网络智能化中的 AI 工程化技术方案

朱明伟
（中国移动通信集团设计院有限公司，北京 100080）

摘　要：网络智能化是通信行业借助 AI 技术，对外增强网络赋能能力，对内实现降本增效的重要举措。从 AI 工程化的视角系统分析网络智能化应用落地的难点，提出了包括数据采集处理、训练计算资源的管理与任务调度、推理部署优化在内的面向生产环境的 AI 工程化技术方案，探讨网络智能化生态发展的策略。

关键词：网络智能化；人工智能；云原生；模型压缩；推理服务
中图分类号：TP181；TN929.5
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022016

AI engineering technology solutions in network intelligence

ZHU Mingwei
China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: Depending on AI technology, network intelligence is becoming an important initiative for communication industry to enhance network empowerment externally, and to achieve cost reduction and efficiency internally. The difficulties implementing network intelligence applications from the perspective of AI engineering were analyzed. The industrial grade AI engineering technical solutions were proposed, including data collection and processing, computing resources management and task scheduling, and inference deployment optimization. The strategies of network intelligence's ecosystem development were studied.

Key words: network intelligence, artificial intelligence, cloud native, model compression, inference service

0　引言

2012 年神经网络在图像识别领域的成功应用拉开了本轮人工智能热潮的帷幕，并相继在语音识别、自然语言处理等领域取得了突破性发展，改变了行业生态。目前，人工智能技术正向纵深方向发展，从互联网领域的推荐、广告、搜索，到科学计算领域的分子动力学模拟[1]和蛋白质结构预测[2]，人工智能在越来越多的场景中推广和落地。

在通信领域，人工智能的影响力同样与日俱增。标准化组织、设备商、运营商等产业各界立足自身，积极探索人工智能在通信网络的应用，为网络智能化的落地做出了大量有益的工作。ITU、3GPP、ETSI、CCSA 等国内外标准组织设立相关课题，从网络智能化的体系框架、互通接口、网元、流程、用例等方面进行研究[3]；华为、中兴等通信设备提供商则以解决方案为突破口，

收稿日期：2021−09−08；修回日期：2022−01−12

推出诸如华为自动驾驶网络[4]、中兴自主进化网络[5]等网络智能化产品；运营商则以采购建设厂商 AI 设备/服务与自建 AI 能力平台相结合的方式，不断提升自身网络智能化水平，如中国移动自主开发的九天人工智能平台的智能交互、智慧稽核、网络自服务等 AI 服务已大规模商用[6]。

1 网络智能化应用落地的难点

运营商网络是一个按照不同地域、不同专业领域和不同层级进行分布式部署的半自治网络，网络结构极为复杂，难以对全网进行统一建模。网络智能化需要从局部入手，实现场景化的网络智能，并逐步拓展智能化应用范围。

网络智能化应用场景大致可以分为基础网络、网络管理和对外服务 3 类[7]。其中，基础网络智能化的关注点在网络设备层面，场景包括基站的无线资源管理和流量预测、核心网的用户策略管理和移动性管理、承载网的路由调度等。这类场景的 AI 算法往往紧密嵌入网元逻辑功能或业务流程中，对于处理时延和厂商互通性要求较高。网络管理智能化的关注点在网络规划运维层面，通过分析网络的规划、建设、维护、优化、运营等各方面的海量数据，实现智能化的网络/切片参数配置、故障定位、根因分析等。对外服务智能化的关注点在产品和服务层面，主要为对用户提供智能客服、产品推荐等运营商特色 AI 产品，以及图像分类、语音识别、自然语言处理（natural language processing，NLP）、光学字符识别（optical character recognition，OCR）等通用化 AI 产品，此类场景对于通信的领域知识要求相对较低。

由上文可知，网络智能化应用与图像识别、NLP 等通用 AI 在业务场景上既有共性又有差异。随着场景广度和深度的不断发展，网络智能化需要借鉴通用 AI 的技术和经验构建一套生产级别 AI 工程化系统作为各类网络智能化应用的技术底座。但由于领域特殊性，网络智能化系统的生产级别落地在数据采集和特征抽取、模型训练、模型部署等 AI 工程化环节还存在以下难点。

