ICS 号
中国标准文献分类号

T/CES

团 体 标 准

T/CES XXX-XXXX

电力人工智能平台总体架构及技术要求

The overall architecture and technical requirements of the artificial intelligence platform in power industry

（征求意见稿）

XXXX-XX-XX 发布 XXXX-XX-XX 实施

中国电工技术学会发布

目　　录

前　言

本标准按照 GB/T1.1—2009《标准化工作导则 第 1 部分 标准的结构与编写》给出的规则起草。

请注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别这些专利的责任。

本标准由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本标准起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司、北京国网信通埃森哲信息技术有限公司、国网重庆市电力公司电力科学研究院、四川大学、安徽继远软件有限公司、四川中电启明星信息技术有限公司、国网重庆市电力公司、中国电力科学研究院有限公司。

本标准主要起草人：李强、赵峰、邱镇、刘迪、庄莉、李炳森、廖逍、黄晓光、刘永清、谢可、向辉、谭洪恩、苏少春、杨迎春、周孔均、王晓东、钟加勇、彭舰、王秋琳、黄飞虎、王金策、田鹏、吕小红、厉仄平、宋卫平、苏江文、费长顺、邢国用、丘志强、禹国印、杨成、王蓓、张琳瑜、崔迎宝、刘璟、刘晓飞、阎誉榕、宫晓辉、尹玉、周伟、梁翀、李温静、王乖强、伍臣周、王晓辉。

本标准为首次发布。

人工智能平台架构标准规范

1 范围

本标准规定了人工智能平台建设架构、技术要求以及规定了电力人工智能算法模型在共享应用中所涉及的文件、描述文档和使用方式的基本要求。

本标准适用于人工智能平台的规划、设计、开发、运维和算法模型应用。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1—2020 标准化工作导则 第1部分：标准的结构和编写

GB/T 5271.1—2000 信息技术 词汇 第1部分：基本术语

GB/T 5271.28—2001 信息技术 词汇 第28部分：人工智能基本概念与专家系统

GB/T 5271.29—2006 信息技术 词汇 第29部分：人工智能语音识别与合成

GB/T 5271.31—2006 信息技术 词汇 第31部分：人工智能机器学习

GB/T 5271.34—2006 信息技术 词汇 第34部分：人工智能神经网络

DA/T 77-2019 纸质档案数字复制件光学字符识别OCR工作规范

TAF—WG7—AS0043—V1.0.0:2019 智能产品语音识别测评方法 第二部分：智能音箱

3 术语和定义

下列术语和定义适用于本文件。

3.1 人工智能 artificial intelligence

一门交叉学科，通常视为计算机科学的分支，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统。

[GB/T 5271.28—2001,定义28.01.01]

3.2 训练（在神经网络中） training（in neural network）

教会神经网络在输入值的样本和正确输出值之间作出结合的步骤。

[GB/T 5271.34—2006,定义34.03.18]

3.3 样本数据 sample data

其具备的特征能够正确反映总体数据情况的一部分个体数据。

3.4 推理 inference

从已知前提导出结论的推理方法。

注1：在人工智能领域中，前提是事实或规则。

注2：术语“推理”既指过程也指结果。

[GB/T 5271.28—2001,定义28.03.01]

3.5 深度学习框架 deep learning framework

一种支持深度学习模型设计、训练和推理的软件。

3.6 资源 resource

执行所要求的操作而必需的数据处理系统的任何组成部分。
[GB/T 5271.1—2000,定义 01.01.23]

3.7 配置 configuration

信息处理系统中的硬件和软件组织和互连起来的方式。
[GB/T 5271.1—2000,定义 01.01.26]

3.8 接口 interface

两个功能单元共享的边界，它由各种特征（如功能、物理互连、信号交换等）来定义。
[GB/T 5271.1—2000,定义 01.01.38]

3.9 计算机视觉 computer vision

功能单元获取、处理和理解可视数据的能力。
[GB/T 5271.28—2001,定义 28.01.19]

3.10 语音识别 speech recognition

利用功能单元进行的，从语音信号到语音内容的某一表示的转换。
[GB/T 5271.29—2006,定义 29.01.30]

3.11 数据标注 data annotation

通过分类、画框、注释等方式对数据进行标记，形成可供计算机分析识别的数据。

3.12 模型训练 model training

基于一系列数据集、学习框架等，并通过最优的建模方法和参数得到一个算法模型的过程。

3.13 OCR 光学字符识别 optical character recognition

将图片、照片上的文字内容直接转换为可编辑文本的一种技术。
[DA/T 77—2019,定义 3.3]

