ALIBABA CLOUD

阿里云

机器学习PAI

最佳实践

文档版本：20220527

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
`Courier字体`	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.概览

本文汇总了PAI的最佳实践。

热点文章

- 推荐解决方案综述
- NLP意图智能识别解决方案
- 新用户流失召回
- 图像内容风控解决方案
- 文本内容风控解决方案
- 相似图像匹配与图像检索解决方案

PAI-Studio

- 推荐解决方案
  - 使用ALS算法实现音乐评分预测
  - 基于对象特征的推荐
  - 使用FM-Embedding实现推荐召回
  - 基于Alink框架的FM推荐
  - 使用协同过滤实现商品推荐
  - 基于二部图GraphSAGE算法实现推荐召回
  - 使用EasyRec构建推荐模型
- 智能风控解决方案
  - 基于外卖评论实现舆情风控
  - 基于图算法实现金融风控
  - 评分卡信用评分
  - 异常指标监控
  - 用户流失预警风控
- 分类
  - 使用TensorFlow实现图片分类
  - 相似标签自动归类
  - 基于文本分析算法实现新闻分类
- 心脏病预测
- 基于回归算法实现农业贷款发放预测
- 预测学生考试成绩
- 雾霾天气预测
- 发电场输出电力预测
- 用户窃电识别

PAI-DSW

- 使用WebIDE在线调试代码

- 使用EasyVision进行目标检测
- 使用EasyTransfer进行文本分类
- 使用EasyASR进行语音识别
- 使用EasyASR进行语音分类

PAI-DLC

- 快速提交单机PyTorch迁移学习任务
- 使用NAS提交单机PyTorch迁移学习任务
- 使用paiio读写MaxCompute表数据

Blade敏捷版

- RetinaNet优化案例1：使用Blade优化RetinaNet（Detectron2）模型
- RetinaNet优化案例2：结合Blade和Custom C++ Operator优化模型
- RetinaNet优化案例3：结合Blade和TensorRT Plugin优化模型
- ResNet50优化案例：使用Blade优化基于TensorFlow的ResNet50模型
- BERT模型优化案例：使用Blade优化基于TensorFlow的BERT模型

PAI-EAS

部署PS-LR模型

SDK

PAI-TF

- 使用TensorFlow实现分布式DeepFM算法
- 使用TensorFlow实现图片分类
- 模型仓库（FastNN）

2.NLP意图智能识别解决方案

针对App内大量问题咨询场景，PAI提供NLP意图智能识别解决方案。本文介绍如何基于BERT模型，快速构建符合业务的NLP文本分类模型，从而实现智能客服QA平台。

前提条件

- 已开通PAI（Studio、DSW、EAS）后付费，详情请参见开通。
- 已开通MaxCompute，用于存储训练数据和测试数据。关于如何开通MaxCompute，请参见通过查询编辑器使用MaxCompute。

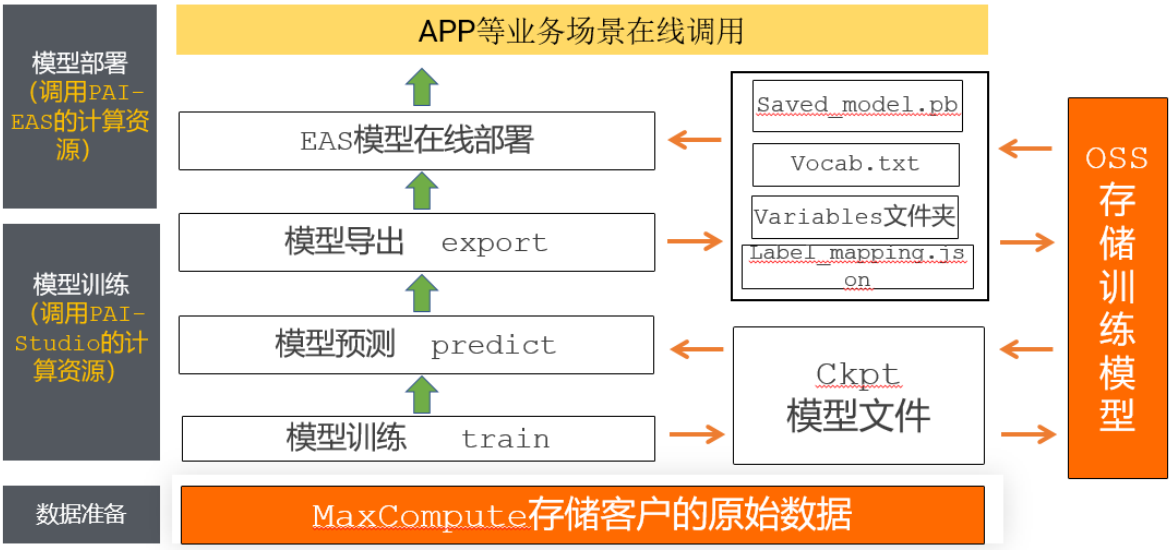
  > **说明** 由于本文的解决方案需要使用PAI-Studio计算资源，而PAI-Studio使用MaxCompute存储结构化数据，因此您必须将训练数据和测试数据导入至MaxCompute中。

- 已创建OSS存储空间（Bucket），用于存储标签文件和训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 在DataWorks中，已创建ODPS SQL节点，用于执行PAI命令。关于如何创建ODPS SQL节点，请参见创建ODPS SQL节点。
- 已配置好EASCMD客户端工具，本文使用该工具部署模型。关于如何配置EASCMD客户端工具，请参见下载并认证客户端。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创建专属资源组。

背景信息

App内每天有大量用户咨询问题，问题种类不仅多，而且类别相似度高。如果采用人工解答，则人力成本过高。针对该问题，阿里云机器学习PAI提供如下解决方案，助力您有效降低人工服务成本。您仅需执行几段命令，即可在极小开发量下快速搭建智能客服QA平台：

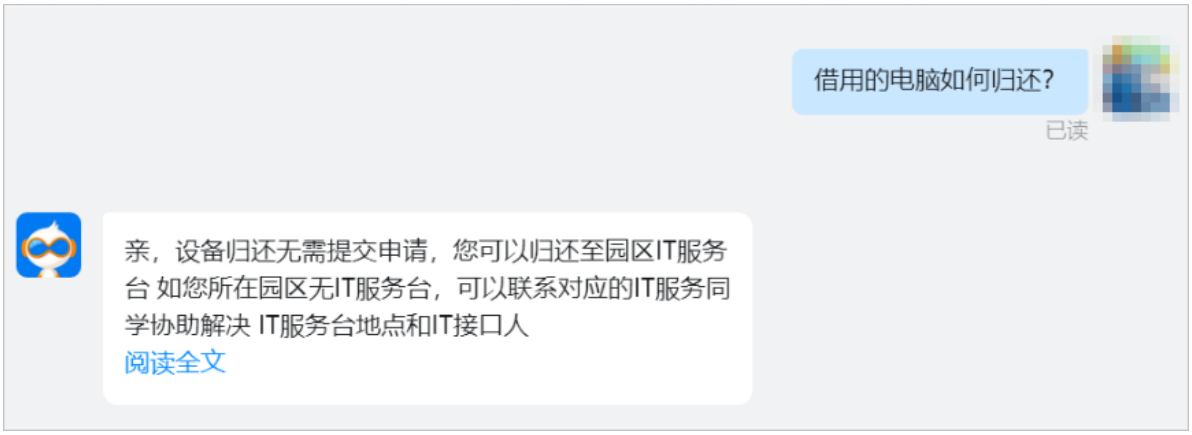
- **解决方案**：在App内自建NLP智能客服QA平台，将数据集中的用户问题分为多种不同类别的QA问题。当用户在线提问时，首先通过基于NLP领域的预训练BERT模型的文本分类算法，将用户问题定位到某一类别。然后通过人机对话形式，将提前准备的答案反馈给用户。
- **方案架构**：NLP智能客服的架构图如下所示。



- **BERT模型**：利用Wikipedia和BooksCorpus的海量大数据语料预训练获得的NLP迁移学习模型，在具体任

务中仅需要根据自己的业务数据进行Fine-tune即可。

- **场景示例**：智能客服在真实场景下的示例如下图所示。



- **效果提升**：该方案无需人工客服解答用户问题，在通常场景下，训练的NLP智能客服模型解答问题的准确率高达94%左右，响应时间在30 ms左右。如果使用更小的预训练模型，响应时间甚至降低到了10 ms，极大满足App内用户的问题咨询需求。

操作流程

基于阿里云机器学习PAI平台，搭建NLP智能客服的流程如下：

1. 步骤一：准备数据

   使用的训练集和测试集需要导入至MaxCompute，标签文件需要上传至OSS。

2. 步骤二：训练模型

   基于BERT进行业务数据的模型训练，您只需要输入自己的业务相关数据，并根据业务配置训练任务的相关参数，即可完成模型训练，从而获得训练好的NLP文本分类模型。

3. 步骤三：评估模型

   对训练好的NLP文本分类模型进行评估，查看模型效果（例如，预测准确率）。

4. 步骤四：模型预测

   对训练好的NLP文本分类模型输入测试数据集或评估数据集，获得预测的分类结果。

5. 步骤五：导出模型

   将训练满意的模型导出至文件夹中，以便后续将模型部署为在线服务。

6. 步骤六：部署模型

   通过PAI-EAS的EASCMD工具，将导出的模型部署为RESTful API。

7. 步骤七：调试模型服务

   您可以通过PAI-EAS控制台在线调用，或Python脚本批量调用的方式调式模型服务。

8. 步骤八：监控服务指标

   调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

步骤一：准备数据

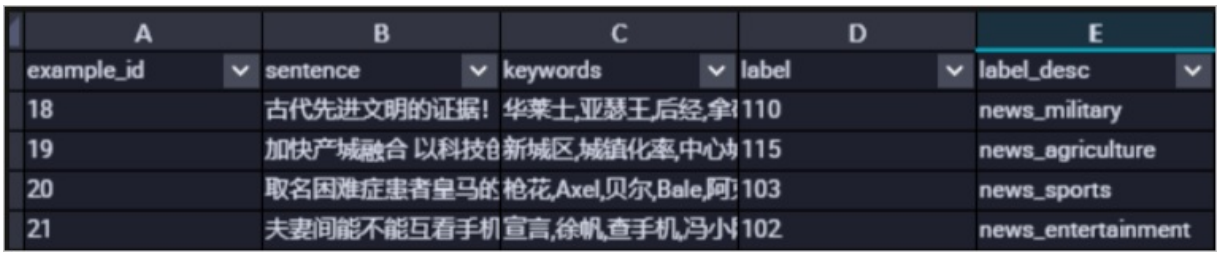
本文使用公开数据集TNews的数据进行实验，具体数据集如下：

- 训练集和测试集：已经上传到MaxCompute的公共项目pai_online_project中，训练表名为

tnews_bert_et_train，测试表名为tnews_bert_et_dev。您可以在后续的PAI命令中直接查看或调用这两张表。在实际应用中，您需要将自己的数据上传至MaxCompute中，详情请参见导入数据。

- 标签文件：下载标签文件tnews_labels_ext.txt，并将其上传至您的OSS目录中。关于如何上传文件至OSS，请参见上传文件。

本文使用的数据集包含五个字段，如下图所示。



	A	B	C	D	E
	example_id	sentence	keywords	label	label_desc
18		古代先进文明的证据！	华莱士,亚瑟王后经,拿	110	news_military
19		加快产城融合 以科技创	新城区,城镇化率,中心	115	news_agriculture
20		取名困难症患者皇马的	枪花,Axel,贝尔,Bale,阿	103	news_sports
21		夫妻间能不能互看手机	宣言,徐帆,查手机,冯小	102	news_entertainment

