人工智能 异构人工智能加速器统一接口

1 范围

本文件定义异构人工智能加速器的统一接口及其语义和使用方法，描述了各加速器为实现此接口所需的接入方法和试验方法。

本文件适用于人工智能加速器接口的设计和实现，也可为人工智能加速器应用提供参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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

IEEE 754－2019 IEEE浮点数计算标准 （IEEE Standard for Floating-Point Arithmetic）

3 术语和定义

GB/T 41867－2022界定的以及下列术语和定义适用于本文件。

3.1

[应用编程]接口 [application programming] interface

<人工智能>用来使用人工智能加速器功能的语法和语义定义。

[来源：ISO/IEC/IEEE 9945:2009，3.19，有修改]

3.2

人工智能加速[处理]器 artificial intelligence **accelerating processor**

人工智能加速芯片 artificial intelligence accelerating chip

具备适配人工智能算法的运算微架构，能够完成人工智能应用运算处理的集成电路元件。

注：在本文件中，在不引起误解的语境中，将人工智能加速器简称为加速器。

[来源：GB/T 41867－2022, 3.1.5]

3.3

计算图 computational graph

用来表示数学函数，由节点和连接构成的有向图。

注1：节点表示数学运算，即算子。

注2：连接表示数学运算之间的依赖关系。

注3：一个连接联通起始节点和终止节点。

[来源：ISO/IEC/IEEE 24765:2017，3.1762.1，有修改]

3.4

算子 operator

函数的名字，或表示行动的数学或逻辑符号。

注：算子由硬件定义，由软件封装。

[来源：ISO/IEC TR 17903:2023，3.18，有修改]

3.5

算子端口 operator port

计算图中，指定连接关联算子输入与输出中数据对应关系的索引值。

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）

BF16：布瑞恩半精度浮点数（Brain Floating-point）

CPU： 中央处理单元 (Central Processing Unit)

