腾讯云 TI 平台 TI-EMS

快速入门

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】

©2013-2022 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

快速入门
- 使用流程
- 账号权限说明
- 创建模型服务配置
- 启动和调用模型服务

版权所有：腾讯云计算（北京）有限责任公司

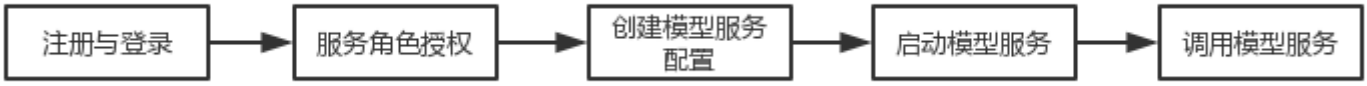
快速入门
使用流程

最近更新时间：2021-12-22 11:40:37

腾讯云 TI 平台 TI-EMS 是具备虚拟化异构算力和弹性伸缩能力的模型服务推理平台。您只需要使用简单几步操作即可完成模型服务的部署。本章节简要介绍了 TI-EMS 平台的使用流程，通过控制台操作，您可以快速了解 TI-EMS 的基本功能。

TI-EMS 使用流程图

TI-EMS 使用基本流程如下图所示：



1. 注册与登录： 任何用户想要使用腾讯云的服务，都需要先注册腾讯云账户。单击进入 腾讯云注册页面，详细指引请参考 注册腾讯云。账号注册完成以后登录 TI-EMS 控制台，开始使用 TI-EMS 服务。
2. 服务角色授权： 主账号首次登录使用 TI-EMS 服务需要进行服务角色授权，完成授权后 TI-EMS 可以访问您的其他腾讯云资源（对象存储、私有网络等）。子账号/协作者账号首次登录使用 TI-EMS 需要完成主账号策略授权，关于账号授权的详细说明请查看 账号权限说明。
3. 创建模型服务配置： 进入 模型服务配置创建页面，确定配置名称，模型运行环境和模型文件，即可创建模型服务配置。
4. 启动模型服务： 单击**在线服务**，进入服务启动配置页面，选择地域和资源组信息，配置实例资源策略，即可启动模型服务。
5. 调用模型服务： 服务启动成功后，您即可在服务列表页面进行服务调用。

