ISSN: 2661-3719(Print); 2661-3727(Online)

面向国产化生态的模型推理优化支持技术及应用适配

傅文琦　朱奕健　顾剑峰
联通（上海）产业互联网有限公司　上海　200050

摘　要：国产化生态持续完善且自主可控需求不断增强，在此情况下人工智能应用落地的重要环节是模型推理的高效适配。以近一年的技术实践为基础针对国产 CPU、GPU、NPU 和本土操作系统分析模型在不同推理场景的性能瓶颈、迁移障碍并提出优化路径且要兼顾精度、时延、能耗和稳定性。研究方法包括构建统一中间表示、算子标准化与内核加速、量化与结构压缩、运行时调度与资源管理以及针对行业应用的端边云协同部署策略，结果表明优化后的推理系统典型任务延迟降低超 20%、能效比提升约 15% 且安全可控与合规性方面满足业务要求。结论是系统化的优化支持技术与应用适配能有效提升国产化生态下模型推理的可用性和工程价值并给后续行业推广提供可复制的技术参考。

关键词：国产化生态；模型推理优化；算子加速；量化压缩；端边云协同；安全合规

近年来，政务、金融、制造、医疗等行业里人工智能模型的部署规模不断扩大，但在国产化软硬件环境下实现高性能推理依旧挑战重重 [1]。硬件架构、指令集特性、系统调度机制、工具链支持度各有差异，直接移植通常难以让模型潜力充分发挥出来，并且端、边、云等多层部署形态对延迟、吞吐和能效有着差异化要求，应用场景多样又进一步加大了优化的复杂程度。企业推进国产化替代时，得在性能、稳定性、成本和安全可控之间达成平衡，这不仅是技术方面的事，还和系统生命周期的可持续性有关。因此在保证业务精度与响应速度的条件下，如何充分挖掘国产硬件和软件栈的潜力成了产业界和技术界共同关注的点。

1 国产化生态下的模型推理需求与挑战

1.1 生态版图与迁移边界

处理器 / 加速器架构、指令集与算子库、编译与运行时、国产 OS 与容器、驱动与中间件、开发框架与工具链等被国产化生态涵盖 [2]。而明确“模型—算子—内核—指令”的映射路径，识别与通用生态在算子覆盖度、动态形状处理能力、内存层次与并行模型等方面的差异面是迁移的关键。通过基线评测和瓶颈画像建立优先级清单，聚焦高影响环节，避免因“大而全”改造而造成成本外溢则能为后续优化确定清晰的边界与目标。

1.2 多目标约束下的性能与稳定性诉求

真实业务面前，推理链路得兼顾延迟分位数、吞吐 / 成本比、能效、可用性和可维护性，由于端、边、云负载形态各有不同，交互式场景对尾时延和抖动敏感，离线批处理着眼于吞吐和单位成本，边缘部署注重能效和资源占用。，要以 SLA 为引导，确定目标 P99、EOPS、内存水位和故障恢复时间，再加上容量基线和弹性策略，从而在负载波动与版本迭代时保障稳定表现和可预期体验。

1.3 安全可控与评测方法的工程化落地

国产化替代重视数据安全与合规，在推理环节要涵盖访问控制、审计追踪、脱敏和本地化处理并且在输出侧设置内容安全与可追溯机制。评测不能仅停留在通用基准上而应构建符合场景的工作负载并将准确率变化、延迟分布、功耗与稳定性联动起来构建灰度对比和回归体系，通过“可观测—定位—验证”的闭环将优化收益固化为标准流程以支撑规模化上线和持续迭代。

2 面向国产化的模型推理优化支持技术体系

2.1 统一中间表示与异构编译

用统一的 IR 来承载图级和算子级的优化，模式识别与子图划分相配合，把算子融合、常量折叠和形状推导提前到编译期以减少运行期的开销。动态形状用区间和约束来编码并结合 AOT/JIT 混合策略与缓存复用从而缩短冷启动时间并稳定尾时延，对于注意力、卷积、矩阵乘这些高频结构提供模板化的代码生成和自动调参来给国产 CPU/GPU/NPU 生成高质量的内核以确保跨后端的一致性。将 IR 当作纽带使工具链统一处理混合精度、内存布局和图重写从而使多平台不用重复适配且移植效率和可维护性显著提高。

