ICS 35.040
CCS L 71

团　　体　　标　　准

T/AI 117—2022

深度学习推理引擎应用编程接口（API）

Application Programming Interface (API) for Deep Learning Inference Engine

T/AI 117-2022

2022 - 12 - 30 发布　　2022 – 12 – 30 实施

中关村视听产业技术创新联盟　发布

目　　次

前言 II
引言 I
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 2
5 推理引擎应用编程接口框架 2
6 推理引擎接口分类及要求 3
6.1 接口分类 3
6.2 接口要求 3
7 模型管理接口 3
7.1 接口概述 3
7.2 网络接口与参数读取 4
8 设备管理接口 7
8.1 能耗管理接口 7
8.2 工作线程数接口 7
9 输入输出接口 8
9.1 推理对象及相应接口 8
9.2 获取输入数据 8
9.3 获得输出数据 9
10 推理接口 10
10.1 创建推理实例 10
10.2 执行 11
附录 A（资料性）模型管理接口 C++语言描述 12
A.1 网络接口、参数读取 12
附录 B（资料性）设备管理接口 C++语言描述 16
B.1 能耗管理接口 16
B.2 工作线程数接口 16
附录 C（资料性）输入输出接口 C++语言描述 18
C.1 推理对象及相应接口 18
C.2 获取输入数据 18
C.3 获得输出数据 19
附录 D（资料性）推理接口 C++语言描述 21
D.1 创建推理实例 21
D.2 执行 21

T/AI 117—2022

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由新一代人工智能产业技术创新战略联盟AI标准工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：北京百度网讯科技有限公司、北京大学、北京大学计算与数字经济研究院、清华大学、北京旷视科技有限公司、上海商汤智能科技有限公司、上海燧原科技有限公司。

本文件主要起草人：胡晓光、杨超、翟季冬、梅敬青、蒋慧、党青青、严春伟、冯海军、李笑如、李雨芮、马艳军、于佃海、高铁柱、赵乔、徐浩、蒋晓琳、张小飞、张艺伯、吴庚、毕然。

T/AI 117-2022

引　言

本文件的发布机构提请注意，声明符合本文件时，可能涉及到6、7、8、9章节如下6项专利的使用。专利名称如下：

CN202111347748.5 一种推理引擎开发平台、方法、电子设备及存储介质；CN202010537231.1 深度学习推理引擎的调度方法、装置、设备和介质；CN202011598792.9 一种深度学习模型的跨平台执行方法、装置及电子设备；CN202110119167.X 推理库的部署方法、装置和电子设备；CN202011061260.1 推理服务的配置方法、装置、电子设备及存储介质；CN202111504826.8 基于深度学习的硬件适配装置和方法。

本文件的发布机构对于该专利的真实性、有效性和范围无任何立场。

该专利持有人已向本文件的发布机构保证，他愿意同任何申请人在合理且无歧视的条款和条件下，就专利授权许可进行谈判。该专利持有人的声明已在本文件的发布机构备案，相关信息可以通过以下联系方式获得：

联 系 人：黄铁军（新一代人工智能产业技术创新战略联盟秘书长）

通讯地址：北京大学理科2号楼2641室

邮政编码：100871

电子邮件：tjhuang@pku.edu.cn

电话：+8610-62756172

传真：+8610-62751638

网址：http://www.aitisa.org.cn请注意除上述专利外，本文件的某些内容仍可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

T/AI 117-2022

深度学习推理引擎应用编程接口（API）

1 范围

本文件规定了深度学习推理引擎应用编程接口，定义了模型管理接口、设备管理接口、输入输出接口、推理接口四类接口规范。

本文件适用于部署在设备上的深度学习推理引擎，包括云侧（服务器端）推理引擎及端侧（移动端、边缘端）推理引擎，可用于指导推理引擎 API 的设计与开发。

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 41867-2022 信息技术 人工智能 术语

3 术语和定义

下列术语和定义适用于本文件。

3.1

推理 inference

从给定的前提进行论证并得出结论。

注1：在人工智能中一个前提是一个事实、一个规则、一个模型、一个特征或原始数据。

注2：术语“推断”既指过程也指结果。

[来源：GB/T 41867-2022，3.2.30]