数据集字段说明

字段	描述
example_id	索引编号。
sentence	用于文本分类的文本内容。
keywords	文本关键词。
label	标签代号。
label_desc	标签详情。

步骤二：训练模型

基于BERT进行业务数据的模型训练，您只需要输入自己业务的相关数据，并根据业务配置训练任务的相关参数，即可完成模型训练，从而获得训练好的NLP文本分类模型。模型训练的输入数据及训练完成后的输出数据分别如下：

- 输入：在PAI命令中，将 `inputTable` 参数配置为训练集*tnews_bert_et_train*和测试集*tnews_bert_et_dev*，将 `labelEnumerateValues` 参数配置为标签文件*tnews_labels_ext.txt*。
- 输出：在PAI命令 `checkpointDir` 参数配置的路径中，输出模型文件和中间结果。

在DataWorks的ODPS SQL节点中，执行如下模型训练的PAI命令。

```
pai -name easy_transfer_app_ext
  -Dmode=train
  -DinputTable="odps://pai_online_project/tables/tnews_bert_et_train,odps://pai_online_proj
ect/tables/tnews_bert_et_dev"
  -DfirstSequence=sentence
  -DsecondSequence=keywords
  -DlabelName=label
  -DlabelEnumerateValues='oss://tongxin-demo1/EasyTransfer/tnews_labels_ext.txt'
  -DsequenceLength=64
  -DcheckpointDir='oss://tongxin-demo1/EasyTransfer/ckp_dir/'
  -DbatchSize=32
  -DnumEpochs=5
  -DdistributionStrategy=ExascaleStrategy
  -DworkerCount=2
  -DworkerGPU=2
  -DworkerCPU=2
  -DoptimizerType=adam
  -DlearningRate=2e-5
  -DmodelName=text_classify_bert
  -DuserDefinedParameters='pretrain_model_name_or_path=google-bert-base-zh'
  -Dbuckets="oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=acs:ram:
:1664081xxxxxxxx:role/aliyunodpspaidefaultrole,oss://tongxin-demo1.oss-cn-beijing-internal.
aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -Darn="acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -DossHost="oss-cn-beijing-internal.aliyuncs.com"
```

命令参数说明

参数	是否必选	描述	类型	示例值	默认值
name	是	PAI命令调用的算法包。本文使用迁移学习算法包，因此该参数配置为固定值 `easy_transfer_app_ext` 。	STRING	easy_transfer_app_ext	无
mode	是	使用模式。进行模型训练时，需要配置为固定值 `train` 。	STRING	train	无
inputTable	是	输入数据，即MaxCompute中的训练集表和测试集表，两张表之间使用英文逗号（,）分隔。	STRING	"odps://<your_project_name>/tables/<your_train_table_name>,odps://<your_project_name>/tables/<your_test_table_name>"	无

参数	是否必选	描述	类型	示例值	默认值
firstSequence	是	进行文本分类的内容在输入表中对应的列名。	STRING	sentence	无
secondSequence	否	进行辅助分类的文本内容在输入表中对应的列名。如果您需要提高文本分类的准确率，可以添加另外的文本相关字段进行辅助分类。例如，文本关键词。	STRING	keywords	空
labelName	是	标签在输入格式中对应的列名。	STRING	label	无
labelEnumerateValues	是	标签枚举值，支持以下两种方式： • 以英文逗号（,）分隔。 • 一个TXT文件的OSS路径，该TXT文件中的枚举值以换行符分隔。	STRING	• 1001,1003,1005,1007,1009 • "oss://<your_bucket_name>/<path_to_your>/labels.txt"	无
sequenceLength	否	序列整体最大长度，范围为1~512。	INT	128	128
checkpointDir	是	模型存储路径。	STRING	"oss://<your_bucket_name>/<path_to_your>/ckp_dir/"	无
batchSize	否	训练过程中的批处理大小。如果使用多机多卡，则表示每个GPU上的批处理大小。	INT	32	32
numEpochs	否	训练总Epoch的数量。	INT	1	1
distributionStrategy	否	分布式策略，支持以下取值： • MirroredStrategy：单机多卡。 • ExascaleStrategy：多机多卡。	STRING	"MirroredStrategy"	"MirroredStrategy"
workerCount	否	分布式服务器的数量，默认值表示1个Worker。	INT	2	1

参数	是否必选	描述	类型	示例值	默认值
workerGPU	否	每个Worker下的GPU卡数量。	INT	2	1
workerCPU	否	每个Worker下的CPU卡数量。	INT	2	1
optimizerType	否	优化器类型，支持以下取值： • adam • lamb • adagrad • adadeleta	STRING	"adam"	"adam"
learningRate	否	学习率。	FLOAT	2e-5	2e-5
modelName	否	模型名称。	STRING	text_classify_bert	text_match_bert
userDefinedParameters	是	用户自定义参数，可以配置预训练模型pretrain_model_name_or_path。常用如下三种预训练模型： • pai-bert-base-zh • google-bert-base-zh • cro-roberta-tiny-zh	STRING	"google-bert-base-zh"	无

参数	是否必选	描述	类型	示例值	默认值
buckets	是	OSS Buckets信息，格式为 `oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=<your_arn>,oss://<your_bucket>.<your_host>/?role_arn=<your_arn>` 。您需要将如下字段替换为实际值，其他部分使用上述的固定值： • <your_arn>：替换为您的ARN，取值与arn参数相同。 • <your_bucket>：替换为您创建的OSS存储空间（Bucket）的名称。 • <your_host>：替换为您OSS地域的Endpoint，取值与ossHost参数相同。	STRING	oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole,oss://tongxin-demo1.oss-cn-beijing-internal.aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"	无
arn	是	权限认证。您可以在PAI-Studio项目空间的左侧导航栏，选择**设置 > 基本设置**，即可查看ARN。	STRING	"acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"	无
ossHost	是	OSS地域的Endpoint。OSS地域与Endpoint的对应关系，请参见公共云下OSS Region和Endpoint对照表。	STRING	"oss-cn-beijing-internal.aliyuncs.com"	无

步骤三：评估模型

获得训练好的模型后，您可以通过模型评估的PAI命令，一键查看模型效果（例如模型预测准确率）。模型评估的输入数据及输出数据分别如下：

- 输入：在PAI命令中，将 `inputTable` 参数配置为测试集*tnews_bert_et_dev*，将 `checkpointPath` 参数配置为训练好的模型的路径。
- 输出：在运行日志Logview中可以查看评估结果。

在DataWorks的ODPS SQL节点中，执行如下PAI命令进行模型评估。

```
pai -name easy_transfer_app_ext
  -Dmode=evaluate
  -DinputTable="odps://pai_online_project/tables/tnews_bert_et_dev"
  -DcheckpointPath='oss://tongxin-demo1/EasyTransfer/ckp_dir/model.ckpt-2085'
  -DbatchSize=64
  -Dbuckets="oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=acs:ram:
:1664081xxxxxxxx:role/aliyunodpspaidefaultrole,oss://tongxin-demo1.oss-cn-beijing-internal.
aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -Darn="acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -DossHost="oss-cn-beijing-internal.aliyuncs.com"
```

该命令涉及的主要参数如下表所示，其他参数解释请参见命令参数说明。

参数	是否必选	描述	类型	示例值	默认值
mode	是	使用模式。进行模型评估时，需要配置为固定值 `evaluate` 。	STRING	evaluate	无
inputTable	是	输入数据，即MaxCompute中的测试集表。	STRING	"odps://pai_online_project/tables/tnews_bert_et_dev"	无
checkpointPath	是	训练好的模型的存储路径，对应模型训练PAI命令中的checkpointDir参数。但是这个参数存在如下区别： • checkpointDir：指定至模型文件的上层文件夹。 • checkpointPath：需要指定到模型名称，不能只到上层文件夹。	STRING	“oss://tongxin-demo1/EasyTransfer/ckp_dir/model.ckpt-2085”	无

步骤四：模型预测

对训练好的模型输入测试数据集或评估数据集，可以获得预测的分类结果。模型预测的输入数据和输出数据如下所示：

- 输入：在PAI命令中，将 `inputTable` 参数配置为测试集*tnews_bert_et_dev*，将 `checkpointPath` 参数配置为训练好的模型的路径。
- 输出：在PAI命令 `outputTable` 参数配置的MaxCompute表中，输出预测结果。

在DataWorks的ODPS SQL节点中，执行如下PAI命令进行模型预测。

```
pai -name easy_transfer_app_ext
  -Dmode=predict
  -DinputTable="odps://pai_online_project/tables/tnews_bert_et_dev"
  -DoutputTable="odps://tongxin_demo/tables/tnews_dev_pred"
  -DfirstSequence=sentence
  -DsecondSequence=keywords
  -DappendCols=example_id,label
  -DoutputSchema=predictions,probabilities,logits
  -DcheckpointPath='oss://tongxin-demo1/EasyTransfer/ckp_dir/'
  -Dbuckets="oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=acs:ram:
:1664081xxxxxxxx:role/aliyunodpspaidefaultrole,oss://tongxin-demo1.oss-cn-beijing-internal.
aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -Darn="acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -DossHost="oss-cn-beijing-internal.aliyuncs.com"
```

该命令涉及的主要参数如下表所示，其他参数解释请参见命令参数说明。

参数	是否必选	描述	类型	示例值	默认值
mode	是	使用模式。进行模型预测时，需要配置为固定值 `predict` 。	STRING	predict	无
inputTable	是	输入数据，即MaxCompute中的测试集表。	STRING	"odps://pai_online_project/tables/tnews_bert_et_dev"	无
checkpointPath	是	训练好的模型的存储路径，与模型训练PAI命令中的checkpointDir参数取值相同。	STRING	"oss://tongxin-demo1/EasyTransfer/ckp_dir/"	无
outputTable	是	预测结果输出表。	STRING	"odps://tongxin_demo/tables/tnews_dev_pred"	无
appendCols	是	输入表中需要添加到输出表的列，多个列之间使用英文逗号（,）分隔。	STRING	example_id,label	无

参数	是否必选	描述	类型	示例值	默认值
outputSchema	是	需要输出的预测值，支持输出以下三种预测值： • predictions：数据预测结果。 • probabilities：该数据属于每个Label的概率。 • logits：预测的原始值，即 `probabilities = softmax(logits)` 。 通常需要同时输出以上三种预测值，多个预测值之间使用英文逗号（,）分隔。	STRING	predictions,probabilities,logits	无

步骤五：导出模型

将训练满意的模型导出至文件夹中，以便后续将模型部署为在线服务。模型导出的输入数据和输出数据分别为：

- 输入：在PAI命令中，将 `checkpointPath` 配置为已训练好的模型。
- 输出：在PAI命令 `exportDirBase` 参数配置的路径下，输出导出的模型及相关配置文件。

在DataWorks的ODPS SQL节点中，执行如下PAI命令导出模型。

```
pai -name easy_transfer_app_ext
  -Dmode=export
  -DcheckpointPath='oss://tongxin-demo1/EasyTransfer/ckp_dir/model.ckpt-2085'
  -DexportType=app_model
  -DexportDirBase='oss://tongxin-demo1/EasyTransfer/export/'
  -Dbuckets="oss://atp-modelzoo-sh.oss-cn-shanghai-internal.aliyuncs.com/?role_arn=acs:ram:
:1664081xxxxxxxx:role/aliyunodpspaidefaultrole,oss://tongxin-demo1.oss-cn-beijing-internal.
aliyuncs.com/?role_arn=acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -Darn="acs:ram::1664081xxxxxxxx:role/aliyunodpspaidefaultrole"
  -DossHost="oss-cn-beijing-internal.aliyuncs.com"
```