3.14 人脸识别 face recognition

基于人的脸部特征信息进行身份识别的一种图像识别技术。

3.15 知识图谱 knowledge graph

显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

3.16 自然语言处理 natural language processing

研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。
[TAF—WG7—AS0043—V1.0.0:2019，定义 3.1.7]

3.17 目标检测 target detection

一种确定图像中目标的类别和位置属性信息的技术。

3.18 模型运行脚本 model script

使用脚本语言所写的用于部署算法模型的程序。

3.19 模型文件部署 model file deployment

根据提供的算法模型源文件、模型配置文件等，结合相应的开发语言、深度学习框架、模型配置说明、运行依赖说明，手动完成运行框架、依赖环境安装和配置文件调整，实现算法模型的成功部署，完成相应推理服务。

3.20 容器部署 docker deployment

利用容器封装算法模型源文件、运行框架、依赖环境、配置文件等资源，通过容器方式实现算法模型的成功部署，完成相应推理服务。

3.21 预言模型标记语言(PMML) predictive model markup language

用于呈现数据挖掘模型，支持在不同的应用程序之间共享预测分析模型。

3.22 系统负载（SL） system load

一种用于描述 CPU 当前负荷的指标，具体统计了运行和等待运行的进程/线程数。

3.23 电力人工智能平台 electric artificial intelligence platform

整合多种机器学习计算框架，具备数据管理、模型开发、平台服务、运营管理、运维管理、安全管控及跨域协同等功能的人工智能软件系统。平台支持样本选择、模型创建、模型训练及服务发布的全流程一站式管理，包含面向电力领域的模型库及数据样本库。

3.24 电力专用模型 electricity dedicated model

支撑国家电网有限公司电力生产的专用模型。

3.25 电力专用模型任务 electricity dedicated model task

图像分类、目标检测、图像分割、视频分类、行为检测、单目标跟踪、多目标跟踪、数值分类、数值回归、数值聚类等电力行业专用的模型服务。

4 缩略语

下列缩略语适用于本文件。
OCR：光学字符识别（Optical Character Recognition）
ROC：受试者工作特征曲线(Receiver Operating Characteristic Curve)
AUC：ROC曲线下的面积（Area Under Curve）
BLEU：双语互译质量评估辅助工具(bilingual evaluation understudy)
FRR：拒识率（False Rejection Rate）

SER：句子识别错误率（Sentence Error Rate）
FAR：误识率（False Acceptance Rate）
FPR：伪正类率(False Positive Rat,)
TPR：真正类率(True Positive Rate)
F1值：精确率和召回率的调和均值(F-Measure)
TP：真阳性(True Positive)
FP：假阳性(False Positive)

5 架构要求

5.1 概述

人工智能平台架构要求包括总体架构、技术架构、数据架构。

1) 总体架构：规定平台总体结构以及和其它平台的关系；
2) 技术架构：规定平台主体组件的技术选型和技术范围；
3) 数据架构：规定平台数据的架构。

5.2 总体架构

人工智能平台架构应包括：训练环境、模型库、样本库、运行环境、管理中心和统一服务门户 6 部分。生产环境中的模型库仅可包括验证通过的模型，生产环境的样本库仅包括生产数据相关样本，训练环境样本库包括训练验证用的样本，人工智能平台总体架构如下图 1：

