智能钛弹性模型服务

最佳实践

产品文档

腾讯云

【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

最佳实践
运行 TensorFlow 镜像模型批处理任务

最近更新时间：2020-12-22 16:51:22

本示例提供了一个图像分类应用的样例，使用 tfserving 运行环境，帮助您快速熟悉使用 TI-EMS 发布模型服务的过程。此示例针对牛津花卉的图像数据集，然后使用经典深度学习 inception 的图像分类模型，将此模型部署为离线批处理服务，部署完成后，用户可通过离线批处理服务识别输入图片的花卉种类。

开始使用示例前，请仔细阅读准备内容罗列的要求，提前完成准备工作。

准备内容

1. 经典深度学习 inception 模型：inception_model.zip

> **我们已经为您准备好了上述 inception 模型的 COS 访问地址：** `cos://ti-ems-1255502019.cos.ap-beijing.myqcloud.com/models/tfserving/inception/` **。您可以输入该 COS 地址，也可以将模型文件夹下载下来，解压上传到自己的 COS 存储桶中，并在【创建模型服务配置】页面选择相应的模型文件夹。**

2. 测试图片：

我们已经为您准备好了测试数据集。您可以直接下载 花卉数据集。

base64 编码

将上述测试花朵图片按照 inception 模型定义的 JSON 数据格式 `{"instances":[{"b64": "图片 base64 编码"}]}` 进行编码，将 jpg 转换成 base64。`flowers.json` 为经过编码的测试图片数据，或者您可以直接下载已经编码完成的 JSON 文件 flowers.json，跳过该图片编码步骤，直接进行下一步。

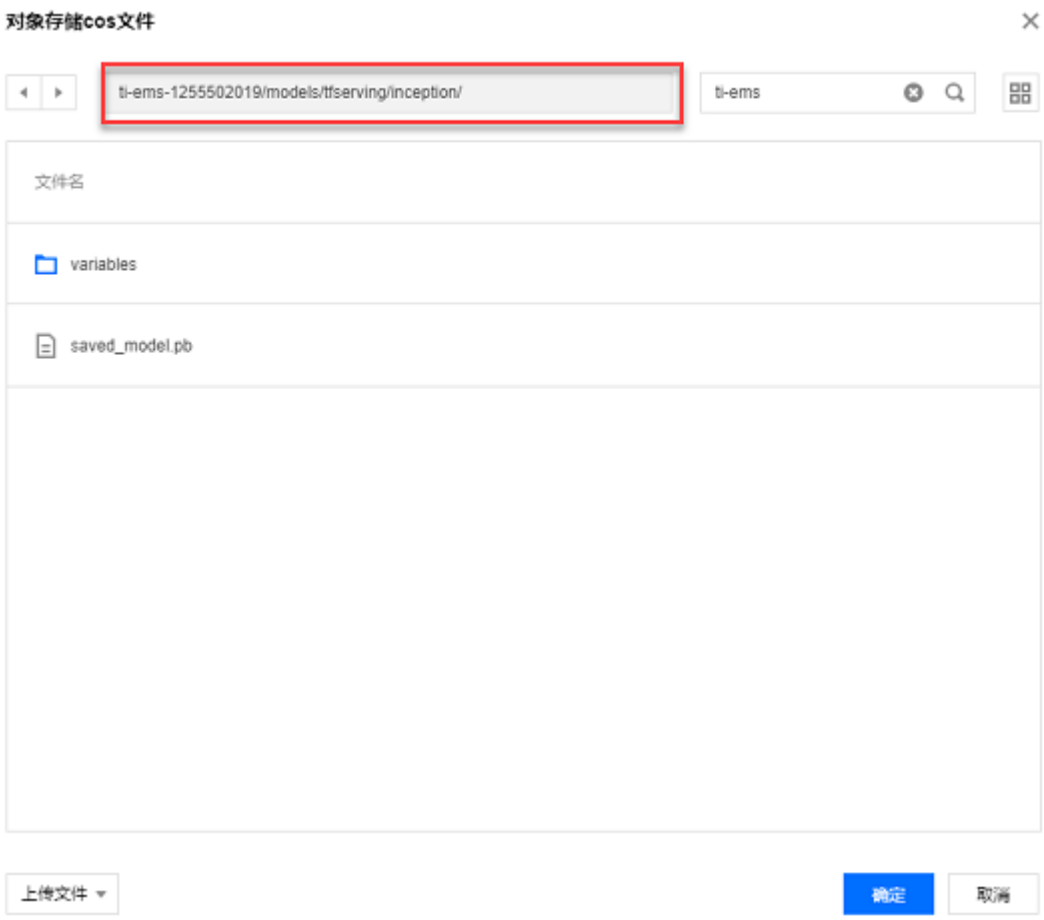
步骤1：创建模型服务配置

1. 在【模型服务配置】页面单击【新建】，进入【新建模型服务配置】页面。
2. **输入配置名称：**demo_tfserving。
3. **选择地域：**地域为模型文件夹所在 COS 地域。