该命令涉及的主要参数如下表所示，其他参数解释请参见命令参数说明。

参数	是否必选	描述	类型	示例值	默认值
mode	是	使用模式。进行模型导出时，需要配置为固定值 `export` 。	STRING	export	无

参数	是否必选	描述	类型	示例值	默认值
checkpointPath	是	训练好的模型的存储路径，对应模型训练PAI命令中的checkpointDir参数。但是这个参数存在如下区别： • checkpointDir：指定至模型文件的上层文件夹。 • checkpointPath：需要指定到模型名称，不能只到上层文件夹。	STRING	'oss://tongxin-demo1/EasyTransfer/ckp_dir/model.ckpt-2085'	无
exportType	是	导出类型，支持以下类型： • app_model：导出Finetune模型自身。 • ez_bert_feat：导出文本向量化组件所需模型。	STRING	app_model	无
exportDirBase	是	导出模型的目录。	STRING	"oss://<your_bucket_name>/<path_to_your>/export/"	无

步骤六：部署模型

1. 打包导出的模型文件。

   将OSS输出目录export（导出模型PAI命令中的 `exportDirBase` 参数配置的目录）下的*saved_model.pb*、*vocab.txt*、*variables*文件夹及模型目录ckp_dir下的*label_mapping.json*文件打包为*bert_model.tar.gz*。打包文件的详细信息如下图所示。

2. 使用PAI-EAS的在线部署工具EASCMD，将*bert_model.tar.gz*包上传至OSS。关于EASCMD工具中的命令使用方法，请参见命令使用说明。

   上传命令如下所示。

   ```
   $ eascmd64 upload bert_model.tar.gz
   ```

该命令执行成功后，系统返回如下类似的模型地址。

```
oss://eas-model-beijing/16640818xxxxxxxxx/bert_model.tar.gz
```

3. 创建如下服务描述文件*service_desc_json.json*。

```
{
  "name": "easynlp_tongxin_demo1",
  "generate_token": "true",
  "model_path": "oss://eas-model-beijing/16640818xxxxxxxxx/bert_model.tar.gz",
  "processor": "easy_nlp_gpu_tf1.12",
  "model_config": "{\"type\":\"text_classify_bert\"}",
  "metadata": {
    "resource": "eas-r-a47xufufi6yxxxxx",
    "cuda": "9.0",
    "instance": 1,
    "memory": 4000,
    "gpu": 1,
    "cpu": 4,
    "rpc.worker_threads" : 5
  }
}
```

该文件的核心参数如下（其他参数的含义请参见创建服务）：

- name：在线服务的名称，您可以自定义。
- model_path：上传打包的模型文件后，系统返回的模型地址。
- processor：本文的模型需要使用PAI-EAS提供的EasyNLP Processor进行部署，该Processor Code为`easy_nlp_gpu_tf1.12`。关于PAI-EAS提供的预置Processor及其Processor Code对应关系请参见预置Processor使用说明。
- resource：推荐使用PAI-EAS专属资源组部署模型服务，该参数为部署服务的专属资源组名称。

4. 在EASCMD工具中，执行以下命令部署模型。

```
$ eascmd64 create service_desc_json.json
```

步骤七：调试模型服务

您可以通过如下任何一种方式调式模型服务：

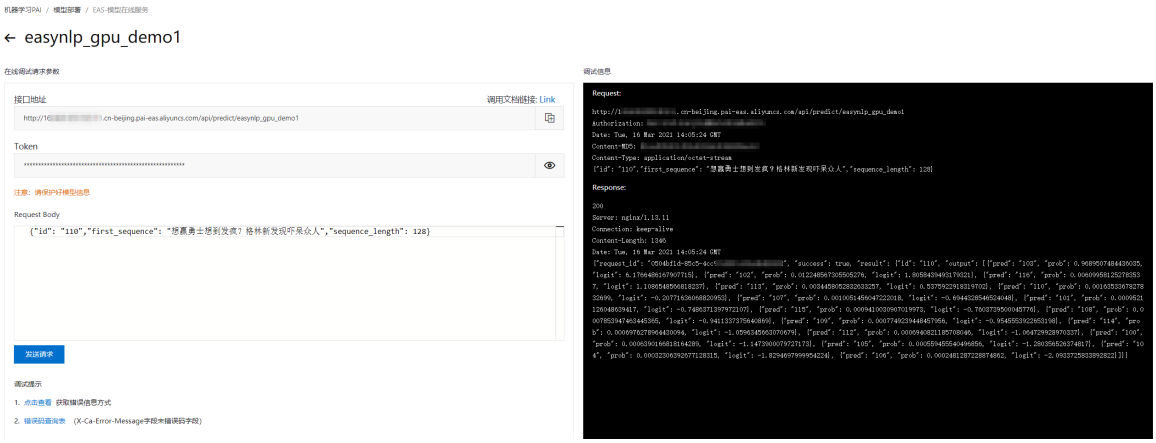
- 通过PAI-EAS控制台在线调用
- 通过Python脚本批量调用

通过PAI-EAS控制台在线调用服务的详细步骤如下。

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。
2. 在**PAI EAS模型在线服务**页面，单击上一步中已部署服务**操作**列下的**在线调试**。
3. 在调试页面的**在线调试请求参数**区域，配置参数。

参数	描述
接口地址	系统自动填入，无需手动配置。
Token	系统自动填入，无需手动配置。
Request Body	请求数据的样式如下所示。 `{"id": "110","first_sequence": "想赢勇士想到发疯？格林新发现吓呆众人","sequence_length": 128}`

4. 单击**发送请求**，即可在**调试信息**区域查看预测结果。



编写Python脚本，通过公网地址调用模型服务，详情请参见公网地址调用。通过Python脚本批量调试服务的详细步骤如下。

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入EAS-模型在线服务页面。
2. 在**PAI EAS模型在线服务**页面，单击已部署服务**服务方式**列下的**调用信息**。
3. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的**访问地址**和**Token**。
4. 创建如下Python脚本*eas_bert_demo.py*。

```
#!/usr/bin/env python
from eas_prediction import PredictClient
from eas_prediction import StringRequest
if __name__ == '__main__':
    #下面的client = PredictClient()入参源于公网调用的访问地址。
    client = PredictClient('http://1664081855xxxxxx.cn-beijing.pai-eas.aliyuncs.com/api/predict/easynlp_gpu_demo1')
    #Token可以在公网地址调用信息中获取。
    client.set_token('MGJmODcxY2Q4ZDc2NGNlZDIxMjJhNGE0Zmxxxxxxxxxxxxxxx')
    client.init()
    request = StringRequest('[{"id": "110","first_sequence": "想赢勇士想到发疯？格林新发现吓呆众人","sequence_length": 128},{"id": "112","first_sequence": "互金协会成立个人信用信息共享平台","sequence_length": 128}]')
    for x in range(0, 1):
        resp = client.predict(request)
        print(resp)
```

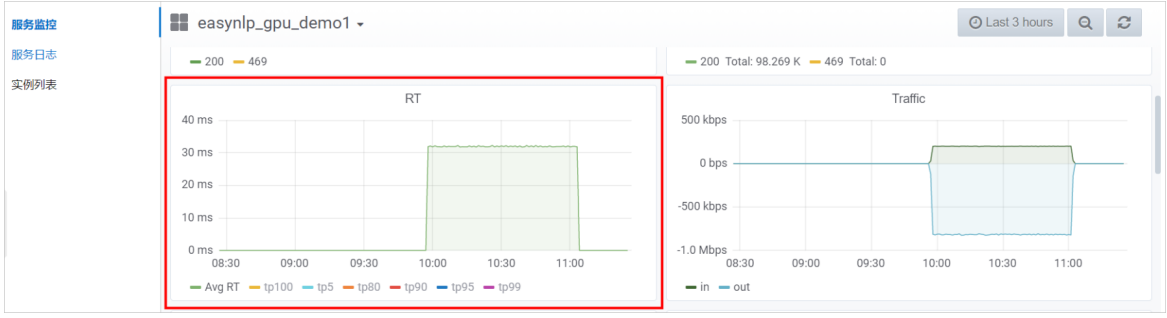
5. 将*eas_bert_demo.py*脚本上传至您的任意环境，并在脚本上传后的当前目录执行如下调用命令。

```
$ python3 eas_bert_demo.py
```

步骤八：监控服务指标

调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。
2. 在**PAI EAS模型在线服务**页面，单击已调用服务**服务监控**列下的图标。
3. 在**服务监控**页签，即可查看模型调用的指标水位。
   从服务监控的水位图中可以看到部署的BERT文本分类模型的时延在30 ms左右。



如果您需要在准确率损失较小的情况下，提升调用效率，建议模型训练时，配置pretrain_model_name_or_path参数为cro-roberta-tiny-zh模型，可以将时延降低到10 ms。

3.图像内容风控解决方案

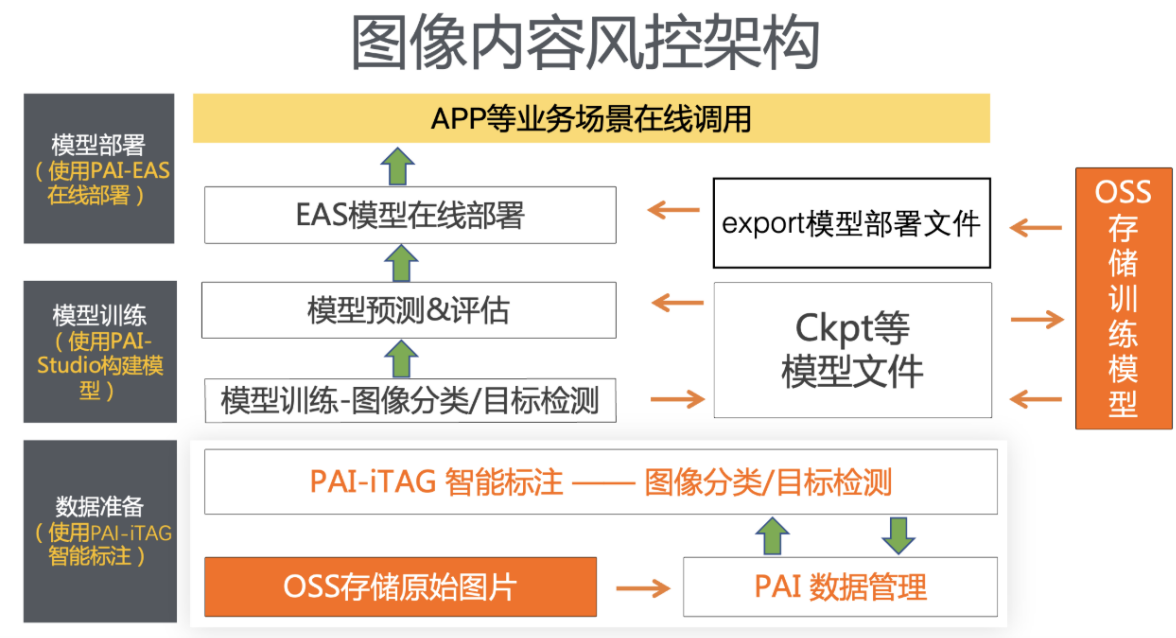
在线上业务的内容生产过程中，为了及时识别其中的高风险内容，PAI提供了图像内容风控解决方案。本文介绍如何基于人工智能算法，快速构建符合业务场景的风控模型，助力您快速识别高风险内容，进而对其进行拦截。

背景信息

在诸多生产内容的场景（例如使用图像进行评论、发布短视频、直播等）中，由于生产内容的范围不受限，因此难免出现高风险内容，您需要识别这些高风险内容，并及时拦截。针对该问题，阿里云机器学习PAI提出了如下解决方案，借助人工智能算法，帮助您快速判断风险内容：