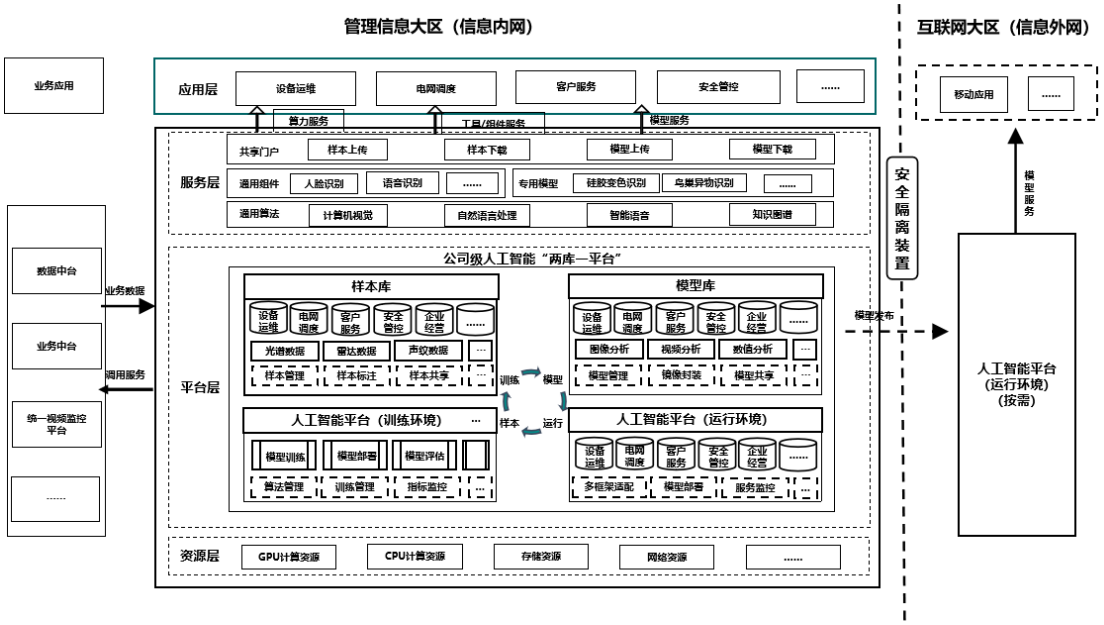

图 1 人工智能技术应用总体架构

人工智能平台总体架构要求为：

1) 人工智能平台的数据样本应直接来源于数据中台；
2) 人工智能平台服务层中的通用组件，专用模型应部署在资源层，为应用层提供算力服务，中间件和模型服务；
3) 人工智能平台应统一服务门户主要分为服务层和平台层，服务层提供通用算法(如：计算机视觉，自然语言处理，智能语音，知识图谱)，平台层分为样本库与模型库，提供人工智能训练与人工智能运行平台，服务层供平台层进行调用；
4) 样本库应支撑算法训练，训练结果应输出至模型库；

5) 模型库应支持模型管理，镜像封装，模型共享，模型发布；
6) 人工智能平台主要在管理信息大区部署，在互联网大区仅部署模型管理组件和业务模型服务组件，用于支撑互联网大区人工智能业务应用；
7) 人工智能平台互联网大区模型从管理信息大区模型库中获取，模型同步通过隔离装置穿透管理 信息大区和互联网大区。