FP16：16位半精度浮点 （16-bit half-precision Floating Point）

FP32：32位单精度浮点 （32-bit single-precision Floating Point）

INT8：8位四分之一精度整型 （8-bit quarter-precision INTeger）

JPG： 联合图像专家组 （Joint Photographic experts Group）

NCHW：批通道高宽模式 （Batch-Channel-Height-Width mode）

ND： numpy ndarry对象 （Numpy nDarry object）

NHWC：批高宽通道模式 （Batch-Height-Width-Channel mode）

ONNX：开放神经网络交换 （Open Neural Network Exchange）

TF32：张量单精度浮点数 （Tensor Floating-point）

5 概述

5.1 接口功能

异构人工智能加速器的使用，以调用接口的方式完成。接口定义应支持以下功能：

a） 供以下组件使用加速器，或运行在加速器上：

1） 机器学习框架；

2） 模型；

3） 直接使用接口编制的领域业务应用；

b） 用户自定义的操作及其注册、运行；

注：如人工智能加速器提供的功能不在本文件所提供接口的范围内，可不实现本文件的接口。

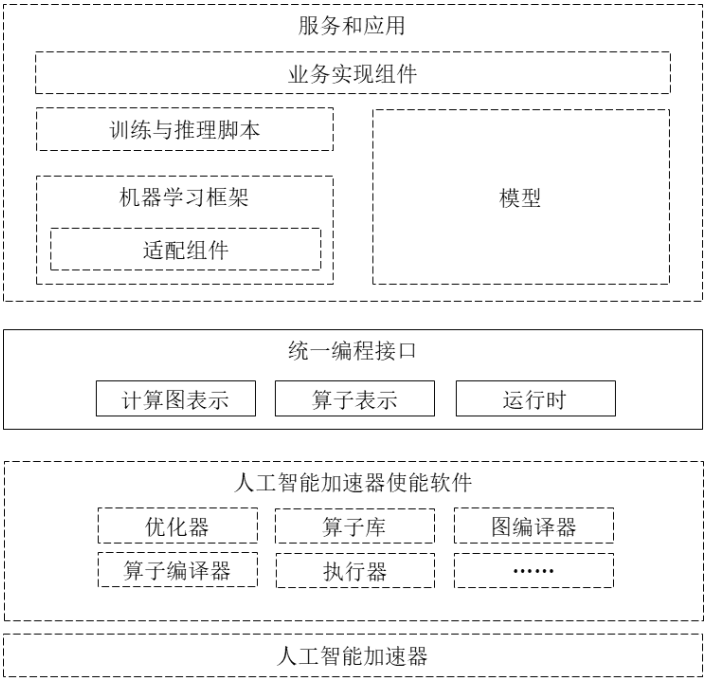
5.2 架构

异构人工智能加速器统一编程接口架构及组件关系如图1。其中：

a） 计算图表示：定义计算流程构建，解析，编译和运行等相关模型操作；

b） 算子表示：定义数据抽象、存储格式、数据布局形状信息和运算操作；

c） 运行时：定义异构人工智能加速器设备管理、事件管理、内存管理、执行控制、上下文管理及模型和算子的加载、执行、卸载等功能。



注 1：图中实线框起的部分对应本文件规定的范畴，虚线框起的部分仅表明本文件定义的统一编程接口所处的位置及与其它部件或组件的关系，不属于本文件规定的范畴。

图 1　异构人工智能加速器统一接口架构

5.3　基本概念

5.3.1　同步和异步

本文件定义的接口包含同步接口和异步接口，其中：

a)　标注为 “异步”的接口，是异步接口；

b)　未做标注的接口，是同步接口；

c)　对同步接口，调用者进程或线程在接口执行完毕后才能执行后续计算；

d)　对异步接口，调用者进程或线程在接口调用后立即执行后续计算，并接受事件通知。

5.3.2　上下文

上下文是人工智能加速器执行相关对象（数据结构）的封装，管理对象的生命周期，上下文的性质包含：

a)　1 个上下文可对应 1 个计算设备，也可对应在 1 个系统中的多个计算设备；

b)　用户线程使用上下文管理人工智能加速器执行的对象,通过接口切换线程当前使用的上下文；

c)　1 个进程可创建多个上下文，在某些加速器上 1 个线程同一时刻使用 1 个上下文；

d)　线程中创建多个上下文时，在某些加速器上线程缺省使用最后创建的上下文；

e)　执行相关对象，包含：

1)　流；

2)　事件；

3)　内存。

注 1：特定加速器实现本文件定义的接口时，可按需构造上下文数据结构。

注 2：不采用上下文封装、管理时，也应构造、保存、使用、管理执行相关的对象（数据结构）

5.3.3　流

流是由特定人工智能加速器执行的操作序列，流的性质包含：

a) 流是加速器上的实际执行体；

b) 1 个线程中可包含多个流，不同的流上的计算任务可并行执行；

c) 多个线程的流在人工智能加速器上相互独立。

注 1：特定加速器实现本文件定义的接口时，可按需构造并管理流的结构。

注 2：不采用流时，异步操作维护，操作保序等，应由该实现自行按需实现并管理。

5.3.4 事件

事件用于同步多个任务的执行，包含：

a) 不同流的任务；

b) 主机 CPU 与人工智能加速器间的任务；

c) 不同人工智能加速器间的任务。

注 1：特定人工智能加速器实现本文件定义的接口时，可按需构造并管理事件的数据结构。

注 2：不采用基于事件的同步机制时，应由该实现自行按需实现同步机制。

注 3：当同步两个任务的执行顺序时，在先执行的任务后，触发事件，通知另一个任务执行。

5.3.5 内存

本文件中的内存指用于人工智能任务执行的内存，因人工智能服务器的体系机构差异，包含：

a) 人工智能加速器的内存；

b) 人工智能加速器的片外内存；

c) 主存。

注：本文件并不特别规定特定内存的使用或共享机制，该机制由实现本文件定义的接口的人工智能加速器及其适配的计算系统定义和实现。

6 接入方法

6.1 加速器

加速器研制机构，为通过统一接口使用异构人工智能加速器，接入统一接口，应实施以下工作：

a) 基于已有的加速器模型描述实现库，实现本文件规定的模型操作接口（见 8.1）；

b) 基于已有的数据预处理、集合通信和运行时库等，实现本文件规定的运行时接口（见 8.2）；

c) 基于已有的算子库，实现本文件规定的运算接口（见 8.3）。

6.2 机器学习框架

机器学习框架，如使用本文件规定的统一接口时，应实现框架产出模型格式到本文件定义的统一接口的解析器。

6.2.1 运行过程说明

运行时接口为调用硬件加速设备功能而设计，调用运行过程如下：

a) 异步运行单算子，流程宜为：

1) 初始化系统以及申请资源（如计算设备、上下文和流）；

2) 申请计算设备输入、输出内存，并完成计算数据准备；

3) 开始算子异步执行，下发任务至计算设备后，异步函数返回；

4) 同步接口等待处理结果，然后将数据从输出内存中复制到调用者（如应用程序的某组件、服务器中其它部件等）；

5） 释放资源（如计算设备、上下文和流等）；

b） 运行计算图时，流程宜为：

1） 初始化系统以及申请资源（如计算设备、上下文和流）；

2） （从文件或内存中）加载模型；

3） 申请计算设备输入、输出内存，并完成计算数据准备；

4） 数据预处理（如解码 JPG 图片后再缩放图片）；

5） 推理（以预处理后的数据作为输入）；

6） 后处理（如从推理结果中查找最大置信度的类别标识）；

7） 如有多项数据待处理，则循环执行上述步骤 3）-5）；

8） 全部执行结束后，释放资源。

7 接口要求

7.1 精度

各类人工智能加速器在实现、对接本文件定义的接口时，所使用的精度（含附录B中的测试负载）：

a） 应符合 IEEE 754－2019 的定义；

b） IEEE 754－2019 定义之外的精度，可由实现者根据实际情况自行选择或定义所支持的精度；

注：接口定义中的数据类型（如表25，表26和表81等）是编程时所使用的类型，在不同加速器上可用不同精度来映射和表示（如布尔在特定加速器上可用INT8精度来承载），由加速器的设计决定。

7.2 接口参数

输入参数的可选或必选约束，含义如下：

a） 必选：为实现接口功能所必须的参数，当无法提供该参数时，接口无法运行或给出期望的结果；

b） 可选：为实现接口功能可能需要的附加参数，当无法提供该参数或使用其它方式提供参数值时，接口仍能运行并给出期望的结果。

7.3 接口执行状态

人工智能加速器实现本文件定义的接口，执行状态应满足以下要求：

a） 定义超时的时限值；

b） 调用执行超时，返回超时码；

c） 根据调用语义，实现处理方法；

d） 接口执行返回错误码（非执行成功码）时，由使用接口的程序定义和实现处理策略。

注：本文件定义的接口的执行返回码，应符合附录A的要求。

8 接口定义

8.1 计算图操作接口

8.1.1 计算图构建接口

8.1.1.1 模型解析接口

加速器执行特定格式的模型文件，应使用表1规定的模型解析接口。格式包含机器学习软件框架的训练结果模型格式或使用特定中间表示的模型的格式。

表 1　解析特定格式模型文件

名称	功能语义描述	参数	输入/输出	可选/必选
解析特定的格式模型文件	解析特定格式模型文件，以便提供对后端硬件的部署、优化能力	模型文件内存（首）地址 [a]	输入	可选
		模型文件内存长度 [a]	输入	可选
		模型文件路径 [b]	输入	可选
		解析后的模型文件	输出	必选
		名字对照表	输入	可选
		参数数组	输入	必选
		框架类型	输入	可选
[a] 适用于模型在内存中的情况。 [b] 适用于模型文件在硬盘中的情况。				

8.1.1.2　算子原型构建接口

算子原型规定了在加速器上可运行算子的约束，主要体现算子的数学含义，包含定义算子输入、输出、属性。通过算子原型构建计算图的流程如下：

1）　注册算子原型。定义算子的要素声明，完整的算子包含算子名称、输入、输出、属性等信息。

2）　创建算子实例，设置算子信息。创建算子实例，并设置算子的输入、输出及属性信息。

3）　创建计算图实例。完成算子定义后，需创建计算图实例，并设置计算图的输入、输出算子，完成计算图构建。

算子原型构建计算图过程包含表2~表23规定的接口。

表 2　注册算子原型

名称	功能语义描述	参数	输入/输出	可选/必选
注册算子原型	定义算子的要素声明：完整的算子包含算子名称、输入、输出、属性等信息	算子类型名称	输入	必选
		必备输入的名称和类型	输入	必选
		可选输入的名称和类型	输入	可选
		动态输入的名称和类型	输入	可选
		输出的名称和类型	输入	必选
		动态输出的名称和类型	输入	可选
		必备属性的名称和类型	输入	必选
		可选属性的名称和类型	输入	可选
		算子中包含的子图信息	输入	可选
		动态算子子图信息	输入	可选

表 3　创建算子实例对象

名称	功能语义描述	参数	输入/输出	可选/必选
创建算子实例对象	创建算子实例对象	算子名称	输入	必选
		算子实例对象	输出	

表 4　设置算子输入

名称	功能语义描述	参数	输入/输出	可选/必选
设置算子输入	设置算子输入，即设置由哪个（源）算子的输出连到当前（目标）算子输入	目标算子输入的名称	输入	必选
		源算子对象	输入	必选
		源算子输出的名称	输入	可选
		源算子输出的索引值	输入	可选

表 5　创建算子动态输入

名称	功能语义描述	参数	输入/输出	可选/必选
创建动态多输入算子的动态输入	创建动态多输入算子的动态输入	算子动态输入的名称	输入	必选
		算子动态输入的个数	输入	必选

表 6　设置算子动态输入

名称	功能语义描述	参数	输入/输出	可选/必选
设置动态多输入算子的动态输入	设置动态多输入算子的动态输入，即设置由哪个（源）算子的输出连到当前（目标）动态多输入算子的输入	目标算子动态输入的名称	输入	必选
		目标算子动态输入的索引值	输出	必选
		源算子对象	输入	必选

表 7　创建算子动态输出

名称	功能语义描述	参数	输入/输出	可选/必选
创建动态多输出算子的动态输出	创建动态多输出算子的动态输出	算子动态输出的名称	输入	必选
		算子动态输出的个数	输入	必选

表 8　设置算子属性

名称	功能语义描述	参数	输入/输出	可选/必选
设置算子属性	设置算子属性的属性值，算子可以包括多个属性，初次设置值后，算子属性值的类型固定，算子属性值的类型包括布尔、整型、浮点型等	算子属性名称	输入	必选
		属性值	输入	必选

表 9　添加算子的控制边

名称	功能语义描述	参数	输入/输出	可选/必选
添加算子的控制边	计算图中某个算子的执行依赖于其他算子执行完，用于控制算子的执行顺序	控制边对应的源算子	输入	必选

表 10　删除算子连接

名称	功能语义描述	参数	输入/输出	可选/必选
删除算子连接	删除当前算子与前后算子之间的所有连接关系	算子名称	输入	必选

表 11　构建张量对象

名称	功能语义描述	参数	输入/输出	可选/必选
构建张量对象	张量对象的构造函数	张量描述对象，需要设置张量描述符	输入	必选
		需设置的数据	输入	可选
		数据的长度，单位为字节	输入	可选
		张量类型的对象	输出	

表 12　构建张量描述对象

名称	功能语义描述	参数	输入/输出	可选/必选
构建张量描述对象	张量描述对象的构造函数	张量形状	输入	必选
		张量格式	输入	必选
		张量数据类型	输入	必选
		张量描述类型的对象	输出	

表 13　更新算子输入的张量描述

名称	功能语义描述	参数	输入/输出	可选/必选
更新算子输入的张量描述	根据算子输入名称更新算子输入的张量描述	算子输入名称	输入	必选
		待设置的张量描述对象	输入	必选

表 14　更新算子输出的张量描述

名称	功能语义描述	参数	输入/输出	可选/必选
更新算子输出的张量描述	根据算子输出名称更新算子输出的张量描述	算子输出名称	输入	必选
		待设置的张量描述对象	输入	必选