2.2 算子标准化与压缩加速协同

高频算子周围建立语义一致的接口规范以减少不同后端的行为差异，内核方面引入向量化、循环分块、寄存器重用与预取策略并通过缓存友好布局降低访存瓶颈。量化采取 PTQ/QAT 混合方案且注意力和全连接优先选择对称 int8、嵌入和归一化保留 fp16 从而将精度波动控制在正负 0.5% 以内，结构化剪枝把重点放在通道、头数与 FFN 宽度上并配合蒸馏与重参数化将训练期的冗余折进推理期等价计算图以减少内核调用次数与中间张量物化。以“算子标准化—低比特—结构压缩”协同路径可兼顾精度稳定、吞吐提升与能效改进。

2.3 系统级优化与服务化运行

从端到端的角度对张量生命周期加以分析并进行缓冲复用，用激活检查点和就地更新的方式把峰值内存降下来，通过通信 / 计算重叠以及流水化执行来减轻带宽压力。运行的时候有动态批处理和并发隔离，并且批量上限根据 P99 阈值自适应调整以防止吞吐优化对交互体验造成挤占。对 NUMA 进行亲和处理、绑核并治理 DMA 通道，减少跨节点的抖动，在容器这边用 cgroups 细致地设置限额让 SLA 保持稳态。可观测性涉及请求链路、算子和设备这三层，且 Tracing/Profiling 和异常检测构成“定位—修复—回归”的闭环。在统一的实验环境里把优化前后的关键指标作个对比，结果看表 1。

表 1 代表性任务在国产化环境中的示范性评测结果（实验室环境）

任务 / 场景	优化前 P99(ms)	优化后 P99(ms)	优化前 QPS	优化后 QPS	优化前 EOPS/W	优化后 EOPS/W	精度 Δ(%)
LLM-7B 对话（国产 GPU/NPU）	260	180	35	52	0.9	1.3	–0.1
CV-ResNet50 批推理（批 =64，国产 CPU）	95	70	1200	1650	3.1	4.2	0.0
RAG 检索增强问答（端边云协同）	420	280	18	27	0.6	0.9	+0.3

2.4 国产硬件 / 操作系统适配与质量保障

国产 CPU 的向量扩展和缓存层级方面要建立起算子到指令级原语的映射范式以及自动调参流程 [3]。而国产 GPU/NPU 的张量核心与片上存储要提供块大小、并行度和访存策略的微基准搜索从而形成可复用配置集。操作系统层需校准计时器精度、线程模型和调度器参数，容器靠设备插件和驱动治理保持稳定隔离。质量体系通过“基线—灰度—回归”控制变更风险，其中基线评测把 P99、QPS、EOPS/W 和内存水位作为锚定，灰度发布对实际负载下的分位数和抖动进行监测，回归用一致的数据分布和种子重现实验条件。配套的性能预算和异常门限让优化收益能够被度量、追踪和复制，以此支撑规模化上线和持续演进。

3 应用适配与工程落地实践路径

3.1 行业需求映射与模型选型、数据—知识适配

业务目标牵引着落地路径，准确率、时延和成本被量化为可度量指标，模型家族、规模以及推理优先级据此确定 [4]。7B 到 13B 量级的指令微调模型适合通用对话与检索增强问答，金融风控更多依赖结构化规则与小模型级联，制造质检则偏向视觉主干的高吞吐推理。数据与知识侧通过“高质量知识库＋稳健检索重排＋轻量 SFT”进行闭环治理。构建多源知识向量索引和域内术语表以减少幻觉和跨域漂移。

国产化栈迁移政务热线问答，选用 13B 指令对齐模型并与 RAG 相结合。在业务流程不变的情况下建立 2.1M 条知识条目、128 维向量索引与 BM25 重排，使 FAQ 命中率从 83.7% 提升到 90.9%、拒答率从 7.5% 降为 3.2%、交互式 P95 时延控制在 320ms 以内、敏感实体屏蔽命中率提高到 99.4%，实现“高可用＋可控输出”双目标。

3.2 端—边—云一体化部署与可观测 SLA 治理

“就近计算、分层协同”原则主导着部署形态，由边缘侧负责预处理和轻量推理，云侧负责复杂模型与知识检索，端侧专注于低功耗与离线容灾。路由会根据数据敏感度与实时性动态选路，靠缓存与请求合并来缓解瞬时流量。运行时引入动态批处理、并发隔离和自适应限流，P99 阈值绑定批量上限以维持交互和批处理的平衡，可观测体系覆盖请求链路、图级和设备级，Tracing/Profiling 与异常检测构建起“定位—修复—回归”的闭环，SLA 持续评估 P95/P99、失败率、抖动指数和资源水位。