快速入门

您可以通过如下操作示例进行一步完整体验 TI-EMS 的功能：

- 账号权限说明
- 创建模型服务配置
- 启动和调用模型服务

版权所有：腾讯云计算（北京）有限责任公司

账号权限说明

最近更新时间：2021-12-22 11:40:45

主账号登录方式

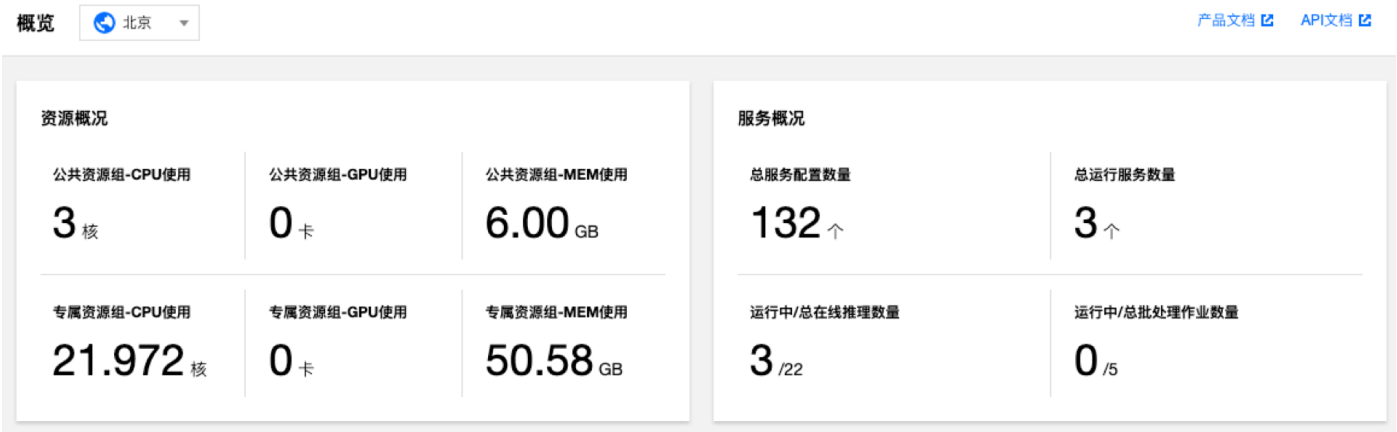
如果您当前是主账号用户，则操作步骤如下：

1. 主账号首次登录 TI-EMS 您将会看到服务角色授权页面，单击**确定授权**，您将跳转到腾讯云 CAM 控制台进行角色登录授权。

2. 单击**同意授权**，将创建服务角色并授予 TI-EMS 相关权限。

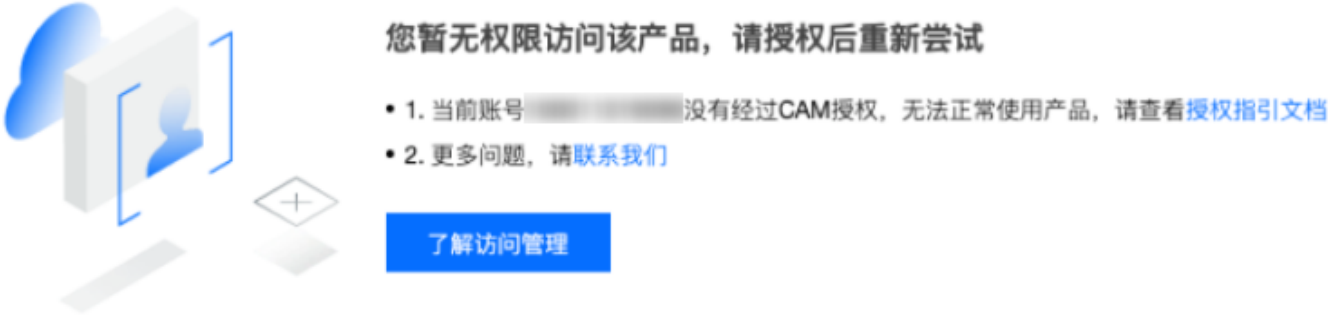
3. 确认授权后，将会为您自动创建角色 TI-EMS_QCSRole。如下图所示：

4. 返回腾讯云 TI 平台 TI-EMS 控制台，此时，您可以正常使用产品功能。



子账号/协作者登录方式

如果您当前是子账号/协作者用户，首次登录 TI-EMS 控制台时，您将会看到登录限制页面，这是由于 TI-EMS 已接入腾讯云访问管理 CAM 服务级策略，子账号/协作者若要使用 TI-EMS 产品，需要主账号授予该子账号/协作者相关预设策略，更多 CAM 相关介绍请查看 授权指南。