- 解决方案
  i. 基于智能标注（iTAG）平台和PAI数据集管理，对目标场景的图像进行快捷标注和样本管理。
  ii. 基于PAI提供的预训练模型，针对自己的图像风控场景，在可视化建模平台PAI-Studio上进行模型Fine-Tune，从而基于Resnet50构建图像分类模型或基于YoloV5构建目标检测模型。
  iii. 将模型进行PAI-EAS在线部署，形成完整的端到端流程，从而自动识别生产过程中的风险内容。
- 方案架构

  图像内容风控解决方案的架构图如下所示。



前提条件

- 已开通PAI（Studio、DSW、EAS）后付费，详情请参见开通。
- 已开通AI工作空间，并添加MaxCompute计算资源或DLC计算资源，详情请参见AI工作空间（旧版）。
- 已开通MaxCompute，用于存储预测数据。关于如何开通MaxCompute，请参见通过查询编辑器使用MaxCompute。
- 已创建OSS存储空间（Bucket），用于存储原始数据、标签列表文件及训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创建专属资源组。

使用限制

由于OSS不能跨地域访问数据，因此存储数据的OSS Bucket与构建模型的PAI-Studio实验必须在同一地域。

操作流程

基于阿里云机器学习PAI平台，构建图像内容风控解决方案的流程如下：

1. 准备数据

   首先将原始图片数据存储在OSS，然后利用PAI数据集管理将原始数据扫描生成索引文件，最后通过智能标注（iTAG）进行数据标注，从而获得标注结果数据集，用于后续的模型训练。

   PAI提供了原始数据集，您可以直接使用它进行数据准备。关于数据集的下载方式，请参见内容风控领域的图像分类数据集或内容风控领域的目标检测数据集。

2. 构建用于图像内容风控的模型。

   在可视化建模平台PAI-Studio中，基于预训练模型，您可以构建Resnet50图像分类模型或YoloV5目标检测模型进行风险识别。实际应用中，您需要结合业务场景选择构建合适的模型：

   - 构建图像分类模型

     如果您的业务场景是将图像整体进行风险类别分类，则构建图像分类模型。

   - 构建目标检测模型

     如果您的业务场景是将图像中的某些高风险的实体进行框选检测，则构建目标检测模型。

3. 部署及调用模型服务

   通过模型在线服务PAI-EAS，您可以将训练好的内容风控模型部署为在线服务，并在实际的生产环境中调用，从而进行推理实践。

准备数据

1. 将原始图片分为训练数据集和测试数据集，分别上传至已创建的OSS Bucket。关于如何将文件上传至OSS，请参见上传文件。
2. 利用PAI数据集管理将原始数据扫描生成.manifest索引文件，详情请参见创建数据集：扫描文件夹创建数据集。
3. 通过智能标注（iTAG）管理员控制台，创建标注任务，详情请参见创建标注任务。
4. 通过智能标注（iTAG）标注员控制台，进行数据标注，详情请参见处理标注任务。
5. 在**智能标注（iTAG）**页面的**任务中心**页签，单击标注完成的任务操作列下的**获取标注结果**，得到存放在指定OSS目录下的标注结果数据集。

构建图像分类模型

1. 进入**可视化建模（Studio）**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏，选择**模型开发和训练 > 可视化建模（Studio2.0）**。
2. 基于实验模板，创建图像分类实验。
   i. 在**可视化建模（Studio）**页面，单击**实验模板**页签。
   ii. 单击CV页签。

iii. 在**图像分类**区域，单击**创建**。

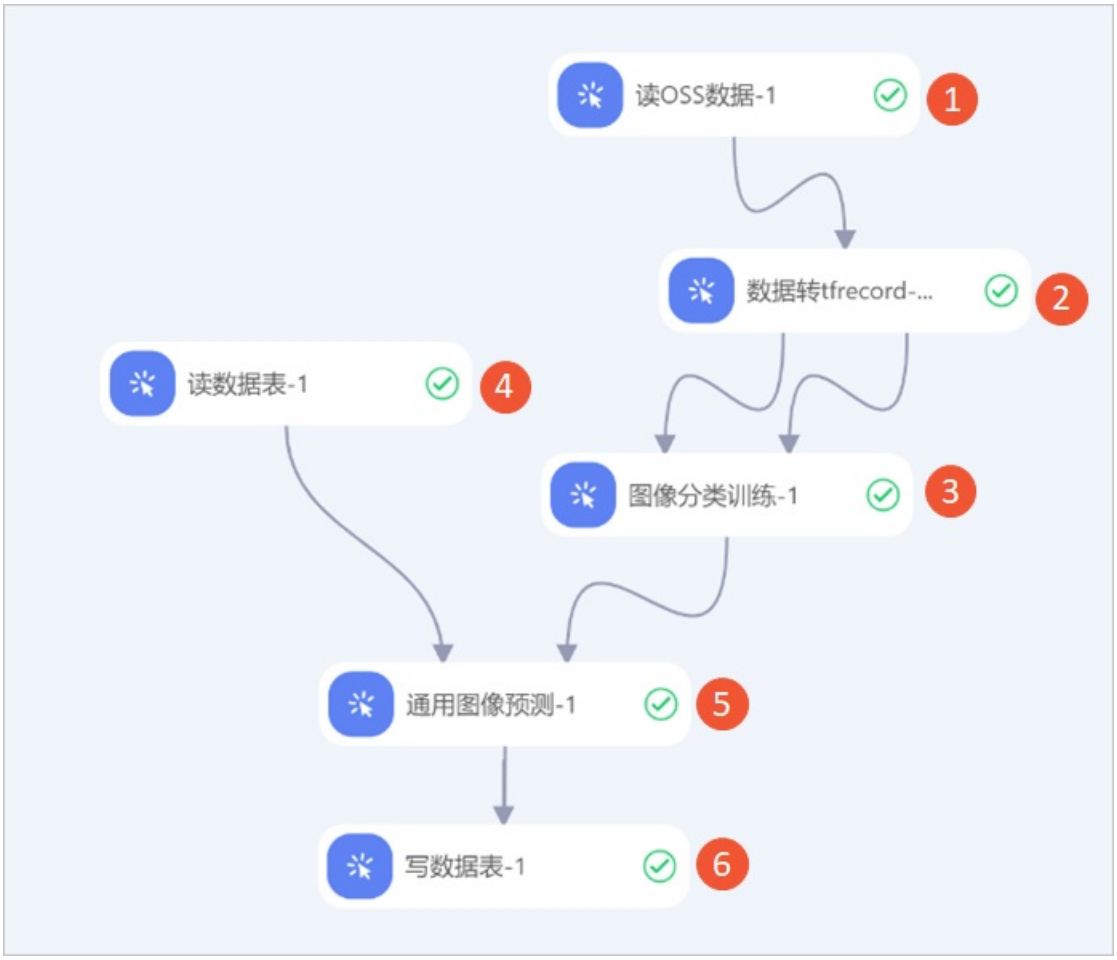
iv. 在**新建模板**对话框，先输入实验**名称**和**描述**、选择绑定的**AI工作空间**和实验存放的**位置**，再单击**确定**即可创建实验。

3. 进入实验，并配置组件参数。

i. 在**可视化建模（Studio）**页面，单击**实验列表**页签。

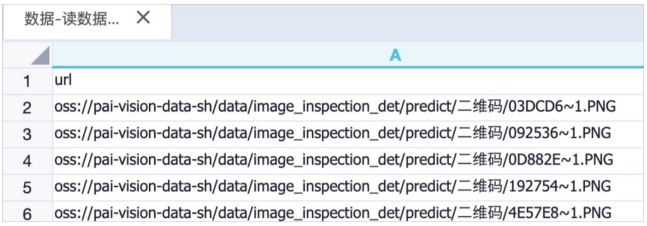
ii. 在**实验列表**页面，选中刚才创建好的模板实验，并单击**进入实验**。

iii. 系统根据预置的模板，自动构建实验，如下图所示。



区域	描述
①	配置实验的数据集，即配置**读OSS数据**组件的**OSS数据路径**参数为标注结果数据集的OSS路径。例如 `oss://pai-online-shanghai.oss-cn-shanghai.aliyuncs.com/ev_demo/图像智能审核检测标注_1626960686929.manifest` ，该数据集是PAI在华东2（上海）提供的标注好的数据集，您可以直接使用。 **说明** 由于OSS不能跨地域访问，因此使用该数据集构建模型时，您必须在华东2（上海）构建相关实验。
②	将数据集拆分为图像分类所需的训练集和测试集。**数据转tfrecord**组件的配置详情请参见下文的数据转tfrecord组件配置。
③	配置图像分类模型训练的参数。**图像分类训练**组件的配置详情请参见下文的图像分类组件配置。

区域	描述

④	配置用于模型预测的输入数据集。
	您需要先通过MaxCompute客户端的Tunnel命令将预测的输入数据集上传至MaxCompute，再将**读数据表**组件的**表名**参数配置为该MaxCompute表。关于MaxCompute客户端的安装及配置请参见MaxCompute客户端（odpscmd），关于Tunnel命令详情请参见Tunnel命令。本案例使用的示例命令如下所示。
	`# 建表语句。`
	`CREATE TABLE cv_risk_train(url STRING);`
	`# 上传语句。`
	`tunnel upload /Users/tongxin/xxx/cv_risk_train.csv cv_risk_train;`
	PAI提供了公开可访问的预测数据表pai_online_project.pascal_predict_hz，您可以直接使用。该预测数据集的内容如下图所示。
	 

⑤	使用训练好的图像分类模型对预测数据集进行预测。对于**通用图片预测**组件，您需要配置如下参数，其他参数使用实验模板中配置的默认值即可： ■ 选择**数据来源**为**Table**。 ■ 选择**模型类型**为**classifier** ■ 选择**图片列名**为区域④中MaxCompute表中存储图片地址的列，本案例选择为url列。 ■ 选择**保留列名**为**图片列名**，便于展示预测结果。
⑥	将预测的结果写入数据表中。例如写入test_image_inspection_cls中，即将**写数据表**组件的**数据表名**参数配置为test_image_inspection_cls。

数据转tfrecord组件配置

页签	参数	描述	本案例的示例值
字段设置	**转换配置文件路径**	转化配置文件的OSS路径。在PAI-Studio中无需使用该配置文件。	无需填写
	输出tfrecord路径	组件运行成功后，系统会自动在该路径下生成训练集和测试集。	`oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/test/convert_img_inspect_cls`
	输出tfrecord前缀	自定义输出TFRecord文件名称的前缀。	img_cls

页签	参数	描述	本案例的示例值
参数设置	**转换数据用于何种模型训练**	**数据转tfrecord**组件的输出数据可以用于以下类型的模型训练： ■ CLASSIFICATION：图像分类或多标签分类 ■ DETECTION：物体检测 ■ SEGMENTATION：语义分割 ■ POLYGON_SEGMENTATION：多边形语义分割 ■ INSTANCE_SEGMENTATION：实体分割 ■ TEXT_END2END：端到端OCR ■ TEXT_RECOGNITION：单行文字识别 ■ TEXT_DETECTION：文字检测 ■ VIDEO_CLASSIFICATION：视频分类	CLASSIFICATION
	类别列表文件路径	类别列表文件的OSS路径，类别列表文件的示例请参见图像分类类别列表文件。	`oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/ev_demo/`**图像审核-分类标注**`_class.txt`
	测试数据分割比例	测试数据分割比例。如果设置为0，则所有数据转换为训练数据。示例值0.1表示10%的数据作为验证集。	0.1
	图片最大边限制	如果配置了该参数，则大图片会被Resize后存入TFRecord，从而节省存储、提高数据读取速度。	无需填写
	测试图片最大边限制	同**图片最大边限制**，用于配置测试数据。	无需填写
	默认类别名称	默认类别名称。对于在类别列表文件中未找到的类别，系统将其映射到默认类别。	无需填写
	错误类别名称	含有该类别的物体和Box会被过滤，不参与训练。	无需填写