（1）数据采集和特征抽取

运营商的无线接入网、核心网、传输网、云资源池、运维系统、业务支撑系统中散布着的海量指标、日志、调用链（trace）数据及用户使用网络留下的数据，这些数据符合大数据的大量（volume）、价值稀疏（value）、多元（variety）、高速（velocity）、真实（veracity）的 5V 特性，同时具备多维、多边、多粒度、个性化等特点[8]。采集这些散布且异构的海量数据并统一处理存储的难度很大、成本很高：首先，数据质量不高，存在多源数据易缺失、统计口径不一致、大量数据没有标注或标注质量不一致等问题；其次，同一主题的网络数据具有高维的特点，这种高噪声增加了网络智能化特征选择的难度；第三，不同于图像识别和 NLP 利用神经网络进行特征抽取的方式，网络智能化需要根据场景不同而手工进行特征选择和抽取，增加了特征向量的构造难度。

（2）模型训练

网络智能化在不同场景下所用的模型各不相同，根据训练数据的不同，总的来说表格类数据、小数据应优先采用以集成学习模型为代表的机器学习模型进行训练，而非结构化数据、大数据则一般采用深度学习算法，其中，网络流量预测等时间序列场景常用循环神经网络（recurrent neural network，RNN）、长短期记忆（long short-term memory，LSTM）网络等深度学习算法，以及梯度提升决策树（gradient boosting decision tree，GBDT）、极限梯度提升（extreme gradient boosting，XGBoost）算法、支持向量机（support vector machine，SVM）等有监督机器学习算法，也会用到线性回归、移动平均等传统算法；故障检测等分类场景中主要使用轻量梯度提升机（light gradient boosting machine，LightGBM）、XGBoost 等

集成学习算法和多层感知机（multilayer perceptron，MLP）等深度学习算法，也会用到 k 均值（k-means）聚类算法、基于密度的噪声应用空间聚类（density-based spatial clustering of applications with noise，DBSCAN）算法等聚类算法，差分、核密度估计等传统算法和决策树等简单机器学习算法在数据分布较为匹配时也能取得不错效果；对于智能客服、图像识别等通用化 AI 场景则必须使用卷积神经网络（convolutional neural network，CNN）、基于变换器的双向编码表示（bidirectional encoder representation from transformers，BERT）算法、视觉变换器（vision transformer，ViT）算法等大型深度学习模型。上述各类模型在训练过程中对算力资源消耗的差异极大，而且大量训练和推理任务在云上混合部署，对云的算力资源管理、多训练任务调度能力提出了高要求。

（3）模型部署上线

无线资源调度等部分网络智能化应用对于模型推理时延要求高，可低至毫秒乃至微秒级别，导致此类网络智能化应用的模型只能部署在网络设备侧，而网络设备计算资源高度异构，设备的算力有限，现网网络设备的系统较为封闭，芯片不支持 AI 模型编译或者适配效果不好。此外，模型部署上线需要解决模型的持久化、模型服务的构建及模型与网元系统的集成或服务调用等一系列问题。

面对以上问题，本文按照数据准备—模型训练—模型部署的 AI 工作流顺序，研究和提出一套网络智能化的 AI 工程化技术方案，并阐述对于打造网络智能化生态的思考。

2　网络智能化中的数据准备

高维、海量、治理良好的训练数据和实时准确的推理数据是网络智能化应用落地的前提，因此，需要建立包括采集、处理、存储等功能在内的 AI 数据管理平台，AI 数据管理平台架构如图 1 所示，统一管理全网与网络智能化相关的历史数据与实时数据。

2.1　数据采集

不同于通用 AI 模型训练中常使用数据集，网络智能化领域的数据大多需要从网元和系统运行的各类指标、埋点日志、接口信令监控中采集。目前运营商普遍通过数据中台或数据仓库对网络数据进行统一管理。为避免重复建设，以降低成

