TI-ACC 加速工具

产品简介

产品文档

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2023 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【商标声明】

腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100。

文档目录

产品简介

产品概述

最近更新时间：2022-06-27 11:35:44

TI-ACC 加速工具（TI Acceleration Service，TI-ACC）为企业提供 AI 模型训练、推理加速服务，支持多种框架和场景，显著提高模型训练推理效率，降低用户成本。

产品功能

TI-ACC 训练加速

TI-ACC 训练加速基于腾讯云帆 Light 常年内外部项目打磨验证，在推荐、CV、NLP 等模型训练场景中，实现数据 IO 优化、计算加速、通信加速、并行训练、显存优化等能力，能力在 腾讯云 TI 平台-训练工坊 通过统一的加速库以及简单易用的函数/类的形式提供，并很好的兼容原生 PyTorch、TensorFlow 框架和 DDP、PS 工具，可帮助客户以较低的使用门槛在进行模型训练时，显著节省训练时间和计算成本。

TI-ACC 推理加速

TI-ACC 推理加速基于腾讯云帆 TNN 常年内外部项目打磨验证，在推荐、CV、NLP 等模型推理场景中，实现计算优化、低精度加速、内存优化等能力，能力通过统一的加速库和优化函数的形式提供，用户可通过 腾讯云 TI 平台-模型优化 通过创建简单的一个优化任务即可进行推理加速优化，支持多种模型输入格式、多种优化级别、固定&动态输入维度、输出测试报告以及对模型进行保存输出，并很好的兼容原生 PyTorch等框架，无需进行模型转换，可帮助客户以较低的使用门槛在进行模型推理时，显著节省推理时间和计算成本。

版权所有：腾讯云计算（北京）有限责任公司

产品优势

最近更新时间：2022-06-27 11:35:48

性能优越

基于业界领先的 AI 加速技术，提供高性能模型训练、推理加速服务，可显著提升性能，其中训练加速能力基于腾讯云帆 Light 在内部游戏 AI、小红书、虎牙等多个项目验证，推理加速能力基于腾讯云帆 TNN 在手Q、微视等多个项目落地。

推理加速实测数据

硬件环境	模型	Batchsize	torchscript（ms）	TI-ACC（ms）	加速比
腾讯云TI 平台 32C128G T4 * 1	resnet50(torchvision)224x224	1	5.4622	1.1482	4.8x
		8	27.062	4.5707	5.9x
	resnet50(mmcls)224x224	1	7.7667	4.3958	1.8x
		8	36.806	14.1152	2.6x
	centernet640x640	1	20.9992	4.7775	4.4x
		8	170.5488	34.3523	5.0x
	yolov3(ultralytics)640x640	1	47.19	10.3671	4.5x
		8	302.983	82.6971	3.7x
	Cascade Mask R-CNN(mmdet)2016x3008	1	600.0671	165.8467	3.6x
	Faster R-CNN(mmdet)1088x800	1	107.3483	35.5021	3.0x

	Vision Transformer224x224	8	28.887	10.53	2.7x
	Wide & Deep(NVIDIA DeepLearningExamples)	512	15.7	4.436	3.5x
	DeepFM(NVIDIA DeepLearningExamples)	512	12.91	4.51	2.9x

训练加速-DDP 通信优化实测效果

硬件环境	模型	GPU 卡数	原生 DDP (examples/sec per V100)	TI-ACC 通信优化 (examples/sec per V100)
腾讯云 TI 平台 80C320G V100 * 8	resnext50_32x4d	1（单机）	227	227
		8（单机）	215	215
		16（双机）	116	158.6

训练加速-数据 IO 优化实测效果

硬件环境	模型	GPU 卡数	原生 DDP（examples/sec per V100）	TI-ACC 数据 IO 优化（examples/sec per V100）
腾讯云 TI 平台 80C320G V100 * 8	resnet50 mmcls	8（单机）	70.8	350.5
	centernet mmdet	8（单机）	26.4	28.6

