模型推理服务化框架Triton保姆式教程（一）：快速入门

吃果冻不吐果冻皮

收起
理部署/服务化方式
种模型推理服务化工具
简介
的优势
能
量且高可扩展性
量
展性
有管理服务的模型编排
语言模型推理
用模型分析器优化模...
orest Inference Librar...

背景

近年来，随着人工智能的快速发展，AI模型如雨后春笋般涌现。而整个AI模型的生命周期如下图所示，主要包括数据采集、数据预处理、模型训练、模型评估、模型部署/服务化、模型监控等环节。

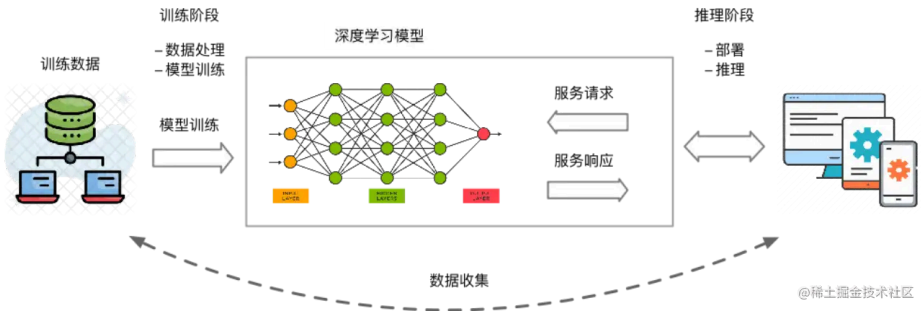

image.png

而模型推理部署/服务化是抽象的算法模型触达具体的实际业务的最后一公里。因此，模型部署是实际AI应用落地非常重要的环节，工程师辛苦训练出来的模型到落地部署应用有大量的工作，包括模型提取、模型压缩、模型加速以及模型服务化等等。

模型推理部署/服务化方式

常见的模型部署方式有以下几种：

- 服务器端部署：模型推理服务部署在服务器上，从而进行高性能完成推理任务；
- 边缘设备端部署：模型部署在手机或者其他端侧设备，利用端侧算力完成推理任务；
- 云端部署：模型部署在云端提供线上服务，用户可以使用客户端发送数据和请求，云端响应请求，完成推理任务并返回推理结果；
- Web 端部署：模型部署在网页端，网页端完成推理任务；

如何选择模型推理服务化工具

常见的模型服务化工具如下图所示，主要分为三大类：

- 第一类：通过WEB框架封装AI模型提供服务，如：Sanic、Flask、Tornado等。
- 第二类：使用深度学习框架自带的Serving封装。如：TensorFlow Serving、TorchServe、MindSpore Serving等。
- 第三类：支持多种框架的统一推理服务化工具。如：Triton Inference Server、BentoML等。

知乎 首发于 AI工程（MLOps）

image.png

从上图可知。开源模型服务框架的选择非常广泛。为了缩小范围，可以从以下几个因素进行考虑：

- **对机器学习库的支持。**任何模型都将使用 TensorFlow、PyTorch 或 scikit-learn 等 ML 库进行训练。但是一些服务化工具支持多个 ML 库，而另一些服务化工具可能仅支持 TensorFlow。
- **模型是如何打包的。**一个典型的模型由原始模型资产和一堆代码依赖组成。比如：通过将模型 + 依赖项打包到 Docker 容器中来进行工作。 Docker 是将软件打包、分发和部署到现代基础设施的行业标准方式。
- **模型运行的地方。**一些服务框架只是为您提供了一个 Docker 容器，您可以在任何支持 Docker 的地方运行该容器。而一些服务框架则建立在 Kubernetes 之上，通过 Kubernetes 进行自动化部署、扩展和管理容器。

本文给大家讲解模型服务化部署框架Triton基本概念、特性等。

Triton 简介

Triton 是 Nvidia 发布的一个高性能推理服务框架，可以帮助开发人员高效轻松地在云端、数据中心或者边缘设备部署高性能推理服务。

Triton Server 可以提供 HTTP/gRPC 等多种服务协议。同时支持多种推理引擎后端，如：TensorFlow, TensorRT, PyTorch, ONNXRuntime 等。Server 采用 C++ 实现，并采用 C++ API 调用推理计算引擎，保障了请求处理的性能表现。

在推理计算方面，Triton 支持多模型并发，动态batch等功能，能够提高GPU的使用率，改善推理服务的性能。Triton不仅支持单模型部署，也支持多模型集成（ensemble），可以很好的支持多模型联合推理的场景，构建起视频、图片、语音、文本整个推理服务过程，大大降低多个模型服务的开发和维护成本。