> **说明：**
> 为您提供的 COS 访问地址地域为北京，如已将模型文件夹上传至到自己的 COS 存储桶，可选择自己 COS 存储桶所在地域。

4. **选择运行环境**：单击【运行环境】，在弹出页面的【公共运行环境】栏选择 tfserving。

5. **提供模型文件地址**：直接输入 COS 访问地址或单击【对象存储 COS 文件】，弹出【对象存储 COS 文件】选择页面，选择 inception 模型文件夹所在的路径，单击【确定】。



6. **完成模型服务配置**：单击【确定】。

步骤2：购买创建专用资源组

此步骤为可选步骤，如已有可用资源组可跳过本步骤，模型服务可部署在专用资源组和公共资源组，公共资源组和专用资源组的计费方式请详见 计费概述。

1. 在【资源组管理】页面单击【新建资源组】，进入【TI-EMS 资源组】购买页面。
2. **选择地域**：地域与模型文件夹所在 COS 地域保持一致。
3. **选择节点规格**：下拉菜单中选择24核48G。
4. **选择节点数量**：保持默认1。
5. **选择计费模式**：单击【按量计费】。
6. 单击【开通】。

步骤3：启动批处理任务

1. 在【模型服务配置】页面找到 demo_tfserving 配置，单击配置卡片的【批处理作业】，进入【启动批处理作业】页面。
2. **输入作业名称**：输入启动的服务名称。
3. **选择资源组**：选择将要启动的资源组，这里选择已购买的专用资源组。
4. **选择实例配置**：选择【CPU 配置】，实例配置填写为2核4G。
5. **选择输入数据**：直接输入 COS 访问地址或单击【选择 COS 文件】，弹出【对象存储 COS 文件】选择页面，选择测试图片所在的路径，单击【确定】。
6. **选择数据类型**: 选择 JSON。
7. **选择Batchsize**: 选择64。
8. **选择输出数据**：直接输入 COS 访问地址或单击【选择 COS 文件】，弹出【对象存储 COS 文件】选择页面，选择期望输出推理结果的路径，单击【确定】。
9. **选择实例数量**：实例数量设置为1。
0. 全部设置完成后，单击【启动作业】，进入【批处理任务】页面。

步骤4：验证推理结果

1. 在【批处理任务】页面找到创建的批处理任务，点击列表上方搜索框右侧的刷新按钮刷新服务状态。
2. 服务状态由【运行中】变为【运行成功】时，前往所配置的输出数据 COS 路径查看推理结果。

TI-EMS 使用过程中如遇任何问题，欢迎加入 智能钛 AI 开发者社区，与腾讯云 AI 专家和众多 AI 爱好者交流技术。

调用 TensorRT 镜像模型服务

最近更新时间：2020-12-22 16:50:15

本示例提供了一个图像分类应用的样例，使用 TensorRT 运行环境，帮助您快速熟悉使用 TI-EMS 发布模型服务的过程。此示例针对 ImageNet 标签为230的图像数据集，然后使用经典深度学习 inception-v3 的图像分类模型，将此模型部署为在线服务，部署完成后，用户可通过在线服务识别输入图片的图像种类。

开始使用示例前，请仔细阅读准备内容罗列的要求，提前完成准备工作。

准备内容

1. 经典深度学习 inception 模型： inception_v3.tar

> 说明：
> 我们已经为您准备好了上述 inception 模型的 COS 访问地址： `cos://ti-ems-1255502019.cos.ap-beijing.myqcloud.com/models/tensorRT/inception_v3/1/` 。您可以输入该 cos 地址，也可以将模型文件夹下载下来，解压上传到自己的 COS 存储桶中，并在【新建模型服务配置】页面选择相应的模型文件夹。

