ModelArts

服务公告

文档版本 01

发布日期 2024-05-22

华为云计算技术有限公司

商标声明


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云计算技术有限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，华为云计算技术有限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 下线公告 1
1.1 【下线公告】华为云 ModelArts 服务旧版训练管理下线公告 1
1.2 【下线公告】华为云 ModelArts 服务模型转换下线公告 2
1.3 【下线公告】华为云 ModelArts 旧版自动学习下线公告 2
1.4 【下线公告】华为云 ModelArts 服务旧版数据集下线公告 3
2 产品发布记录 4
2.1 昇腾云服务 6.3.T051 版本发布说明 4
2.2 昇腾云服务 6.3.904 版本发布说明 6
2.3 昇腾云服务 6.3.T041 版本发布 9
2.4 ModelArts 6.5.0 版本配套关系表 11

1 下线公告

1.1 【下线公告】华为云 ModelArts 服务旧版训练管理下线公告

华为云ModelArts服务旧版训练管理在2023年6月30日 00:00(北京时间)正式退市。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，用户将无法再使用旧版训练管理的功能，包括旧版训练作业、训练参数管理、可视化作业功能，建议将相关作业迁移到新版训练管理。

如您有任何问题，可随时通过工单或者服务热线（4000-955-988或950808）与我们联系。

常见问题

为什么要下线旧版训练管理?

ModelArts旧版训练全面上线以后为众多开发者提供了AI训练能力，其中训练服务作为基础服务之一，经过持续迭代已经无法完全满足众多开发者的新特性需求。基于服务演进，ModelArts团队已于2021年上线新版训练，力求解决存在的历史问题，并为新特性提供高性能、高易用、可扩展、可演进的底座，给用户提供更好的AI训练体验，打造易用、高效的AI平台。

下线旧版训练管理对现有用户的使用是否有影响?

正在使用的训练作业不受影响，但是用户无法使用旧版训练创建新的作业。

旧版训练管理是否停止新购?

是的，旧版训练管理将于2023年6月30日 00:00(北京时间)正式退市。

旧版训练管理如何升级到新版训练?

请参考新版训练指导文档（**模型训练**）来体验新版训练。

旧版训练迁移至新版训练需要注意哪些问题?

新版训练和旧版训练的差异主要体现在以下3点，详细内容请参见**旧版训练迁移至新版训练注意事项**。

- 新旧版创建训练作业方式差异
- 新旧版训练代码适配的差异
- 新旧版训练预置引擎差异

1.2 【下线公告】华为云 ModelArts 服务模型转换下线公告

华为云ModelArts服务模型转换在2024年4月30日 00:00(北京时间)正式下线。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，用户将无法再使用模型转换的功能，包括创建和删除模型转换任务、查询模型转换任务列表和详情功能。

如您有任何问题，可随时通过工单或者服务热线（4000-955-988或950808）与我们联系。

常见问题

为什么要下线模型转换?

ModelArts模型转换向AI开发者提供了便捷的模型转换页面，将Tensorflow和Caffe框架的模型格式转换为MindSpore的模型格式，即模型后缀为.om，使之能在昇腾硬件中进行推理。由于产品演进规划，后续昇腾硬件推理时主要使用后缀为.mindir的模型格式，因此ModelArts下线.om格式的模型转换能力，在ModelArts中逐步增加.mindir格式的支持能力。

下线模型转换后是否有替代功能?

您可以通过链接下载**ATC模型转换工具**，按照指导，在线下转换成.om格式模型。

ModelArts中是否还会增加模型转换的能力?

ModelArts开发环境中在贵阳一Region，支持将ONNX或PyTorch模型转换到.mindir格式。其它能力在持续增加中。若您暂时无法在该region中使用该能力，您可以通过链接下载**MindSpore Lite离线转换模型工具**，线下将其转换为.mindir格式模型。

1.3 【下线公告】华为云 ModelArts 旧版自动学习下线公告

华为云ModelArts在2024年5月15日 00:00（北京时间）用新版自动学习全面替代旧版自动学习，旧版自动学习正式下线。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，用户将无法再使用旧版自动学习的功能，且无法找回旧版自动学习的作业记录。

如您有任何问题，可随时通过工单或者服务热线（4000-955-988或950808）与我们联系。

常见问题

为什么要下线旧版自动学习?

ModelArts自动学习是帮助用户实现AI应用的低门槛、高灵活、零代码的定制化模型开发工具。ModelArts团队对自动学习模块进行了架构与前端页面的升级，新版自动学习已于2023年6月上线，并已作为主入口面向用户开放，用户可实现在租户账号下管理个人的作业与资源。

下线旧版自动学习对现有用户的使用是否有影响?