页签	参数	描述	本案例的示例值
	忽略类别名称	仅用于检测模型，含有该类别的Box会在训练中被忽略。	无需填写
	转换类名称	标注数据的来源类型，支持以下取值： ■ **PAI标注格式** ■ **亲测标注格式** ■ **自监督标注格式**	PAI标注格式
	分隔符	用于标记内容的分隔符。	无需填写
	图片编码方式	TFRecord中图片的编码方式。	jpg
执行调优	**读取并发数**	训练过程读取并发数。	10
	写tfrecord并发数	训练过程写TFRecord并发数。	1
	每个tfrecord保存图片数	训练过程每个TFRecord保存的图片数量。	256
	worker个数	训练过程中的Worker数量。	3
	CPU Core个数	训练过程中的CPU Core数量。	800
	memory大小	训练过程中的内存大小，单位为MB。	20000

图像分类组件配置

页签	参数	描述	本案例的示例值
字段设置	**训练所用oss目录**	存储训练模型的OSS目录。	`oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/test/test_image_inspection`
	训练数据文件oss路径	训练数据集的OSS路径。如果通过上游组件传递训练数据，则无需指定该参数。	无需填写
	评估数据oss路径	评估数据文件的OSS路径。如果通过上游组件传递评估数据，则无需指定该参数。	无需填写
	label_map_path文件oss路径	如果不指定类别列表文件，则使用TFRecords自动识别的类别。	无需填写

页签	参数	描述	本案例的示例值
	是否使用预训练的模型	建议使用预训练模型，以提高训练模型的精度。	是
	预训练模型oss路径	如果有自己的预训练模型，则将该参数配置为自己预训练模型的OSS路径。 如果没有配置该参数，则使用PAI提供的默认预训练模型。	无需填写
	识别模型网络	识别模型的网络名称，系统支持如下主流的识别模型： ■ vgg_16 ■ vgg_19 ■ inception_v1 ■ inception_v2 ■ inception_v3 ■ inception_v4 ■ mobilenet_v1 ■ mobilenet_v2 ■ resnet_v1_50 ■ resnet_v1_101 ■ resnet_v1_152	resnet_v1_50
	分类类别数目	类别标签的数量。	4
	图片resize大小	图片Resize后的大小。	224
	是否使用crop进行数据增强	是否对输入图片进行裁剪。通常建议进行裁剪，以提升模型的训练精度和效率。	是
	是否针对每个类别单独做评估	在日志中，是否对每个类别进行单独评估，不影响训练结果。	否

页签	参数	描述	本案例的示例值
参数设置	**学习率调整策略**	系统支持以下调整策略： ■ **exponential_decay**：指数衰减，详细请参见exponential_decay。 ■ **polynomial_decay**：多项式衰减，详情请参见polynomial_decay。其中num_steps自动设置为总的训练迭代次数，end_learning_rate为initial_learning_rate的千分之一。 ■ **manual_step**：人工指定各阶段的学习率，建议进阶用户使用。通过decay_epochs指定需要调整学习率的迭代轮数，通过learning_rates指定对应迭代轮数使用的学习率。	exponential_decay
	初始学习率	初始学习率。	0.001
	decay_epoches	学习率迭代过程中，连续两次衰减之间的间隔轮数。	10
	decay_factor	学习率迭代过程的衰减比率。	0.9
	staircase	如果选中该复选框，则按照离散间隔进行学习率衰减，否则按照连续间隔进行学习率衰减。	是
	训练batch_size	训练的批大小，即单次模型迭代或训练过程中使用的样本数量。	32
	评估batch_size	评估的批大小，即评估时单次模型迭代或训练过程中使用的样本数量。	32
	训练迭代轮数	训练的迭代轮数。	80
	评估数据条目数	训练过程中评估数据条目数。	200
	评估过程可视化显示的样本数目	评估过程可视化显示的样本数量。	64
	保存checkpoint的频率	保存Checkpoint的频率，以Epoch为单位。取值为1表示每完成一次训练就保存一次Checkpoint。	5

页签	参数	描述	本案例的示例值
执行调优	**优化方法**	模型训练的优化方法，支持以下取值： ■ momentum：指sgd ■ adam	momentum
	读取训练数据线程数	读取训练数据的线程数量。	4
	单机或分布式	组件运行的引擎，您可以结合实际情况选择。系统支持以下计算引擎： ■ MaxCompute：使用您在对应的AI工作空间中绑定的MaxCompute实例进行计算。关于如何添加计算资源请参见配置工作空间。关于计费，请参见深度学习组件定价。 ■ DLC：使用您在对应的AI工作空间中绑定的DLC实例进行计算。关于如何添加计算资源请参见配置工作空间。关于计费，请参见PAI-DLC公共资源组计费。 结合对训练模型的单机或分布式要求，系统支持以下取值： ■ **单机maxCompute** ■ **分布式maxCompute** ■ **单机DLC** ■ **分布式DLC**	单机maxCompute
	是否使用GPU	训练使用GPU卡的数量，取值100表示使用1张GPU卡。	100

构建目标检测模型

1. 进入**可视化建模（Studio）**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏，选择**模型开发和训练 > 可视化建模（Studio2.0）**。
2. 基于实验模板，创建目标检测实验。
   i. 在**可视化建模（Studio）**页面，单击**实验模板**页签。
   ii. 单击**CV**页签。

iii. 在**基于Yolov5模型的目标检测**区域，单击创建。

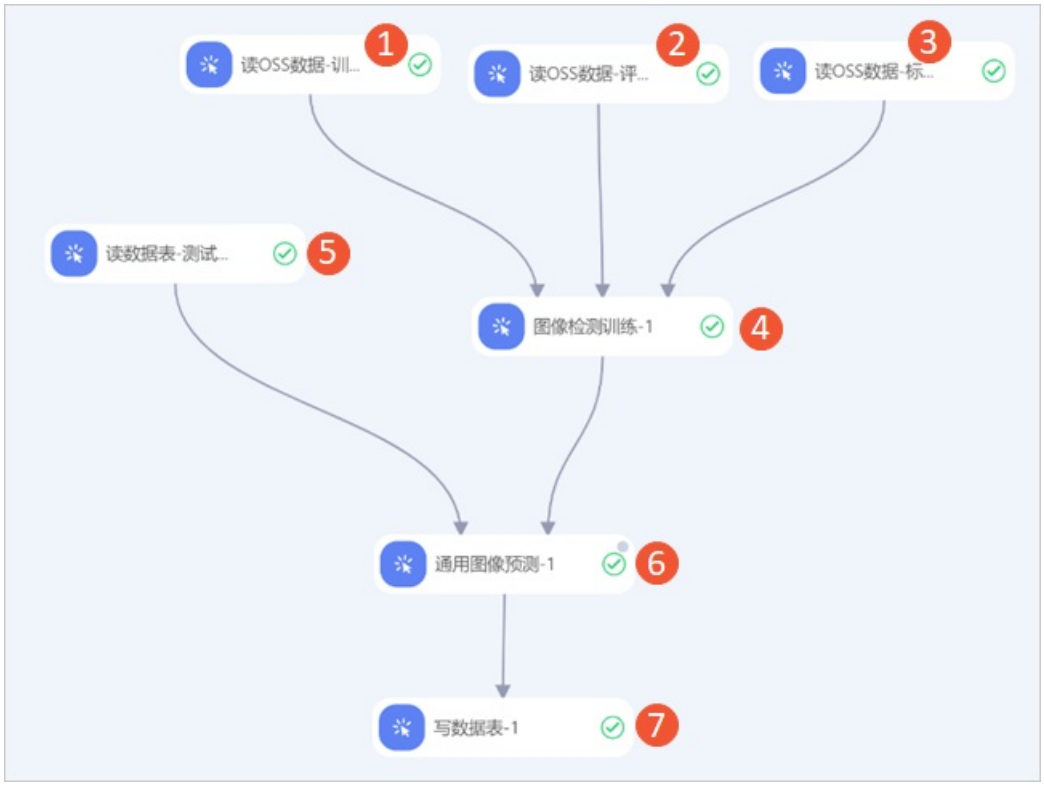
iv. 在**新建模板**对话框，先输入实验**名称**和**描述**、选择绑定的**AI工作空间**和实验存放的**位置**，再单击**确定**即可创建实验。

3. 进入实验，并配置组件参数。

i. 在**可视化建模（Studio）**页面，单击**实验列表**页签。

ii. 在**实验列表**页面，选中刚才创建好的模板实验，并单击**进入实验**。

iii. 系统根据预置的模板，自动构建实验。您查看的实验如下图所示。



区域	描述
①	配置实验的训练数据集，即配置**读OSS数据-训练集**组件的**OSS数据路径**参数为标注结果数据集中训练集的OSS路径。例如 `oss://pai-online-shanghai.oss-cn-shanghai.aliyuncs.com/ev_demo/risk_det_train.manifest` ，该数据集是PAI在华东2（上海）提供的标注好的训练集，您可以直接使用。 **说明** 由于OSS不能跨地域访问，因此使用该数据集构建模型时，您必须在华东2（上海）构建相关实验。
②	配置实验的评估数据集，即配置**读OSS数据-评估集**组件的**OSS数据路径**参数为标注结果数据集中评估集的OSS路径。例如 `oss://pai-online-shanghai.oss-cn-shanghai.aliyuncs.com/ev_demo/risk_det_dev.manifest` ，该数据集是PAI在华东2（上海）提供的标注好的评估集，您可以直接使用。 **说明** 由于OSS不能跨地域访问，因此使用该数据集构建模型时，您必须在华东2（上海）构建相关实验。
③	配置实验数据的标签列表文件，即配置**读OSS数据-标签列表**组件的**OSS数据路径**参数为标签列表文件的OSS路径。例如 `oss://pai-online-shanghai.oss-cn-shanghai.aliyuncs.com/ev_demo/图像审核-检测标注_class.txt` 。您可以参考示例标签列表文件（下载图像检测的标签列表文件示例），构建自己的标签列表文件，
④	配置目标检测模型训练的参数。**图像检测训练**组件的配置详情请参见下文的图像检测组件配置。

⑤	配置模型预测需要的输入数据集。
	您需要先通过MaxCompute客户端的Tunnel命令将预测的输入数据集上传至MaxCompute，再将**读数据表**组件的**表名**参数配置为该MaxCompute表。关于MaxCompute客户端的安装及配置请参见MaxCompute客户端（odpscmd），关于Tunnel命令详情请参见Tunnel命令。本案例使用的示例命令如下所示。
	`# 建表语句。`
	`CREATE TABLE cv_risk_train(url STRING);`
	`# 上传语句。`
	`tunnel upload /Users/tongxin/xxx/cv_risk_train.csv cv_risk_train;`
	PAI提供了公开可访问的预测数据表pai_online_project.pascal_predict_hz，您可以直接使用。该预测数据集的内容如下图所示。
	数据-读数据... ×
	A
	1 url
	2 oss://pai-vision-data-sh/data/image_inspection_det/predict/二维码/03DCD6~1.PNG
	3 oss://pai-vision-data-sh/data/image_inspection_det/predict/二维码/092536~1.PNG
	4 oss://pai-vision-data-sh/data/image_inspection_det/predict/二维码/0D882E~1.PNG
	5 oss://pai-vision-data-sh/data/image_inspection_det/predict/二维码/192754~1.PNG
	6 oss://pai-vision-data-sh/data/image_inspection_det/predict/二维码/4E57E8~1.PNG

