ICS 35.240
L 70

团　体　标　准

T/CESA 1037—2019

信息技术 人工智能 面向机器学习的系统框架和功能要求

Information technology- Artificial intelligence- Framework and functional requirements of system for machine learning

2019－04－01 发布　　2019－04－01 实施

中国电子工业标准化技术协会　　发 布

目　　次

前　言

本标准按照GB/T　1.1－2009《标准化工作导则　第1部分：标准的结构和编写》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。

本标准由中国电子技术标准化研究院提出并归口。

本标准起草单位：中国电子技术标准化研究院、华为技术有限公司、第四范式（北京）技术有限公司、浪潮软件集团有限公司、重庆邮电大学、广州中科凯泽科技有限公司、深圳前海微众银行股份有限公司、曙光信息产业股份有限公司、中电莱斯信息系统有限公司、海尔优家智能科技（北京）有限公司、山东省计算中心（国家超级计算济南中心）、深圳市和讯华谷信息技术有限公司、中国电信集团有限公司、深圳云天励飞技术有限公司、广州广电银通金融电子科技有限公司、上海拟合智能科技有限公司、成都四方伟业软件股份有限公司、中国医学科学院生物医学工程研究所、华夏芯(北京)通用处理器技术有限公司、浪潮（北京）电子信息有限公司、北京航空航天大学、重庆中科云从科技有限公司、北京深醒科技有限公司、广州广电运通金融电子股份有限公司、万达信息股份有限公司、江苏中堃数据技术有限公司、北京市商汤科技开发有限公司、北京眼神科技有限公司、北京航天自动控制研究所、深圳区块大陆科技有限公司、上海孚恩电子科技有限公司等。

本标准主要起草人：代红、董建、张群、王燕妮、光亮、符海芳、汪小娟、马珊珊、张冠一、杜宁、王一鹤、王功明、黄先芝、张焱、黄庆卿、庞宇、黄启军、张栋栋、阳马生、宋怀明、郑少秋、蒋锴、陆保国、孙雨新、赵志刚、武鲁、陈宇、洪晶、杨震、李洁、胡文泽、王孝宇、程冰、梁添才、赵清利、王挺、曾理、龙吟、徐圣普、蒲江波、刘军、王超、李洪革、李军、刘君、田永会、林冠辰、陈良旭、陈诚、魏清、张海静、蒋慧、尚可、宋方方、王丽娜、徐颂、杨扬、杨杰、徐坚强等。

信息技术　人工智能　面向机器学习的系统框架和功能要求

1　范围

本标准给出了面向机器学习的系统框架，规定了系统整体及各组件的功能要求。

本标准适用于各领域人工智能系统及解决方案的规划、设计，可作为评估、选型及验收的依据。

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　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

3　术语和定义

GB/T　5271.31-2006，　GB/T　5271.34-2006界定的以及下列术语和定义适用于本文件。

3.1

机器学习模型　machine　learning　model

采用机器学习方法建立的输入与目标输出联系的计算模型。

3.2

机器学习引擎　machine　learning　engine

提供机器学习开发及运行的计算组件。

3.3

算法服务　algorithm　service

算法在推理部署后的运行态。

注：算法服务接受用户的应用请求，对输入数据进行处理，返回处理结果。

3.4

作业　job

机器学习训练或推理任务的逻辑组合。

注：一个作业属于且仅属于某一个资源池，一个作业包括一个或多个任务。

3.5

任务 task

被调度的训练/推理对象。

注：任务用于完成一个相对独立的业务功能。一个任务属于且仅属于一个作业。

3.6

资源池 resource pool

各类资源的集合。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能(Artificial Intelligence)

ARM：高级精简指令集机器(Advanced RISC Machine)

ASIC：专用集成电路(Application-Specific Integrated Circuit)

CPU：中央处理器(Central Processing Unit)

DAG：有向无环图（Directed Acyclic Graph）

FPGA：现场可编程逻辑门阵列(Field Programmable Gate Array)

GPU：图形处理器(Graphic Processing Unit)

IDE：集成开发环境 (Integrated Development Environment)