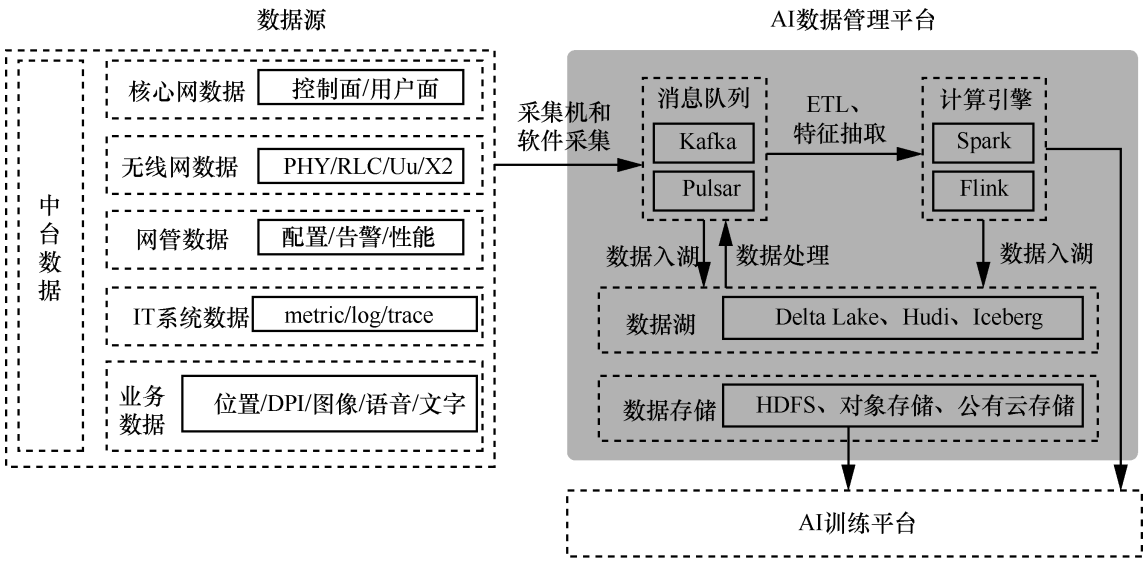

图 1　AI 数据管理平台架构

本，网络智能化的数据采集应充分利用已有数据，所需原始数据如果已纳入数据中台或数仓的管理范围则通过采集工具从数据中台或数仓采集；对于未纳入数据中台或数仓的数据，IT 系统中的指标、日志、调用链数据通过消息队列推送至平台，网络设备中的数据通过在网络设备侧的信令采集机进行数据采集。训练数据并非越多越好，模型的精度在达到一定程度后，训练数据的增长将越来越难以带来精度的进一步提升；另一方面，增加特征参数不一定能提升模型性能，特征太多反而更易出现过拟合。因此，为了控制数据采集成本，在实际应用中应控制采集数据的指标维度、单指标的数据采集区间，从而减少信令采集机和 IT 系统监控模块的改造量。

2.2 数据处理和特征抽取

AI 数据管理平台针对采集数据的特点，需要对多源异构海量的数据进行包括数据提取、缺失值补齐、多源数据的合并/去重/归一化在内的各种处理，对于无标签数据的标注通过少量标签数据结合算法进行标签生成[9]，保证数据质量还需要通过 schema 规则等方法进行数据校验[10]。目前网络智能化多为离线训练、在线推理模式，因此，数据管理平台既需要对离线数据进行批量计算，也需要对采集到的实时数据进行增量计算或流式计算。为了在保证批量数据和实时数据处理时效性的同时简化平台架构，平台需要采用流批一体的大数据计算引擎进行数据处理，将消息队列中或已入湖的数据导入 Spark 或 Flink 的工作节点进行计算。如果部分场景计算规则较为简单，则可以在消息队列中直接计算，以降低通信时延。数据特征的提取紧接着数据处理之后进行，选择和构造特征向量和特征向量维数直接影响模型精度、泛化性、计算量，目前自动机器学习（auto-machine learning，AutoML）方式的特征搜索尚不足以取代人工方法，因此，使用非深度学习算法的情况下，网络智能化的特征提取主要依靠领域专家知识和特征工程经验。

2.3 数据存储

海量网络数据经过处理后需要在 AI 数据管理平台中存储，由于网络智能化应用具有场景化、碎片化、长尾化特点，只有在网络智能化场景需求明确时才能确定对应数据处理和特征抽取的方式，难以提前建立数据处理模型，因此，平台的数据存储应采用数据湖架构，使各网络智能化场景的数据能够快速入湖存储，以供后续 AI 训练或推理任务读取。为降低云上部署的在线推理任务的数据读取时延，数据处理完成后的数据不落盘直接从消息队列推送至推理模型。此外，为降低数据存储成本，平台的底层存储应综合采用自有的分布式文件系统、对象存储，并根据数据治理要求将部分低密级的冷数据转存公有云存储。

3 网络智能化中的模型训练

生产级别的网络智能化模型训练需要处理好大规模计算资源的管理、训练任务算力需求异构、大量训练任务的编排调度等问题。将 AI 计算框架与云原生技术结合，构建面向网络智能化各场景的统一 AI 训练平台，能够很好地解决这些问题。

