

MindStudio
5.0.RC1

术语和缩略语

文档版本 01
发布日期 2022-06-02



版权所有 © 华为技术有限公司 2022。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

1 术语和缩略语.....	1
---------------	---

1 术语和缩略语

表 1-1 术语/缩略语

术语/缩略语	全称	含义
A		
AI	Artificial Intelligence	人工智能 研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。
AIPP	Artificial Intelligence Pre-Processing	AI预处理 AIPP用于在AI Core上完成图像预处理，包括改变图像尺寸、色域转换（转换图像格式）、减均值/乘系数（改变图像像素），数据处理之后再行真正的模型推理。
Ascend EP	Ascend Endpoint	昇腾AI处理器作为终端节点（从控节点），主要功能是配合主设备（X86，ARM等各种Server），快速高效的处理推理、训练、图像识别等工作，例如PCIe加速卡。
Ascend RC	Ascend Root Complex	昇腾AI处理器作为根组件（主控节点），提供主机控制功能，主要应用于移动端侧，例如Atlas 200 DK。

术语/缩略语	全称	含义
AscendCL	Ascend Computing Language	昇腾编程语言 提供Device管理、Context管理、Stream管理、内存管理、模型加载与执行、算子加载与执行、媒体数据处理等C语言的API库供用户开发深度神经网络应用，用于实现目标识别、图像分类等功能。
ASHA	Asynchronous Successive Halving Algorithm	异步连续减半算法 一种基于动态资源分配的超参优化算法。基础思想为：并行训练多组超参数，每轮进行少量的训练迭代。对所有超参数进行评估和排序，所有超参数排列在下半部分的训练都会提前停止。对剩余的超参数执行下一轮评估。评估再次减半，直到达到优化目标。
ATC	Ascend Tensor Compiler	昇腾张量编译器 <ul style="list-style-type: none">通过ATC，可以将开源框架的网络模型（如Caffe、TensorFlow等）转换成昇腾AI处理器支持的离线模型。模型转换过程中可以实现算子调度的优化、权值数据重排、内存使用优化等通过ATC，可以进行算子编译。
AutoML	Automated machine learning	自动机器学习 是特征提取、模型选择、参数调优等一系列自动化算法，可以实现自动训练有价值的模型。
B		

术语/缩略语	全称	含义
BOHB	Hyperband with Bayesian Optimization	在Hyperband基础上结合贝叶斯进行超参优化 BOHB依赖Hyperband来决定每次跑多少组参数和每组参数分配多少资源，Hyperband在每个循环开始时依赖之前的数据建立模型（贝叶斯优化）进行参数选择。
BOSS	Bayesian Optimization via Sub-Sampling	基于下采样的贝叶斯优化是基于贝叶斯优化框架下的一种针对计算资源受限，需要高效搜索的，具有普适性的超参优化算法。
BP	BP Point	训练网络迭代轨迹反向算子的结束位置
C		
CPU	Central Processing Unit	中央处理单元 计算机的主要设备之一，其功能是解释计算机指令以及处理计算机软件中的数据，与内部存储器、输入及输出设备成为现代电脑的三大部件。
D		
DDR	Double Data Rate	双倍数据速率 与传统的单数据速率相比，DDR技术实现了一个时钟周期内进行两次读/写操作，即在时钟的上升沿和下降沿分别执行一次读/写操作。
DiffThd	Different Threshold	误差阈值
DSL	Domain-Specific Language	基于特性域语言 算子开发方式之一，用户仅需要使用DSL接口完成计算过程的表达，后续的算子调度、算子优化及编译都可通过已有的接口一键式完成。

术语/缩略语	全称	含义
DVPP	Digital Vision Pre-Processing	数字视觉预处理 提供对特定格式的视频和图像进行解码、缩放等预处理操作，以及对处理后的视频、图像进行编码再输出的能力。
F		
FP	FP Point	训练网络迭代轨迹正向算子的开始位置
FpDiff	Floating-point Different	浮点误差
G		
GDB	GNU debugger	GNU调试器 GNU操作系统的标准调试器。
GE	Graph Engine	图引擎 提供了Graph/Operator IR作为安全易用的构图接口集合，用户可以调用这些接口构建网络模型，设置模型所包含的图、图内的算子、以及模型和算子的属性。
GPU	Graphics Processing Unit	图形处理器 GPU是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。
H		
HBM	High Bandwidth Memory	高带宽存储器 高带宽存储器是超微半导体和SK Hynix发起的一种基于3D堆栈工艺的高性能DRAM（dynamic random access memory），适用于高内存带宽需求的应用场合，像是图形处理器、网络交换及转发设备（如路由器、交换器）等。