3.2

深度学习 deep learning

深度神经网络学习 deep nerual network learning

通过训练具有许多隐层的神经网络来创建丰富层次表示的方法。

[来源：GB/T 41867-2022，3.2.27]

3.3

深度学习推理引擎 deep learning inference engine

一种专门用于深度学习预测或推理的软件框架，提供一系列高层接口用于加载深度学习模型，搭载并适配硬件执行模型中涉及的所有运算，获得推理计算结果，可应用于图像识别、自然语言处理等场景。

T/AI 117—2022

3.4

模型 model

一种基于输入数据生成推理或预测的计算结构。

3.5

计算图 computational graph

一种由一系列节点和边构成的有向无环图，节点分为变量和函数，变量节点通过边相连作为函数节点的输入与输出。

3.6

张量 tensor

多维数组，数组中的每个元素称为张量的元素（tensor element），在深度学习中特指计算图中用于存储计算结果的一个变量节点。

4 缩略语

下列缩略语适用于本文件。

API：应用编程接口（Application Programming Interface）

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

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphics Processing Unit）

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

5 推理引擎应用编程接口框架

深度学习推理引擎是一种提供深度学习预测或推理功能的软件框架，应用编程接口（API）位于推理引擎与应用层之间，上层应用可通过API进行深度学习模型的调用，如图1所示。

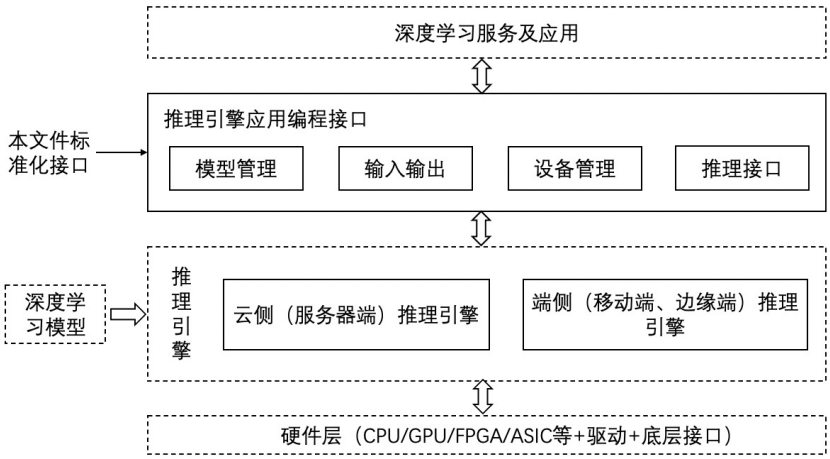

图1 推理引擎应用编程接口框架

T/CCSA 117-2022

深度学习推理引擎应用编程接口涉及的内容包括模型管理、输入输出、设备管理和推理接口。图1中的深度学习服务及应用、深度学习模型、推理引擎、硬件层不在本文件标准范围内。推理引擎应用编程接口框架各功能模块描述如下：

a) 硬件层：从硬件方面，为深度学习提供计算能力支持，其范围包括芯片、驱动及底层接口。

b) 深度学习模型：用于解决某一特定任务的深度学习算法，通常特指用于表示深度学习算法的计算图结构信息和参数信息。

c) 推理引擎：一种专门用于深度学习预测或推理的软件框架，提供一系列高层接口用于加载深度学习模型，搭载并适配硬件执行模型中涉及的所有运算，获得推理计算结果，可应用于图像识别、自然语言处理等场景。深度学习推理引擎可以运行在云侧（IDC），也可以运行在端侧（移动端、边缘侧）。

d) 推理引擎应用编程接口：以深度学习推理引擎作为后端运行，通过预先定义的函数，使得深度学习应用及服务开发者可通过调用函数，在无需访问原码，或理解内部工作机制的细节的条件下进行应用开发。