区域	描述
⑥	使用训练好的图像分类模型对预测数据集进行预测。对于**通用图片预测**组件，您需要配置如下参数，其他参数使用实验模板中配置的默认值即可： ■ 选择**数据来源**为Table。 ■ 选择**模型类型**为**yolo5预测** ■ 选择**图片列名**为区域⑤中MaxCompute表中存储图片地址的列，本案例选择为url列。 ■ 选择**保留列名**为**图片列名**，便于展示预测结果。
⑦	将预测的结果写入数据表中。例如写入yolov5_predict中，即将**写数据表**组件的**数据表名**参数配置为yolov5_predict。

图像检测组件配置

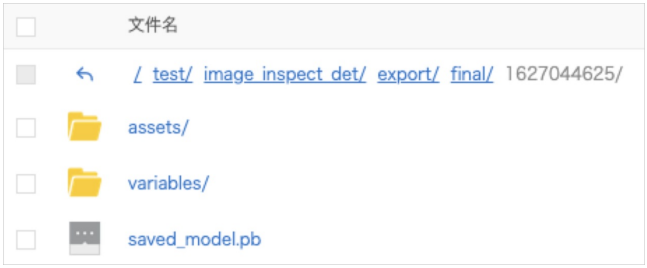
页签	参数	描述	本案例的示例值
字段设置	**训练模型类型**	训练所选的模型类型，支持以下取值： ■ SSD ■ FasterRCNN ■ RFCN ■ YOLOV5：最新推出的模型网络，拥有精度高、训练快的优势。	YOLOV5
	训练所用oss目录	存储训练模型的OSS目录。	`oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/test/risk_det_yolov5/ckpt/`
	训练数据文件oss路径	训练数据文件的OSS路径。如果通过上游组件传递训练数据，则无需指定该参数。	无需填写
	评估数据oss路径	评估数据文件的OSS路径。如果通过上游组件传递评估数据，则无需指定该参数。	无需填写
	类别列表文件OSS路径	如果不指定类别列表文件，则使用TFRecords自动识别的类别。	无需填写
	YOLOV5数据源格式	标注数据的来源类型，支持**PAI标注格式**、**Pascal VOC格式**及**YOLOV5原生格式**三种检测数据，您需要结合实际情况选择。	PAI标注格式

页签	参数	描述	本案例的示例值
	预训练模型oss路径	如果您有自己的预训练模型，则填入预训练模型的OSS路径，以提高训练模型的精度。	无需填写
参数设置	**YOLOV5检测模型使用的模型大小**	选择YOLOV5检测模型使用的模型大小和网络结构。	yolov5s
	检测类别数目	数据中类别标签的数目。如果没有配置该参数，则默认从数据集中分析得到类别数量。	3
	yolov5输入图片大小	图片大小等于图片的高度乘以宽度，取值为整数。	640
	初始学习率	网络训练的学习率。	0.01
	训练batch_size	训练的批大小，即单次模型迭代或训练过程中使用的样本数量。目	32
	评估batch_size	评估的批大小，即单次模型迭代或训练过程中使用的样本数量。	30
	总的训练迭代epoch轮数	总的训练迭代轮数。	50
	保存checkpoint的频率	保存模型文件的频率。取值1表示对所有训练数据都进行一次迭代。	1
执行调优	**读取训练数据线程数**	读取训练数据的线程数。	4
	单机或分布式(maxCompute/DLC)	运行模式，**YOLOV5**模型仅支持**分布式DLC**。	分布式DLC
	worker个数	用于计算的Worker数量。	1
	cpu机型选择	计算资源的CPU机型。	32vCPU+128GB Mem-ecs.g6.8xlarge
	gpu机型选择	计算资源的GPU机型。	48vCPU+368GB Mem+4xv100-ecs.gn6e-c12g1.12xlarge

部署及调用模型服务

通过模型在线服务PAI-EAS，您可以将训练好的内容风控模型部署为在线服务，并在实际的生产环境中调用，从而进行推理实践。

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。
2. 部署模型服务。
   i. 在**PAI EAS模型在线服务**页面，单击**模型上传部署**。
   ii. 在**资源和模型**面板，配置参数（此处仅介绍与本案例相关的核心参数配置方法，其他参数的解释请参见控制台上传部署），并单击**下一步**。

参数	描述
自定义模型名称	模型的名称，建议结合实际业务进行命名，以区分不同的模型服务。
资源组种类	建议使用专属资源组部署模型服务，从而避免公共资源组资源有限时导致的服务排队。关于如何创建专属资源组，请参见专属资源组。
资源种类	如果您选择**资源组种类**为**公共资源组**，则需要选择**资源种类**为GPU。
Processor种类	选择EasyVision。
模型类型	如果部署图像分类模型，则选择**图像分类**。 如果部署目标检测模型，则选择**目标检测和实例分割**。
模型文件	本案例中训练好的模型均存储在OSS中，因此选择**OSS文件导入**。 如果部署图像分类模型，则选择**训练所用oss目录**下export文件夹中的模型文件夹（其中包含assets文件夹、variables文件夹及saved_model.pb）。例如，对于下图中的模型目录结构，您需要将OSS路径选择到/1627044625/目录。  如果部署目标检测模型，则选择**训练所用oss目录**下的模型文件即可。例如选择epoch_50_export.pt。

iii. 在**部署详情及配置确认**面板，配置模型服务占用资源的相关参数。

参数	描述
实例数	本案例配置为1。实际应用时建议配置多个服务实例，以避免单点部署带来的风险。
卡数	本案例配置为1。实际应用时根据情况配置。
核数	本案例配置为1。实际应用时根据情况配置。
内存数（M）	本案例配置为16384 MB。实际应用时根据情况配置。

iv. 单击**部署**，等待一段时间即可完成模型部署。

3. 查看模型服务的公网地址和访问Token。
   i. 在**PAI EAS模型在线服务**页面，单击目标服务**服务方式**列下的**调用信息**。
   ii. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的访问地址和**Token**。
4. 使用脚本进行批量调用。
   i. 创建调用模型服务的Python脚本。
      - 图像分类的Python脚本cv_risk_cls.py

```
import requests
import base64
import sys
import json
resp = requests.get('https://tongxin-lly.oss-cn-hangzhou.aliyuncs.com/iTAG/pic_public/1.jpg')
ENCODING = 'utf-8'
datas = json.dumps( {
            "image": base64.b64encode(resp.content).decode(ENCODING),
            })
head = {
   "Authorization": "YjdhYWRhYWZhYzNjZTFlMDZlNjAxxxxxxxxxxxxxxxxxxxx" # 服务的访问Token需要替换为实际值。
}
for x in range(0,1):
    # 下面的服务公网访问地址需要替换为实际值。
    resp = requests.post("http://1664081855183111.cn-shanghai.pai-eas.aliyuncs.com/api/predict/cv_risk_cls001", data=datas, headers=head)
    print(resp.text)
print("test endding")
```

- 目标检测的Python脚本cv_risk_det.py

```
import requests
import base64
import sys
import json
img_file = './xxx.jpg'
ENCODING = 'utf-8'
datas = json.dumps( {
            "image": base64.b64encode(open(img_file, 'rb').read()).decode(ENCODIN
G),
            })
head = {
   "Authorization": "NGVkMTVmZjNlNzA3ZGVlNWIzxxxxxxxxxxxxxx" # 服务的访问Token需要
替换为实际值。
}
# 下面的服务公网访问地址需要替换为实际值。
r = requests.post("http://1664081xxxxxxxxx.cn-shanghai.pai-eas.aliyuncs.com/api/p
redict/cv_risk_obj001", data=datas, headers=head)
print(r.text)
```

ii. 将图像分类或目标检测的Python脚本上传至您的任意环境，并在脚本上传后的当前目录执行如下调用命令。

```
$ python3 <cv_risk_xxx.py>
```

其中<cv_risk_xxx.py>需要替换为实际的Python脚本名称。

调用服务后，得到的预测结果示例如下图所示。

{"request_id": "7352db33-1cd7-4e5b-[redacted]", "success": true, "class": 0, "class_name": "火灾", "pro
bs": [0.9999966621398926, 1.8512555470806547e-06, 2.8873432711407077e-07, 1.178896241071925e-06], "class_probs
": {"火灾": 0.9999966621398926, "暴恐": 1.8512555470806547e-06, "游行": 2.8873432711407077e-07, "其他": 1.1788
96241071925e-06}}

5. 监控服务指标。

调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

i. 在**PAI EAS模型在线服务**页面，单击已调用服务**服务监控**列下的图标。

ii. 在**服务监控**页签，即可查看模型调用的指标水位。

从下面的服务监控水位图中可以看到，本案例部署的模型时延在50 ms左右。您自己的模型时延以实际为准。

4.文本内容风控解决方案

在线上业务的内容生产过程中，为了及时识别其中的高风险内容，PAI提供了文本内容风控解决方案。本文介绍如何基于人工智能算法，快速构建符合业务场景的风控模型，助力您快速识别高风险内容，进而对其进行拦截。

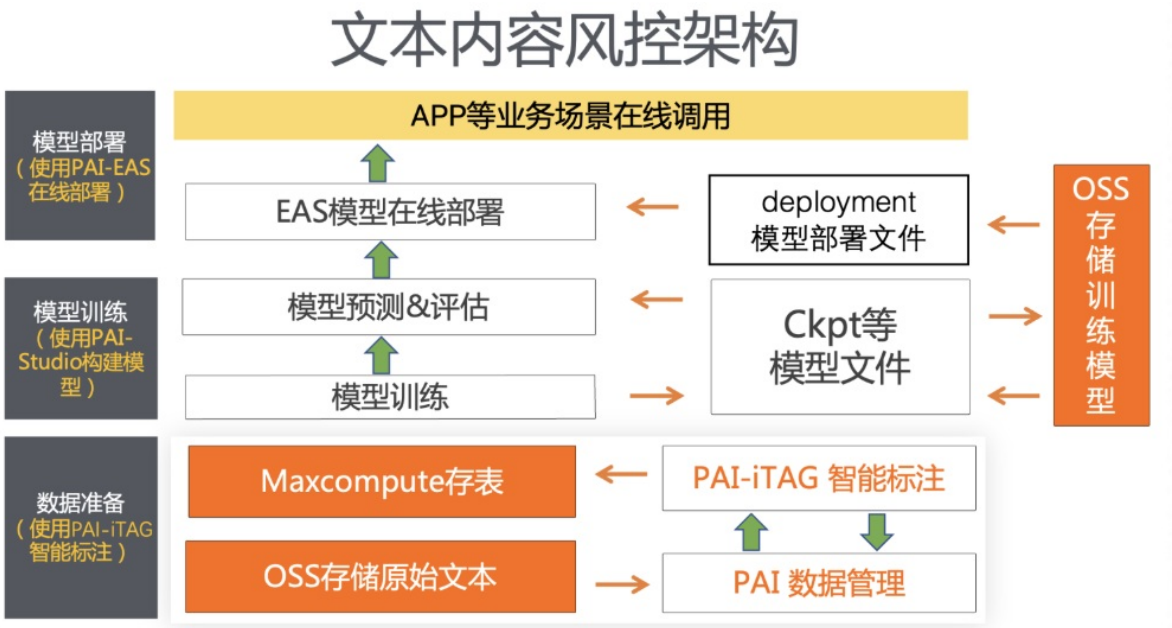
您可前往阿里云AI体验馆中的内容安全审核页面，体验文本内容风控的智能服务。

背景信息

在诸多生产内容的场景（例如评论、博客、商品介绍）中，由于生产内容的范围不受限，因此难免出现高风险内容，您需要识别这些高风险内容，并及时拦截。针对该问题，阿里云机器学习PAI提出了如下解决方案，借助人工智能算法，帮助您快速判断风险内容：