3.1 模型训练与云原生的结合

云原生基于容器轻量、快速、易于迁移的特点，依托 k8s 对各类资源进行定义、调度、控制和编排，实现了算力的精细化管理与高弹性伸缩；同时，k8s 的系统设计拥有很强的开放性，易于扩展。随着近年来云原生的快速发展，以其为中心形成了包括容器运行时、存储、网络、安全、服务网络、无服务器（serverless）、持续集成/持续部署（CI/CD）工具链、可观测性工具链在内的云原生生态圈[11]。云原生极大地提高了 IT 产业的开发、测试、部署、运维的工作效率，包括数据库、消息队列在内的各类基础中间件软件和各种业务软件都已经或正在重构以适应云原生环境。目前，TensorFlow、PyTorch 等主流 AI 计算框架都已支

持云原生，具备分布式部署能力。

AI 训练平台架构如图 2 所示，网络智能化的 AI 训练平台采用 k8s 作为资源管理和任务调度的枢纽，硬件方面支持 x86、ARM、NVIDIAGPU 及专用 AI 芯片（ASIC），满足各类机器学习和深度学习算法训练的算力需求。平台基于云原生 AI 领域的事实标准 KubeFlow[12]，通过创建操作器（operator）的方式，对 TensorFlow、PyTorch 等机器学习框架的训练任务进行自定义资源声明和资源状态控制，并将训练的工作负载部署在 k8s 集群的节点中；平台从 AI 数据管理平台的存储中拉取训练数据集、测试数据集，写入训练任务所在 k8s 节点的内存，并通过分布式缓存和远程直接内存访问（remote direct memory access，RDMA）加速数据写入速度；平台同时支持面向模型研发和面向生产部署的模型训练任务，提供各类算法库进行模型训练和验证，提供推理模型部署工具和超参搜索工具，并通过工作流编排工具将 AI 训练流程串联起来。此外，平台可通过 Spark 操作器同网络智能化的 AI 数据管理平台的大数据计算体系打通，将网络智能化的模型训练任务与数据处理阶段的批量计算任务混合部署，提高资源池整体利用率。

3.2　多训练任务调度

网络智能化的 AI 训练平台需要针对训练任务特点对资源调度器进行功能增强。k8s 原生的调度器是针对微服务架构设计的，适合对小颗粒度、长时间运行的互联网业务进行资源调度，但机器学习训练属于批处理任务，存在作业、任务队列、流水线等概念，k8s 原生调度器对此是不支持的；另一方面，为防止死锁发生，满足多租户、二次调度等需求，机器学习训练的调度需要满足批量 pod 调度、多队列调度、动态调度、任务间公平性等能力，原生 k8s 调度器同样不支持。AI 训练平台架构如图 2 所示，解决方案有两个：在 k8s 集群中部署一套专用批处理调度器，如 volcano 调度器[13]，此方案存在一个集群中的两套调度器冲突的问题，虽然最新版本中支持了多调度器混合部署，但实际部署时仍建议按调度作用范围把集群拆开；采用 k8s 原生的调度器框架（scheduling framework）方案[14]，即把 k8s 原生调度器插件化，将批处理调度算法以插件形式整合进 k8s 调度器。