训练加速-自适应混合精度优化实测效果

硬件环境	模型	GPU 卡数	原生 DDP（examples/sec per V100）	TI-ACC 数据 IO 优化（examples/sec per V100）	TI-ACC 数据 IO + 自适应混合精度优化（examples/sec per V100）

版权所有：腾讯云计算（北京）有限责任公司

腾讯云 TI 平台 80C320G V100 * 8	resnet50 mmcls	8（单机）	70.8	350.5	379.2
	centernet tmmdet	8（单机）	26.4	28.6	30.6

训练加速-PS 相关优化实测效果

硬件环境	模型	GPU 卡数	原生 TensorFlow (global_step/sec)	TI-ACC 优化后 (global_step/sec)
腾讯云 TI平台 80C320G V100 * 8	DeepFM	16（双机）	41.9-56	96.1-103.3
	Wide & Deep	16（双机）	49.9-69	120-128

功能丰富

- 训练加速底层通过接口提供数据 IO 优化、自适应FP16、通信加速等功能。
- 推理加速底层通过接口支持多种模型输入格式、多种优化级别、固定&动态输入维度、自定义测试数据输出测试报告以及对模型进行保存输出等功能。

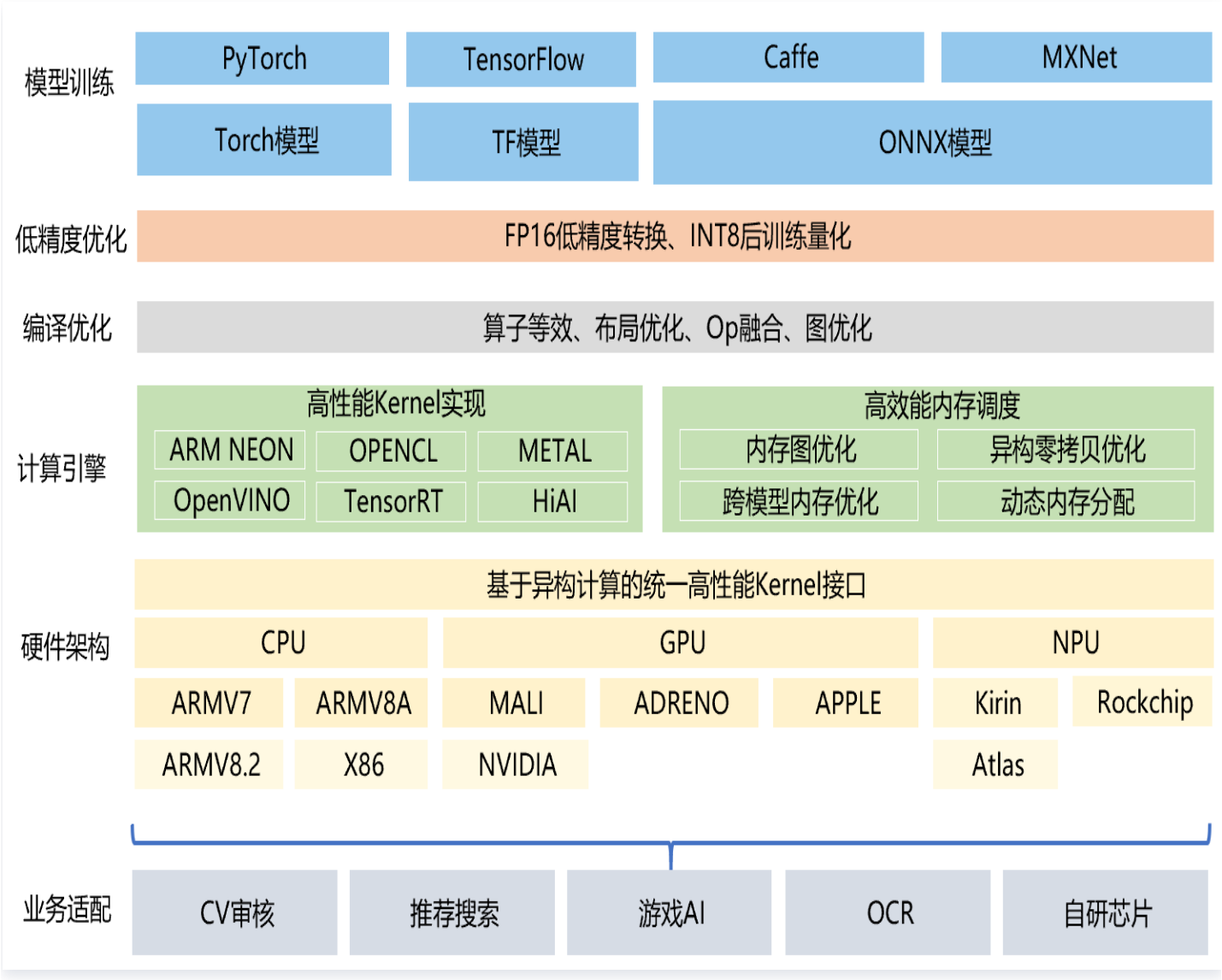
接入方便

- 训练加速和推理加速已支持原生的 Pytorch 框架等框架，支持 TensorFlow 等框架，用户可直接在原生框架下使用 TI-ACC 的加速能力，无需进行额外的模型格式转换等适配工作。
- 训练加速中的通信加速能力通过兼容原生的 DDP 工具提供，用户无需修改原生的使用代码可直接进行使用，数据 IO 优化、自适应 FP16 都通过封装好的简单函数/类进行提供，用户仅需增加几行代码便可使用。
- 推理加速整体能力通过一个函数提供，用户可通过这个函数使用到所有推理加速的能力。
- 推理加速整体能力通过一个新建优化任务使用，用户即可使用到所有推理加速的能力。