用户将无法再使用旧版自动学习的功能，且因旧版自动学习文件均存储于ModelArts统一管理账号下，用户无法找回旧版自动学习的作业记录。

旧版自动学习如何升级到新版自动学习?

请参考**新版自动学习指导文档**来体验新版自动学习。

1.4 【下线公告】华为云 ModelArts 服务旧版数据集下线公告

华为云计划于2024/10/31 00:00（北京时间）用AI开发平台ModelArts的新版数据集全面替代旧版数据集，旧版数据集正式下线。

下线范围

下线区域：华北-北京四（其他区域已下线）

受影响服务

ModelArts旧版数据集。

下线影响

正式下线后，所有用户将无法使用旧版数据集。为了避免影响您的业务，建议您在2024/10/30 23:59:59（北京时间）前备份数据或切换至新版数据集。

如您有任何问题，可随时通过工单或者服务热线（4000-955-988或950808）与我们联系。

2 产品发布记录

2.1 昇腾云服务 6.3.T051 版本发布说明

昇腾云服务6.3.T051版本发布支持的软件包和能力说明如下，软件包获取路径：**Support网站**。

此版本仅支持部分客户的beam-search、AWQ量化和SmoothQuant量化特性使用。

发布包	软件包特性说明	配套说明	备注
昇腾云模型代码	包名：AscendCloud-3rdLLM **三方大模型**，包名：AscendCloud-3rdLLM PyTorch框架下支持如下模型训练： 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B PyTorch框架下支持如下模型推理： 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b	配套CANN8.0.RC2镜像（非商发） 其中Llama/Llama2/Llama3系列、Qwen系列、Qwen1.5系列推理支持AWQ（W4A16），SmoothQuant(W8A8)量化 所有推理请求均支持beam-search短期方案。	无

发布包	软件包特性说明	配套说明	备注
	17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.baichuan2-7b 26.baichuan2-13b 27.chatglm2-6b 28.chatglm3-6b 29.gemma-2b 30.gemma-7b 31.mistral-7b		
	算子，包名：AscendCloud-OPP 配套W4A16和W8A8的算子	配套CANN8.0.RC2镜像(非商发）	无
基础镜像	**PyTorch**： 西南-贵阳一 swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc2-py_3.9-hce_2.0.2312-aarch64-snt9b-20240518201626-e439695	镜像发布到SWR（非商发）	无

2.2 昇腾云服务 6.3.904 版本发布说明

昇腾云服务6.3.904版本发布支持的软件包和能力说明如下，软件包获取路径：**Support-E网站**。

发布包	软件包特性说明	配套说明	备注
昇腾云模型代码	**三方大模型**，包名：AscendCloud-3rdLLM PyTorch框架下支持如下模型训练： 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b PyTorch框架下支持如下模型推理： 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.baichuan2-7b 26.baichuan2-13b 27.chatglm2-6b	配套CANN8.0.RC1镜像	训练参考文档： **LLama2系列（PyTorch）基于DevServer训练指导** **Qwen系列（PyTorch）基于DevServer训练指导** **GLM3-6B（PyTorch）基于DevServer训练指导** **Baichuan3-13B（PyTorch）基于DevServer训练指导** 推理参考文档： **主流开源大模型（PyTorch）基于DevServer推理部署**

发布包	软件包特性说明	配套说明	备注
	28.chatglm3-6b 29.gemma-2b 30.gemma-7b 31.mistral-7b		
	AIGC，包名：ascendcloud-aigc 1. Controlnet插件支持NPU推理（适配ComfyUI） 2. Open-Clip模型昇腾适配 3. SD1.5 Finetune 高性能训练 4. moondream2推理适配昇腾 5. BERT、YOLO等8个常用模型适配	配套CANN8.0.RC1镜像	参考文档 **SDXL文生图ComfyUI插件基于DevServer适配NPU推理指导** **Open-Clip基于DevServer适配PyTorch NPU 训练指导** **SD1.5文生图Finetune高性能训练适配NPU指导** **moondream2 基于DevServer适配PyTorch NPU 推理指导** **BERT和YOLO等常用小模型适配NPU推理指导**
	大模型推理框架，包名：ascendcloud-llmframework 1. VLLM调度层适配ATB、pybind 2. 支持LLAMA7B/13B/65B 3. 支持单机多卡推理 4. ATB模式支持w8a16量化，推理性能提升	配套CANN8.0.RC1镜像	无