3.3　并行弹性训练

近年来各类 AI 的模型规模普遍增大，并出现了以预训练为主要目的超大通用 AI 模型，以 NLP 领域的第三代生成式预训练变换器（generative

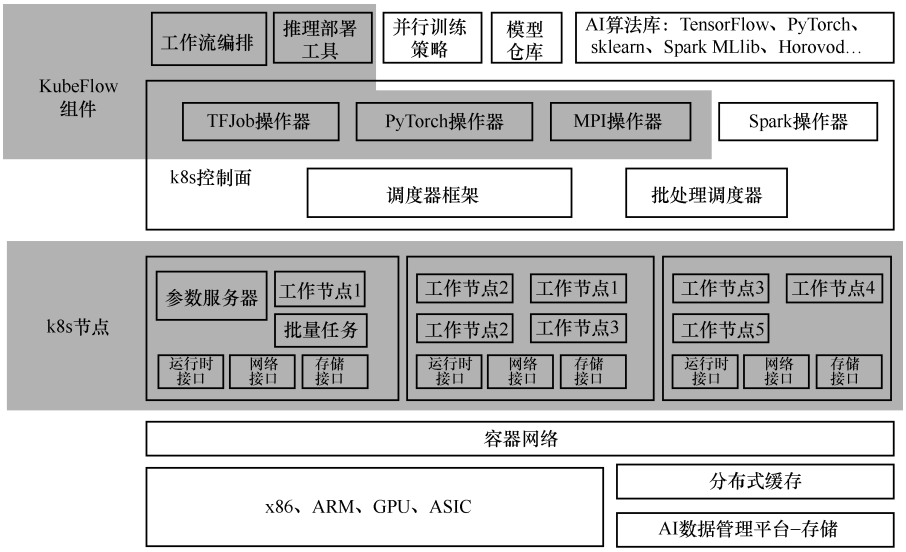

图 2　AI 训练平台架构

pretrained transformer 3，GTP-3）模型为例，其最大支持 1 750 亿参数[15]，单张 GPU 卡的内存和算力已远无法满足此类训练需求，需要多张 GPU 卡或多个 GPU 服务器进行并行训练。网络智能化领域中同样存在大模型训练需求，如无线网时序数据预测的 LSTM 模型，又如智能客服语义理解和对话生成的变换器（transformer）模型。为处理好网络智能化领域的大模型并行训练问题，首先 AI 训练平台需要在支持主流 AI 训练框架的数据并行、模型/流水线并行训练功能的基础上，结合网络智能化领域知识，不断完善和丰富混合并行、自动并行等复杂并行的训练策略。其次，平台需要优化并行训练的节点弹性伸缩能力，弹性需求在很多并行训练场景中存在，如在 GPU 资源池资源空闲时增加训练的 GPU 卡数量，又如在部分训练节点宕机时保持训练不中断，与一般的无状态微服务不同，并行训练属于复杂的有状态任务，GPU 之间存在大量参数传递，以数据并行的环形全局规约（ring allreduce）算法为例，其将模型存储在各个 GPU 上，每张 GPU 卡只对部分数据进行训练，节点之间有严格的前后次序，ring allreduce 算法中的 GPU 结构如图 3 所示。AI 训练平台应引入 Horovod 等支持 ring allreduce 弹性训练[16]的分布式训练框架，为各个训练任务的 operator 设置动态可调的训练节点数量，进一步提高资源池的 GPU 利用率。

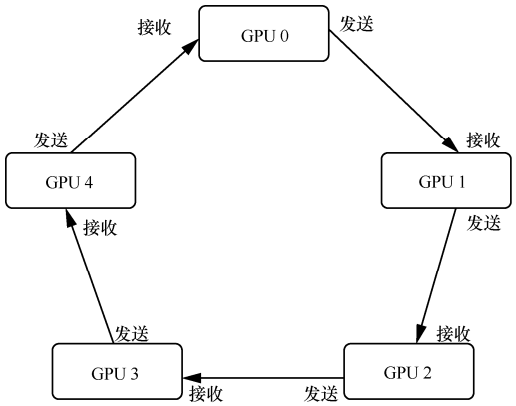

图 3　ring allreduce 算法中的 GPU 结构[17]

4　网络智能化中的模型部署

网络智能化的模型部署方案需要考虑以下几个因素。

- 时延，即模型推理环节消耗的时间，包括推理的计算时延，以及业务与推理模型之间的通信时延。
- 模型服务上线，即在网络中部署模型推理服务。
- 数据隐私，即推理是否在本地进行，或数据加密传递。
- 精度，即推理模型的精度与训练模型相比是否降低。

4.1　通信时延压缩

网络管理类的网络智能化应用一般对时效性要求不高，时效性要求在小时级/天级，此类应用一般采用在云上部署推理模型，充分利用云资源池算力。对于毫秒级和微秒级的实时类网络智能化场景，如空口物理层智能算法、核心网的移动性管理，必须采用在网元侧部署模型的方案以消除长距离通信时延，具体而言，网元侧需要预先部署推理框架，模型在 AI 训练平台完成训练后下发至相关网元，经过编译后加载到计算芯片内存，在本地对数据进行预测。需要说明的是，部分网络智能化场景的数据概率分布会随时间推移或地域不同发生偏移，影响推理模型效果，因此，模型需要周期性重新训练和部署，或按地域部署对应的模型。

4.2　模型加速

网络智能化推理模型部署在网元侧时，需要解决有限的网元计算资源影响模型推理速度的问题。模型计算时延与模型结构和模型大小直接相关，机器学习模型一般对算力要求不高，计算时延较低，无须专门进行模型加速处理；如果部署的是深度学习模型，则通过选择 MoblieNet 等专门为资源受限场景设计的小模型[18]、模型压缩、