2. 测试图片： imagenet_230.tar（ImageNet label 为230的图片）

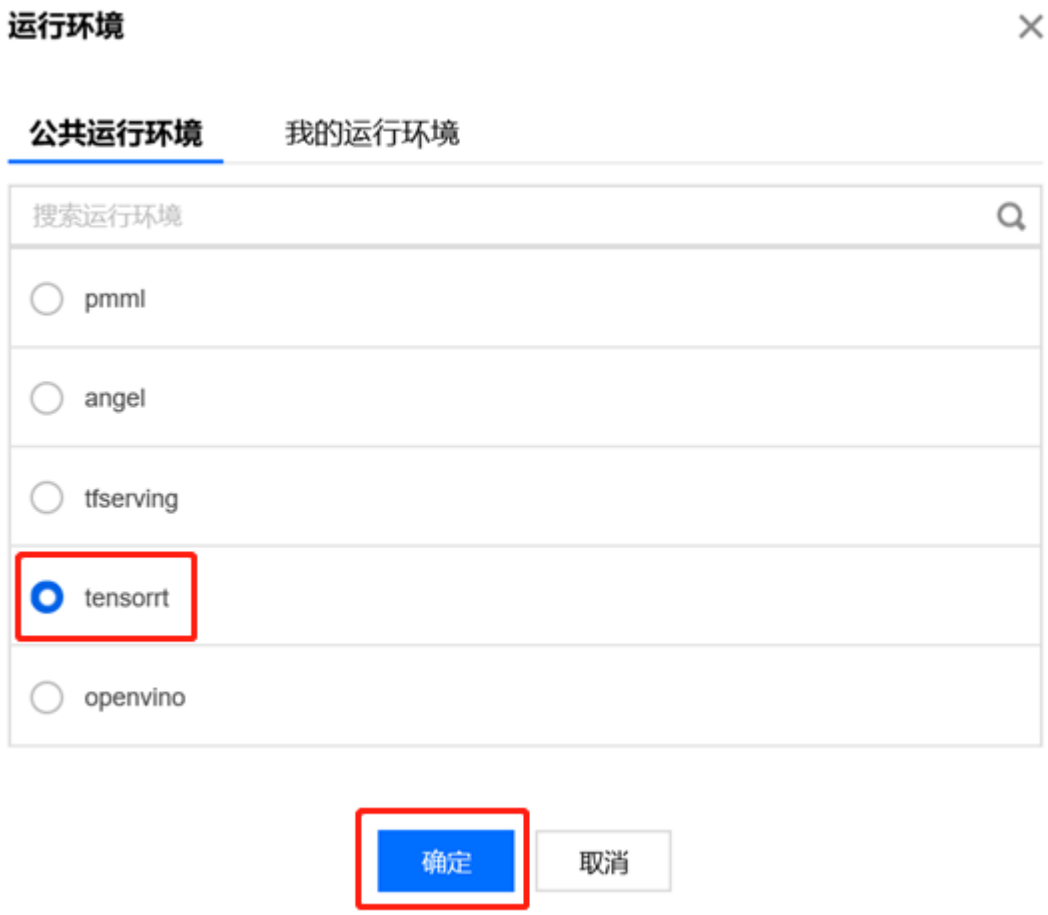
> 说明：
> 我们已经为您准备好了测试图像数据集，您可以下载后在本地运行测试脚本时使用。

步骤1：创建模型服务配置

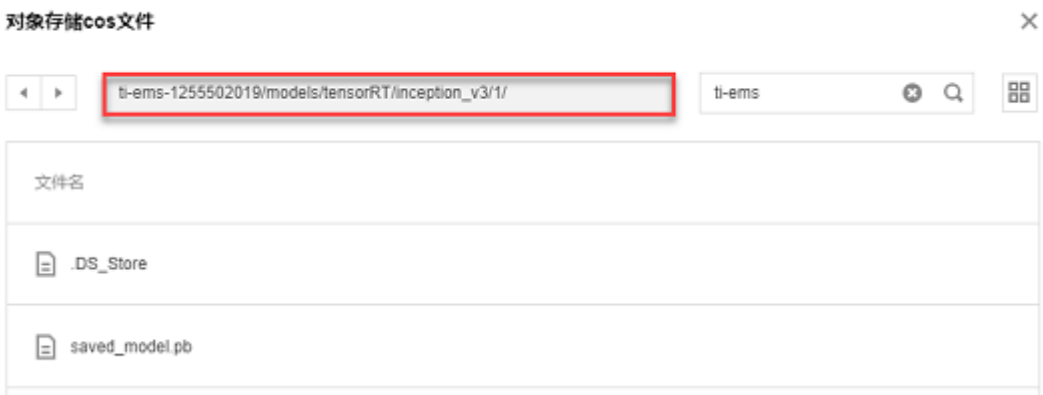
1. 在【模型服务配置】页面单击【新建】，进入【新建模型服务配置】页面。
2. **输入配置名称**：demo_tensorrt。
3. **选择地域**：地域为模型文件夹所在 COS 地域。

