人工智能视域下工业大模型优化的路径研究

安士才

（山东捷瑞数字科技股份有限公司 山东 烟台 264003）

摘 要 人工智能技术的快速发展为工业领域带来了深刻变革，工业大模型作为新一代人工智能技术的重要载体，在推动工业智能化转型中发挥关键作用。针对工业大模型在实际应用中存在的数据质量参差、模型泛化能力不足、算力资源受限等问题，文中从数据层面、模型架构、计算优化和部署应用4个维度探讨了优化路径。通过分析工业场景特征，提出了基于领域知识的数据增强方法、轻量化模型裁剪策略、分布式训练加速方案和边缘智能部署框架，为提升工业大模型性能和应用效果提供了新思路。

关键词： 工业大模型；模型优化；数据增强；轻量化；边缘部署

中图分类号 TP391

Research on the Path of Industrial Model Optimization from the Perspective of Artificial Intelligence

AN Shicai

（Shandong Jierui Digital Technology Co., Ltd., Yantai, Shandong 264003, China）

Abstract The rapid development of artificial intelligence technology has brought profound changes to the industrial field, and industrial large-scale models, as an important carrier of the new generation of artificial intelligence technology, play a key role in promoting the transformation of industrial intelligence. In response to the problems of uneven data quality, insufficient model generalization ability, and limited computing resources in the practical application of industrial large-scale models, this paper explores the optimization path from four dimensions: data level, model architecture, computational optimization, and deployment application. By analyzing the characteristics of industrial scenarios, domain knowledge based data augmentation methods, lightweight model pruning strategies, distributed training acceleration schemes, and edge intelligence deployment frameworks have been proposed, providing new ideas for improving the performance and application effectiveness of industrial large-scale models.

Key words Industrial large-scale model, Model optimization, Data augmentation, Lightweight, Edge deployment

0 引言

工业大模型凭借其强大的知识表征和任务处理能力，能有效解决传统工业智能化过程中的技术瓶颈。然而，工业场景的特殊性对大模型的适应性提出了更高的要求，如何针对性地优化工业大模型，提升其在实际应用中的表现，成为亟待解决的关键问题。深入研究工业大模型的优化路径，对推动工业智能化转型具有重要的理论价值和实践意义。

1 工业大模型状态分析

1.1 工业大模型的发展概况

工业大模型经历了从单一任务处理到多场景融合的演进历程，在智能制造、预测性维护和质量控制等领域展现出显著优势[1]。近年来，基于深度学习的工业大模型在参数规模和架构设计上取得突破性进展，模型参数量从百万级跃升至百亿级，具备更强大的特征提取和知识表征能力。国内领先企业相继推出面向工业领域的预训练大模型，通过海量工业数据训练和领域知识注入，实现了对设备状态、生产工艺和质量指标的精准建模。工业大模型在生产过程优化、设备健康管理和产品质量预测等应用场景中的实践表明，其模型性能和泛化能力仍在持续提升。

1.2 工业应用中的主要挑战

当前，工业数据呈现出高维稀疏、类别不均衡和标注困难等问题，导致模型训练数据质量难以保证。模型结构复杂度与实时性要求之间存在矛盾，庞大的参数规模增加了计算开销，影响模型在边缘设备上的部署效果。知识获取和迁移存在障碍，工业场景的专业性和封闭性制约了模型对新环境的适应能力。异构数据源的融合与协同处理机制尚不完善，跨域知识迁移效果欠佳，模型在复杂工况下表现不稳定。

1.3 优化需求分析

工业大模型亟需在多个层面实现性能优化和功能提

作者简介：安士才（1980—），硕士，工程师，研究方向为人工智能。

升[2]。数据层面需要构建高质量的工业数据集，提升训练数据的代表性。模型层面需要设计轻量化网络结构，在保证精度前提下提高推理速度。算法层面需要增强模型的鲁棒性和泛化性。部署层面需要优化资源调度机制，满足工业现场的实时性要求。为深入阐述这些优化路径，本文选取钢铁、化工、汽车制造、半导体等典型行业案例进行分析。这些案例涵盖了离散制造和流程制造，能全面展示工业大模型在不同场景下的优化策略和实施效果。