5.3 技术架构

人工智能平台技术架构要求主要包括：应用层、服务层、能力层和资源层的要求。管理中心处于服务层和能力层。人工智能平台技术架构如下图：

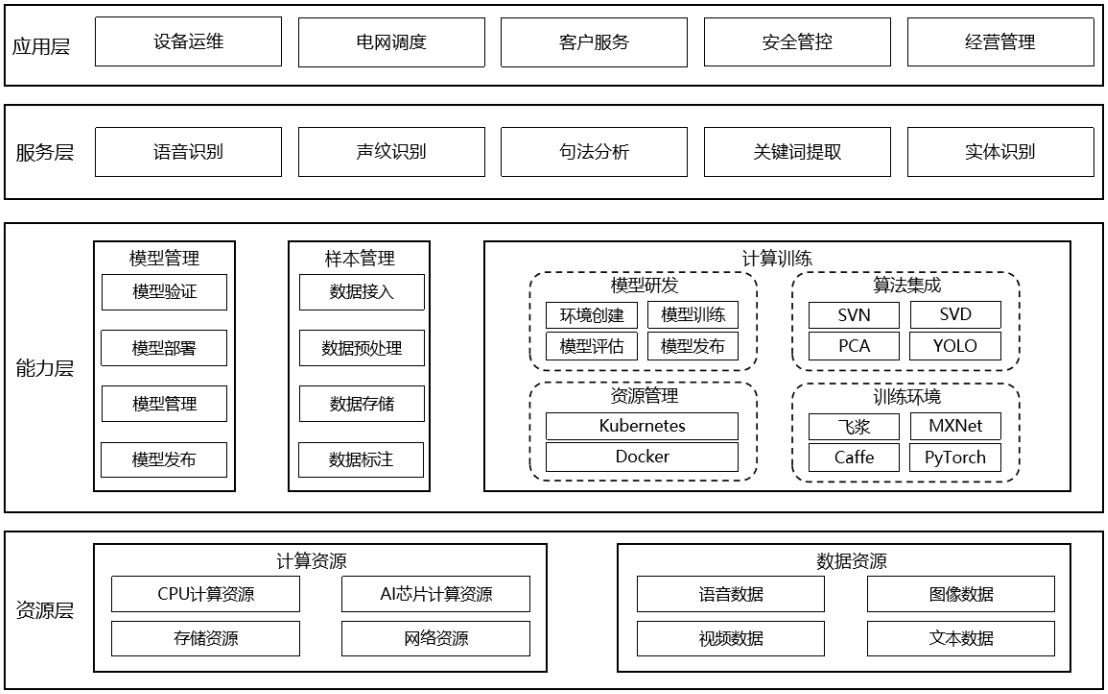

图 2 人工智能平台技术架构

人工智能平台技术架构要求为：

1) 应用层通过 API 和 SDK 两种方式使用平台的服务，通过 GUI 的方式访问能力层提供的能力；
2) 模型管理中的模型部署应采用容器化部署方式，应使用 Kubernetes 和 Docker 组件；
3) 算法集成应包括算法模块：SVM、SVD、PCA、YOLO 等；
4) 学习框架应包括主流的开源深度学习框架；
5) 数据接入宜采用 Kettle 和 Sqoop 等组件；
6) 特征预处理中的特征提取宜支持 Numpy、Scikit-learn 等组件；
7) 数据存储宜支持 Ceph 等组件；
8) 平台应支持多租户；
9) 配置管理模块宜采用配置引擎的方式开发；
10) AI 芯片计算资源包括但不限于：FPGA、ASIC 等；

5.4 数据架构

将数据按照图像、视频、语音、文本等类型存储，经过数据标注后形成样本库。人工智能平台数据架构如下图：

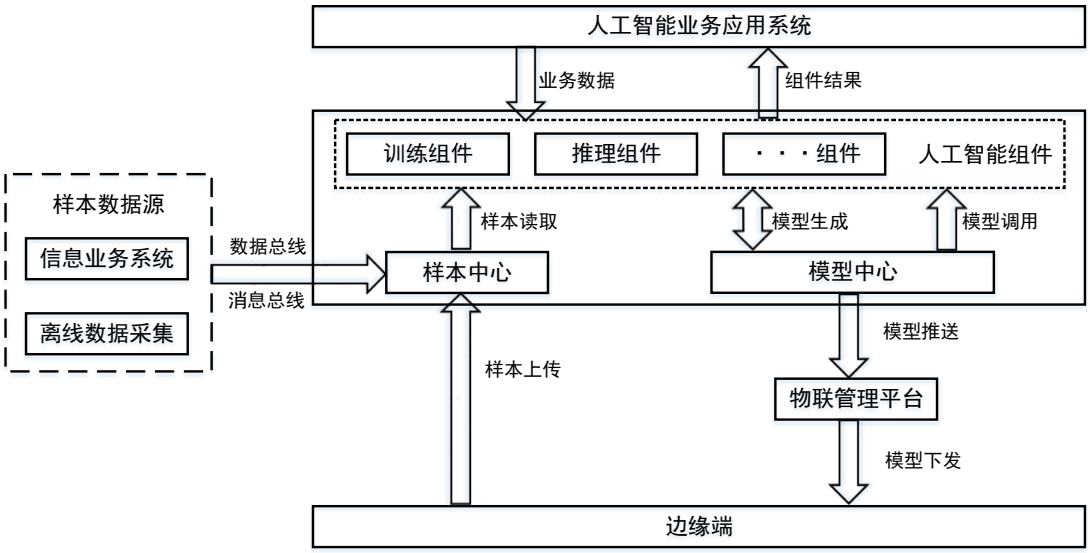

图 3 人工智能平台数据架构

人工智能平台数据架构要求为:

1) 样本数据源来源于信息业务系统或通过离线数据采集，经数据总线、消息总线推送至样本中心；
2) 样本中心通过对样本数据源进行数据标注等预处理操作，形成用于训练模型的样本数据，提供给训练组件、推理组件、回归测试组件等；
3) 人工智能组件读取样本中心的数据，调用模型中心的模型，并将训练好的模型入库存储到模型中心；调用模型中心的模型，使用业务应用推送的业务数据，将组件训练结果反馈给业务应用系统；
4) 样本中心可实现样本同步；
5) 模型中心可实现模型同步，并提供模型调用服务；
6) 模型中心的模型推送至物联管理平台，通过物联管理平台下发到边缘端；
7) 边缘端的数据样本可通过物联管理平台上传至样本中心；
8) 边缘端使用模型后产生的拒真/存伪或其他异常结果，可作回测试样本上传至样本中心，用于迭代优化模型。