e) 深度学习应用及服务：基于深度学习实现图像分类、语音识别、语义表示等应用或服务功能。深度学习应用及服务可以部署在云侧，也可以部署在端侧。

6 推理引擎接口分类及要求

6.1 接口分类

按照推理引擎应用编程接口功能进行分类，具体如下：

a) 模型管理接口：用于网络接口与参数读取；

b) 设备管理接口：用于设备能耗及线程管理；

c) 输入输出接口：用于输入数据传入及输出数据获取；

d) 推理接口：用于推理实例创建、推理执行以及推理计算。

6.2 接口要求

推理引擎应用编程接口要求具体如下：

a) 接口函数应开放给用户开发；

b) 接口函数可用多种语言进行实现，C++语言示例见附录；

c) 支持在现有接口的基础上，添加新的接口类型和函数。

7 模型管理接口

7.1 接口概述

模型管理接口应支持实现网络接口与参数读取等功能，主要包括网络接口与参数读取接口。

T/AI 117-2022

7.2 网络接口与参数读取

7.2.1 设置模型文件夹路径

该接口用于设置模型文件夹路径，当需要从磁盘加载非组合模型时使用。

表 1 设置模型文件夹路径接口

参数	输入/输出类型	是否必选	参数说明
文件夹路径	输入参数	是	设置的模型文件夹路径

返回值：调用成功，返回NULL；调用失败，返回非NULL。

7.2.2 获取设置模型文件夹路径

该接口用于返回设置的模型文件夹路径，加载非组合模型时使用。

表 2 获取设置模型文件夹路径接口返回的字段

参数	输入/输出类型	是否必选	参数说明
文件夹路径	输出参数	是	返回设置的文件夹路径

7.2.3 设置模型文件路径

该接口用于设置模型文件路径，加载组合形式模式时使用。

表 3 设置模型文件路径接口

参数	输入/输出类型	是否必选	参数说明
文件路径	输入参数	是	设置的模型文件路径

返回值：调用成功，返回NULL；调用失败，返回非NULL。

7.2.4 返回模型文件路径

该接口用于获取设置模型文件路径，加载组合形式模式时使用。

表 4 获取设置模型文件夹路径接口返回的字段

参数	输入/输出类型	是否必选	参数说明
文件路径	输出参数	是	获取的模型文件路径

T/AI 117-2022

7.2.5 设置模型参数文件路径

该接口用于设置模型参数文件路径，加载组合形式模型时使用。

表 5 设置模型参数文件路径接口

参数	输入/输出类型	是否必选	参数说明
模型参数文件路径	输入参数	是	加载组合形式模型时使用

返回值：调用成功，返回NULL；调用失败，返回非NULL。

7.2.6 获取模型参数文件路径

该接口用于获取模型参数文件路径，加载组合形式模型时使用。

表 6 获取模型参数文件路径接口返回的字段

参数	输入/输出类型	是否必选	参数说明
模型参数文件路径	输出参数	是	加载组合形式模型时使用

7.2.7 设置可用的计算模式列表

该接口用于设置可用的计算模式列表。

表 7 设置可用的计算模式列表接口

参数	输入/输出类型	是否必选	参数说明
计算模式列表	输入参数	是	可用的计算模式列表

返回值：调用成功，返回NULL；调用失败，返回非NULL。

7.2.8 设置模型、参数的内存地址

该接口用于设置模型、参数的内存地址，需要从内存加载模型时使用。