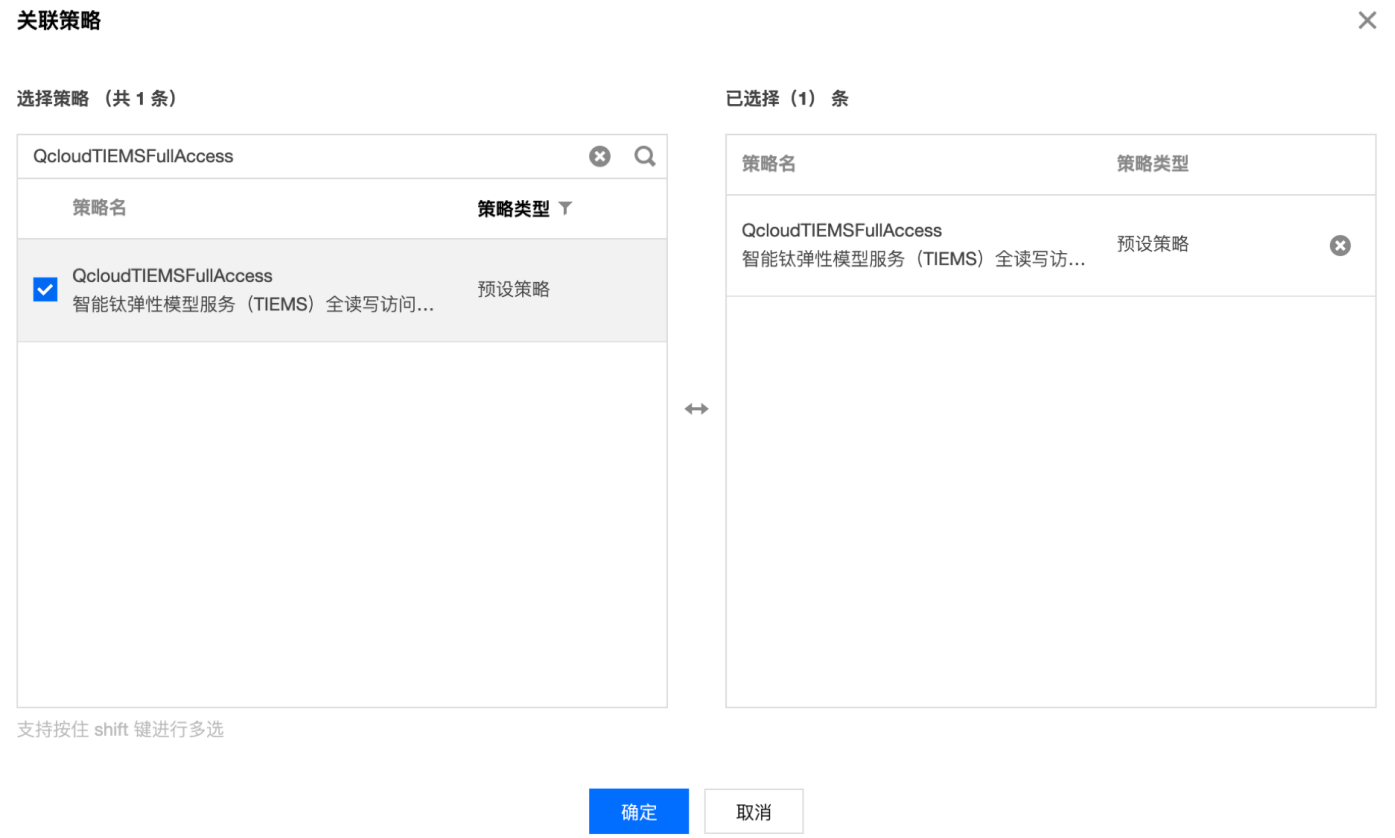
以下是子账号/协作者使用 TI-EMS 产品的授权指南：

步骤一： 授权 TI-EMS 预设策略

联系主账号或者拥有 QcloudCamFullAccess 权限的子账号为您的当前子账号/协作者用户授予 QcloudTIEMSFullAccess 策略。具体操作方式为：

1. 登录 CAM 控制台，进入 >**用户**>**用户列表**，找到需要使用 TI-EMS 的子账号/协作者账号。单击**授权**，开始进行策略添加。

2. 打开关联策略对话框，输入 QcloudTIEMSFullAccess ，勾选策略，单击确定，便可成功添加策略。至此，该子账号可以拥有 TI-EMS 产品资源的操作权限。



此时您可以正常进入 TI-EMS 页面。

步骤二： 服务角色授权

1. 若您的主账号已经进行过服务角色授权，则可跳过此步骤，进入步骤三。
2. 若您的主账号没有进行过服务角色授权，则子账号/协作者账号登录后依然会提示进行角色登录授权，即您将会看到如下页面：

