天数智芯

PaddlePaddle 框架功能说明

版本：V4.0.0-MR

日期：2024.4.25

适用产品：智铠 50| 智铠 100

1 声明

1.1 版权声明

1.2 免责声明

天数智芯可以随时对本文档或本文档中描述的产品进行改进和/或更改。本文档包括与天数智芯产品有关的信息，作为说明典型应用的一种方式, 因此，不一定提供足以进行生产设计的完整信息。对于本文档中内容的准确性或完整性，天数智芯不做任何陈述或保证。

1.3 联系方式

地址：上海闵行区陈行公路 2168 号 3 幢

电话：021-68886607

网址：www.iluvatar.com

Contents

2 PaddlePaddle 框架功能说明

2.1 PaddlePaddle 框架简介

天数智算软件栈提供天数智芯适配版的 PaddlePaddle v2.5.2 深度学习编程框架。

飞桨 PaddlePaddle 以百度多年的深度学习技术研究和业务应用为基础，是中国首个自主研发、功能完备、开源开放的产业级深度学习平台，集深度学习核心训练和推理框架、基础模型库、端到端开发套件和丰富的工具组件于一体。更多使用说明，您可参考官方主页 飞桨 PaddlePaddle-源于产业实践的开源深度学习平台 和 GitHub 。

2.2 支持的特性

天数智芯适配版 PaddlePaddle 支持 x86 架构，已在一个或多个模型训练推理中验证以下特性。

以下两个环境变量为保证性能，默认值与 PaddlePaddle 官方文档中不同：

- FLAGS_cudnn_exhaustive_search，默认值为 True，PaddlePaddle 官方文档中为 False
- FLAGS_cudnn_batchnorm_spatial_persistent，默认值为 True，PaddlePaddle 官方文档中为 False

2.2.1 分布式训练

天数智芯适配版 PaddlePaddle 目前支持单机多卡训练，后端为 ixCCL，详细说明请参考 分布式训练-使用文档-PaddlePaddle 深度学习平台 。

2.2.2 AMP 自动混合精度训练

天数智芯适配版 PaddlePaddle 目前支持 PaddlePaddle 混合精度训练 AMP O1 和 AMP O2 级别的优化运行。

混合精度训练是指在训练过程中，同时使用单精度（FP32）和半精度（FP16），其目的是相较于使用单精度（FP32）训练模型，在保持精度持平的条件下，能够加速训练。使用飞桨框架提供的 API，能够在原始训练代码基础上快速开启自动混合精度训练（Automatic Mixed Precision，AMP），即在相关 OP 的计算中，根据一定的规则，自动选择 FP16 或 FP32 计算。

依据 FP16 在模型中的使用程度划分，飞桨的 AMP 分为两个等级：

- level = 'O1'：采用黑白名单策略进行混合精度训练，黑名单中的 OP 将采用 FP32 计算，白名单中的 OP 将采用 FP16 计算，训练过程中框架会自动将白名单 OP 的输入参数数据类型从 FP32 转为 FP16，使用 FP16 与 FP32 进行计算的 OP 列表可参考paddle.amp 官方文档。
- level = 'O2'：该模式采用了比 O1 更为激进的策略，除了框架不支持 FP16 计算的 OP，其它全部采用 FP16 计算，框架会预先将网络参数从 FP32 转换为 FP16，相比 O1，训练过程中无需做 FP32 转为 FP16 的操作，训练速度会有更明显的提升，但可能会存在精度问题，为此，框架提供了自定义黑名单，您可通过该名单指定一些存在精度问题的 OP 执行 FP32 运算。

2.2.3 推理部署

天数智芯适配版 PaddlePaddle 支持 PaddlePaddle 框架原生的推理库 Paddle Inference。您可以使用 Paddle Inference 进行云端、服务器端的推理部署加速。

由于能力直接基于飞桨的训练算子，因此 Paddle Inference 可以通用支持飞桨训练出的所有模型。

Paddle Inference 功能特性丰富，性能优异，针对不同平台不同的应用场景进行了深度的适配优化，做到高吞吐、低时延，保证了飞桨模型在服务器端即训即用，快速部署。

详细说明文档参考：推理部署-使用文档-PaddlePaddle 深度学习平台

2.2.4 FlashAttention 支持

天数智芯适配版 PaddlePaddle 支持通过自有算子库加速的 FlashAttention 算法，大幅提高了大模型训练的效率。

2.2.5 版本更新

天数智芯适配版 PaddlePaddle 从 v2.4 版本升级到 v2.5 版本，支持动静统一新架构，大幅提升了调度性能；PHI 算子库算子架构统一，提升了架构统一性，降低了框架开发的理解成本。

详细的更新信息，请参考 PaddlePaddle v2.5.0 Release Note。

2.3 已验证通过的网络模型

天数智芯适配版 PaddlePaddle 框架已验证对如下网络模型的支持：

- BERT_finetune
- ResNet50
- ResNet50_dist

3 商标声明

- 天数智芯、天数智芯 logo、Iluvatar CoreX 等商标、标识、组合商标为上海天数智芯半导体有限公司之注册商标或商标，受法律保护。
- 除了天数智芯的注册商标外，本内容中使用的其他产品名称及标志仅用于识别目的，该等名称及标志可能是归属于其各自公司的商标。我们否认对该等名称及标志的所有权利。
- CentOS 标识为 Red Hat 公司的商标。
- Docker 为 Docker 公司在美国和其他国家的商标或注册商标。
- Linux 为 Linus Torvalds 在美国和其它国家的注册商标。
- NVIDIA 和 CUDA 为 NVIDIA 公司在美国和/或其它国家的商标和/或注册商标。
- PyTorch 为 Facebook 公司的商标。
- TensorFlow 为 Google 公司的商标。
- Ubuntu 为 Canonical 公司的注册商标。