表 15　构建计算图

名称	功能语义描述	参数	输入/输出	可选/必选
构建计算图	构建给定名称的计算图	计算图名称	输入	必选

表 16　设置计算图的输出算子

名称	功能语义描述	参数	输入/输出	可选/必选
设置计算图的输出算子	设置计算图的输出算子	成对的序列，每1对包含输出算子及其名字	输入	必选

表 17　设置计算图的终止节点列表

名称	功能语义描述	参数	输入/输出	可选/必选
设置计算图的终止节点[a]列表	设置计算图的终止节点列表	终止节点序列	输入	必选
[a] 终止节点需要被执行，但不返回其输出。				

表 18　判断计算图合法性或能否被执行

名称	功能语义描述	参数	输入/输出	可选/必选
判断计算图的合法性，是否能被执行	判断计算图是否能被执行，返回布尔值	TRUE：合法，能被执行；FALSE：不合法，不能被执行	输出	

表 19　按算子名字在计算图中查找算子

名称	功能语义描述	参数	输入/输出	可选/必选
基于算子名称，查找计算图中的相应对象	按算子名称，查找计算图中的对象	算子名字	输入	必选
		操作对象	输出	
		执行返回码（见附录A）	输出	

表 20　按算子类型在计算图中查找算子

名称	功能语义描述	参数	输入/输出	可选/必选
基于算子类型，查找计算图中的相应对象序列	按算子类型，查找计算图中的对象序列	算子类型	输入	必选
		操作对象序列	输出	
		执行返回码（见附录A）	输出	

表 21　获取计算图中所有已注册的算子

名称	功能语义描述	参数	输入/输出	可选/必选
返回计算图中所有已注册的算子	返回计算图中所有已注册的算子。	算子名字序列	输出	
		执行返回码（见附录A）	输出	

表 22　返回计算图的名字

名称	功能语义描述	参数	输入/输出	可选/必选
返回计算图的名字	返回计算图的名字	计算图的名字	输出	
		执行返回码（见附录A）	输出	

表 23　拷贝计算图

名称	功能语义描述	参数	输入/输出	可选/必选
拷贝计算图	拷贝计算图，填充一个新的计算图数据结构	新计算图数据结构的引用	输出	必选

8.1.2　计算图修改接口

图修改过程直接优化图的结构，比如将某些特定子图替换成一个大算子，以减少计算步骤、外存访问、调度时间等，或者在某些算子中间添加一个算子，或者删除计算图中某个算子，以及增、删计

算图中算子间的连接等，此时可以通过计算图修改接口直接将图直接修改成期望的结构。计算图修改包含表24~表47的接口。

表 24　获取算子类型

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子类型	获取算子类型	算子	输入	必选
		算子类型字符串	输出	
		执行返回码（见附录A）	输出	

表 25　获取算子名字

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子参数名字	获取算子参数名称	算子	输入	必选
		算子类型字符串	输出	
		执行返回码（见附录A）	输出	

表 26　获取算子指定输入端口对应的对端算子及其输出端口

名称	功能语义描述	参数	输入/输出	可选/必选
获取对端算子及其输出端口	获取对端算子及其输出端口	算子	输入	必选
		算子输入端口号	输入	必选
		对端算子及其输出端口号值对的序列	输出	

表 27　获取算子指定输出端口对应的对端算子及其输入端口

名称	功能语义描述	参数	输入/输出	可选/必选
获取对端算子及其输入端口	获取对端算子及其输入端口	算子	输入	必选
		算子输出端口号	输入	必选
		对端算子及其输入端口号值对的序列	输出	

表 28　获取算子的控制输入节点

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子的控制输入节点	获取算子的控制输入节点	算子	输入	必选
		输入控制节点序列	输出	

表 29　获取算子的控制输出节点

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子的控制输出节点	获取算子的控制输出节点	算子	输入	必选
		输出控制节点序列	输出	

表 30　获取算子输入个数

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子输入个数	返回节点的有效输入个数，即算子的实际输入个数	算子	输入	必选
		算子的输入个数	输出	

表 31　获取算子输出个数

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子输出个数	返回有效输出个数，即算子定义时的输出个数	算子	输入	必选
		算子的输出个数	输出	

表 32　获取算子中指定输入端口的张量格式

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子中指定输入端口的张量格式	获取算子中指定输入端口的张量格式	算子	输入	必选
		端口号	输入	必选
		张量格式数据接口的引用	输出	
		执行返回码（见附录A）	输出	

表 33　获取算子中指定输出端口的张量格式

名称	功能语义描述	参数	输入/输出	可选/必选
获取算子中指定输出端口的张量格式	获取算子中指定输出端口的张量格式	算子	输入	必选
		端口号	输入	必选
		张量格式数据结构的引用	输出	
		执行返回码（见附录A）	输出	

表 34　更新算子中指定输入端口的张量格式

名称	功能语义描述	参数	输入/输出	可选/必选
更新算子中指定输入端口的张量格式	更新算子中指定输入端口个的张量格式，其中形状信息的某个维度可以为-1，支持运行时动态形状输入	算子	输入	必选
		端口号	输入	必选
		形状信息	输入	可选
		新张量格式数据结构的引用	输出	
		执行返回码（见附录A）	输出	

表 35　更新算子中指定输出端口的张量格式

名称	功能语义描述	参数	输入/输出	可选/必选
更新算子中指定输出端口的张量格式	更新算子中指定输出端口的张量格式，其中形状信息的某个维度可以为-1，支持运行时动态形状输入	算子	输入	必选
		端口号	输入	必选
		形状信息	输入	可选
		新张量格式数据结构的引用	输出	

		执行返回码（见附录A）	输出	

表 36　设置节点属性的属性值

名称	功能语义描述	参数	输入/输出	可选/必选
设置节点属性的属性值	设置节点属性的属性值	算子	输入	必选
		属性名	输入	必选
		属性值，属性值的类型，包含：双精度整型，单精度整型，单精度无符号整型，单精度浮点，字符串，布尔，张量，以上类型数据的数组等	输入	必选
		执行返回码（见附录A）	输出	

表 37　设置节点属性的属性名的值

名称	功能语义描述	参数	输入/输出	可选/必选
设置节点属性的属性名的值	设置节点属性的属性名的值	算子	输入	必选
		属性名	输入	必选
		属性名的值，类型包含：双精度整型，单精度整型，单精度无符号整型，浮点，字符串，布尔，张量，以上类型数据的数组等	输入	必选
		执行返回码（见附录A）	输出	

表 38　查询属性是否存在

名称	功能语义描述	参数	输入/输出	可选/必选
查询属性是否存在	查询属性是否存在	算子	输入	必选
		属性名	输入	必选
		TRUE：存在；FALSE：不存在	输出	

表 39　向计算图增加算子

名称	功能语义描述	参数	输入/输出	可选/必选
向计算图增加算子	向计算图增加算子，所增加的算子在所描述模型的尾部出现	原计算图	输入	必选
		要增加的算子	输入	必选
		新计算图	输出	

表 40　从计算图删除节点

名称	功能语义描述	参数	输入/输出	可选/必选
从计算图删除节点	从计算图删除节点，删除后，所关联的连接消除。节点的删除，可使计算图无效	特定待删除的节点	输入	必选
		新计算图	输出	

表 41　从计算图删除连接

名称	功能语义描述	参数	输入/输出	可选/必选
从计算图中删除连接	从计算图删除连接	连接的起始节点	输入	必选
		连接的起始节点使用的端口	输入	必选
		连接的终止节点	输入	必选
		连接的终止节点使用的端口	输入	必选
		新计算图	输出	

表 42　向计算图增加算子作为节点

名称	功能语义描述	参数	输入/输出	可选/必选
向计算图增加算子，作为节点	向计算图增加算子，作为节点。增加后，新节点在计算图的尾部	待增加算子	输入	必选
		新计算图	输出	

表 43　向计算图增加数据连接

名称	功能语义描述	参数	输入/输出	可选/必选
向计算图增加数据连接	向计算图增加数据连接	待增加连接的起始节点	输入	必选
		待增加连接的起始节点使用的端口	输入	必选
		待增加连接的终止节点	输入	必选
		待增加连接的终止节点使用的端口	输入	必选
		新计算图	输出	

表 44　向计算图增加控制连接

名称	功能语义描述	参数	输入/输出	可选/必选
向计算图增加控制连接	向计算图增加控制连接	待增加连接的起始节点	输入	必选
		待增加连接的终止节点	输入	必选
		新计算图	输出	