此时您有两种方式完成服务角色授权：

（1）联系主账号登录 TI-EMS 控制台，按照**主账号登录方式**指引进行服务角色授权；

（2）联系主账号或者拥有 QcloudCamFullAccess 权限的子账号为您当前的子账号/协作者用户授予 QcloudCamSubaccountsAuthorizeRoleFullAccess 策略（操作步骤请详见步骤1），使您的子账号/协作者用户拥有创建服务角色授权的权限。授权完成后您使用子账号/协作者账号登录 TI-EMS 的时候，便可以直接进行服务角色授权。

步骤三： 其他腾讯云资源的访问权限

您在使用 TI-EMS 控制台时，可能会需要操作 COS 或者部分其他产品的权限。例如需要读取存储在 COS 中的模型文件数据等，因此主账号或者拥有 QcloudCamFullAccess 权限的子账号需要授予子账号/协作者账号 QcloudCOSDataFullControl、QcloudCLSFullAccess、QcloudAPIGWFullAccess 、QcloudCCRFullAccess 等策略（操作步骤请详见步骤1）。

完成以上步骤后，子账号/协作者账号便可以正常使用腾讯云 TI 平台 TI-EMS 产品。

版权所有：腾讯云计算（北京）有限责任公司

创建模型服务配置

最近更新时间：2022-01-10 17:14:26

进入模型仓库>模型服务配置，单击新建，开始创建模型服务配置。

创建 TI-EMS 模型服务配置，需要：

1. 输入基本配置信息（模型服务配置的名称、地域、模型服务的运行环境）
2. 配置模型文件

基本配置信息

名称：要创建的模型服务配置的名称，不超过20个字符。

地域：选择服务所运行的地域。

版本：新建模型服务配置版本自动生成，为1.0，后续每新增一个版本，版本号自动加1。

运行环境： 单击**运行环境**，可选择模型运行的公共环境。TI-EMS 提供 TF-Serving、PMML、TensorRT、

OpenVINO 和 Angel 五种模型服务运行环境，相关介绍请详见 模型运行环境。

模型文件配置

TI-EMS 支持三种模型文件输入方式：

1. 直接输入 COS 路径

2. 从 COS 存储桶选择模型文件

3. 本地上传模型文件

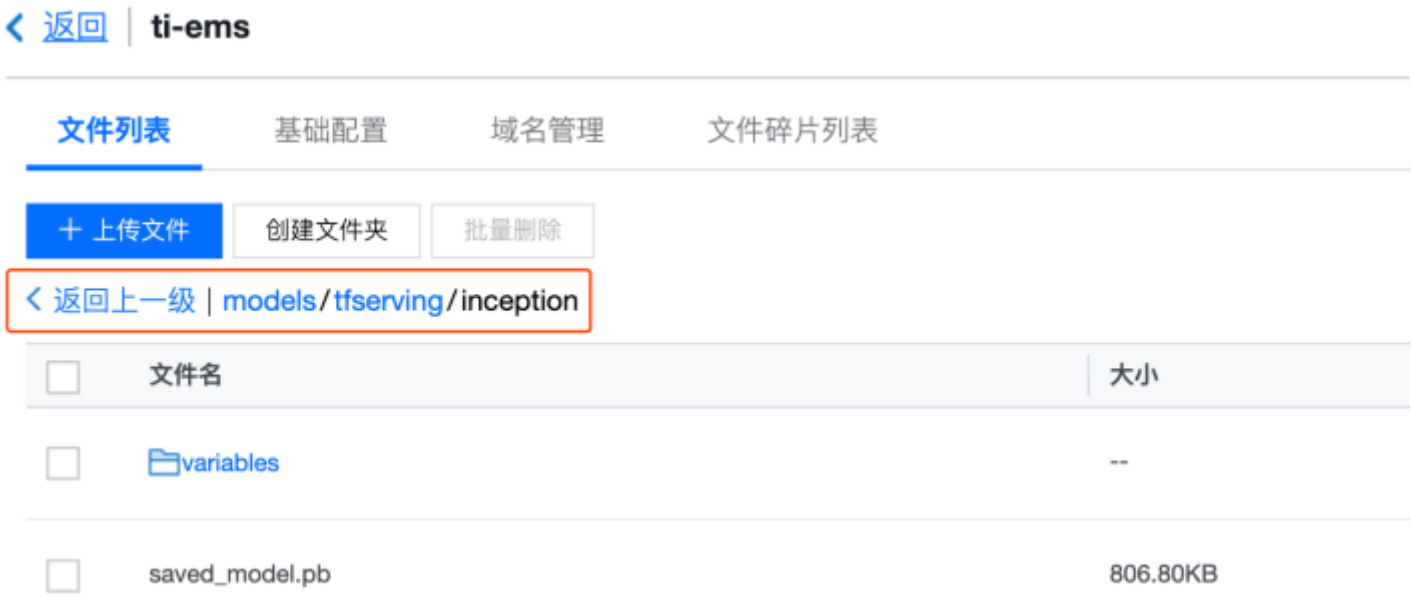
直接输入 COS 路径

您可以直接在模型文件输入框输入模型文件夹 COS 路径，其中模型地址格式为：

cos://${bucket_name}-${appid}.cos.{region}.myqcloud.com/模型文件夹路径/

示例

假设用户通过主账号（AppID 为1255502019）登录 COS 控制台进行创建存储桶，存储桶的配置信息为：地域选择为北京地域，存储桶名称自定义为 ti-ems，模型文件夹为 inception（单击下载模型文件夹，上传至存储桶 ti-ems 的 models/tfserving 路径下，如下图所示：



则存储桶默认域名如下所示：

```
cos://ti-ems-1255502019.cos.ap-beijing.myqcloud.com/models/tfserving/inception/
```