表 8 设置模型、参数的内存地址接口

参数	输入/输出类型	是否必选	参数说明
模型结构数据	输入参数	是	内存中模型结构数据
模型结构数据大小	输入参数	是	内存中模型结构数据的大小

T/AI 117—2022

表 8　设置模型、参数的内存地址接口（续）

参数	输入/输出类型	是否必选	参数说明
模型参数数据	输入参数	是	内存中模型参数数据
模型参数数据大小	输入参数	是	内存中模型参数数据大小

返回值：调用成功，返回NULL；调用失败，返回非NULL。

7.2.9　从内存加载模型接口

该接口用于判断是否从内存中加载模型。

表 9　从内存加载模型接口的返回的字段

参数	输入/输出类型	是否必选	参数说明
是否从内存加载模型	输出参数	是	当调用设置模型、参数的内存地址函数接口时，返回true，否则，返回flase

7.2.10　获取内存中模型结构数据

该接口用于获取内存中模型结构数据。

表 10　获取内存中模型结构数据接口返回的字段

参数	输入/输出类型	是否必选	参数说明
模型结构数据	输出参数	是	内存中模型结构数据

7.2.11　获取内存中模型参数数据

该接口用于获取内存中模型参数数据。

表 11　获取内存中模型参数数据接口返回的字段

参数	输入/输出类型	是否必选	参数说明
模型参数数据	输出参数	是	内存中模型参数数据

T/AI 117-2022

8 设备管理接口

8.1 能耗管理接口

8.1.1 设置能耗模式

该接口用于设置设备的能耗模式，该接口为可选接口。

表 12 设置能耗模式接口

参数	输入/输出类型	是否必选	参数说明
能耗模式	输入参数	是	设置的设备能耗模式

返回值：调用成功，返回NULL；调用失败，返回非NULL。

8.1.2 获取能耗模式

该接口用于获取设备的能耗模式，该接口为可选接口。

表 13 获取能耗模式接口返回的字段

参数	输入/输出类型	是否必选	参数说明
能耗模式	输出参数	是	获取的设备能耗模式

8.2 工作线程数接口

8.2.1 设置工作线程数

该接口用于设置设备的工作线程数，该接口为可选接口。

表 14 设置工作线程数接口

参数	输入/输出类型	是否必选	参数说明
工作线程数	输入参数	是	设置的设备线程模式。若不设置，则默认使用单线程

返回值：调用成功，返回NULL；调用失败，返回非NULL。

8.2.2 获取工作线程数

该接口用于获取设备的线程数，该接口为可选接口。

T/AI 117-2022

表 15　获取工作线程数接口返回的字段

参数	输入/输出类型	是否必选	参数说明
工作线程数	输出参数	是	获取设置的工作线程数

9　输入输出接口

9.1　推理对象及相应接口

该接口用于创建推理对象，包含了获取输入、获取输出两大类接口及执行函数接口，其中执行接口描述见本标准9.2小节。

表 16　推理对象接口

参数	输入/输出类型	是否必选	参数说明
输入张量的指针	输入函数	是	用来设置模型的输入数据
输出张量的指针	输入函数	是	用来获取模型的输出结果
获取输入张量的指针	输入函数	是	根据名称获取模型输入张量的指针
获取输出张量的指针	输入函数	是	根据名称获取模型输出张量的指针
执行函数	输入函数	是	执行推理