编译优化 3 种方法减少模型推理对网络设备有限算力和内存的需求，从而降低模型计算时延。

模型压缩主要方法如图 4 所示，有以下 3 类。

（1）模型剪枝

深度学习模型中存在权重接近 0 的参数，这些参数对模型输出影响不大，因此可对训练模型中特定层内的权重低于设定阈值的参数进行裁减[19]，减少层间参数的连接和网络规模，剪枝之后需要对裁剪后的模型进行重新训练。

（2）参数量化

参数量化即降低模型参数精度（如参数精度从 fp32 降到 int8 将使模型大小降为原模型的 1/4），减少模型计算过程中对内存的访问，提升计算芯片每条指令中的数值数量，从而加速模型计算速度[20]。

（3）知识蒸馏

通过调整算法中的“温度”超参、进行 softmax 变换[21]，并将完成训练的大模型同另一个小的模型进行联合训练，较小模型能在大模型的监督下获得大模型的泛化能力，从而得到可用于推理的小模型。

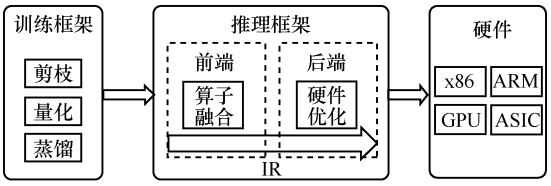

图 4　模型压缩主要方法

编译优化一方面通过推理框架中的前端（优化器）对训练后的模型进行优化[22]，如图 4 所示，通过算子融合合并模型中的部分层或相同结构降低模型计算量和冗余度，从而降低计算过程中因访问内存导致的计算芯片等待时间；另一方面通过推理框架的后端（编译器）对网元侧的专用计算芯片进行适配，优化模型在专用硬件上的计算效率，实现降低计算时延的效果。

4.3　模型服务上线

模型推理需要与网络系统深度融合，推理模型服务上线运行图如图 5 所示，推理模型与数据预处理和后处理功能模块一起打包为软件开发工具包（software development kit，SDK）或镜像上传到模型仓库，云上集中推理部署工具或网元侧推理部署工具从模型仓库拉取所需要的 SDK 或镜像，并以 SDK 软件集成、容器、serverless 等形式将推理服务部署上线[23-24]，推理服务对外提供 API，用户通过 HTTP/RPC 向其发起服务请求。推理服务监控模型运行状态，模型精度低于设定阈值将触发新一轮模型训练[25]。

网络管理类、对外服务类等非时延敏感型网络智能化推理服务宜采用容器方式在云上集中部署，其中，低频、突发、定时启动等非长期运行的推理服务可采用 serverless 的弹性容器或函数方式部署。与网元业务处理逻辑强相关的时延敏感