Triton的优势

与其他一些模型服务化工具相比，Triton具备如下的优势：

- **支持多种框架**：Triton 支持几乎所有主流的训练和推理框架，例如： TensorFlow、NVIDIA TensorRT、PyTorch、Python、ONNX、XGBoost、scikit-learn RandomForest、OpenVINO、自定义 C++ 等。
- **高性能模型推理**：Triton 支持所有基于 NVIDIA GPU、x86、Arm CPU 和 AWS Inferentia 的推理。 它提供动态batching、并发执行、最佳模型配置、模型集成（ensemble）和流式音频/视频输入，以最大限度地提高吞吐量和利用率。
- **专为 DevOps 和 MLOps 而设计**：Triton 可以与 Kubernetes 集成以进行模型服务编排和扩展，支持导出用于监控的 Prometheus 指标，支持实时模型更新，并可用于所有主流的公有云 AI 和 Kubernetes 平台。 它还被集成到了许多 MLOps 软件解决方案中。

各个部分，并可以在集成（ensemble）中使用多个框架。

- **具备企业级的安全性及 API 稳定性**：用于生产环境推理的 NVIDIA Triton，通过企业级的安全性和 API 稳定性加速企业走向 AI 的前沿，同时降低开源软件⁺的潜在风险。

主要功能

Triton的主要功能有支持大模型推理、具备高吞吐量和高可扩展性、支持使用模型分析器优化模型配置等等。

高吞吐量且高可扩展性

高吞吐量

Triton 可在单个 GPU 或 CPU 上并行的指定相同或不同框架下的多个模型。在多 GPU 的情况下，Triton 会自动为基于每个 GPU 的每个模型创建一个实例，以提高利用率。

它还可在严格的延迟限制条件下实时优化推理服务，通过支持批量推理来更大限度地提高 GPU 和 CPU 利用率，并内置对音频和视频流输入的支持。对于需要使用多个模型来执行端到端推理（例如：对话式 AI）的场景，Triton 支持多模型集成。

除此之外，模型可在生产环境中实时更新，无需重启 Triton 或应用。Triton 支持对单个 GPU 显存无法容纳的超大模型进行多 GPU 以及多节点推理。

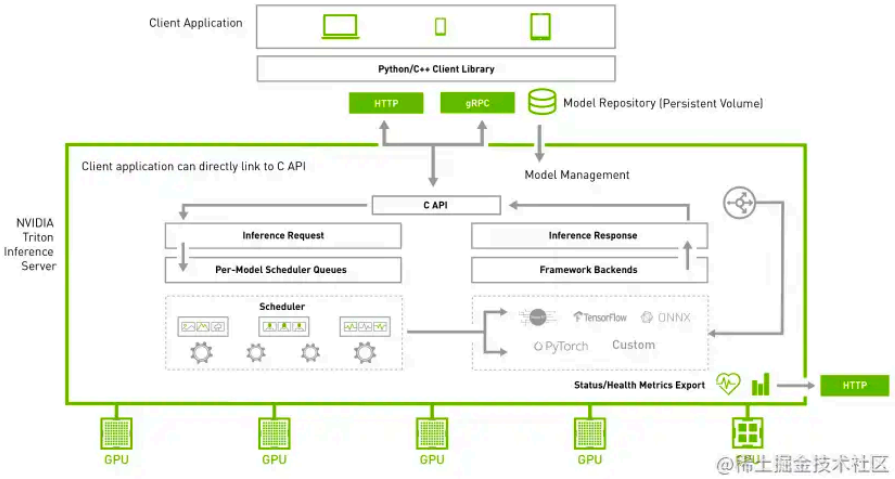

image.png

高可扩展性

作为一个 Docker 容器，Triton 可以与 Kubernetes 轻松集成，用于编排、metrics 和 autoscaling 等。Triton 还与 Kubeflow 和 KServe 集成以实现端到端的 AI 工作流，并导出 Prometheus 指标以监控 GPU 利用率、延迟、内存使用和推理吞吐量。 它支持标准的 HTTP/gRPC 接口来连接⁺负载均衡器（load balancer）等其他应用程序，并且可以轻松扩展到任意数量的服务器，来为任意模型处理日益增长的推理负载。

Triton 可以服务数十个、甚至上百个模型。 模型可以根据需求变化加载到推理服务中或从推理服务中卸载，以适应 GPU 或 CPU 的内存。 同时，支持具有 GPU 和 CPU 的异构集群以实现跨平台推理，并可以动态扩展到任何 CPU 或 GPU 以处理峰值负载。