NLP：自然语言处理（Natural Language Processing）

REST：表现层状态转换（Representational State Transfer）

5 系统框架

5.1 概述

面向机器学习的系统框架见图1，包括机器学习、多算法管理、异构资源调度等核心能力，提供数据预处理、特征工程、模型开发、模型训练、模型推理服务发布的端到端能力。

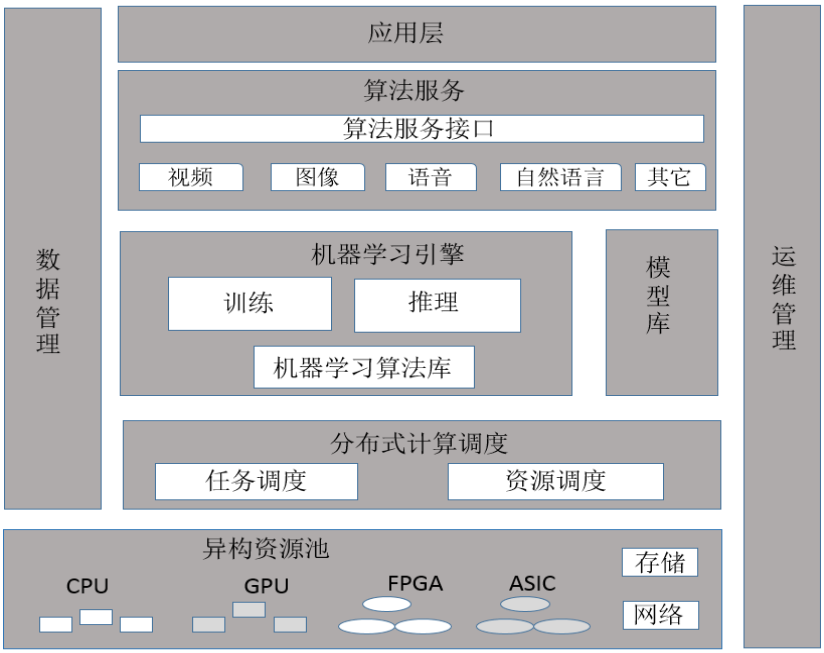

图1　面向机器学习的系统框架

系统提供应用场景所需的各类基础设施，包括各种异构计算单元（例如CPU、GPU、FPGA、ASIC等），存储（例如分布式云存储），网络等资源，结合实际任务进行分布式调度，提供按需分配、统一智能编排、动态调度、弹性伸缩及自动管理等能力。通过统一管理、动态更新模型库，提供机器学习算法的开发、训练、部署、运行和管理能力。各类机器学习算法通过有效组合，构成视频、图像、语音、自然语言处理等领域的算法服务，支持各领域AI应用。

系统具备数据管理（数据输入、输出、元数据管理、数据生命周期管理等）和运维管理等能力（多用户管理、多租户管理、监控告警等）。

系统在应用层、算法服务、机器学习引擎、模型库、分布式计算调度、异构资源池、数据管理、运维管理等模块间及模块内提供接口支持信息传递及互操作。

5.2　数据管理

数据的全生命周期管理，包含收集、预处理、分析、可视化及访问过程。数据管理包括各类数据源（结构化、半结构化、非结构化数据）的接入，中间数据的管理、最终数据的管理、元数据的管理、数据质量管理、数据的标注，并提供统一的数据管理工具等。

5.3　异构资源池

异构资源池统筹管理机器学习所需的各类计算、存储和网络资源。计算资源可包含不同类别的计算硬件，如CPU、GPU、FPGA、ASIC等，提供适合与应用场景的运算资源（如高效节能的处理器）。存储资源包括不限于缓存、主存、辅存等各级存储。网络资源包括但不限于异构单元间、计算节点间或集群间的互联网络。异构资源可以不同形态，如服务器、一体机、边缘计算节点、计算集群和云基础设施等方式提供。

异构资源池支持资源的动态调度、按需（数据规模、算法模型、实时性要求等）分配，满足计算任务的资源需求。资源池能够灵活集成各类计算单元并发挥其各自特性，构建异构计算节点，节点内实现

多粒度并行和任务调度，节点间实现协同调度，基于网络提供系统扩展并优化互联性能，支撑部署大规模AI计算框架和相关算法。

5.4 分布式计算调度

分布式计算调度分为任务调度和资源调度。任务调度是根据应用特性和运算需求将应用负载分解成任务，并配置任务执行的顺序和优先级。资源调度依据应用对资源的需求，将合适的异构资源分配给特定任务，满足任务对运算性能和时间的要求。多任务应共享资源以优化资源使用率。

分布式计算调度是人工智能系统的重要能力，在对计算、存储、网络等异构资源的统一纳管的基础上，系统根据资源标签将任务优化调度，以容器形式支持任务的大规模部署。

5.5 机器学习引擎

机器学习引擎基于各类机器学习算法（例如统计机器学习、深度学习、强化学习、迁移学习等）进行模型训练与推理。机器学习引擎支持开源的计算框架、算法库，兼容开源的主流接口，可根据商用的要求在企业版本中增强或优化。

