性能和安全性，可根据用户的业务需求进行扩展。借助内置的加速器，用户可以在 AI、分析、云和微服务、网络、数据库、存储等类型的工作负载中获得优化的性能。通过与强大的生态系统相结合，第四代英特尔® 至强® 可扩展处理器能够帮助用户构建更加高效、安全的基础设施。

第四代英特尔® 至强® 可扩展处理器在 AI 性能上更进一步。该处理器内置了创新的英特尔® AMX 加速引擎。英特尔® AMX 针对广泛的硬件和软件优化，通过提供矩阵类型的运算，显著增加了人工智能应用程序的每时钟指令数（IPC），可为 AI 工作负载中的训练和推理上提供显著的性能提升。

第四代英特尔® 至强® 可扩展处理器与 OpenVINO™ 工具套件的结合可以进一步提升 AI 推理性能。OpenVINO™ 工具套件支持从边缘到云的深度学习推理，可在包括英特尔 CPU、iGPU 和 FPGA 在内的英特尔硬件平台（包括加速器）上部署并加速神经网络模型，能够在保持精度的同时提高推理速度。OpenVINO™ 工具套件支持开发人员使用行业标准人工智能框架、标准或自定义层，将深度学习推理轻松集成到应用中。

东软医保 OCR 票据识别解决方案在智能 OCR 应用中采用了 OpenVINO™ 工具套件作为 AI 框架，并验证了 OCR 算法在第三代/第四代英特尔® 至强® 可扩展处理器上的代际性能对比，以及 OCR 算法在不同数据精度（FP32/INT8）下的性能对比。

首先，东软对比了第三代/第四代英特尔® 至强® 可扩展处理器的 OCR 模型推理性能。测试数据如图 2 所示，在数据精度同为 FP32 时，相比未采用矢量神经网络指令（VNNI）的第三代英特尔® 至强® 可扩展处理器，第四代英特尔® 至强® 可扩展处理器实现了 1.42 倍的性能提升[2]。

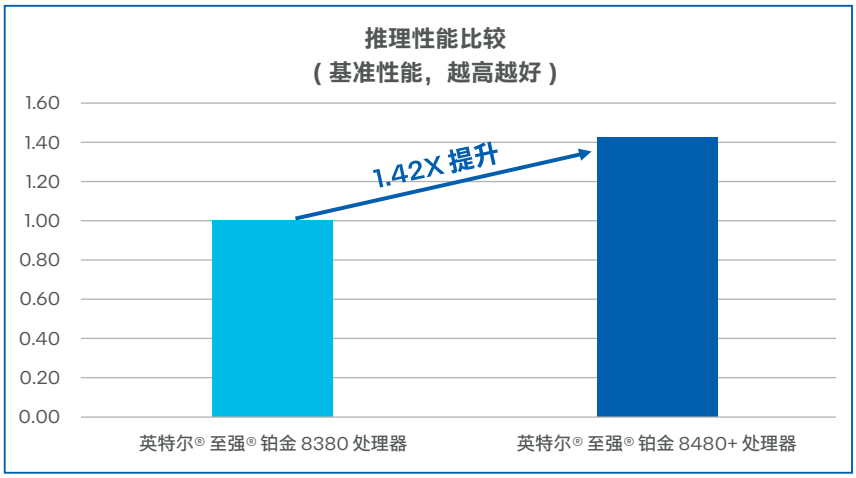

图 2. OCR 模型在第三代/第四代英特尔® 至强® 可扩展处理器上的推理性能对比[3]

随后，东软利用第四代英特尔® 至强® 可扩展处理器的英特尔® AMX 加速器，将模型转换为 INT8 数据精度。转化后的模型推理性能结果与采用 VNNI 的第三代英特尔® 至强® 可扩展处理器相比，实现了 2.29 倍的性能提升，与第四代英特尔® 至强® 可扩展处理器 + FP32 数据精度相比实现了 4.66 倍的性能提升[4]。

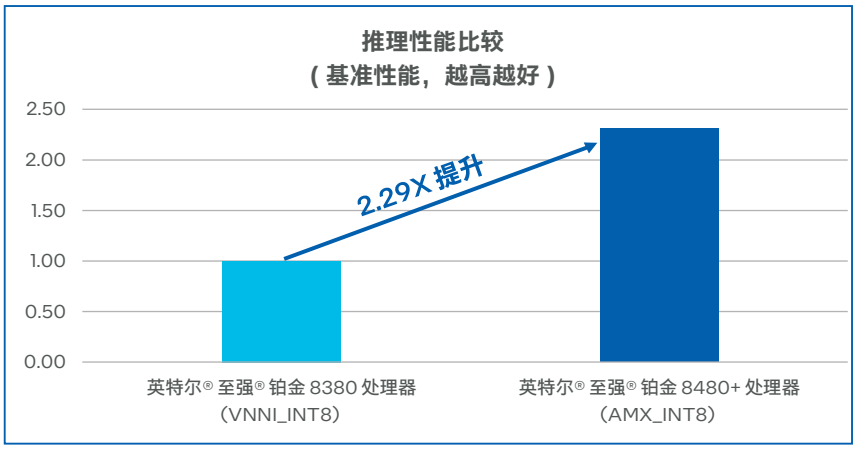

图 3. 第三代英特尔® 至强® 可扩展处理器 + VNNI_INT8 与第四代英特尔® 至强® 可扩展处理器 + AMX_INT8 性能对比[5]