术语/缩略语	全称	含义
HCCL	Huawei Collective Communication Library	华为集合通信库 提供了深度学习训练场景中服务器间高性能集合通信的功能。
HCCS	High Confidence Computing Systems	高性能计算系统 提供多卡场景下的高性能片间（Device间）数据通信能力。
HPO	Hyperparameter Optimization	超参数优化 是指用自动化的算法来优化原机器学习/深度学习算法中无法通过训练来优化的超参数，如学习率、激活函数、优化器等。
HWTS	Hardware Task Scheduler	硬件任务调度 提供对AI Core任务的硬件调度能力，减少调度时延。
I		
IR	Intermediate Representation	中间表示 IR是一种数据结构，可将输入的资料建构为一个计算机程序，也可以将一部分或是所有输出的程式反推回输入资料。
J		
JDK	Java Software Development Kit	Java软件开发包 基于Java的软件开发工具集合。
K		
KL散度	Kullback Leibler Divergence	KL散度算法 取值范围为0到无穷大。 KL散度越小，真实分布与近似分布之间的匹配越好。
L		
L2 Cache	Second Level Cache	二级缓存 在访问内存之前调用的共享第二级别缓存通常称为二级缓存。

术语/缩略语	全称	含义
LLC	Last Level Cache	最后一级缓存 在访问内存之前调用的共享最高级别缓存通常称为最后一级缓存（LLC）。
M		
msproftx	msprof tool extension	MindStudio系统调优工具扩展
MTE1	Memory Transfer Engine 1	内存传输引擎1 从L1 Buffer拷贝内存。
MTE2	Memory Transfer Engine 2	内存传输引擎2 从DDR或者L2 Buffer拷贝内存。
MTE3	Memory Transfer Engine 3	内存传输引擎3 从UB拷贝内存。
N		
NAS	Neural Architecture Search	神经架构搜索 一种自动设计神经网络的技术，可以通过算法根据样本集自动设计出高性能的网络结构，可以有效的降低神经网络的使用和实现成本。
NIC	Network Interface Controller	网络接口控制器 也称为网络接口卡、网络适配器、LAN适配器以及类似术语。是将计算机连接到计算机网络的计算机硬件组件。
NPU	Neural-Network Processing Unit	神经网络处理器单元 采用“数据驱动并行计算”的架构，特别擅长处理视频、图像类的海量多媒体业数据，专门用于处理人工智能应用中的大量计算任务。
O		
OP	Operator	算子 操作运算，比如AI的ReLU、Conv、Pooling、Scale、Softmax等。

术语/缩略语	全称	含义
OPP	Operator Package	算子库
OS	Operating System	操作系统
P		
PCIe	Peripheral Component Interconnect Express	快捷外围部件互连标准 PCIe属于高速串行点对点双通道高带宽传输。所连接的设备分配独享通道带宽，不共享总线带宽。主要支持主动电源管理、错误报告、端对端的可靠性传输、热插拔以及服务质量（QOS）等功能。
PctRlt	Percent Result	实际百分比
PctThd	Percent Threshold	百分比阈值
R		
RateDiff	Rate Different	误差比
RoCE	RDMA over Converged Ethernet	部署在以太网上RDMA的网络协议 RDMA是一种远程内存管理能力，允许不同服务器上应用的内存直接移动数据，不需要CPU的干预。RoCE是一种机制，提供了通信接口带宽数据。
RUNTIME	-	Runtime运行于APP进程空间，为APP提供了针对昇腾AI处理器的Memory管理、Device管理、Stream管理、Event管理和Kernel执行等功能。
S		
Sample-based	-	Profiling的AICore数据以固定的时间周期（AI Core-Sampling Interval）进行性能数据采集。
SDK	software development kit	软件开发工具包 一般都是一些软件工程师为特定的软件包、软件框架、硬件平台、操作系统等建立应用软件时的开发工具的集合。

术语/缩略语	全称	含义
Step Trace	-	迭代轨迹 包含迭代的正、反向计算 开始结束时间、梯度更新 以及数据增强拖尾阶段。
T		
Task-based	-	Profiling的AI Core数据以 task为粒度进行性能数据 采集。
TBE	Tensor Boost Engine	张量加速引擎 提供通过Python语言实现 算子的接口，能够编译生 成CCE算子。
Tensor	-	张量是TensorFlow程序中 的主要数据结构。张量是 N维（其中N可能非常大） 数据结构，最常见的是标 量、向量或矩阵。张量的 元素可以包含整数值、浮 点值或字符串值。
TIK	Tensor Iterator Kernel	张量嵌套内核 算子开发方式之一，调用 TIK提供的API基于Python 语言编写自定义算子，TIK 编译器会将其编译为适配 昇腾AI处理器应用程序的 二进制文件。
TransData	-	格式转换算子
TS	Task Scheduler	任务调度 通过Task Schedule分发不 同的kernel到AI CPU/AI Core执行。
V		
VECTOR	-	向量运算