机器学习算法库为算法提供安全可靠的管理功能，包括算法的注册、存储、下载、评价、优化以及用户鉴权、多版本管理、升级维护、运行监控等。

按算法需求，机器学习引擎提供特征数据的选择、提取、构建等功能。

5.6 模型库

模型库提供对机器学习模型开发和存储管理能力。

模型管理包括预置常用的AI模型，以及支持模型导入、导出、更新、发布、迁移、版本控制等功能。模型开发通过可视化辅助开发工具、多模型融合开发、模型二次训练等方式支持模型的开发与部署。

5.7 算法服务

算法服务是AI应用访问、利用机器学习能力和资源的主要方式。为满足应用场景的需求，系统提供各类通用算法服务（例如视频、图像、语音、自然语言处理等）。系统提供统一算法服务框架，进行服务管理、服务运行状态监控、服务上线等，并提供一致性的服务接口，供各领域上层应用调用。

5.8 运维管理

运维管理提供系统所需的基本运维（例如安装部署、扩展、监控、告警、健康检查、问题及故障定位、升级和补丁、备份恢复、操作审计等）及管理功能（例如资源管理、权限管理、用户管理、日志管理、配置管理、安全管理等）。

5.9 应用层

面向机器学习的AI系统可为各类应用（例如智慧交通、智能制造、智慧家庭、智慧城市、车联网等）提供支持，按应用需求提供系统资源，支持企业级、商业级的AI应用。

6 功能要求

6.1 总体要求

面向机器学习的AI系统应支持各领域不同场景AI应用对机器学习引擎、模型库、数据管理、异构资源池、分布式计算调度、算法服务、运维管理及接口等方面的要求。

系统从功能设计上应符合开放性（分层解耦、各层级可独立演进）、高可靠和可用性（避免单点故障、保证服务等级协议要求等）、统一性（统一的算法服务框架和接口框架）、可扩展性（从算法服务、机器学习引擎、资源供给、接口等各层面支持业务的灵活部署与弹性扩展）、易管理及运维、安全等核心要求。应支持业界主流的AI算法、编程模型、计算框架，针对各类使用场景设计应用层，并提供符合用户习惯的分析、开发和交互接口和开发文档支持。各模块间的接口应遵循业界常见的架构和协议（例如REST），兼容主流开源框架的接口。

6.2 数据管理

数据管理的要求包括：

a) 应支持各类数据源，包括结构化数据（例如传统关系型数据库），半结构化数据，非结构化数据（例如图片、音频、视频等）；
b) 应支持引入和解析常见文件和数据格式（例如 parquet、carbondata 等）；
c) 应支持对数据进行标注；
d) 应提供数据生命周期管理，可以对中间数据及产出数据进行增删改查及数据检索等操作；
e) 应提供数据访问及权限控制；
f) 应提供数据 IDE 工具，支持数据可视化；
g) 宜支持多种元数据管理方法（例如数据元信息生成、增删改查、元数据分类、血缘管理等）；
h) 宜支持多种数据预处理手段（例如数据的聚合、过滤、排序等）；
i) 宜支持常见的多媒体文件格式的元数据信息获取与管理。

6.3 异构资源池

异构资源池的要求包括：

a） 应支持 CPU 加异构计算单元的架构,通过异构计算显著提升计算性能；
b） 应支持异构资源池化，对异构资源模块进行统一管理、配置、编排，提升资源利用率；
c） 应支持以容器化提供资源，利用容器技术对异构资源提供统一调度和管理，支持对接主流深度学习计算框架；
d） 应支持资源池内 CPU 和异构计算单元的不同配比；
e） 应支持中心集群与边缘节点的统一管理；
f） 宜支持大规模高性能计算集群的资源管理；
g） 宜支持异构资源的高性能互联；
h） 宜支持高效节能处理器架构（例如 ARM 架构）。

6.4 分布式计算调度

分布式计算调度的要求包括任务调度的要求与资源调度的要求。

a） 任务调度的要求包括：
  1） 应支持模型训练和推理的任务调度，支持基于主流开源框架的计算任务；
  2） 应支持大规模任务容器化调度，支持系统在物理机或虚拟机上的部署；
  3） 应支持任务跨集群调度，本地任务可调度到另一个集群中计算；
  4） 应支持基于任务的有向无环图进行计算调度；
  5） 应提供任务调度及资源使用的视图；
  6） 宜支持定义作业的优先级，支持定时作业、超时作业、重试作业设置。
j) 资源调度的要求包括：
  1） 应支持对异构资源池统一调度，支持资源池的动态伸缩；