9.2　获取输入数据

9.2.1　获取输入张量指针

该接口可获取输入张量的指针，用来设置模型的输入数据。

表 17　获取输入张量指针接口

参数	输入/输出类型	是否必选	参数说明
张量索引	输入参数	是	输入张量的索引

表 18　获取输入张量指针返回的字段

参数	输入/输出类型	是否必选	参数说明
张量指针	输出参数	是	根据索引确认输入张量的指针

T/AI 117-2022

9.2.2 获取所有输入张量名称

该接口可获取所有输入张量的名称。

表 19 获取所有输入数据接口返回的字段

参数	输入/输出类型	是否必选	参数说明
张量名称	输出参数	是	所有输入张量的名称

9.2.3 根据名称获取输入张量指针

该接口可根据名称获取输入张量的指针。

表 20 根据名称获取输入张量指针的接口

参数	输入/输出类型	是否必选	参数说明
张量名称	输入参数	是	输入张量的名称

表 21 根据名称获取输入张量指针的接口返回的字段

参数	输入/输出类型	是否必选	参数说明
张量指针	输出参数	是	输入张量的指针

9.3 获得输出数据

9.3.1 获取输出张量指针

该接口可获取输出张量的指针，用来获取模型的输出结果。

表 22 获取输出张量指针接口

参数	输入/输出类型	是否必选	参数说明
张量索引	输入参数	是	输出张量的索引

T/AI 117-2022

表 23　获取模型输出张量接口返回的字段

参数	输入/输出类型	是否必选	参数说明
张量指针	输出参数	是	根据索引确认输出张量的指针

9.3.2　获取所有输出张量名称

该接口用于获取所有张量的名称。

表 24 获取所有输出张量名称接口返回的字段

参数	输入/输出类型	是否必选	参数说明
张量名称	输出参数	是	所有输出张量的名称

9.3.3　根据名称获取输出张量指针

该接口用于根据名字获取输出张量的指针，用来获取模型的输出结果。

表 25　根据名称获取输出张量指针接口

参数	输入/输出类型	是否必选	参数说明
张量名称	输入参数	是	输出张量的名称

表 26　根据名称获取输出张量指针接口返回的字段

参数	输入/输出类型	是否必选	参数说明
张量指针	输出参数	是	根据张量名称返回的输出张量指针

10　推理接口

10.1　创建推理实例

该接口可用于创建推理实例，根据配置构建推理器。

T/AI 117-2022

表 27　创建推理实例接口

参数	输入/输出类型	是否必选	参数说明
配置信息	输入参数	是	构建推理器的配置信息

表 28　获取模型输出结果接口返回的字段

参数	输入/输出类型	是否必选	参数说明
推理器	输出参数	是	构建的推理器

10.2　执行

接口用于执行模型预测，需在设置输入数据后调用，无输入输出参数。

T/AI 117-2022

附　录　A
（资料性）
模型管理接口 C++语言描述

A.1　网络接口、参数读取

A.1.1　设置模型文件夹路径

接口名称：

set_model_dir(model_dir)

功能描述：

设置模型文件夹路径，当需要从磁盘加载非组合模型时使用。

参数说明：

set_model_dir (string)，模型文件夹路径

返回：

None

返回类型：

None

A.1.2　获得设置模型文件夹路径

接口名称：

model_dir()

功能描述：

获取设置的模型文件夹路径。

参数说明：

None

返回：

设置的文件夹路径。

返回类型：

string

A.1.3　设置模型文件路径

接口名称：

set_model_file(model_file)

功能说明：

设置模型文件路径，加载组合形式模型时使用。

参数说明：

model_file(string)，模型文件路径

返回：

None

T/AI 117-2022

返回类型：

None

A.1.4 返回模型文件路径

接口名称：

model_file()

功能说明：

获取设置模型文件路径，加载组合形式模型时使用。

参数说明：

None

返回：

模型文件路径

返回类型：

string

A.1.5 设置模型参数文件路径

接口名称：

set_param_file(param_fie)

功能说明：

获取设置模型参数文件路径，加载组合形式模型时使用。

参数说明：

param_fie(str)，模型参数文件路径

返回：

None

返回类型：

None

A.1.6 获取模型参数文件路径

接口名称：

param_file()

功能说明：

获取设置模型参数文件路径，加载组合形式模型时使用。

参数说明：

None

返回：

模型参数文件路径

返回类型：

String

A.1.7 设置可用的计算模式列表

T/AI 117-2022

接口名称：

set_valid_places(valid places)

功能说明：

设置可用的模型计算模式列表。

参数说明：

valid_places(list)，可用计算模式列表

返回：

None

返回类型：

None

A.1.8 设置模型、参数的内存地址

接口名称：

set_model_buffer(model_buffer, model_buffer_size, param_buffer, param_buffer_size)

功能说明：

设置模型、参数的内存地址，当需要从内存加载模型时使用。

参数说明：

model_buffer(const char*)，内存中模型结构数据；

model_buffer_size(size _t)，内存中模型结构数据的大小；

param_buffer(const char*)，内存中模型参数数据；

param_buffer_size(size_t)，内存中模型参数数据的大小。

返回：

None

返回类型：

Void

A.1.9 从内存加载模型接口

接口名称：

model_from_memory()

功能说明：

是否从内存中加载模型，当调用set_mode_buffer接口时返回为true。

参数说明：

None

返回：

是否从内存加载模型

返回类型：

Bool

A.1.10 获取内存中模型结构数据

T/AI 117-2022

接口名称：

model_buffer()

功能说明：

获取内存中模型结构数据。

参数说明：

None

返回：

内存中模型结构数据

返回类型：

string

A.1.11 获取内存中模型参数数据

接口名称：

param_buffer()