表 45　从文件读取计算图到内存

名称	功能语义描述	参数	输入/输出	可选/必选
从文件读取计算图到内存	从文件读取计算图到内存	文件路径	输入	必选
		计算图对象	输出	

表 46　将内存中的计算图保存到文件

名称	功能语义描述	参数	输入/输出	可选/必选
将内存中的计算图保存到文件	将内存中的计算图保存到文件	保存文件的路径	输入	必选
		保存后的计算图对象	输出	

表 47　切分计算图 （可选）

名称	功能语义描述	参数	输入/输出	可选/必选
切分计算图	将整个计算图切分为若	整个计算图的指针	输入	必选

	干子图	切图结果中每个子图包含算子的最小数量	输入	可选
		切图结果中包含的最小子图数量	输入	可选
		切图规则名字字符串指针数组	输入	可选
		切分得到的子图的指针数组	输出	必选

8.1.3 计算图编译接口

编译过程，将计算图编译至能在特定硬件后端运行的模型文件，包含表48~表51的接口。

表 48 编译为特定硬件后端模型

名称	功能语义描述	参数	输入/输出	可选/必选
编译为后端模型	将计算图编译至能在特定硬件后端运行的模型文件	计算图 [a]	输入	必选
		名字对照表	输入	可选
		输出模型文件（指针和模型大小）	输出	
[a] 在某些加速器上，可输入多个计算图（切片模型），将多个计算图按名字对照表中的接续关系（通信算子可包含在某些计算图中），编译为1个模型。此时，此输入是一个由多个计算图（切片模型）组成的向量或集合，而在多个计算设备上执行。				

表 49 保存特定硬件后端模型

名称	功能语义描述	参数	输入/输出	可选/必选
保存模型	保存特定硬件后端模型	模型文件（内存）地址	输入	必选
		文件路径	输入	必选

表 50 编译指定算子

名称	功能语义描述	参数	输入/输出	可选/必选
编译指定算子	编译指定算子	算子类型名称的指针	输入	必选
		算子的输入张量的数量	输入	必选
		算子的输入张量的描述指针数组	输入	必选
		算子的输出张量的数量	输入	必选
		算子的输出张量的描述指针数组	输入	必选
		算子属性指针	输入	必选
		算子执行环境类型	输入	必选
		编译标识，分为注册过的算子，和未注册过的算子，两类	输入	必选
		算子实现文件路径指针（不含文件名）	输入	必选
		算子编译的高级选项	输入	可选
		执行返回码（见附录A）	输出	

注：在某些加速器上，本接口的输入可简化为要编译的算子，输出可简化为编译后的算子。在这种情况下，算子的数据结构由加速器自行给出。

表 51　编译指定子图 （可选）

名称	功能语义描述	参数	输入/输出	可选/必选
编译指定子图	编译指定子图	子图的指针	输入	必选
		子图输入算子的输入张量的数量	输入	必选
		子图输入算子的输入张量的描述指针数组	输入	必选
		子图输出算子的输出张量的数量	输入	必选
		子图输出算子的输出张量的描述指针数组	输入	必选
		编译后的子图指针（内部可能已经进行算子融合生成了新的自定义算子）	输出	必选
		执行返回码（见附录A）	输出	必选
注1：子图输入算子是指该算子的输入张量不会是子图中任何算子的输出张量，子图输出算子是指该算子的输出张量不会是子图中任何算子的输入张量。 **注2：**某些加速器上，本接口的输入可简化为要编译的计算图，输出可简化为编译后的计算图。在这种情况下，计算图的数据结构由加速器自行给出。				

8.1.4　计算图运行接口

针对特定输入，执行计算图代表的运算过程，获得模型推理结果，应实现表52~表58规定的接口。

表 52　计算图运行环境初始化

名称	功能语义描述	参数	输入/输出	可选/必选
计算图运行环境初始化	计算图运行环境初始化	参数列表，每1个参数是1个字符串对	输入	必选
		执行返回码（见附录A）	输出	

表 53　计算图运行环境去初始化

名称	功能语义描述	参数	输入/输出	可选/必选
计算图运行环境去初始化	计算图运行环境去初始化	执行返回码（见附录A）	输出	

表 54　计算图运行会话初始化

名称	功能语义描述	参数	输入/输出	可选/必选
计算图运行会话初始化	计算图运行会话初始化	参数列表，每1个参数是1个字符串对	输入	必选
		执行返回码（见附录A）	输出	

表 55　添加要运行的计算图

名称	功能语义描述	参数	输入/输出	可选/必选
添加计算图以备运行	添加计算图以备运行	待添加的计算图标识	输入	必选
		原有计算图	输入	必选
		运行选项	输入	必选

表 56　运行计算图

名称	功能语义描述	参数	输入/输出	可选/必选
运行计算图	针对给定的输入，运行计算图表示的计算过程，输出结果	待运行的计算图标识	输入	必选
		输入数据	输入	必选
		输出数据	输出	

表 57　编译特定的计算图，生成可运行的代码

名称	功能语义描述	参数	输入/输出	可选/必选
编译特定的计算图，生成可运行的代码序列	编译特定的计算图，生成可运行的代码序列	计算图标识	输入	必选
		可运行代码数据结构的引用	输出	
		执行返回码（见附录A）	输出	

表 58　运行计算图（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
异步运行计算图	针对给定的输入，异步运行计算图表示的计算过程，输出结果。在输出结果时，将调用给定的回调函数，实现（后）处理	待运行的计算图标识	输入	必选
		输入数据	输入	必选
		输出数据	输出	
		回调函数签名	输入	必选

8.2　运行时接口

8.2.1　预处理接口

为使用对接硬件的数据预处理功能（如支持），应实现表59~表61规定的可选接口。

表 59　预处理所用内存分配

名称	功能语义描述	参数	输入/输出	可选/必选
为预处理分配内存	为预处理分配内存，以容纳待处理数据	申请内存的大小	输入	必选
		已分配内存的指针	输出	
		执行返回码（见附录A）	输出	

表 60　释放为预处理所分配的内存

名称	功能语义描述	参数	输入/输出	可选/必选
释放为预处理分配的内存	释放为预处理分配的内存	待释放内存的指针	输入	必选
		执行返回码（见附录A）	输出	

表 61　执行模型后处理部分

名称	功能语义描述	参数	输入/输出	可选/必选
模型后处理部分	获取模型推理的结果，执行后处理操作	模型推理结果	输入	可选
		模型后处理结果	输出	

8.2.2　运算执行接口

为执行模型定义的运算过程，应实现表62~表139定义的运算执行接口。

8.2.2.1　系统配置接口

表 62　统一接口子系统初始化

名称	功能语义描述	参数	输入/输出	可选/必选
统一接口子系统初始化	用于初始化子系统内相关资源，包含实现该接口的组件	配置文件路径指针	输入	必选
		执行返回码（见附录A）	输出	

表 63　统一接口子系统去初始化

名称	功能语义描述	参数	输入/输出	可选/必选
统一接口子系统去初始化	用于释放子系统内相关资源，包含实现该接口的组件	执行返回码（见附录A）	输出	

表 64　获取统一接口版本号

名称	功能语义描述	参数	输入/输出	可选/必选
获取统一接口版本号	用于查询统一接口的版本号	主版本号	输出	必选
		次版本号	输出	可选
		补丁版本号	输出	可选
		执行返回码（见附录A）	输出	必选

表 65　获取当前运行环境中芯片版本

名称	功能语义描述	参数	输入/输出	可选/必选
获取当前运行环境中芯片版本	查询当前运行环境的人工智能加速器型号	人工智能加速器芯片版本号	输出	必选
		执行返回码（见附录A）	输出	

8.2.2.2　设备管理接口

表 66　获取当前正使用的设备

名称	功能语义描述	参数	输入/输出	可选/必选
获取当前正使用的设备	获取当前正使用的设备	设备标识符	输出	
		执行返回码（见附录A）	输出	

表 67　获取可用设备的数量

名称	功能语义描述	参数	输入/输出	可选/必选
获取可用设备的数量	获取可用设备的数量	可用设备数量	输出	
		执行返回码（见附录A）	输出	

表 68　指定计算设备

名称	功能语义描述	参数	输入/输出	可选/必选
指定计算设备	指定当前进程或线程中用于运算的设备，同时创建计算上下文	设备标识符	输入	必选
		执行返回码（见附录A）	输出	