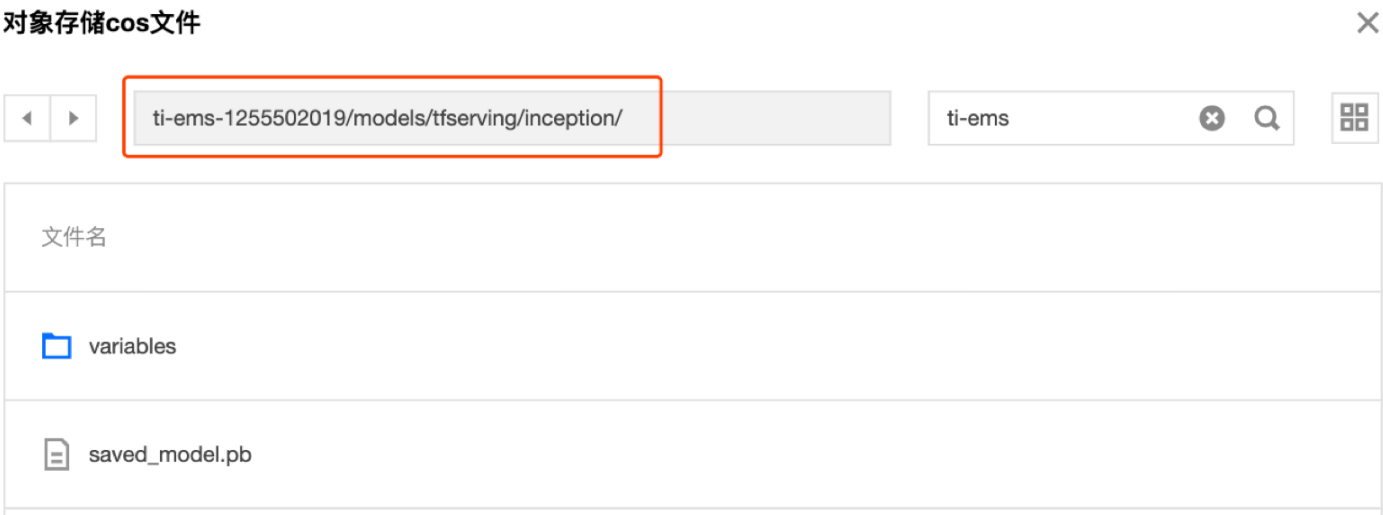
其中：

- ti-ems-1255502019：表示该存储桶归属于 AppID 为1255502019的用户。AppID 是您在成功申请腾讯云账户后所得到的账号，由系统自动分配，具有固定性和唯一性，可在 账号信息 中查看。
- cos：指对象存储 COS。
- ap-beijing：指存储桶的地域简称。
- myqcloud.com：腾讯云域名，固定字符。
- models/tfserving/inception/：模型文件夹路径。

从 COS 存储桶选择

TI-EMS 控制台可直接选择 COS 模型文件夹，在模型服务配置新建页面，单击**对象存储 COS 文件**，进入 COS 文件选择弹窗页面。

选中模型文件夹所在的路径，单击**确定**，完成模型文件配置。



本地上传模型文件

TI-EMS 还可支持在控制台直接上传本地模型文件夹，无须跳转至对象存储页面进行操作。在模型服务配置新建页面，点击**对象存储 COS 文件**，在 COS 文件选择弹窗页面，进入对应的目录，单击**上传**，即可将本地模型文件夹上传至 COS 存储桶中，上传完成后，选中模型文件夹路径，单击**确定**，完成模型文件配置。

> ⚠ 注意：
>
> 1. 模型服务部署所在的地域与模型文件所在 COS 存储桶地域必须一致。例如，您需要将服务部署在北京地域，请将模型文件上传至北京地域存储桶。
> 2. TI-EMS 所约定的模型文件路径，是模型文件所在的文件夹的路径。