边缘网关的制造质检产线部署了 ResNet50 和轻量检测头且瑕疵库检索与复核保留在云侧。上线之后单工位 P99 从 92 毫秒降到了 63 毫秒、误检率从 2.8% 降到 1.6% 且产线

OEE 提高了 1.9 个百分点，容量基线设定为 1.4× 高峰 QPS 且拥塞时会自动降级为低分辨率检测、违规帧溯源花费的时间不到 200 毫秒。在这个场景里与表 1 相同的图融合和内存复用策略让 QPS 提升了大概 38%。

3.3 安全可信与合规工程、成本—能耗治理

将安全合规前置到设计阶段，构建起“数据分级—最小权限—全链路审计”的标准化流水线，对涉敏字段在采集、传输与落地环节进行脱敏或本地化处理，并且在输出侧添加水印与可追溯机制以支撑责任界定。让内容安全规则与模型对齐策略联动起来从而减少灰区输出。以 EOPS/W 作为成本与能耗治理的核心指标，融合混合精度、量化、批处理与资源共享来防止低效空转。对服务开展碳账核算与能耗画像，按照“峰—谷—闲”分层弹性收敛以保持 TCO 与体验双优。

某城商行票据影像与对话辅助审核在国产 GPU/NPU 混部环境上线。账号和密钥被集中托管且短期凭证用于跨域访问，日志和审计保留 180 天。推理侧用激活对称 int8 与权重量化 int8 ＋ fp16 混合方案，票据 OCR 准确率维持在 98.8% ± 0.2% 且交互式客服 P99 从 340ms 降到 230ms，能效 EOPS/W 提高了约 31% 且夜间低峰自动缩容到日间实例数的 35%，月度算力费用降低 22.4%。

3.4 迁移方法论、标准化与生态共建

工程实践要遵循“基线—瓶颈—优化—灰度—回归”的闭环。工作负载靠基线用业务真实数据构建且要拉齐 P95/P99、QPS、EOPS/W 和内存水位，热段与热点算子是瓶颈定位聚焦对象，优化采取小步快跑和可回退策略且变更由特征开关控制发布，灰度期按分区或者人群逐步增加流量并观测分位数与抖动，回归用一致的随机种子和数据切片重现实验以确保收益可证得。在标准化方面，算子接口、IR 约定、运行时 API 和观测指标周围形成统一规范，从而沉淀可复用的适配资产和测试集，降低后续项目的人力和时间成本。

某三甲医院影像辅助筛查从通用栈迁移至国产 GPU。基线阶段 CT 二分类 AUC 为 0.962、吞吐达 420img/s、P99 为 118ms，定位时发现片上存储受限情况下卷积核与归一化出现访存抖动，于是采取卷积—归一化—激活融合、块大小自动搜索和张量重排措施并配合 INT8 感知量化与小规模校准。灰度在两周内按 20%—40%—80% 放量且关键指标没有负漂，最终 AUC 为 0.960、吞吐提升到 615img/s、P99 降至 81ms、EOPS/W 提升约 28%、误报率降低 0.4 个百分点，项目产生的算子模板与微基准被纳入内部仓库，后续胸片与超声场景可直接复用且单项目平均交付周期缩短近 35%。

结语：国产化生态下的模型推理优化不仅关乎技术性能的提升，更体现了产业自主可控与持续创新的能力。通过软硬件协同、算法与系统融合、工程化治理与生态共建，能够在确保安全合规的前提下，充分释放国产硬件潜力，形成高效、稳健、可持续的智能服务体系。这一过程推动了本土技术标准的完善与产业链的协同升级，为人工智能在更多关键领域的深度落地奠定了坚实基础，也为未来面向更大规模、更高复杂度的智能应用创造了广阔空间。

参考文献：

[1] 段柳成 , 肖巧玲 , 金怡 , 等 . 大模型时代国产大算力 GPU 的关键挑战与发展路径 [J]. 人工智能 ,2025,(03):8-21.

[2] 刘磊 , 卢国彬 , 杜科 , 等 . 大模型运管平台建设实践与技术优化 [J]. 信息技术与标准化 ,2025,(06):71-77.

[3] 刘磊 , 卢国彬 , 杜科 , 等 . 大模型运管平台建设实践与技术优化 [J]. 信息技术与标准化 ,2025,(06):71–77.

[4] 李俊伟 , 林龙 , 王力华 . 基于国产化生态系统的医院信息系统智能运维助手构建与应用 [J]. 中国卫生信息管理杂志 ,2025,22(01):14–19+25.

作者姓名: 傅文琦(1992—), 汉族, 山东省, 中级工程师, 硕士研究生，研究方向为模型的应用适配能力。