表 69　复位计算设备

名称	功能语义描述	参数	输入/输出	可选/必选
复位计算设备	复位特定计算设备，释放设备上的资源，包含默认的计算上下文和流。当计算上下文或流相关的任务未完成时，系统会等待任务完成后，再释放资源	设备标识符	输入	必选
		执行返回码（见附录A）	输出	

表 70　获取 AI 软件栈运行模式

名称	功能语义描述	参数	输入/输出	可选/必选
获取AI软件栈运行模式	获取AI软件栈运行位置，包括AI设备上或者主机设备上等	运行模式	输出	必选
		执行返回码（见附录A）	输出	

8.2.2.3　上下文管理接口

表 71　创建计算上下文

名称	功能语义描述	参数	输入/输出	可选/必选
创建计算上下文	在当前进程或线程中显式创建一个计算上下文	待创建上下文的设备标识符	输入	必选
		创建的计算上下文的指针	输出	
		执行返回码（见附录A）	输出	

表 72　销毁给定计算上下文

名称	功能语义描述	参数	输入/输出	可选/必选
销毁计算上下文	销毁一个经显式创建的计算上下文，释放其关联的资源	待销毁的计算上下文指针	输入	必选
		执行返回码（见附录A）	输出	

表 73　设置线程的计算上下文

名称	功能语义描述	参数	输入/输出	可选/必选
设置线程的计算上下文	设置线程的计算上下文	待设置的计算上下文指针	输入	必选
		执行返回码（见附录A）	输出	

表 74　获取线程的计算上下文

名称	功能语义描述	参数	输入/输出	可选/必选
获取线程的计算上下文	获取线程的计算上下文	计算上下文指针	输出	必选
		执行返回码（见附录A）	输出	

8.2.2.4　流管理接口

表 75　创建流

名称	功能语义描述	参数	输入/输出	可选/必选
创建流	在当前进程或线程中创建一个流	创建的流的指针	输出	
		执行返回码（见附录A）	输出	

表 76　销毁流

名称	功能语义描述	参数	输入/输出	可选/必选
销毁流	销毁指定的流，只能销毁显式创建的流	待销毁的流的指针	输入	必选
		执行返回码（见附录A）	输出	

8.2.2.5　内存管理接口

表 77　申请特定计算设备上的内存

名称	功能语义描述	参数	输入/输出	可选/必选
申请计算设备上的内存	在特定计算设备上申请线性内存，返回所分配内存的指针	设备地址	输入	必选
		所分配内存的大小	输入	必选
		内存分配策略	输入	必选
		内存首地址	输出	
		执行返回码（见附录A）	输出	

表 78　释放设备上的内存

名称	功能语义描述	参数	输入/输出	可选/必选
释放设备上的内存	释放特定计算设备上的内存	内存首地址	输入	必选
		执行返回码（见附录A）	输出	

表 79　申请主机上的内存

名称	功能语义描述	参数	输入/输出	可选/必选
申请主机上的内存	在主机上申请线性内存，返回所分配内存的指针	设备地址	输入	必选
		所分配内存的大小	输入	必选
		内存分配策略	输入	必选
		内存首地址	输出	
		执行返回码（见附录A）	输出	

表 80　释放主机上的内存

名称	功能语义描述	参数	输入/输出	可选/必选
释放主机上的内	释放主机上的内存	内存首地址	输入	必选

存		执行返回码（见附录A）	输出	

表 81　内存拷贝（同步）

名称	功能语义描述	参数	输入/输出	可选/必选
内存拷贝（同步）	人工智能计算设备（加速器）宿主机服务器内，或其与加速器设备之间，加速器设备之内，加速器设备之间的同步内存复制	目标内存地址	输入	必选
		目标内存容量	输入	必选
		源内存地址	输入	必选
		源内存大小	输入	必选
		内存复制模式	输入	必选
		关联的流	输入	必选
		执行返回码（见附录A）	输出	

表 82　内存拷贝（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
内存拷贝（异步）	人工智能计算设备（加速器）宿主机服务器内，或其与加速器设备之间，加速器设备之内，加速器设备之间的异步内存复制	目标内存地址	输入	必选
		目标内存容量	输入	必选
		源内存地址	输入	必选
		源内存大小	输入	必选
		内存复制模式	输入	必选
		关联的流	输入	必选
		执行返回码（见附录A）	输出	

表 83　初始化内存（同步）

名称	功能语义描述	参数	输入/输出	可选/必选
初始化内存（同步接口）	初始化内存，将内存中的内容设置为指定的值，同步接口	设备指针	输入	必选
		初始化的内存最大数量	输入	必选
		初始化的内存数量	输入	必选
		初始化的值	输入	必选
		执行返回码（见附录A）	输出	

表 84　初始化内存（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
初始化内存（异步接口）	初始化内存，将内存中的内容设置为指定的值，异步接口	设备指针	输入	必选
		初始化的内存最大数量	输入	必选
		初始化的内存数量	输入	必选
		初始化的值	输入	必选
		流	输入	必选
		执行返回码（见附录A）	输出	

表 85　获取设备内存信息

名称	功能语义描述	参数	输入/输出	可选/必选
获取设备内存信息	获取特定计算设备上可用内存的空闲大小和总大小。	设备指针	输入	必选
		需要查询的内存的属性值	输入	必选
		对应属性内存空闲大小	输出	
		对应属性内存总大小	输出	
		执行返回码（见附录A）	输出	

8.2.2.6 事件管理接口

表 86 创建事件

名称	功能语义描述	参数	输入/输出	可选/必选
创建事件	创建一个事件，可用于流之间的同步，主机和设备之间同步等场景	事件标识	输入	可选
		事件指针	输出	必选
		执行返回码（见附录A）	输出	

表 87 销毁事件

名称	功能语义描述	参数	输入/输出	可选/必选
销毁事件	销毁之前创建的一个事件	事件指针	输入	可选
		执行返回码（见附录A）	输出	

表 88 复位事件

名称	功能语义描述	参数	输入/输出	可选/必选
复位事件	等待流中的任务都完成后复位事件	事件指针	输入	可选
		流	输入	
		执行返回码（见附录A）	输出	

表 89 记录事件

名称	功能语义描述	参数	输入/输出	可选/必选
记录事件	在流中记录一个事件	事件指针	输入	可选
		流	输入	
		执行返回码（见附录A）	输出	

表 90 应用同步事件

名称	功能语义描述	参数	输入/输出	可选/必选
应用同步事件	阻塞应用程序进行，直到指定的事件完成	事件指针	输入	可选
		执行返回码（见附录A）	输出	

表 91 流同步事件

名称	功能语义描述	参数	输入/输出	可选/必选
流同步事件	阻塞流运行，直到指定的事件完成，支持多个流等待同一个事件场景	事件指针	输入	可选
		执行返回码（见附录A）	输出	

表 92　应用同步设备

名称	功能语义描述	参数	输入/输出	可选/必选
应用同步设备	阻塞应用程序，直到正在运算中的设备完成运算	执行返回码（见附录A）	输出	

表 93　应用同步流

名称	功能语义描述	参数	输入/输出	可选/必选
应用同步流	阻塞应用程序，直到指定流中的所有任务都完成	流	输入	可选
		执行返回码（见附录A）	输出	

表 94　设置等待事件完成的超时时间

名称	功能语义描述	参数	输入/输出	可选/必选
设置等待事件完成的超时时间	设置等待事件完成的超时时间，用于事件无法处理完成的长期阻塞场景	超时时间	输入	可选
		执行返回码（见附录A）	输出	

表 95　获取事件执行状态

名称	功能语义描述	参数	输入/输出	可选/必选
获取事件执行状态	获取记录的事件是否执行完成	事件指针	输入	可选
		状态指针	输出	
		执行返回码（见附录A）	输出	

表 96　获取事件接收状态

名称	功能语义描述	参数	输入/输出	可选/必选
获取事件接收状态	获取流中是否接收到制定的事件	事件指针	输入	可选
		状态指针	输出	
		执行返回码（见附录A）	输出	