> 3. TI-EMS 目前支持 TensorFlow SavedModel、PMML、Angel、OpenVINO IR 四种格式的模型文件。
> 4. 对象存储 COS 相关操作请详见 COS 产品文档 .

版权所有：腾讯云计算（北京）有限责任公司

启动和调用模型服务

最近更新时间：2021-11-17 10:45:55

TI-EMS 可将模型服务配置启动为在线推理或批处理作业，下面以启动和调用 Tf-Serving 在线推理服务为例说明使用方法：

内容准备

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

测试图片：

flower.jpg

base64 编码

将上图测试花朵图片按照 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。flower.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 flower.json，跳过该图片编码步骤，直接进行下一步。

```
echo "{\"instances\":[{\"b64\": \"$(base64 -i flower.jpg)\"}]}" | tee flower.json
```

创建模型服务配置

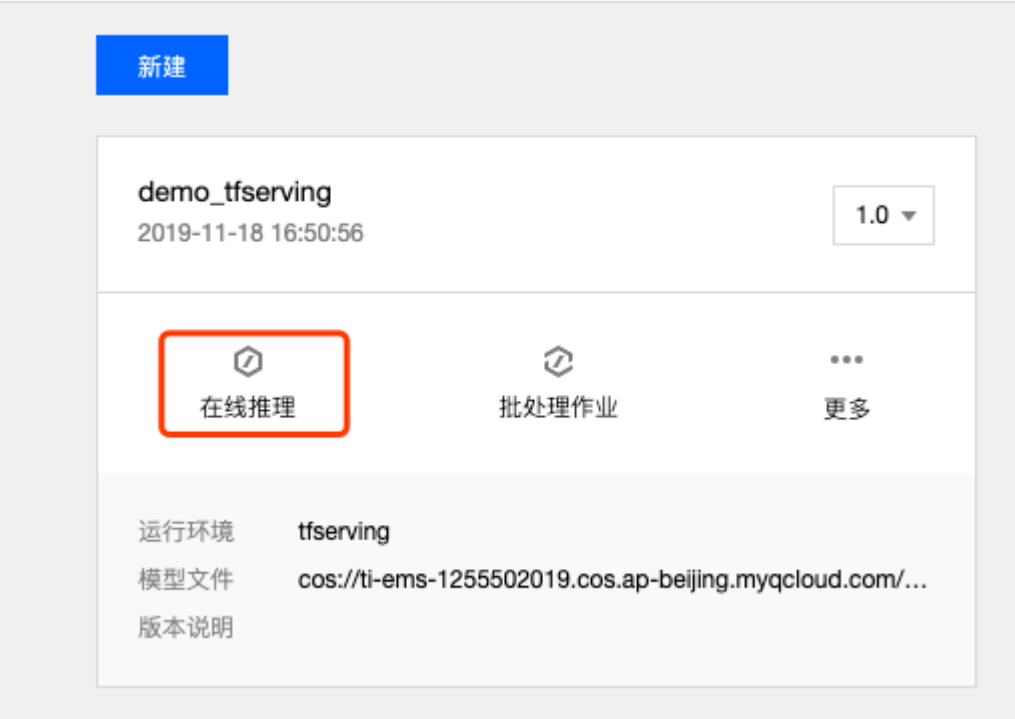
在模型服务配置页面单击**新建**，进入模型服务配置新建页面，输入配置名称：demo_tfserving，单击**运行环境**，在弹出页面的**公共运行环境**栏选择 tfserving。