2） 应支持根据资源标签调度及下发任务；
3） 应支持统一的调度接口，调度不同类型的异构资源；
4） 应支持多级资源池灵活调度和共享；
5） 应提供 GPU 池化，支持分时复用 GPU 资源。

6.5 机器学习引擎

机器学习引擎的要求包括训练和推理的要求与机器学习算法库的要求。

a） 训练与推理的要求包括：
6） 应支持主流开源计算框架（例如 Tensorflow，Caffe，PyTorch）；
7） 应支持多种类型的统计机器学习算法：监督学习算法（例如逻辑回归，支持向量机，梯度提升决策树）、非监督学习算法（例如聚类算法，关联规则学习）；
8） 应支持多种类型的深度学习算法 （例如卷积神经网络,递归神经网络等)；
9） 应支持主流深度学习框架模型镜像的发布管理、版本管理，以及服务实例、资源的动态伸缩调度；
10） 宜根据算法需求，支持特征的选择、提取和构建。

k) 机器学习算法库的要求包括：
1） 应支持算法的统一注册和管理；
2） 应提供算法训练的日志及中间结果分析功能；
3） 应提供多种形式的建模方式（例如拖拽式 DAG 图、Notebook 等）；
4） 应为集成提供标准接口（例如 REST）；
5） 宜支持算法的分布式训练，提供高性能运算能力；
6） 宜支持算法参数调节功能，提供推荐参数帮助用户进行调参。

6.6 模型库

模型库提供适用于应用场景的模型功能，包括：

a） 应具备模型的导入导出、更新、版本管理、权限控制等基础功能；
b） 应预置常用 AI 模型，集成典型机器学习模型，支持根据输入数据的重新训练，提升模型在应用场景下的效果；
c） 工作流应支持多模型的融合开发；
d） 应提供可视化开发和管理界面；
e） 应基于多租户的权限控制，实现模型的安全管控；
f） 应提供模型封装和发布的能力，通过统一的接口提供模型服务的调用。

6.7 算法服务

算法服务的要求包括：

a） 应提供一种或多种算法服务（例如图像、视频、语音、自然语言处理等）；
b） 应支持在不影响现有算法服务能力的前提下，部署新的算法服务；
c） 应支持算法服务的增删启停、服务版本管理、服务历史记录、服务当前状态的查询等；
d） 应支持一种或多种离线服务（例如模型自学习服务、批量推理服务等）；
e） 应支持一种或多种在线实时服务（例如实时推理服务等）；
f） 应支持多用户同时使用算法服务；
g） 应支持配置用户权限，控制用户所能调用的算法服务；
h） 应支持同一算法服务的多实例部署；

i） 应支持不同算法服务并发调用，各服务独立运行；
j） 应支持在多用户、高并发情况下的流量负载均衡，保证服务稳定运行；
k） 应提供通用简便的服务上线流程，提供统一服务框架；
l） 应提供统一、易用的算法服务接口框架。

6.8 运维管理

运维管理的要求包括：
a） 应提供多用户管理，支持多用户的权限管理（例如增删改查），及支持常用的认证系统；
b） 应提供多租户管理，支持租户间的应用隔离、数据隔离、资源隔离、运行隔离；
c） 应提供安装与升级能力，支持分发安装包、数据或模型参数文件，进行安装、升级、扩展和回滚；
d） 应提供备份与恢复能力，支持安装包、数据或模型参数文件的备份，以供故障后的系统恢复；
e） 应提供运行环境的监控能力，包括底层资源的统一监控（例如 CPU 利用率、系统负载等）；
f） 应提供日志管理，可以根据日志进行故障定位及排查；
g） 应提供针对监控指标及日志的告警能力；
h） 宜提供主要监控指标的可视化展示功能。

中国电子工业标准化技术协会（CESA）是全国电子信息产业标准化组织和标准化工作者自愿组成的社会团体。广泛联系全国电子信息产业标准化机构和标准化工作者，协助政府部门搞好电子信息产业标准化工作，开拓信息技术领域的标准化工作是中国电子工业标准化技术协会的主要工作内容之一。中国境内从事科研开发、制造、营销和服务的企事业单位、高等院校、社会组织和个人均可随时向中国电子工业标准化技术协会团体标准工作部提出团体标准项目建议。

中国电子工业标准化技术协会标准按照《电子工业标准化技术协会协会团体标准管理办法》进行制定和管理。

在本标准实施过程中，如发现需要修改或补充之处，请将意见和有关资料寄至中国电子工业标准化技术协会，以便修订时参考。

中国电子工业标准化技术协会地址：北京市海淀区万寿路27号
电话：010 - 64102952 电子邮箱： standards@cesa.cn
网址：www.cesa.cn