> 说明：
> 为您提供的 COS 访问地址地域为北京，如已将模型文件夹上传到自己的 COS 存储桶，可选择自己 COS 存储桶所在地域。

4. **选择运行环境**：单击【运行环境】，在弹出页面的【公共运行环境】栏选择 tensorrt。



5. **提供模型文件地址**：直接输入 COS 访问地址或单击【对象存储 COS 文件】，弹出【对象存储 COS 文件】选择页面，选择 inception_v3 模型文件夹所在的路径，单击【确定】。



6. **完成模型服务配置**：单击【确定】。

步骤2：购买创建专用资源组

此步骤为可选步骤，如已有可用资源组可跳过本步骤，模型服务可部署在专用资源组和公共资源组，公共资源组和专用资源组的计费方式请详见 计费概述。

1. 在【资源组管理】页面单击【新建资源组】，进入【TI-EMS 资源组】购买页面。
2. **选择地域**：地域与模型文件夹所在 COS 地域保持一致。
3. **选择节点规格**：下拉菜单中选择8核32G1T4。

> 说明：
> 本示例需要使用带有 GPU 的节点 。

4. **选择节点数量**：保持默认1。
5. **选择计费模式**：单击【按量计费】。
6. 单击【开通】。

步骤3：启动服务

1. 在【模型服务配置】页面找到 demo_tensorrt 配置，单击配置卡片的【在线推理】，进入【启动模型服务】页面。
2. **输入服务名称**：demo_tensorrt。
3. **选择资源组**：选择将要启动的资源组，这里选择已购买的专用资源组。
4. **选择实例配置**：选择【GPU 配置】，实例配置填写为4核8G1卡。

步骤5：调用模型服务获取模型元数据

以 Linux 系统为例，使用如下命令获取模型元数据：

```
curl -H "X-Auth-Token: TOKEN" IP:80/v1/models/m/metadata
```

调用参数说明：

TOKEN：通过模型服务页面的【服务调用】获取的密钥地址 token。

IP：通过单击模型服务页面的【服务调用】获取的服务访问地址。

步骤6：调用模型服务接口

TI-EMS 模型服务支持以 HTTP 访问，本示例通过 HTTP 客户端脚本访问模型服务。

1. 下载服务调用示例脚本

```
git clone https://github.com/tencentyun/ti-ems-client-examples
```

```
cd ti-ems-client-examples
```

2. 安装测试脚本依赖

测试脚本需要在 Python 环境下运行，运行前需要配置环境，requirements.txt 是运行测试脚本所需要的依赖库清单：

```
tensorflow-serving-api==1.13.0
tensorflow==1.13.1
grpcio==1.22.0
requests==2.22.0
numpy==1.16.3
opencv-python==4.1.0.25
preprocessing
```

3. 使用如下命令行一键安装测试脚本所依赖的运行环境，请确保以上依赖安装成功。

```
pip install -r requirements.txt
```

4. 运行客户端脚本

因为需要动态生成优化内核，TensorRT 镜像首次调用模型服务，根据模型大小不同可能需要等待0.5 – 5分钟。

```
python http_client --server IP --token TOKEN --data_dir DATA_DIR
```

调用参数说明：

IP：通过单击模型服务页面的【服务调用】获取的服务访问地址。

TOKEN：通过模型服务页面的【服务调用】获取的密钥地址 token。

DATA_DIR：测试数据集所在路径。

不同模型输入的数据类型、数据 shape 可能不同，或对数据预处理要求不同。请根据具体模型，设计相应的访问程序。了解更多 客户端程序 。

TI-EMS 使用过程中如遇任何问题，欢迎加入 智能钛 AI 开发者社区，与腾讯云 AI 专家和众多 AI 爱好者交流技术。