发布包	软件包特性说明	配套说明	备注
	算子，包名：AscendCloud-OPP 1. Scatter、Gather算子性能提升，满足MoE场景 2. 昇腾随机数生成算子与GPU保持一致 3. 支持GroupNorm+transpose+BMM融合算子 4. FFN推理算子支持geglu激活函数 5. 支持配套pybind推理的10+算子（ matmul 、swiglu、rope等 ）	配套CANN8.0.RC1镜像	无
基础镜像	CANN8.0.RC1商发版本 **MindSpore**： 西南-贵阳一：swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 华北-乌兰察布一：swr.cn-north-9.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 **PyTorch**： 西南-贵阳一：swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 华北-乌兰察布一：swr.cn-north-9.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42	商发镜像发布到SWR	无

2.3 昇腾云服务 6.3.T041 版本发布

昇腾云服务6.3.T041版本发布支持的软件包和能力如下。

发布包	软件包特性说明	镜像配套说明	对应操作指导
昇腾云模型代码	包名：AscendCloud-3rdLLM-6.3.T041-20240424144057.zip 包含大语言模型，具体如下： 1.Qwen-7b 2.Qwen-14b 3.Qwen-72b 4.Llama2-7b 5.Llama2-13b 6.Llama2-70b 7.GLM3-6b	配套CANN7.0的商发基础镜像，地址： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b-20240411153110-ca68771	Qwen系列（PyTorch）基于DevServer训练指导 LLama2系列（PyTorch）基于DevServer训练指导 GLM3-6B（PyTorch）基于DevServer训练指导
	包名：ascendcloud-aigc-6.3.T041-20240425172135.tar.gz 包含AIGC模型，具体如下： 1.SD 1.5 2.SD XL	配套CANN8.0.RC1镜像，见基础镜像Beta包： mindspore_2.3.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b-20240422202644-39b975b.tar.partxx	SD1.5（PyTorch）文生图Finetune高性能训练适配NPU指导 SD1.5（PyTorch）文生图适配MindSpore Lite NPU推理指导 SDXL（PyTorch）文生图适配MindSpore-Lite NPU推理指导
基础镜像Beta包	包含模型代码运行的基础镜像，具体如下： mindspore_2.3.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b-20240422202644-39b975b.tar.partxx	此基础镜像是非商发版本，仅适配了SD1.5和SD XL模型，后续用CANN8.0.RC1的商发版本收编	镜像为分卷压缩，需要合并后使用
临时镜像（生态伙伴用）	包含SD 1.5自研模型，代码以及镜像 anime_vid2vid_code_split.tar.gzxx	此镜像仅提供给伙伴使用。	镜像为分卷压缩，需要合并后使用

昇腾云服务6.3.T041版本目前仅适用于部分企业客户，如需使用请联系您所在企业的华为方技术支持。

2.4 ModelArts 6.5.0 版本配套关系表

当前华为云中国站和国际站所有Region均已上线ModelArts 6.5.0版本。 ModelArts6.5版本中针对Ascend snt9B资源的周边依赖组件配套版本关系如下表所示。

表 2-1 ModelArts 6.5.0 版本配套关系表

强依赖组件	Ascend snt9B配套版本
CCE	1.25/1.23（推荐）/1.21
Volcano插件	1.11.9
Device-Plugin	2.1.5
Lite模式DevServer节点操作系统	EulerOS 2.10
Lite模式Cluster节点操作系统	EulerOS 2.10（CCE标准版）/HCE2.0（CCE Turbo）
Standard模式集群节点操作系统	EulerOS 2.10（CCE标准版）
BMS BMC	3.10.02.49（推荐）/3.10.02.29
BMS BIOS	7.09（推荐）/6.63
BMS CPLD	主板CPLD：3.03 背板CPLD：2.07
NPU MCU	23.3.5
NPU 固件&驱动	7.1.0.7.220-23.0.5（推荐） 7.1.0.5.220-23.0.3
NPU CANN	7.0.1.1（推荐） 7.0.1 8.0.RC1
NPU MindSpore	2.2.12（推荐） 2.2.10 2.3.0rc1
NPU PyTorch	2.1.0（推荐） 2.2.0 1.11.0

强依赖组件	Ascend snt9B配套版本
预置统一镜像	pytorch_2.1.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b mindspore_2.2.12-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b pytorch_1.11.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b
SFS Turbo Client+	23.09.03
AI Turbo SDK	23.12.3