单击**对象存储 cos 文件**，弹出 COS 文件选择页面，选择 inception 模型文件夹所在的路径，单击**确定**。

模型服务配置创建详细操作指引请见**快速入门>创建模型服务配置**。

启动服务

在模型服务配置页面找到 demo_tfserving 配置，单击配置卡片的**在线推理**，进入启动服务页面。



输入服务名称：输入启动的服务名称。
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。资源组详细介绍请看 资源组管理。
选择实例配置：选择**CPU 配置**，实例配置填写为1核2G。

TI-EMS 针对不同的模型服务运行资源需求为您提供了**CPU 配置**和**GPU 配置**，若您选择在公共资源组启动服务，则需要选择 CPU quota 或 GPU quota 并指定对应 quota 的数量，公共资源组的服务计费方式按照 quota 单价 * 数量 * 服务运行时长，若您选择在专用资源组启动服务，则需自定义 CPU 配置和 GPU 配置，用户在预购的专用资源组启动的服务不再另外收费。

实例调节策略：选择**手动调节**，实例数量设置为1。
TI-EMS 支持用户手动设定实例数量或设置实例自动调节策略。选择手动调节实例，可以直接设定启动服务时的实例数量，实例数量最小设置为1。选择自动调节实例，可以设置多种触发策略，当满足任一触发策略时，按照指定的实例伸缩范围自动调节。TI-EMS 目前支持的触发指标如下：

- CPU 利用率
- 内存利用率
- GPU 利用率

当系统实际指标小于触发策略指标设定值时，执行自动缩容，大于设定值时，执行自动扩容。
选择是否生成鉴权：勾选生成鉴权，生成服务访问地址的鉴权密钥。

全部设置完成后，单击**启动服务**，进行在线服务列表页面。

调用测试

单击**模型服务>在线推理**页面选择 tfserving 模型服务，在对应的**操作**列单击**更多>调用**，选择公网地址访问，获得模型服务的公网访问地址和密钥 TOKEN。

使用 curl 工具为例，展示如何调用服务接口：

```
curl -H "Content-Type: application/json" \
-H "x-Auth-Token: TOKEN" \
-X POST IP/v1/models/m:predict -d @flower.json
```

调用参数说明

- TOKEN：通过单击模型服务页面的**调用**获取的密钥地址 token。
- IP：通过单击模型服务页面的**调用**获取的服务访问地址。

返回结果：

```
{
"predictions": [{
"class_idx": 2,
"probabilities": [0.00685295, 0.00298388, 0.965622, 0.00721011, 0.0173306],
```

```
  "classes": "roses"
  }]
  }
```

可以看出模型服务运行正常，并且正确的对发送的花朵进行了分类，具体详情可参考 TensorFlow 模型服务文档。

版权所有：腾讯云计算（北京）有限责任公司