6 技术要求

6.1 功能要求

6.1.1 训练中心

训练中心负责训练全过程管理，包含训练任务项目管理、算法训练、资源调度、算法评估、算法文件管理等功能。

1) 项目管理：应支持训练任务流程管理，包括项目创建、删除、状态监测与查询，以及不同任务之间的切换；
2) 训练框架：应支持兼容 PyTorch、TensorFlow、MXNet、Caffer 等主流深度学习框架；
3) 训练方式：应支持 Notebook 式、命令式、GUI 任务式进行训练；

4) 资源调度：宜支持 GPU 显存分配；
5) 算法评估：不同类型算法应支持不同评价指标的计算与展示，分为以下类别：
   a) 分类算法：计算准确率、召回率、F1 值；
   b) 聚类算法：计算准确率、精确率、召回率、紧密性、间隔性；
   c) 回归算法：计算均方误差、均方根误差、决定系数、校正决定系数；
6) 算法管理：应支持算法文件按类别、来源等分类上传、下载、保存。

6.1.2 模型中心

模型中心负责对模型文件统一管理、提供模型推理服务。

1) 模型统一管理：应支持接收模型、发送模型、删除模型、版本管理、收藏模型；
2) 模型推理服务：服务应支持一键式自动部署发布、引导式发布，支持 GPU 显存分配管理，提供服务测试、接口管理、请求审核等能力和服务。

6.1.3 样本中心

样本中心可对数据文件统一管理、提供数据标注、数据处理服务。

1) 样本统一管理：应支持接入数据、发送数据、数据集管理；
2) 数据导入方式：应支持本地、数据库、HDFS、FTP、NFS 等方式；
3) 数据预处理服务，需提供通用数据预处理方法，缺失处理、异常处理，对不同类型数据应支持不同类型的预处理方法：
   a) 文本数据：标记化、归一化、替换；
   b) 图片、视频数据：特征提取、图片增强、去噪；
   c) 语音数据：特征提取、数据增强、预加重、分帧；
4) 数据标注服务：标注流程管理与服务，支持文本、图像、语音、视频等类型数据标注，宜提供基于已有模型的智能标注。

6.2 性能要求

6.2.1 响应时限

人工智能平台响应实现需符合以下规定：

1) 在系统负载小于 80%时，前端页面响应时间不超过 5 秒；
2) 在系统负载小于 80%时，后端服务响应时间不超过 3 秒；
3) 在系统负载小于 80%时，算法训练任务创建响、推理服务申请应时间不超过 3 分钟；
4) 在系统负载小于 80%时，推理调用响应时间不超过 3 秒。

6.2.2 可靠性

人工智能平台在不能抗力环境下，应满足 7×24 小时服务不中断，提供冗余的网络设备、通信线路、系统硬件，保证容错率和可用性，具体要求如下：

1) 数据完整性：存储节点发生故障时，冗余部分应包含完整数据；
2) 算力完整性：计算节点发生故障时。冗余部分应保证已有训练任务、推理服务结果不变；
3) 消息完整性：消息节点发生故障时，冗余部分应保证消息不丢失、不影响新消息的提交和消费；
4) 任务调度完整性：任务调度节点发生故障时，冗余部分应不影响任务调度管理与执行；
5) 网络完整性：网络故障后，经修复后，系统、系统任务、系统服务应自动继续运行。

6.2.3 可扩展性

人工智能平台应支持硬件、软件、系统扩展升级，且升级过程应不影响现有样本数据、训练任务、模型推理服务。

6.3 安全要求

安全性要求包括基础性安全要求、系统容错安全要求。

1) 基础性要求：
   a) 应符合 GB/T 18336—2015 《信息技术 安全技术 信息技术安全评估准则》的规定；
   b) 应符合 GB/T 18336—2015 《信息技术 安全技术 信息技术安全评估准则》的规定；
   c) 数据部分应符合《中华人民共和国数据安全法》的规定。
2) 容错性安全要求：算法模型支撑业务应用时，应考虑计算结果偏差超出范围、精度降低、响应超时对业务系统的不良影响。

6.4 硬件要求

运行人工智能平台系统、训练任务、推理服务、智能标注等组件的服务器应具备 GPU 算力资源，支持 CUDA 和 CUDNN 加速。各组件具体要求如下：

1) 训练任务：GPU 显存应不低于 16G，宜到达 32G 或以上，宜使用固态硬盘作训练样本的缓存；
2) 推理服务：GPU 显存应不低于 8G，宜到达 16G 或以上；
3) 系统运行：GPU 显存应不低于 8G，宜到达 16G 或以上；