功能说明：

获取内存中模型参数数据。

参数说明：

None

返回：

内存中模型参数数据

返回类型：

string

T/AI 117-2022

附　录　B
（资料性）
设备管理接口 C++语言描述

B.1　能耗管理接口

B.1.1　设置能耗模式

接口名称：

set_power_mode(mode)

功能描述：

设置设备能耗模式。

参数说明：

mode(PowerMode)，设备能耗模式。

返回：

None

返回类型：

void

B.1.2　获取能耗模式

接口名称：

power_mode()

功能说明：

获取设置的设备能耗模式。

参数说明：

None

返回：

设置的设备能耗模式。

返回类型：

PowerMode

B.2　工作线程数接口

B.2.1　设置工作线程数

接口名称：

set_threads(threads)

功能描述：

设置设备工作线程数。若不设置，则默认使用单线程。

参数说明：

threads(int)，工作线程数。

返回：

None

返回类型：

void

T/AI 117-2022

B.2.2 获取工作线程数

接口名称：

threads()

功能说明：

获取设置的工作线程数

参数说明：

None

返回：

工作线程数

返回类型：

int

T/AI 117-2022

附　录　C
（资料性）
输入输出接口 C++语言描述

C.1　推理对象及相应接口

Predictor为推理库执行期的对象，包含了获取输入、获取输出两大类接口。基本定义如下：

```
class Predictor {
public:
    std::shared_ptr<Tensor> GetInput(int idx);
    std::shared_ptr<Tensor> GetOutput(int idx);
    std::shared_ptr<Tensor> GetInputByName(const std::string& name);
    std::shared_ptr<Tensor> GetOutputByName(const std::string& name);
    virtual void Run() = 0;
};
```

C.2　获取输入数据

C.2.1　获取输入张量指针

接口名称：

get_input(index)

功能说明：

获取输入Tensor指针，用来设置模型的输入数据。

参数说明：

index(int)，　输入Tensor的索引

返回：

第index个输入Tensor的指针

返回类型：

unique_ptr<Tensor>

C.2.2　获取所有输入张量名称

接口名称：

get_input_names()

功能说明：

获取所有输入Tensor的名称。

参数说明：

None

返回：

所有输入Tensor的名称

返回类型：

vector<string>

C.2.3　根据名称获取输入张量指针

T/AI 117-2022

接口名称：

get_input_bynames(name)

功能说明：

根据名称获取输出Tensor的指针，用来获取模型的输出结果。

参数说明：

name(string)，输入Tensor的名称

返回：

输入Tensor的指针

返回类型：

unique_ptr<Tensor>

C.3 获得输出数据

C.3.1 获取输出张量指针

接口名称：

get_output(index)

功能说明：

获取输出Tensor的指针，用来获取模型的输出结果。

参数说明：

index(int) ，输出Tensor的索引

返回：

第index个输出Tensor的指针

返回类型：

unique_ptr<Tensor>

C.3.2 获取所有输出张量名称

接口名称：

get_output_names()

功能说明：

获取所有输出Tensor的名称。

参数说明：

None

返回：

所有输出Tensor的名称

返回类型：

vector<string>

C.3.3 根据名称获取输出张量指针

接口名称：

get_output_bynames()

功能说明：

根据名字获取输出张量的指针，用来获取模型的输出结果。

参数说明：

T/AI 117-2022

name(string)，输出Tensor的名称

返回：

输出Tensor的指针

返回类型：

unique_ptr<Tensor>

T/AI 117-2022

附　录　D
（资料性）
推理接口 C++语言描述

D.1　创建推理实例

接口名称：

create_predictor(Config)

功能描述：

创建推理实例，根据配置信息构建推理器。

参数说明：

config(Config)　　构建Predictor的配置信息

返回：

share_ptr<Predictor>返回推理器指针

返回类型：

shared_ptr<Predictor>

D.2　执行

接口名称：

void run()

功能说明：

执行模型推理，需要在设置输入数据后调用。

参数说明：

None

返回：

None

返回类型：

None

T/AI 117-2022