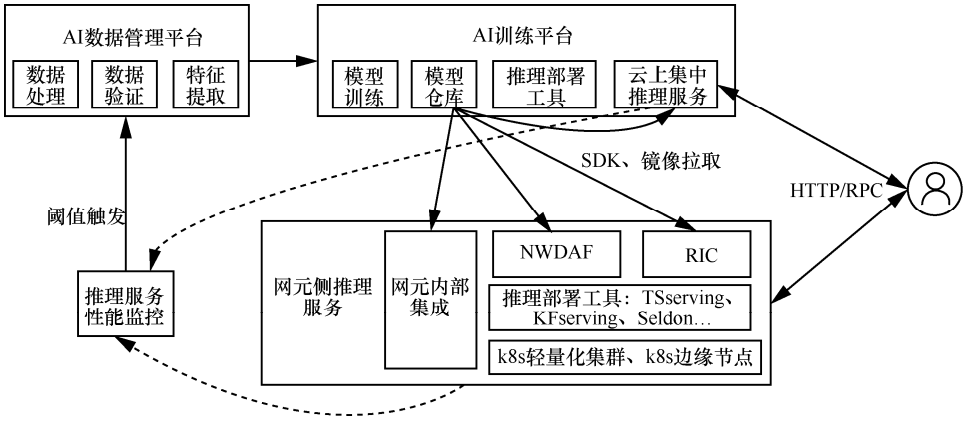

图 5　推理模型服务上线运行图

型网络智能化推理服务需要在网元内部或靠近网元部署，运营商网络设备上的计算芯片既有较为通用的 x86 CPU、ARM CPU、GPU，也有 FPGA、ASIC 等专用芯片，编译器需要针对不同种类的芯片进行定制化适配，使得推理服务能以 SDK 方式与网元软件系统集成，但架构封闭难以适配或适配改造成本较高的网元应将推理服务外置部署，即采用网络数据分析功能（network data analytics function，NWDAF）和无线网智能控制器（RAN intelligent controller，RIC）等标准组织推动的在网元侧构建专用 AI 控制器的方案，专用 AI 控制器的资源管理应采用 k3s 等轻量级 k8s 集群[26]或 KubeEdge[27]、OpenYurt[28]等云原生边缘 k8s 节点方案，使得推理服务能以容器或 serverless 方式部署并使用相关推理部署工具，与云上集中部署保持一致的管理和使用体验。

此外，对于推理模型精度这一问题，由于推理精度同时受网络智能化业务要求、硬件资源、压缩优化方法三者的共同影响，应根据具体情况进行平衡。对于数据隐私问题，目前可以以数据通过 IP 承载网等运营商内网传输的方式解决，如后续有明确需求则需要考虑数据加密或联邦学习等隐私方案，但需要做好安全与性能的平衡。

5 网络智能化的生态构建

网络智能化想要真正成为运营商网络发展的原生动力，就不只需要数据和算力，更需要 AI 算法的支持，三者缺一不可。AI 算法是本轮人工智能热潮的根本推动力，目前其发展呈现 3 个态势：AI 算法的研究主力军从以学术界为主，发展到学术界和产业界并重；AI 算法纷纷开源，极大地提高了研究效率和技术迭代速度；AI 算法与领域知识更加紧密地结合，AI 在更多行业得以落地。

网络智能化要充分认识和借助这种发展趋势，打破 CT-IT 产业之间巨大的思维代沟，填补 AI 产业界与长尾化、碎片化的网络智能化应用场景之间的价值鸿沟，构建一个健康繁荣的产业生态圈。目前，产业界已有美国电话电报公司（AT&T）的 Acumos 开源平台在此方面提供了借鉴[29]，其支持各主流计算框架和编程语言、提供可视化模块化的开发环境、提供模型库和模型商店，降低了 AI 开发的门槛，有利于 CT 行业的技术人员使用平台进行模型训练和数据挖掘。

对于运营商而言，应以自有 AI 平台为基础，为更多用户和产业伙伴提供开放能力：开放网络数据集，并向网络设备供应商、AI 硬件提供商、AI 软件开源组织、高校研究机构、AI 技术服务商等合作伙伴及个人开发者提供真实网络测试环境；提供低代码和图形化开发环境，提供模型市场、特征仓库、详细使用文档、最佳实践，推动公司内部相关员工提升数据分析与 AI 算法的应用能力，推动网络智能化问题以内部研发方式解决；基于模型市场和特征仓库，与产业各方探索中/高价值的网络智能化场景的“需求挖掘/发布——内、外部协同开发测试——上线测试/运行——成果分享”的合理商业模式。通过多种手段，将 AI 产学界的技术能力吸引到网络智能化领域当中，提升 CT 行业内部的 AI 应用能力，在越来越多的网络设备、管理系统、业务系统中集成 AI 能力，最终实现网络的泛在智能。

6 结束语

首先，本文针对网络智能化提出了一套 AI 工程化的技术解决方案，但目前 AI 技术在算法和部署框架等方面的发展日新月异，各种技术百花齐放，未来需要对方案中每个环节进行更加细致的技术方案研究和比选。其次，网络智能化在数据采集侧和部署侧属于典型的分布式架构，和边缘计算的场景基本重合，目前边缘计算也是 ICT 产业重要的研究方向，因此网络智能化的研究工作应充分借鉴边缘智能领域的经验。再次，本文的分析建立在网络智能化训练为集中式训练的基

础上，随着网络智能化的不断发展，网元侧、网元间闭环训练的需求会陆续浮出水面，因此下一步需要对远程分布式训练、联邦学习等技术进行研究。再其次，AI 的加速技术不只限于部署环节，在数据侧和训练侧存在着大量可加速空间。最后，进一步研究 MLOps 等 AI 工程的自动化流水线技术方案，不断提升网络智能化大规模应用的工程效率。从技术、标准、产业、应用、生态等方面来看，网络智能化目前仍处于初级阶段，有大量的问题等待攻克，有大量的需求等待满足，需要 CT 产业界和 AI 产学界更加紧密的合作，不断提升网络智能化水平，为我国构筑更加智能开放的信息基础设施。

参考文献：

[1] JIA W L, WANG H, CHEN M H, et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning[J]. arXiv: 2005. 00223, 2020.