7 算法模型共享要求

7.1 概述

算法模型的共享基础是具有算法模型文件和算法模型描述文件，本章给出对相关文件的要求。

7.2 算法模型文件

算法模型文件应包括但不限于：算法模型源文件； 算法模型配置文件；算法模型运行脚本文件。

7.2.1 算法模型源文件

算法模型源文件是由训练框架和数据集经过模型训练后得到的所有参数存储文件。常见的训练框架包括但不限于 Caffe、PyTorch、TensorFlow、MXNet 等。根据不同描述语言和开发框架，算法模型源文件对应要求如下：

1) Caffe 框架。算法模型源文件应包括存储模型参数 caffemodel 文件和存储模型网络结构的 prototxt 文件
2) PyTorch 框架。算法模型源文件应包括用于存储模型的网络结构和参数 pth 文件，。

3) TensorFlow 框架。模型源文件应包括 meta 文件、data 文件和 index 文件，meta 文件，data 文件存储模型的网络参数，index 文件为张量描述列表或网络结构和参数整合后的 h5/pb 文件。
4) MXNet 框架。算法模型源文件应包括 params 文件和 json 文件组成，params 文件存储模型参数，json 文件存储模型网络结构。
5) 其他框架。算法模型源文件可采用通用 PMML 预言模型标记语言描述，文件格式为 xml，可用于描述和存储算法模型。

7.2.2 算法模型配置文件

模型配置文件所描述的可调参数针对不同框架和算法模型类型具有不同的字段,可调参数应包含但不限于以下字段：

表 1 可调参数说明

序号	字段名称	含义说明
1	BatchSize	描述批处理参数
2	Width	描述输入数据的宽度
3	Height	描述输入数据的高度
4	Channel	描述输入数据通道数
5	GPU	指定 GPU 的 ID

7.2.3 算法模型运行脚本文件

算法模型运行脚本文件应提供对模型加载、训练和推理的代码支持，并以模型配置文件所指定的参数运行模型文件。

7.3 算法模型描述性文档

算法模型描述性文件是对算法模型的量化描述应包括但不限于：数据集描述文档、属性描述文档、性能描述文档。

7.3.1 数据集描述文档

共享算法模型应当提供算法模型在训练、测试、验证各关键环节的数据集描述。描述的内容应当包括但不限于数据来源、数据集类型、数据标注格式、数据集样本量，数据格式等。

7.3.2 属性描述文档

共享算法模型应该提供算法模型的相关属性信息包括但不限于版本信息、研发作者、发布时间、模型框架、编码语言、运行环境、运行硬件配置。

7.3.3 性能描述文档

性能描述文档是对算法模型的性能参数进行描述，根据不同模型类型，提供性能描述核心指标包括但不限于以下类型指标：

1) 目标检测模型的性能描述参数宜包含但不限于准确率 （Accuracy）、精确率（Precision）、召回率（Recall）、平均正确率（AP）、平均精度均值 mean Average Precision(mAP)、交除并（IoU）、ROC、AUC 和 Precision-recall 曲线。
2) 语音识别模型的性能描述宜包括词错误率 WER，句错误率 SER。
3) OCR 识别模型的性能描述宜包括拒识率 FRR、误识率 FAR。

4) 人脸识别类模型的性能描述参数宜包含但不限于准确率 （Accuracy）、精确率（Precision）、召回率（Recall）、平均正确率（AP）、平均精度均值 mean Average Precision(mAP)、F1 值、曲线 ROC、伪正类率 FPR、真正类率 TPR。
5) 自然语言处理模型的性能描述参数宜包含但不限于准确率（Accuracy）、 精确率（Precision）、召回率（Recall）、F1 值、ROC、AUC、BLEU 和偏差与方差。
6) 知识图谱类模型的性能描述参数宜包括但不限于准确率、覆盖率、响应时间。

7.4 算法模型共享应用方式

算法模型应支持以部署方式进行应用，部署方式应至少支持模型文件部署、容器部署两种部署方式中的一种。

算法模型宜支持以二次训练方式应用。根据算法模型的开发语言、深度学习训练框架、数据集和描述文档等内容，应支持对算法模型进行相应参数调优。

算法模型应支持以接口方式对外提供服务，接口方式应至少支持 API 和 SDK 两种接口方式中的一种。