2 数据层面的优化对策

2.1 工业数据特征提取与预处理

针对钢铁生产线实际运行数据，建立特征提取系统捕获轧制过程中的温度、压力、速度等关键参数。通过部署在轧机关键位置的传感器采集高频振动信号，将采样频率设定为1 000 Hz，实时记录轧制过程中的瞬时变化。温度滑动的平均值如式（1）所示：

$$Tt=(1/N)\Sigma Ti,\ i\in[t-300\ s,\ t] \quad (1)$$

其中，N表示5 min内的采样点数，为300 s×1000 Hz；Tt表示在t时刻的滑动平均温度值；$(1/N)$：$1/N$是归一化因子；$\sum$表示将窗口内所有温度值加总，Ti表示第i个时时刻的温度采样值，$i\in[t-300\ s,t]$表示时示时间窗口。

对于温度数据，以5 min为时间窗口进行滑动平均，消除测量过程中的瞬时波动；压力数据结合实际工艺要求，设定有效值范围为300 MPa~800 MPa，对超出范围的数据点进行剔除处理。在板材厚度检测环节，融合激光测厚仪和X射线测厚仪的双源数据，通过加权平均方法提高测量精度，测量误差控制在±0.01 mm以内。厚度双源融合如式（2）所示：

$$H=w1\cdot HX射线；w1+w2=1,\ |H-H实际|\leq 0.01\ mm \quad (2)$$

其中，H表示最终示最终融合后厚度测度测；$w1$和$w2$为权重系数。

质量检测图像数据经过灰度化和边缘增强预处理后，利用128×128像素的滑动窗口扫描提取局部特征，建立缺陷特征库。同时，结合产线工艺专家经验，将轧制速度、道次压下率、乳化液浓度等工艺参数标准化区间设置为[-1，1]，构建多维特征向量，为模型训练提供标准化的数据输入。

2.2 基于领域知识的数据增强方法

在化工生产线的反应釜控制系统中，通过物料平衡原理可以构建数据增强方案。基于实际生产数据，随机扰动80℃~120℃内的温度，扰动幅度控制在±2℃，生成新的温度曲线序列；压力数据区间在0.6 MPa~1.2 MPa，按工艺要求进行波动采样，形成压力变化数据集。针对催化剂投加量，结合化学动力学模型，在标准投料量±5%范围内产生变异数据，并调整反应时间和转化率数据[3]。原料配比数据通过专家规则库指导，在保证化学计量比的前提下，对各组分含量进行±3%的随机偏移，生成多组合格的配方数据。同时，依据设备历史运行工况，分段采样150 rpm~300 rpm内的反应釜搅拌速度，结合流体力学模型生成对应的传热系数和混合指数数据。通过这种方式扩充的训练数据既符合工艺规律，又包含了生产过程的波动特征，有效提升了模型对工况变化的适应能力。

2.3 数据质量评估与筛选机制

在汽车发动机生产线的数字化车间中，建立了三级数据质量评估体系。通过在缸体加工工位部署的在线测量系统，对关键孔径尺寸数据进行实时采集，设定±0.02 mm的公差区间，将测量值偏离该区间的数据标记为疑似异常。针对气缸壁粗糙度数据，采用表面轮廓仪进行检测，将0.8 μm-1.6 μm内的Ra值数据划分为高质量样本，超出此范围则进入人工复检环节。对曲轴箱体配合面的加工数据，结合几何公差链分析，设置平面度0.05 mm、垂直度0.03mm的评估标准，通过自动检测设备筛选合格数据。在数据完整性评估方面，针对加工设备产生的刀具位置、进给速度、主轴转速等过程参数，要求采样间隔不超过100 ms，数据缺失率低于0.1%。同时，对多传感器采集的轴承振动、温度等状态数据建立时间戳对齐机制，将同步误差超过50 ms的数据序列剔除，确保数据的时序一致性[4]。这套机制有效提升了训练数据的质量，为模型优化提供了可靠的数据支撑。