知乎 首发于 AI工程（MLOps）

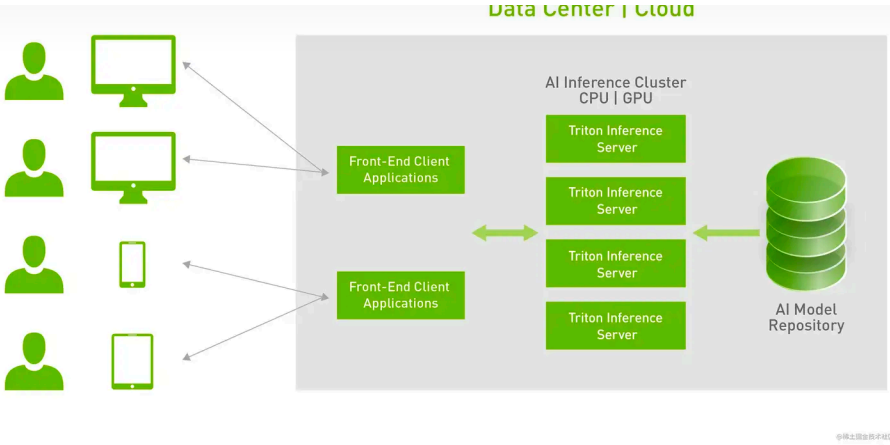

image.png

支持具有管理服务的模型编排

Triton 具有模型编排功能，可实现高效的多模型推理。 此功能作为生产服务运行，可以按需加载模型并在不使用时卸载模型。 它通过在单个 GPU 服务器上放置尽可能多的模型来有效地分配 GPU 资源，并有助于对来自不同框架的模型进行分组以实现高效的内存使用。

支持大语言模型⁺推理

近年来，模型的参数规模正在迅速增长，尤其是在自然语言处理⁺领域，例如：GPT3-175B、Megatron-30B、OPT-175B、Bloom-176B 等模型。 由于这些模型太大，无法放在单个 GPU 上。 因此，Triton 可以将模型划分为多个较小的文件，并在服务器内或跨服务器的单独 GPU 上执行每个文件。 Triton 中的 FasterTransformer 后端支持这种多 GPU、多服务器节点的推理，为当今的 GPT 、T5 、OPT 等大模型提供优化和可扩展的推理。

支持使用模型分析器优化模型配置

Triton 的模型分析器是一个可自动评估 Triton 推理服务中的模型部署配置的工具，例如：目标处理器上的批量大小、精度和并发执行的实例。 它有助于选择最佳配置以满足应用程序服务质量 (QoS) 限制（延迟、吞吐量和内存要求），并将找到最佳配置所需的时间从数周缩短到数小时。该工具还支持模型集成和多模型分析。

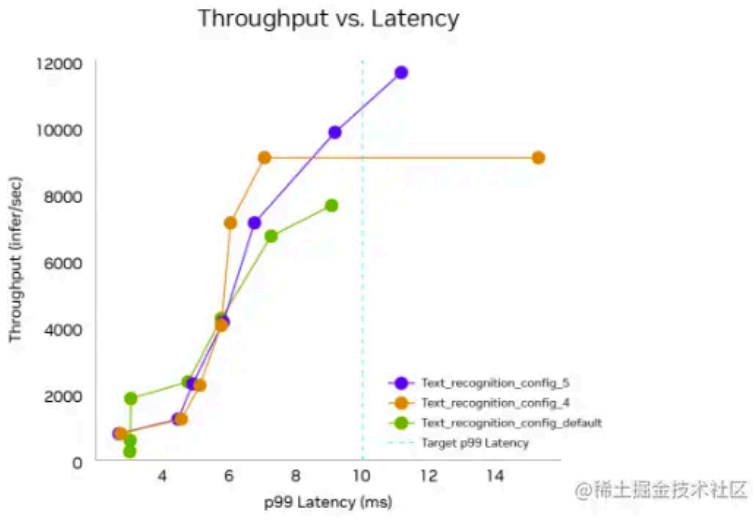

image.png

支持Forest Inference Library (FIL)后端进行基于树的模型推理

Triton 中新的 Forest Inference Library (FIL) 后端支持在 CPU 和 GPU 上对基于树的模型进行高性能推理，并具有可解释性（SHAP 值）。 它支持来自 XGBoost、LightGBM、scikit-learn

知乎 首发于 AI工程（MLOps）

结语

本文给大家简要介绍了模型服务化部署的几种方式以及如何选择一款模型服务化工具，同时，简要介绍了 Triton 主要特征及优势。其中主要特征包括：

- 支持多种深度学习框架
- 支持多种机器学习框架
- 模型并发执行
- 动态批处理(**Dynamic batching**)
- 有状态模型的序列批处理(**Sequence batching**)和隐式状态管理(**implicit state management**)
- 提供允许添加自定义后端和前/后置处理操作的后端 API
- 支持使用 **Ensembling** 或业务逻辑脚本 **(BLS)**进行模型流水线
- HTTP/REST和GRPC推理协议是基于社区开发的**KServe协议**
- 支持使用 C API 和 Java API 允许 Triton 直接链接到您的应用程序，用于边缘端场景
- 支持查看 GPU 利用率、服务器吞吐量、服务器延迟等指标

希望能够给大家带来帮助。

参考文档:

- **NVIDIA Triton**
- **NVIDIA Triton 推理服务**
- **Triton User Guide**

编辑于 2023-06-10 08:49 · IP 属地四川

模型推理 大模型 人工智能

理性发言，友善互动

发布

还没有评论，发表第一个评论吧

文章被以下专栏收录

AI工程（MLOps）
AI System/MLOps