[2] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[3] 欧阳晔, 王立磊, 杨爱东, 等. 通信人工智能的下一个十年[J]. 电信科学, 2021, 37(3): 1-36.
OUYANG Y, WANG L L, YANG A D, et al. Next decade of telecommunications artificial intelligence[J]. Telecommunications Science, 2021, 37(3): 1-36.

[4] 华为技术有限公司. 华为自动驾驶网络解决方案白皮书[R]. 2020.
Huawei Technologies Co., Ltd.. Huawei's white paper on autonomous driving network solutions[R]. 2020.

[5] 中兴通讯股份有限公司. 中兴自主进化网络白皮书[R]. 2020.
ZTE Technology Co., Ltd.. ZTE's white paper on autonomous evolution network[R]. 2020.

[6] 邓超, 王斌, 朱琳, 等. 人工智能在电信运营中的典型应用实践[J]. 信息通信技术与政策, 2019(7): 34-38.
DENG C, WANG B, ZHU L, et al. Typical applications of artificial intelligence in telecom operation[J]. Information and Communications Technology and Policy, 2019(7): 34-38.

[7] 冯俊兰. 5G 自身智能化及赋能智能产业之路[J]. 电信工程技术与标准化, 2020, 33(1): 1-8.
FENG J L. Intelligent 5G network and 5G+AI applications[J]. Telecom Engineering Technics and Standardization, 2020, 33(1): 1-8.

[8] 程强, 刘姿杉. 数据驱动的智能电信网络[J]. 中兴通讯技术, 2020, 26(5): 53-56.
CHENG Q, LIU Z S. Data empowered intelligent communication networks[J]. ZTE Technology Journal, 2020, 26(5): 53-56.

[9] RATNER A, BACH S H, EHRENBERG H, et al. Snorkel: rapid training data creation with weak supervision[J]. The VLDB Journal: Very Large Data Bases: a Publication of the VLDB Endowment, 2020, 29(2): 709-730.

[10] BRECK E, CAI S Q, NIELSEN E, et al. The ML test score: a rubric for ML production readiness and technical debt reduction[C]//Proceedings of 2017 IEEE International Conference on Big Data (Big Data). Piscataway: IEEE Press, 2017: 1123-1132.

[11] CNCF. Cloud Native Interactive Landscape[EB]. 2021.

[12] Kubeflow. Kubeflow Overview[EB]. 2021.

[13] Github. Volcano[EB]. 2021.

[14] Kubernetes. Scheduling Framework[EB]. 2021.

[15] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv: 2005. 14165, 2020.

[16] Horovod. Elastic Horovod[EB]. 2021.

[17] GIBIANSKY A. Bringing HPC techniques to deep learning[EB]. 2021.

[18] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB]. 2017: arXiv: 1704. 04861, 2017.

[19] HAN S, MAO H Z, DALLY W. Deep compression: compressing deep neural network with pruning, trained quantization and huffman coding[J]. ICLR. 2015.

[20] INTEL. Accelerate lower numerical precision inference with Intel® deep learning boost[EB]. 2021.

[21] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7):38-39.

[22] LI M Z, LIU Y, LIU X Y, et al. The deep learning compiler: a comprehensive survey[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(3): 708-727.

[23] Github. KFServing: predict on an InferenceService with TensorFlow model[EB]. 2021.

[24] TensorFlow. TFServing：train and serve a TensorFlow model with TensorFlow serving[EB]. 2021.

[25] Google Cloud. MLOps：continuous delivery and automated pipelines in machine learning [EB]. 2021.

[26] Rancher. K3s[EB]. 2021.

[27] GitHub. KubeEdge[EB]. 2021.

[28] Gitee. OpenYurt[EB]. 2021.

[29] 刘腾飞, 李奥. Acumos: 一种人工智能开放平台[J]. 邮电设计技术, 2018(12): 46-50.
LIU T F, LI A. Acumos—an artificial intelligence open platform[J]. Designing Techniques of Posts and Telecommunications, 2018(12): 46-50.

[作者简介]

朱明伟（1986− ），男，现就职于中国移动通信集团设计院有限公司，主要研究方向为 5G、边缘计算、云原生。