[2,3,4,5] 截止 2022 年 8 月东软联合英特尔开展的测试。测试配置：基准配置/新配置 3 — 单节点，双路英特尔® 至强® 铂金 8380 处理器，40 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/3200 MHz），<SE5C620.86B.01.01.0005.2202160810>，<0xd000375>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>；新配置 1/2 — 单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/4800 MHz），<EGSDCRB1.SYS.0085.D15.2207241333>，<0x2b000070>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>。

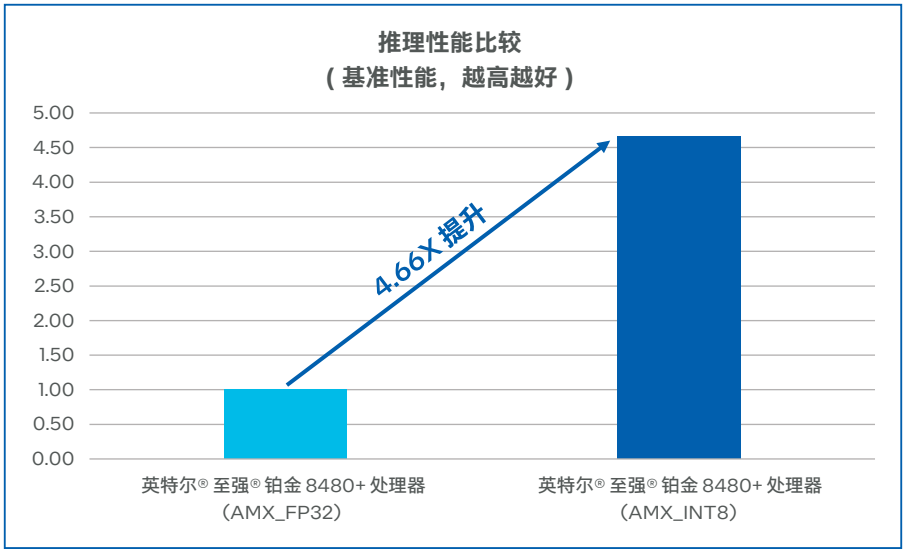

图 4. 不同数据精度在第四代英特尔® 至强® 可扩展处理器上的推理性能比较[6]

收益

东软医保 OCR 票据识别方案能够有效解决单据识别问题，将处理时间缩短为传统手动流程的三分之一[7]。除了效率上的提升之外，该方案还能为客户带来如下收益：

管理规范化

实现单据处理的事务性工作和专业性工作分离，明确责任，落实公平、公正原则；

业务智能化

AI+ 传统业务结合，OCR 识别准确度可达 95% 以上[8]，缩短业务办理周期；

档案电子化

档案业务一体化，减少纸质材料管理成本，提高复查、检索能力；

数据精细化

搭建医疗知识库，目录对照越用越准，提高审计精细化程度，降低医保基金潜在风险。

目前，东软医保 OCR 票据识别方案已经在多家医保部门得到成功落地。以某市医保局为例，自方案正式上线运行以来，日均处理档案袋 20 份，累计处理单据 492 张，积累单据明细比对数据超过 30W，医保定制化目录对照经验库数据累计过百万，有效深化了医保业务的智能化水平[9]。

[6] 截止 2022 年 8 月东软联合英特尔开展的测试。测试配置：基准配置/新配置 3 — 单节点，双路英特尔® 至强® 铂金 8380 处理器，40 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/3200 MHz），<SE5C620.86B.01.01.0005.2202160810>，<0xd000375>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>；新配置 1/2 — 单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/4800 MHz），<EGSDCRB1.SYS.0085.D15.2207241333>，<0x2b000070>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>。

[7] 数据援引自东软内部测试结果，通过对比传统手工报销流程（30 分钟）和新模式下报销流程（10 分钟）计算得出。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。

[8] 数据援引自东软内部测试结果。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。

[9] 数据援引自东软提供的信息。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。

展望

通过使用第四代英特尔® 至强® 可扩展处理器与 OpenVINO™ 工具套件，东软医保 OCR 票据识别方案可帮助用户在显著节约 IT 基础设施投入的前提下，充分挖掘硬件潜力，提升 OCR 识别性能。通过英特尔与东软集团提供的软硬件组件，该方案能够确保可靠的性能、成本效益和扩展能力。

东软在进入医保信息化领域之后，积极参与以政府医保为主要付费方，医保、医疗、医药三医联动，市、县、乡、村四级纵深的大健康生态建设。未来，东软将与英特尔持续深入合作，融合新一代硬件基础设施与“软件定义”创新应用生态，推进 AI 等创新技术在医疗行业的深度应用，促进医疗健康产业高效率、高质量、普惠化发展，助力构建多层次医疗保障体系。