- 解决方案
  i. 基于智能标注（iTAG）平台和PAI数据集管理，对目标场景的文本进行快捷标注和样本管理。
  ii. 基于PAI提供的BERT迁移学习预训练模型，针对自己的文本风控场景，在可视化建模平台PAI-Studio上进行模型Fine-Tune，从而构建具体场景的NLP文本风控模型。
  iii. 将模型进行PAI-EAS在线部署，形成完整的端到端流程，从而自动识别生产过程中的风险内容。
- 方案架构

  文本内容风控解决方案的架构图如下所示。



前提条件

- 已开通PAI（Studio、DSW、EAS）后付费，详情请参见开通。
- 已开通AI工作空间，并添加MaxCompute计算资源，详情请参见AI工作空间（旧版）。
- 已开通MaxCompute，用于存储预测数据。关于如何开通MaxCompute，请参见通过查询编辑器使用MaxCompute。
- 已创建OSS存储空间（Bucket），用于存储原始数据、标签列表文件及训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创

建专属资源组。

操作流程

基于阿里云机器学习PAI平台，构建文本内容风控解决方案的流程如下：

1. 准备数据

   基于智能标注（iTAG）进行原始数据标注，然后将获得的训练集和测试集上传到数据仓库MaxCompute中，用于后续的模型训练。

2. 构建文本分类模型

   在可视化建模平台PAI-Studio中，基于海量大数据语料预训练获得的NLP迁移学习模型，构建文本内容风控场景的文本分类模型。

3. 部署及调用模型服务

   通过模型在线服务PAI-EAS，您可以将训练好的文本风控模型部署为在线服务，并在实际的生产环境中调用，从而进行推理实践。

准备数据

首先基于智能标注（iTAG）进行原始数据标注，然后将获得的训练集和测试集上传到数据仓库MaxCompute中，用于后续的模型训练。

1. 将原始数据分为训练集和测试集，分别按照iTAG的标注数据格式，生成.manifest格式的数据集。文件格式详见示例

   .manifest文件示例请参见textDemo.manifest。

2. 在PAI的数据集管理模块，通过本地上传方式将准备好的.manifest数据集上传，详情请参见创建数据集：本地上传。

3. 通过智能标注（iTAG）管理员控制台，创建标注任务，详情请参见创建标注任务。

4. 通过智能标注（iTAG）标注员控制台，进行数据标注，详情请参见处理标注任务。

5. 在**智能标注（iTAG）**页面的**任务中心**页签，单击标注完成的任务操作列下的**获取标注结果**，得到存放在指定OSS目录下的标注结果数据集。

6. 将获得的标注结果数据集转换为模型构建需要的训练集和测试集的CSV格式数据表。示例训练集请参见nlp_risk_train.csv，示例测试集请参见nlp_risk_dev.csv。

7. 通过MaxCompute客户端的Tunnel命令将训练数据集和测试数据集上传至MaxCompute。关于MaxCompute客户端的安装及配置请参见MaxCompute客户端（odpscmd），关于Tunnel命令详情请参见Tunnel命令。本案例使用的示例命令如下所示。

   ```
   # 建表语句。
   CREATE TABLE nlp_risk_train(content STRING, qince_result STRING);
   # 上传语句。
   tunnel upload /Users/tongxin/xxx/nlp_risk_train.csv nlp_risk_train;
   ```

构建文本分类模型

1. 登录PAI控制台。
2. 在左侧导航栏，选择**模型开发和训练 > 可视化建模（Studio2.0）**。
3. 选择用于创建文本分类的实验模板。

i. 在**可视化建模（Studio）**页面，单击**实验模板**页签。

ii. 单击**NLP**页签。

iii. 在**基于BERT模型的文本分类**区域，单击**创建**。

4. 填写实验基本信息，从而创建文本分类实验。

   在**新建模板**对话框，先输入实验**名称**和**描述**、选择绑定的**AI工作空间**和实验存放的**位置**，再单击**确定**即可创建实验。

5. 进入实验，并配置组件参数。

   i. 在**可视化建模（Studio）**页面，单击**实验列表**页签。

   ii. 在**实验列表**页面，选中刚才创建好的模板实验，并单击**进入实验**。

   iii. 系统根据预置的模板，自动构建实验，如下图所示。

区域	描述
①	配置实验的训练数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的训练表。例如nlp_risk_train。
②	配置实验的评估数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的评估表。例如nlp_risk_dev。

区域	描述
③	配置文本分类模型训练的参数。**文本分类训练**组件的配置详情请参见下文的文本分类组件配置。
④	配置实验的预测数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的预测表。例如nlp_risk_dev。
⑤	使用训练好的文本分类模型对预测数据集进行预测。**通用NLP预测**组件的配置详情请参见下文的预测组件的参数配置。
⑥	对模型预测得到的文本分类结果进行多分类评估。选择原样本结果列的qince_result标签字段和预测分类结果列的predictions进行对比评估。您可以右键单击该组件，在快捷菜单中选择**查看数据**，从而查看Accuracy和F1等评估指标。

文本分类组件配置

页签	参数	描述	本案例使用的示例值
字段设置	**文本列选择**	进行文本分类的内容在输入表中对应的列名。	content
	标签列选择	标签在输入表中对应的列名。	qince_result
	标签枚举值	标签列内容的枚举值，多个值之间使用半角逗号（,）分隔。	正常,低俗,成人,其他风险
	样本权重列	特定列样本的数据增强，非必选。	无需填写
	模型存储路径	训练得到的模型的OSS存储路径。	oss://tongxin-lly.oss-cn-hangzhou-internal.aliyuncs.com/pai/text_spam_rb/
	pretrainModelNameOrPath	选择使用的预训练语言模型。	自定义
	优化器选择	优化器的类型，支持以下取值： ■ adam ■ adagrad ■ lamb	adam
	batchSize	训练过程中的批处理大小。如果使用多机多卡，则表示每个GPU上的批处理大小。	32
	sequenceLength	序列整体最大长度，取值范围为1~512。	64
	numEpochs	训练总Epoch的数量。	1

页签	参数	描述	本案例使用的示例值
参数设置	学习率	模型构建过程中的学习率。	1e-5
	模型额外参数	用户自定义参数，可以配置预训练模型pretrain_model_name_or_path，常用如下四种预训练模型： ■ base-roberta ■ base-bert ■ tiny-roberta ■ tiny-bert 模型精度排序：`base-roberta > base-bert > tiny-roberta > tiny-bert` 。 模型速度排序：`base-roberta = base-bert < tiny-roberta = tiny-bert` 。	`pretrain_model_name_or_path=oss://atp-modelzoo-sh/release/easytexminer_model_zoo/nlu/text_anti_spam_tx/tiny-roberta/model.ckpt`
执行调优	指定Worker数	分布式服务器的数量，默认值表示1个Worker。	1
	指定Worker的CPU卡数	每个Worker下的GPU卡数量。	1
	指定Worker的GPU卡数	每个Worker下的CPU卡数量。	1
	分布式策略	分布式策略，支持以下取值： ■ **MirroredStrategy**：单机多卡。 ■ **ExascaleStrategy**：多机多卡。	MirroredStrategy

预测组件的参数配置

页签	参数	描述	本案例使用的示例值
	第一文本列选择	进行文本分类的内容在输入表中对应的列名。	content
	第二文本列选择	进行辅助分类的文本内容在输入表中对应的列名，非必选。	无需填写
	输出列	预测结果输出列。如果有多个列，则使用半角逗号（,）分隔。	`predictions,probabilities,logits`

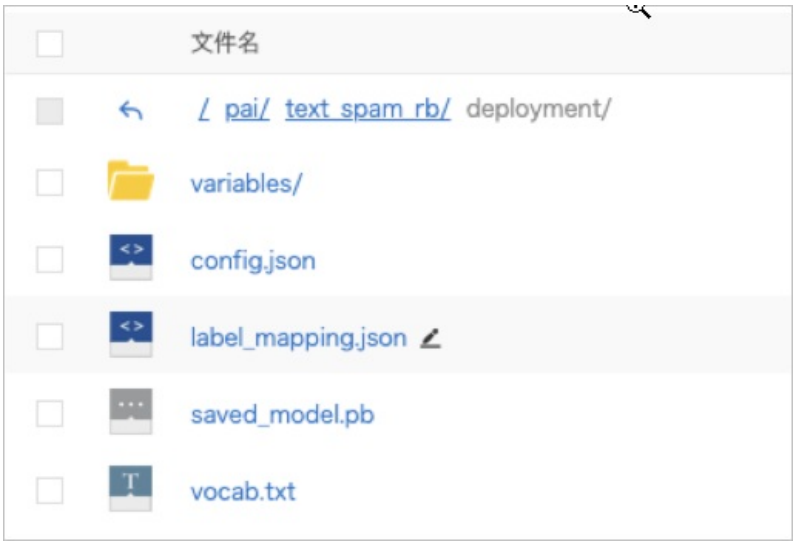
页签	参数	描述	本案例使用的示例值
参数设置	**预测概率截断**	进行预测判断时的概率阈值。	无需填写
	附加列	输入表中需要添加到输出表的列，多个列之间使用半角逗号（,）分隔。	`content, qince_result`
	batch Size	预测过程中的批处理大小。	32
	使用自定义模型	如果没有上游组件，可以自定义一个OSS目录中训练好的模型进行预测。本案例使用上游组件输入训练模型，因此无需填写。	否
执行调优	**指定Worker数**	分布式服务器的数量，默认值表示1个Worker。	1
	指定Worker的CPU卡数	每个Worker下的GPU卡数量。	1
	指定Worker的GPU卡数	每个Worker下的CPU卡数量。	1

部署及调用模型服务

通过模型在线服务PAI-EAS，您可以将训练好的内容风控模型部署为在线服务，并在实际的生产环境中调用，从而进行推理实践。

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。
2. 部署模型服务。
   i. 在**PAI EAS模型在线服务**页面，单击**模型上传部署**。

ii. 在**资源和模型**面板，配置参数（此处仅介绍与本案例相关的核心参数配置方法，其他参数的解释请参见控制台上传部署），并单击**下一步**。

参数	描述
自定义模型名称	模型的名称，建议结合实际业务进行命名，以区分不同的模型服务。
资源组种类	建议使用专属资源组部署模型服务，从而避免公共资源组资源有限时导致的服务排队。关于如何创建专属资源组，请参见专属资源组。
资源种类	如果您选择**资源组种类**为**公共资源组**，则需要选择**资源种类**为GPU。
Processor种类	选择**EasyNLP**。
模型类型	选择**BERT文本分类**。
模型文件	本案例中训练好的模型均存储在OSS中，因此选择**OSS文件导入**。 OSS路径选择文本分类组件**模型存储路径**参数配置的路径下的deployment文件夹中的模型文件夹（其中包含variables文件夹、config.json文件夹、saved_model.pb、vocab.txt及label_mapping.json）。例如，对于下图中的模型目录结构，您需要将OSS路径选择到*/deployment/*目录。 

iii. 在**部署详情及配置确认**面板，配置模型服务占用资源的相关参数。

参数	描述
实例数	本案例配置为1。实际应用时建议配置多个服务实例，以避免单点部署带来的风险。
卡数	本案例配置为1。实际应用时根据情况配置。
核数	本案例配置为1。实际应用时根据情况配置。
内存数（M）	本案例配置为4096 MB。实际应用时根据情况配置。