3 模型架构优化方案

3.1 模型结构轻量化设计

在智能装配生产线的焊接质量检测系统中，通过深度神经网络剪枝实现模型轻量化[5]。对原有含512个神经元的全连接层进行稀疏化处理，基于L1正则化方法计算神经元重要性得分，删除得分低于阈值0.3的神经元，最终将该层压缩至128个神经元。卷积层采用通道级剪枝策略，通过计算特征图的平均激活值，将12个卷积核削减至8个，保留对焊缝特征提取贡献度较高的滤波器。在模型量化环节，将原32位浮点权重转换为8位定点数表示，权重量化比例因子设为0.125，激活值量化区间设定在[-8，7.875]。针对ReLU激活函数，采用分段线性近似方法，将函数曲线划分为4个线性区间，用查找表加速推理计算（见图1）。

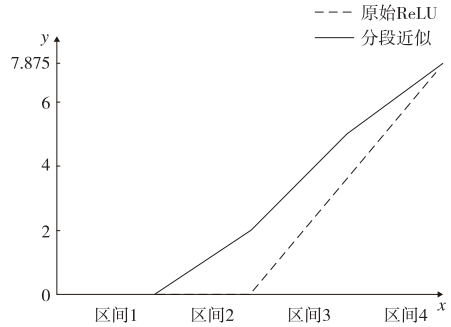

图1 ReLU分段线性近似示意图

经过优化后的模型参数量从15 M降至3.8 M，模型存储空间减少72%，在工业边缘设备上的推理速度提升3.2倍，同时保持焊缝缺陷检测准确率在95%以上，满足生产线

实时检测要求。

3.2 领域适应性迁移学习

在半导体芯片制造线的缺陷检测系统中，针对新产品型号缺乏标注数据的问题，部署领域自适应迁移框架。将已有8寸晶圆缺陷检测模型作为源域，迁移至12寸晶圆检测场景。通过对源域数据中的5 000张缺陷图像进行特征提取，建立包含划痕、颗粒、污点等缺陷的特征库，设置最大均方差对齐损失函数，将源域特征分布映射至目标域。在迁移过程中，采用动态权重调整策略，将源域分类器的权重系数从1.0逐步降低至0.4，目标域分类器权重从0.2提升至0.8，实现特征表示的平滑迁移。针对目标域仅有200张已标注样本的情况，通过对比学习方法增强模型对目标域特征的判别能力，将相似缺陷样本的特征距离控制在0.3以内，不同类别缺陷样本的特征距离扩大至0.8以上。经过500轮迭代训练，模型在新产品缺陷检测任务上的准确率从初始的75%提升至92%，漏检率降低至0.5%以下。

3.3 多任务学习框架构建

在工业燃气轮机运行管理系统中，构建集状态监测、性能预测和故障诊断于一体的多任务学习框架（见图2）。通过共享网络底层提取的压比、转速、振动等设备特征，在顶层分别设计3个任务分支。状态监测分支采用长短期记忆网络结构，输出层对应正常、亚健康、预警和故障状态；性能预测分支基于门控循环单元网络，预测热效率、功率和排气温度指标；故障诊断分支应用图卷积网络，识别轴承故障、叶片裂纹、喷嘴堵塞等故障类型。在损失函数设计中，设置状态监测、性能预测和故障诊断的权重比例分别为0.3、0.4和0.3。优化过程采用梯度归一化技术，使3个任务的梯度幅值保持相近水平。经优化后，状态评估准确率达96%，性能预测平均相对误差在2.5%内，故障诊断准确率达94%。