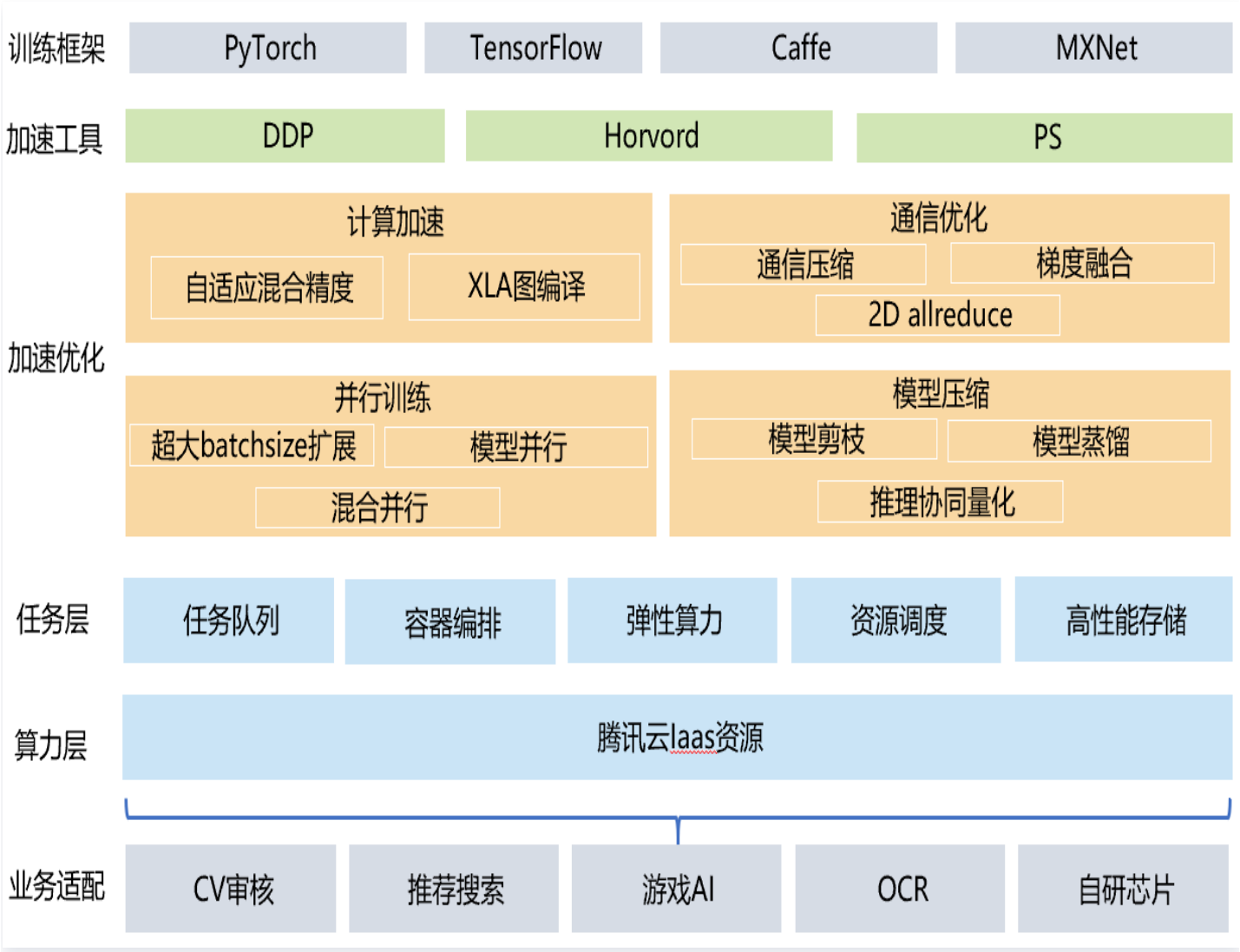
技术强大

TI-ACC 训练加速基于腾讯云帆 Light 常年内外部项目打磨验证，底层提供数据 IO 优化、计算优化、通信加速、并行训练、显存优化等能力；TI-ACC 推理加速基于腾讯云帆 TNN 常年内外部项目打磨验证，底层提供计算优化、低精度加速、内存优化等能力。

推理加速技术架构图


模型训练
PyTorch
TensorFlow
Caffe
MXNet
Torch模型
TF模型
ONNX模型
低精度优化
FP16低精度转换、INT8后训练量化
编译优化
算子等效、布局优化、Op融合、图优化
计算引擎
高性能Kernel实现
ARM NEON
OPENCL
METAL
OpenVINO
TensorRT
HiAI
高效能内存调度
内存图优化
异构零拷贝优化
跨模型内存优化
动态内存分配
硬件架构
基于异构计算的统一高性能Kernel接口
CPU
GPU
NPU
ARMV7
ARMV8A
MALI
ADRENO
APPLE
Kirin
Rockchip
ARMV8.2
X86
NVIDIA
Atlas
业务适配
CV审核
推荐搜索
游戏AI
OCR
自研芯片

训练加速技术架构图


训练框架
PyTorch
TensorFlow
Caffe
MXNet
加速工具
DDP
Horvord
PS
加速优化
计算加速
自适应混合精度
XLA图编译
通信优化
通信压缩
梯度融合
2D allreduce
并行训练
超大batchsize扩展
模型并行
混合并行
模型压缩
模型剪枝
模型蒸馏
推理协同量化
任务层
任务队列
容器编排
弹性算力
资源调度
高性能存储
算力层
腾讯云Iaas资源
业务适配
CV审核
推荐搜索
游戏AI
OCR
自研芯片

应用场景

最近更新时间：2022-06-27 11:35:52

在推荐、CV、NLP 等模型训练和推理场景中，都可以使用到 TI-ACC 的训练和推理加速能力。目前用户可以通过腾讯云 TI 平台提供给用户使用，具体请参考：使用 TI-ACC 推理加速 、使用 TI-ACC 训练加速 。

版权所有：腾讯云计算（北京）有限责任公司