表 97　统计事件之间耗时

名称	功能语义描述	参数	输入/输出	可选/必选
统计事件之间耗时	统计两个事件之间的耗时	起始事件	输入	可选
		中止事件	输入	
		耗时	输出	
		执行返回码（见附录A）	输出	

表 98　注册处理 stream 上回调函数的线程

名称	功能语义描述	参数	输入/输出	可选/必选
注册处理stream上回调函数的线程	注册处理stream上回调函数的线程，用于异步任务场景下	线程ID	输入	可选
		流	输入	
		执行返回码（见附录A）	输出	

表 99　取消注册处理 stream 上回调函数的线程

名称	功能语义描述	参数	输入/输出	可选/必选

取消注册处理stream上回调函数的线程	取消注册处理stream上回调函数的线程，用于异步任务场景下	线程ID	输入	可选
		流	输入	
		执行返回码（见附录A）	输出	

表 100　stream 任务队列上新增回调函数（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
流任务队列上新增回调函数	在流的任务上增加一个需要执行的回调函数	新增的回调函数	输入	可选
		传递给回调函数的用户数据指针	输入	
		指定回调函数调用是否会阻塞设备运行	输入	
		流	输入	
		执行返回码（见附录A）	输出	

表 101　设置回调处理超时时间

名称	功能语义描述	参数	输入/输出	可选/必选
设置回调处理超时时间	等待制定时间后，触发回调处理	超时时间	输入	可选
		执行返回码（见附录A）	输出	

8.2.2.7　模型加载与执行接口

表 102　从文件加载离线模型

名称	功能语义描述	参数	输入/输出	可选/必选
从文件加载离线模型	从文件加载离线模型。加载后模型将与输入的模型标识符关联。	模型文件路径	输入	必选
		模型标识符	输出	可选
		模型执行参数的列表（由参数名与参数值的映射对组成）	输入	可选
		执行返回码（见附录A）	输出	

表 103　从内存加载离线模型

名称	功能语义描述	参数	输入/输出	可选/必选
从内存加载离线模型	从内存加载离线模型	模型地址	输入	必选
		模型体量	输入	必选
		模型执行参数的列表（由参数名与参数值的映射对组成）	输入	可选
		模型标识符	输出	可选
		执行返回码（见附录A）	输出	

表 104　获取模型执行需要的内存大小（模型保存在文件中）

名称	功能语义描述	参数	输入/输出	可选/必选
获取模型执行需要的内存大小	根据模型文件，获取模型执行时所需的内存大小	模型文件路径	输入	必选
		模型执行所需的工作内存大小	输出	必选
		模型执行所需的权值内存大小	输出	
		执行返回码（见附录A）	输出	

表 105　获取模型执行需要的内存大小（模型保存在内存中）

名称	功能语义描述	参数	输入/输出	可选/必选
获取模型执行需要的内存大小	根据模型数据，获取模型执行时所需的内存大小	模型地址	输入	必选
		模型体量	输入	必选
		模型执行所需的工作内存大小	输出	必选
		模型执行所需的权值内存大小	输出	
		执行返回码（见附录A）	输出	

表 106　获取指定算子描述信息

名称	功能语义描述	参数	输入/输出	可选/必选
获取指定算子描述信息	获取指定算子的描述信息，包括算子名，输入张量描述等	设备标识	输入	必选
		流标识	输入	
		任务标识	输入	
		算子名称长度	输入	
		算子名字	输入	
		算子所有输入的张量描述	输出	
		算子输入个数	输出	
		算子所有输出的张量描述	输出	
		算子输出个数	输出	
		执行返回码（见附录A）	输出	

表 107　执行模型定义的（推理）运算

名称	功能语义描述	参数	输入/输出	可选/必选
执行模型定义的运算	执行模型推理，直到返回结果	模型标识符	输入	必选
		数据集（模型的输入）地址	输入	必选
		数据集（模型的输出）地址	输入	必选
		流	输入	必选
		执行返回码（见附录A）	输出	

表 108　执行模型定义的（推理）运算（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
执行模型定义的运算	执行模型推理，调用后马上返回。执行完成后，将自动填充指定的数据结构	模型标识符	输入	必选
		数据集（模型的输入）地址	输入	必选
		数据集（模型的输出）地址	输入	必选
		流	输入	必选
		执行返回码（见附录A）	输出	

表 109　卸载模型

名称	功能语义描述	参数	输入/输出	可选/必选
卸载模型	卸载模型，释放资源	模型标识符	输入	必选
		执行返回码（见附录A）	输出	

表 110　模型加载配置

名称	功能语义描述	参数	输入/输出	可选/必选
模型执行配置	设置模型配置对象的属性值，包含模型执行优先级、模型文件路径或内存地址、内存大小等	模型加载配置对象指针	输入	必选
		模型加载配置对象属性	输入	必选
		模型加载配置对象属性值	输入	必选
		附加值（如内存大小）	输入	可选
		执行返回码（见附录A）	输出	

表 111　指定模型加载配置并加载

名称	功能语义描述	参数	输入/输出	可选/必选
指定模型加载配置并加载	指定模型加载时需要的配置信息，并进行模型加载	模型加载的配置对象的指针	输入	必选
		模型标识符	输出	必选
		执行返回码（见附录A）	输出	

表 112　创建模型配置句柄

名称	功能语义描述	参数	输入/输出	可选/必选
创建模型配置句柄	创建句柄用于模型加载配置	执行返回码（见附录A）	输出	

表 113　销毁模型配置句柄

名称	功能语义描述	参数	输入/输出	可选/必选
销毁模型配置句柄	销毁用于模型加载配置的句柄	执行返回码（见附录A）	输出	

8.2.2.8　算子编译、加载与执行接口

表 114　从内存中加载单算子

名称	功能语义描述	参数	输入/输出	可选/必选
从内存中加载单算子	从内存中加载单算子模型数据	模型地址	输入	必选
		模型体量	输入	
		执行返回码（见附录A）	输出	

表 115　设置算子编译文件加载路径

名称	功能语义描述	参数	输入/输出	可选/必选
设置模型文件目录	设置加载模型文件的目录	文件指针	输入	必选
		执行返回码（见附录A）	输出	

表 116　根据输入推导算子输出 shape

名称	功能语义描述	参数	输入/输出	可选/必选
根据输入推导输出shape	根据输入shape、输入值推导出算子的输出shape	算子类型指针	输入	必选
		算子输入tensor数量	输入	
		算子输入tensor描述的指针数组	输入	

		算子输入tensor的指针数组	输入	
		算子输出tensor数量	输入	
		算子输出tensor的指针数组	输出	
		算子属性	输入	
		执行返回码（见附录A）	输出	

表 117　同步或异步执行特定算子

名称	功能语义描述	参数	输入/输出	可选/必选
同步或异步执行特定算子	同步或异步执行特定算子	算子类型名称的指针	输入	必选
		算子的输入张量的数量	输入	必选
		算子的输入张量的描述指针数组	输入	必选
		算子的输入张量的指针数组	输入	必选
		算子的输出张量的数量	输入	必选
		算子的输出张量的描述指针数组	输入	必选
		算子的输出张量的指针数组	输入	必选
		算子属性指针	输入	必选
		算子需加载的流	输入	必选
		执行返回码（见附录A）	输出	

表 118　编译并执行指定算子

名称	功能语义描述	参数	输入/输出	可选/必选
编译并执行指定算子	编译并执行指定算子	算子类型名称的指针	输入	必选
		算子的输入张量的数量	输入	必选
		算子的输入张量的描述指针数组	输入	必选
		算子的输入张量的指针数组	输入	必选
		算子的输出张量的数量	输入	必选
		算子的输出张量的描述指针数组	输入	必选
		算子的输出张量的指针数组	输入	必选
		算子属性指针	输入	必选
		算子执行环境类型	输入	必选
		编译标识，分为注册过的算子，和未注册过的算子，两类	输入	必选
		算子实现文件路径指针（不含文件名）	输入	必选
		算子需加载的流	输入	必选