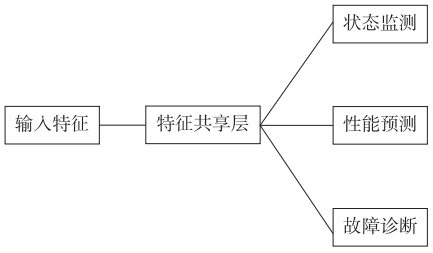

图2 多任务学习框架示意图

4 部署实施优化途径

4.1 分布式训练性能优化

在石化行业大型裂解炉生产线的智能优化系统中，可以通过构建分布式训练架构提升模型训练效率。采用8台GPU服务器组成训练集群，每台服务器配置4块NVIDIA V100显卡，通过InfiniBand网络互联，带宽达到100 Gbps。在参数服务器架构中，将模型参数分散存储在4台参数服务器上，工作节点采用异步更新策略，设置梯度累积步数为16，有效降低通信开销。针对裂解炉温度场预测任务，将3 000万条历史工况数据划分为多个批次，在各工作节点上并行计算梯度。通过引入弹性平均随机梯度下降算法，动态调整学习率从0.01到0.001，将参数更新周期设置为500步。在通信优化方面，实施梯度压缩和稀疏化传输，将梯度值量化为16位浮点数，并设置阈值仅传输绝对值大于0.01的梯度分量。经过优化后的分布式训练系统将模型训练时间从原来的72 h缩短至18 h，训练吞吐量提升3.8倍，加速比达到6.5。

4.2 模型压缩与量化技术

在火电厂锅炉燃烧优化系统部署过程中，通过模型压缩与量化技术降低计算资源占用。针对原有187 MB的深度神经网络模型，采用结构化剪枝方法对卷积层进行通道压缩，将输入通道从256降至64，输出通道从512降至128。在全连接层采用低秩分解技术，将4 096×4 096的权重矩阵分解为512×64和64×512两个小矩阵相乘的形式。在权重量化环节中，将32位浮点数压缩为8位整数表示，量化比例因子设为0.127，在激活值量化中将取值范围限定在[−8，7.875]区间。权重量化如式（3）所示：

$$Q=round\left(F/0.127\right) \tag{3}$$

其中，F为32位浮点数；Q为8位整数。

为保证NOx排放预测精度，在量化过程中采用非均匀量化策略，对数值分布密集区域采用较小量化间隔0.01，稀疏区域采用较大间隔0.1。通过int8指令集加速推理计算，在边缘计算单元上将推理延迟从原来的89 ms降低至23 ms，模型大小压缩至42 MB，内存占用减少65%，满足实时控制需求。经过在线测试验证，优化后模型在NOx预测任务上的平均相对误差维持在3.2%水平。量化间隔如式（4）所示：

$$\Delta=0.01,\ 当\rho(x)>threshold;\ \Delta=0.1,\ 当\rho(x)\leqslant threshold \tag{4}$$

其中，$\rho(x)$表示数值示数值分布的数；threshol表示密度阈度。

4.3 边缘计算部署方案

在工业机器人焊接生产线中，部署3层边缘计算架构，以实现模型本地化推理。生产现场布置8台边缘计算单元，每台配备ARM Cortex-A72处理器和4 GB运行内存，通过工业以太网与中心控制系统互联。在边缘节点上运行轻量化焊缝跟踪模型，将视觉传感器采集的焊缝图像按50 ms间隔进行实时处理，经过目标检测网络输出焊缝位置坐标，控制机器人完成轨迹纠偏。边缘网关层部署完整的焊接工艺模型，融合电流、电压、送丝速度等工艺参数，实时优化调节焊接过程。模型更新采用增量学习机制，边缘服务器每间隔4 h将新增样本传输至云端，由云端重训练后将更新后的模型参数分发至各边缘节点。为应对网络中断情况，边缘节点具备30 min的本地缓存能力，确保断网期间生产连续性。通过边缘计算部署方案，焊接控制响应时间降至15 ms以内，数据处理延迟减少，焊缝跟踪精度达到±0.3 mm，实现生产线智能化升级。

（下转第298页）