智能钛弹性模型服务

产品简介

产品文档

腾讯云

文档目录

产品简介
- 产品概述
- 产品优势
- 应用场景

产品简介
产品概述

最近更新时间：2019-11-18 18:07:15

智能钛弹性模型服务（Tencent Intelligence Elastic Model Service，TI-EMS）是具备虚拟化异构算力和弹性扩缩容能力的在线推理平台，能够帮助客户解决模型部署复杂、资源浪费、手工扩展资源效率低下的问题。客户通过使用 TI-EMS 可以实现模型一键部署，自动调整弹性计算资源。同时，智能钛弹性模型服务具备多模型支持、版本管理和灰度升级等丰富完善的功能，其内置的 CPU/GPU 推理加速镜像为客户提供高性能、高性价比的推理服务。

产品优势

最近更新时间：2019-11-18 18:07:21

异构算力虚拟化

CPU、GPU 算力虚拟化，一键部署不同类型的机器学习模型和深度学习模型，为用户提供最佳推理服务。

自动弹性扩缩容

您可以选择手动或自动调整弹性实例扩展策略，TI-EMS 会根据业务负载情况，动态、实时、自动管理实例数量，帮助您以最合适的实例数量应对业务情况，为您免去人工部署负担。

模型服务 QoS 保障

TI-EMS 可以帮助您及时发掘线上模型服务的瓶颈，并提供可靠的扩展策略，从而保障您的线上服务正常运行。

高性价比

TI-EMS 可以为您提供小至0.25卡级粒度的算力，通过细粒度算力分配，让您随时随地享受高性价比服务体验。

优化加速

TI-EMS 支持模型和框架的优化加速，提升模型服务运行效率，为您提供优质的推理性能。

功能完善

TI-EMS 提供丰富的多模型支持、版本管理和灰度升级等使用功能，为您的各类业务保驾护航。

应用场景

最近更新时间：2019-08-08 15:19:26

实时翻译

实时翻译场景下，线上业务需要应对可能的实时高请求量，TI-EMS 可快速响应并针对性地弹性扩容，高吞吐，低延迟，保障高 QPS 线上业务平稳运行。

图像分类

在大规模图像处理场景（如图像分类业务）中，TI-EMS 可以全面利用异构资源池，结合模型加速优化和框架优化技术，提高大规模图像处理服务在线推理效率。

语音识别

随着语料库的不断更新，语音识别业务面临着服务的快速更迭，TI-EMS 通过多模型支持，版本管理，支持在线灰度升级，高效应对业务的快速稳定迭代。