		执行返回码（见附录A）	输出	

表 119　创建算子执行句柄

名称	功能语义描述	参数	输入/输出	可选/必选
创建算子执行接口	创建句柄用于算子执行	算子类型名称	输入	必选
		算子输入张量数量	输入	
		算子输入张量描述	输入	
		算子输出张量数量	输入	
		算子输出张描述	输入	
		算子属性	输入	
		算子执行句柄	输出	
		执行返回码（见附录A）	输出	

表 120　销毁算子执行句柄

名称	功能语义描述	参数	输入/输出	可选/必选
销毁算子执行句柄	销毁已经创建的算子执行句柄	算子执行句柄	输入	

表 121　句柄方式执行指定算子（异步）

名称	功能语义描述	参数	输入/输出	可选/必选
句柄方式执行指定算子	以句柄方式执行指定算子	算子句柄	输入	必选
		算子的输入张量的数量	输入	必选
		算子的输入张量的指针数组	输入	必选
		算子的输出张量的数量	输入	必选
		算子的输出张量的指针数组	输入	必选
		算子需加载的流	输入	必选
		执行返回码（见附录A）	输出	

8.2.2.9　数据结构接口

表 122　创建模型描述数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
创建模型描述数据结构	创建模型描述数据结构，用于表示模型信息	执行返回码（见附录A）	输出	必选

表 123　销毁模型描述数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
销毁模型描述数据结构	销毁模型描述数据结构，用于表示模型信息	执行返回码（见附录A）	输出	必选

表 124　获取模型描述数据结构

名称	功能语义描述	参数	输入/输出	可选/必选

获取模型描述数据结构	根据模型标识符获取该模型的描述信息	模型标识符	输入	必选
		模型描述数据结构	输出	可选
		执行返回码（见附录A）		

表 125　获取模型描述数据结构信息

名称	功能语义描述	参数	输入/输出	可选/必选
获取模型描述数据结构信息	根据模型数据结构获取该模型数据结构内的信息	模型描述数据结构	输入	必选
		模型的输入个数	输出	可选
		模型的输出个数	输出	可选
		第N个输入的大小（新增输入标识符N为本接口输入）	输出	可选
		第N个输出的大小（新增输出标识符N为本接口输入）	输出	可选
		第N个输入的张量维度信息（新增输入标识符N为本接口输入）	输出	可选
		第N个输出的张量维度信息（新增输出标识符N为本接口输入）	输出	可选
		第N个输入的名称（新增输入标识符N为本接口输入）	输出	可选
		第N个输出的名称（新增输出标识符N为本接口输入）	输出	可选
		第N个输入的类型（新增输入标识符N为本接口输入）	输出	可选
		第N个输出的类型（新增输出标识符N为本接口输入）	输出	可选
		第N个输入的格式（新增输入标识符N为本接口输入）	输出	可选
		第N个输出的格式（新增输出标识符N为本接口输入）	输出	可选
		执行返回码（见附录A）	输出	

注：实现中可拆分成多个接口实现，如获取模型描述数据结构的输入数据名称、输出输出格式。

表 126　创建模型输入输出数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
创建模型输入输出数据结构	用于创建描述模型的输入数据、输出数据，每个模型可能存在多个输入输出数据	执行返回码（见附录A）	输出	必选

表 127　销毁模型输入输出数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
销毁模型输入输出数据结构	用于销毁已经创建的模型的输入数据、输出数据，每个模型可能存在多	模型输入输出数据指针	输入	必选
		执行返回码（见附录A）	输出	

	个输入输出数据			

表 128　模型输入输出数据结构中增加数据

名称	功能语义描述	参数	输入/输出	可选/必选
模型输入输出数据结构中增加数据	用于在已经创建的模型的输入数据、输出数据中增加输入输出数据	模型输入输出数据指针	输入	必选
		输入输出数据指针	输入	必选
		执行返回码（见附录A）	输出	

表 129　获取输入输出数据结构中数据个数

名称	功能语义描述	参数	输入/输出	可选/必选
获取模型输入输出数据结构中的数据个数	用于获取已经创建的模型的输入数据、输出数据个数	模型输入输出数据指针	输入	必选
		执行返回码（见附录A）	输出	

表 130　创建模型单个输入输出数据类型

名称	功能语义描述	参数	输入/输出	可选/必选
创建模型单个输入输出数据类型	用于创建描述模型的单个输入数据、输出数据	存放数据内存地址指针	输入	必选
		内存大小	输入	
		执行返回码（见附录A）	输出	

表 131　销毁模型单个输入输出数据类型

名称	功能语义描述	参数	输入/输出	可选/必选
销毁模型单个输入输出数据类型	用于销毁已经创建的模型的单个输入数据、输出数据	模型单个输入输出数据指针	输入	必选
		执行返回码（见附录A）	输出	

表 132　获取模型单个输入输出数据信息

名称	功能语义描述	参数	输入/输出	可选/必选
获取模型单个输入输出数据信息	用于获取已经创建的模型的单个输入数据、输出数据信息	模型单个输入输出数据指针	输入	必选
		数据内存地址	输出	可选
		数据内存大小	输出	可选
		执行返回码（见附录A）	输出	

注：实现中可拆分成多个接口实现，如获取模型单个输入输出的数据内存地址接口、获取模型单个输入输出的数据内存大小接口。

表 133　创建张量描述数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
创建张量描述数据结构	张量数据结构用于描述张量的数据类型、格式等	张量数据类型	输入	必选
		张量维度个数	输入	
		张量维度指针	输入	
		张量格式	输入	
		执行返回码（见附录A）	输出	

表 134　销毁张量描述数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
销毁张量描述数据结构	用于销毁已经创建的张量描述数据结构	张量描述数据指针	输入	必选
		执行返回码（见附录A）	输出	

表 135　获取张量描述数据结构信息

名称	功能语义描述	参数	输入/输出	可选/必选
获取张量描述数据结构信息	用于获取张量结构的信息	张量描述数据指针	输入	必选
		张量描述名称	输出	可选
		张量数据的数据类型	输出	可选
		张量数据的格式	输出	可选
		张量数据占用的内存空间	输出	可选
		张量的元素个数	输出	可选
		张量shape的维度个数	输出	可选
		张量的指定维度的大小（新增第N个维度为本接口输入）	输出	可选
		张量的指定下标的张量描述信息（新增第N个下标为本接口输入）	输出	可选
		张量的指定下标的张量维度范围（新增第N个下标为本接口输入）	输出	可选
		执行返回码（见附录A）	输出	

注：实现中可拆分成多个接口实现，如获取张量的数据类型接口、获取张量的元素个数接口等。

表 136　设置张量描述数据结构信息

名称	功能语义描述	参数	输入/输出	可选/必选
设置张量描述数据结构信息	用于设置张量描述结构的信息	张量描述数据指针	输出	必选
		张量描述名称	输入	可选
		张量维度	输入	可选
		张量格式	输入	可选
		张量每个维度的取值范围	输入	可选
		算子const信息，指定存放输入tensor数据的内存地址和地址长度	输出	可选
		张量放置属性,用于标识存在张量数据的内存类型		
		执行返回码（见附录A）	输出	

注：实现中可拆分成多个接口实现，如设置张量的数据类型接口、设置张量的元素个数接口等。

表 137　创建算子属性数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
创建算子属性数据结构	用于保存算子的属性	执行返回码（见附录A）	输出	必选

表 138　销毁算子属性数据结构

名称	功能语义描述	参数	输入/输出	可选/必选

销毁算子属性数据结构	用于销毁已经创建的算子属性数据结构	算子属性数据指针	输入	必选
		执行返回码（见附录A）	输出	

表 139　设置算子属性数据结构

名称	功能语义描述	参数	输入/输出	可选/必选
设置算子属性数据结构	设置算子属性数据结构，包括布尔、整形、字符串等	算子属性数据指针	输出	必选
		算子的属性名	输入	可选
		算子的属性值	输入	可选
		算子的属性个数	输入	可选
		算子的属性值数组	输入	可选
		执行返回码（见附录A）	输出	

注：实现中可拆分成多个接口实现，如设置算子布尔属性值接口、设置算子字符串属性值接口等。

8.2.3　集合通信接口

为实施基于人工智能集群的分布式训练或推理，应实现表75~表89定义的通信接口。集合通信的范围是通信域，即参与集合通信操作的通信进程组及相关描述信息。

表 140　使用配置文件创建通信域

名称	功能语义描述	参数	输入/输出	可选/必选
使用配置文件创建通信域	使用配置文件（如通信进程表）初始化通信域	配置文件路径(含文件名)字符串。具体实现应限制路径字符串的最大长度（如4096字节），含结束符	输入	必选
		本通信进程的标识	输入	必选
		初始化后的通信域数据结构首地址指针	输出	
		执行返回码（见附录A）	输出	

表 141　使用根通信进程信息创建通信域

名称	功能语义描述	参数	输入/输出	可选/必选
使用根通信进程信息创建通信域	使用根通信进程信息初始化通信域。此操作一般被收到根通信进程信息（见表X）的进程调用，以与该根通信进程组成通信域	集群中，通信进程的数量 [a]	输入	必选
		根通信进程信息（如通信地址，标识等）	输入	必选
		本通信进程的标识	输入	必选
		初始化后的通信域数据结构首地址指针	输出	
		执行返回码（见附录A）	输出	
[a] 所有通信进程使用的通信进程数量和根通信进程信息应相同。				

表 142　获得根通信进程信息

名称	功能语义描述	参数	输入/输出	可选/必选
获得根通信进程信息	返回本通信进程的信息，如地址，设备标识等。在	根通信进程信息	输出	
		执行返回码（见附录A）	输出	

	组成通信域时，此信息广播到所有通信进程，以便完成集合通信初始化			

表 143　销毁通信域

名称	功能语义描述	参数	输入/输出	可选/必选
销毁通信域	销毁通信域	要销毁的通信域	输入	必选
		执行返回码（见附录A）	输出	

表 144　获得通信进程数量

名称	功能语义描述	参数	输入/输出	可选/必选
获得通信域通信进程的数量	获得通信域中通信进程的数量	通信域	输入	必选
		通信进程数量值地址的指针	输出	
		执行返回码（见附录A）	输出	

表 145　获得本通信进程数的标识

名称	功能语义描述	参数	输入/输出	可选/必选
获得本通信域通信进程的标识	获得本通信域中通信进程的标识	所在的通信域	输入	必选
		通信进程标识值地址的指针	输出	
		执行返回码（见附录A）	输出	

表 146　all-reduce

名称	功能语义描述	参数	输入/输出	可选/必选
all-reduce	将所有通信进程的发送缓冲区中的数据执行特定操作（如相加），再把结果发送到所有通信进程的接收缓冲	本通信进程源数据输入缓冲地址	输入	必选
		本通信进程输出数据缓冲地址，操作结果输出至此缓冲	输出	
		参与此操作的数据个数，按数据类型为单位计数 [a]	输入	必选
		参与此操作的数据类型，如四分之一精度整型，二分之一精度整型，整型，二分之一精度浮点，浮点等（可按需选择支持的数据类型）[a]	输入	必选
		操作类型，如加、乘、取最大值、取最小值等（可按需选择支持的操作类型）[a]	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的流	输入	可选
		执行返回码（见附录A）	输出	
[a] 所有通信进程的数据个数、数据类型和操作应相同。				

表 147　广播

名称	功能语义描述	参数	输入/输出	可选/必选
广播	将发起广播通信进程[b]的特定数据发送到其它通信进程	发送数据的缓冲（数据源）	输入	必选
		接收数据的缓冲	输出	
		参与此操作的数据个数，按数据类型为单位计数 [a]	输入	必选
		参与此操作的数据类型，如四分之一精度整型，四分之一精度无符号整型，二分之一精度整型，二分之一无符号整型，整型，无符号整型，双精度整型，无符号双精度整型，二分之一精度整型，浮点，双精度浮点等（可按需选择支持的数据类型）[a]	输入	必选
		此操作的根通信进程标识	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的流	输入	可选
		执行返回码（见附录A）	输出	
[a] 所有通信进程的数据个数、数据类型和操作应相同。 [b] 1次操作只允许存在1个通信进程发送数据。				

表 148　all-gather

名称	功能语义描述	参数	输入/输出	可选/必选
all-gather	将所有通信进程的特定数据，按通信进程标识的顺序拼接起来，将结果发送到所有通信进程的接收缓冲。	本通信进程的源数据输入缓冲地址	输入	必选
		本通信进程的输出数据缓冲地址，结果输出至此缓冲	输出	
		发送数据尺寸。接收数据缓冲的尺寸是数据尺寸与参与此操作的通信进程数量的积	输入	必选
		参与此操作的数据类型，如四分之一精度整型，四分之一精度无符号整型，二分之一精度整型，二分之一无符号整型，整型，无符号整型，双精度整型，无符号双精度整型，二分之一精度整型，浮点，双精度浮点等（可按需选择支持的数据类型）	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的流	输入	可选

		执行返回码（见附录A）	输出	

表 149 reduce-scatter

名称	功能语义描述	参数	输入/输出	可选/必选
reduce-scatter	对所有通信进程中发送缓冲的数据，实施特定操作（如相加），再把结果按照通信进程标识分散到各通信进程的接收缓冲。	本通信进程的源数据输入缓冲地址	输入	必选
		本通信进程的输出数据缓冲地址，结果输出至此缓冲	输出	
		接收数据尺寸。发送数据缓冲的尺寸是接收数据尺寸与参与此操作的通信进程数量的积	输入	必选
		参与此操作的数据类型，如四分之一精度整型，二分之一精度整型，整型，二分之一精度整型，浮点等（可按需选择支持的数据类型）[a]	输入	必选
		操作类型，如加、乘、取最大值、取最小值等（可按需选择支持的操作类型）[a]	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的流	输入	可选
		执行返回码（见附录A）	输出	
[a] 所有通信进程的接收数据尺寸、数据类型和操作应相同。				

表 150 reduce

名称	功能语义描述	参数	输入/输出	可选/必选
reduce	对所有通信进程中发送缓冲的数据，实施特定操作（如相加），再把结果发送到根进程的接收缓冲	本通信进程的源数据输入缓冲地址	输入	必选
		根通信进程的输出数据缓冲地址，结果输出至此缓冲	输出	
		参与此操作的数据个数，按数据类型为单位计数[a]。	输入	必选
		参与此操作的数据类型，如四分之一精度整型，二分之一精度整型，整型，二分之一精度整型，浮点等（可按需选择支持的数据类型）[a]	输入	必选
		操作类型，如加、乘、取最大值、取最小值等（可按需选择支持的操作类型）[a]	输入	必选
		此操作的根通信进程标识	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的	输入	可选

		流		
		执行返回码（见附录A）	输出	
[a] 所有通信进程的数据个数、数据类型和操作应相同。				

表 151　获得通信进程总数

名称	功能语义描述	参数	输入/输出	可选/必选
获得通信进程总数	返回当前通信域中所有通信进程的总数	此操作所在的通信域	输入	必选
		通信进程总数地址指针	输出	
		执行返回码（见附录A）	输出	

表 152　发送数据

名称	功能语义描述	参数	输入/输出	可选/必选
发送数据	发送当前通信进程的发送缓冲中的数据到目标通信进程的接收缓冲	本通信进程的源数据输入缓冲地址	输入	必选
		参与此操作的数据个数	输入	必选
		参与此操作的数据类型，如四分之一精度整型，四分之一精度无符号整型，二分之一精度整型，二分之一无符号整型，整型，无符号整型，双精度整型，无符号双精度整型，二分之一精度整型，浮点，双精度浮点等（可按需选择支持的数据类型）[a]	输入	必选
		数据接收通信进程的标识	输入	必选
		此操作所在的通信域	输入	必选
		此操作的发起通信进程所使用的流	输入	可选
		执行返回码（见附录A）	输出	

表 153　接收数据

名称	功能语义描述	参数	输入/输出	可选/必选
接收数据	从特定通信进程接收数据，存在本通信进程的接收缓冲中	接收数据通信进程的缓冲地址	输入	必选
		参与此操作的数据个数	输入	必选
		参与此操作的数据类型，如四分之一精度整型，四分之一精度无符号整型，二分之一精度整型，二分之一无符号整型，整型，无符号整型，双精度整型，无符号双精度整型，二分之一精度整型，浮点，双精度浮点等（可按需选择支持的数据类型）[a]	输入	必选

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/676132122231010135