iv. 单击**部署**，等待一段时间即可完成模型部署。

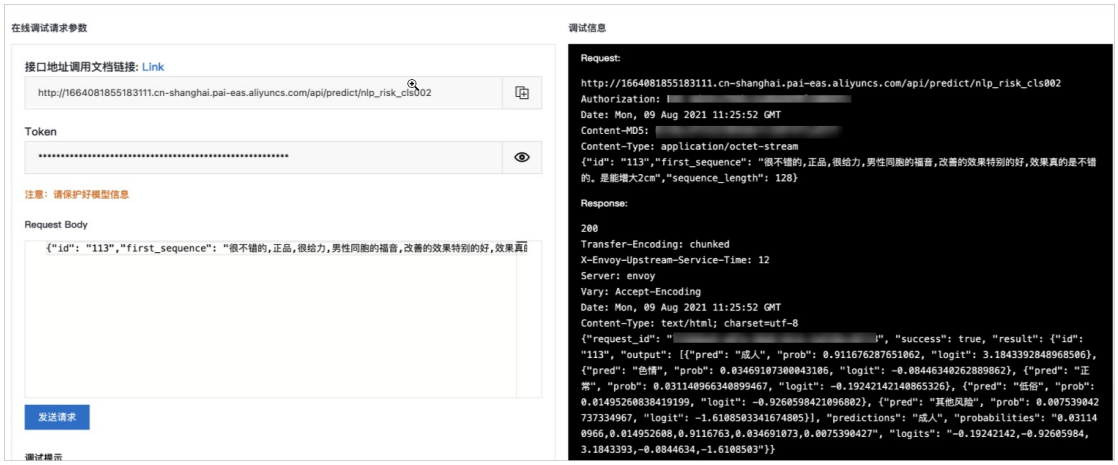
3. 调试模型服务。

i. 在**PAI EAS模型在线服务**页面，单击目标服务**操作**列下的**在线调试**。

ii. 在调试页面的**在线调试请求参数**区域的Request Body处填写如下内容。

```
{"id": "113","first_sequence": "很不错的,正品,很给力,男性同胞的福音,改善的效果特别的好,效果真的是不错的。是能增大2cm","sequence_length": 128}
```

iii. 单击**发送请求**，即可在**调试信息**区域查看预测结果，如下图所示。



4. 查看模型服务的公网地址和访问Token。

i. 在**PAI EAS模型在线服务**页面，单击目标服务**服务方式**列下的**调用信息**。

ii. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的访问地址和Token。

5. 使用脚本进行批量调用。

i. 创建调用模型服务的Python脚本eas_nlp_risk.py。

```
#!/usr/bin/env python
#encoding=utf-8
from eas_prediction import PredictClient
from eas_prediction import StringRequest
if __name__ == '__main__':
    #下面的client = PredictClient()入参需要替换为实际的访问地址。
    client = PredictClient('http://1664xxxxxxxxxx.cn-hangzhou.pai-eas.aliyuncs.com', 'nlp_risk_cls002')
    #下面的Token需要替换为实际值。
    client.set_token('MTgxNjE1NGVmMDdjNDRkY2Q5NWE4xxxxxxxxxxxxxxxxxxxxxxxx')
    client.init()
    #输入请求需要根据模型进行构造，此处仅以字符串作为输入输出的程序为例。
    request = StringRequest('[{"id": "110","first_sequence": "想赢勇士想到 发疯？格林新发现吓呆众人","sequence_length": 128},{"id": "112","first_sequence": "骗人的,千万别买,谁买谁后悔?商家就是欺诈。垃圾商家。买了之后想退货门都没有,以各种手段不退货。买者慎重。","sequence_length": 128},{"id": "113","first_sequence": "很不错的,正品,很给力,男性同胞的福音,改善的效果特别的好,效果真的是不错的。是能增大2cm","sequence_length": 128}]')
    for x in range(0, 50000):
        resp = client.predict(request)
        # print(str(resp.response_data, 'utf8'))
print("test endding")
```

ii. 将eas_nlp_risk.pyPython脚本上传至您的任意环境，并在脚本上传后的当前目录执行如下调用命令。

```
python3 <eas_nlp_ris.py>
```

其中<eas_nlp_ris.py>需要替换为实际的Python脚本名称。

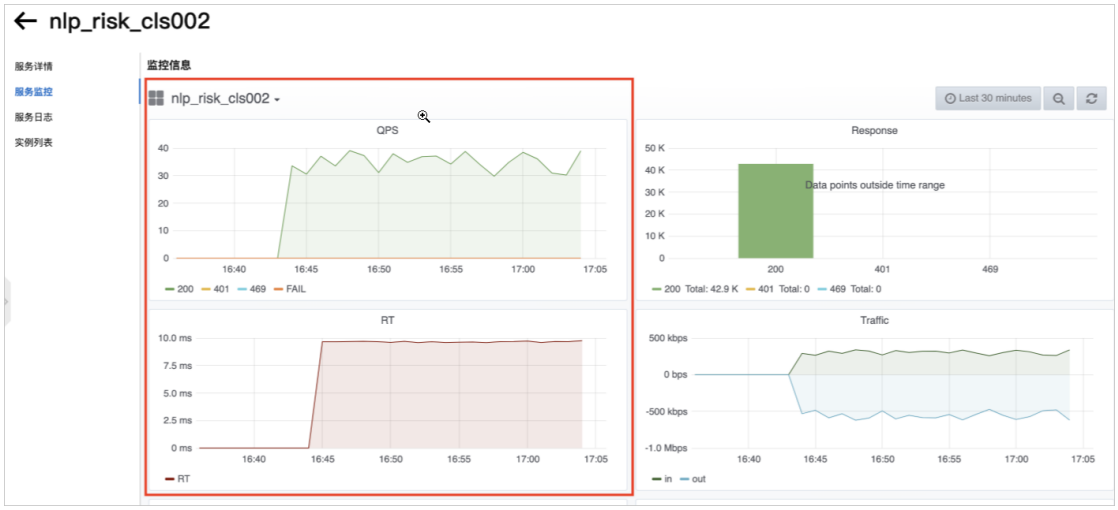
6. 监控服务指标。

调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

i. 在**PAI EAS模型在线服务**页面，单击已调用服务**服务监控**列下的图标。

ii. 在**服务监控**页签，即可查看模型调用的指标水位。

从下面的服务监控水位图中可以看到，本案例部署的模型时延在150 ms左右。您自己的模型时延以实际为准。



5.通用文本打标解决方案

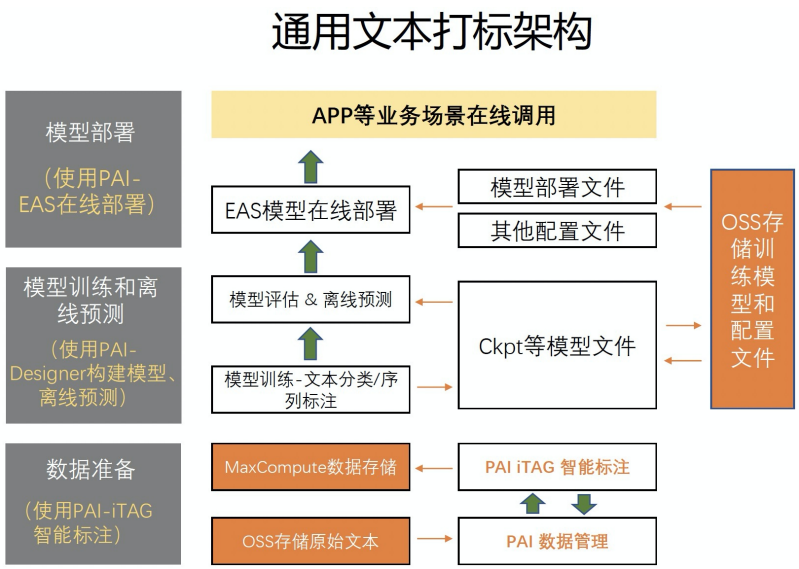
随着UGC等用户生成内容不断涌现，对从文本提取标签用于细粒度分析的需求不断涌现，PAI提供了通用文本打标解决方案。本文为您介绍如何基于人工智能算法，快速构建符合业务场景的文本打标模型和算法，助力您快速从文本中识别各种类型的文本标签，服务于下游推荐或其他应用场景。

背景信息

诸多生产内容的场景（例如评论、博客、商品介绍）中都包含大量具有丰富语义的标签，对理解文本语义、精确建模用户兴趣和需求有重要作用。针对该问题，阿里云机器学习PAI借助人工智能算法，提出了如下解决方案，帮助您快速抽取文本蕴含的标签。

- 解决方案
  1. 基于智能标注（iTAG）平台和PAI数据集管理，对目标场景的文本进行快捷标注和样本管理。
  2. 基于PAI提供的文本打标模型和算法，针对自己的文本打标场景，在可视化建模平台PAI-Designer上进行模型Fine-Tune，从而构建具体场景的NLP文本打标模型。
  3. 基于PAI提供，或者用户自行Fine-Tune的文本打标模型，在可视化建模平台PAI-Designer上进行批量离线推理。
  4. 将模型进行PAI-EAS在线部署，形成完整的端到端流程，从而自动抽取文本中的各种标签。
- 方案架构

通用文本打标解决方案的架构图如下所示。



前提条件

- 已开通PAI（Designer、DSW、EAS）后付费，详情请参见开通。
- 已开通AI工作空间，并添加MaxCompute计算资源，详情请参见创建工作空间。
- 已开通MaxCompute，用于存储预测数据。关于如何开通MaxCompute，请参见通过查询编辑器使用MaxCompute。
- 已创建OSS存储空间（Bucket），用于存储原始数据、标签列表文件及训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创建资源组。

操作流程

基于阿里云机器学习PAI平台，构建文本打标解决方案的流程如下：

1. 准备数据

基于智能标注（iTAG）进行原始数据标注，然后将获得的训练集和测试集上传到数据仓库MaxCompute中，用于后续的模型训练。

2. 构建文本NER模型

在可视化建模平台PAI-Designer中，基于海量大数据语料预训练获得的NLP预训练模型，构建文本NER模候检构建文本NER模型。

3. 构建文本细粒度标签分类模型

在可视化建模平台PAI-Designer中，基于海量大数据语料预训练获得的NLP预训练模型，构建文本细粒度标签分类模型。

4. 离线批量预测

在可视化建模平台PAI-Designer中，对海量用户产生的文本数据进行离线批量文本打标。

5. 部署及调用模型服务

通过模型在线服务PAI-EAS，您可以将训练好的文本打标模型部署为在线服务，并在实际的生产环境中调用，从而进行推理。

准备数据

首先基于智能标注（iTAG）进行原始数据标注，然后将获得的训练集和测试集上传到数据仓库MaxCompute中，用于后续的模型训练。在文本打标解决方案中，支持训练文本NER模型和文本细粒度标签分类模型，您可以根据实际需求，训练需要的模型，得到定制化的文本打标结果，您也可以使用PAI默认提供的模型直接进行预测，在这种情况下，您无需准备数据。

根据您的定制化需求，需要准备的数据如下：

定制模块	需要准备的数据	数据格式示例
关键词抽取-关键词	自定义的词表	user_dict.txt
NER-人名、地名、机构等	文本NER模型的训练集和测试集	ner_train.csv ner_dev.csv
文本细粒度分类-文本主题	文本细粒度分类模型的训练集和测试集	cate_train.csv cate_dev.csv
	自定义文本标签体系	sample_category.json

数据准备的步骤如下：

1. 根据格式要求，准备数据。

i. 如果您需要在关键词抽取模块中使用自定义的词表，则需要进行数据准备，并将数据存放在OSS目录下。

示例请参见user_dict.txt，其中，文件的每一行都是一个关键词。

ii. 如果您需要定制NER或文本细粒度分类模型，需要准备对应的训练集和测试集。

iii. 将原始数据分为训练集和测试集，分别按照iTAG的标注数据格式，生成.manifest格式的数据集。文件格式请参见textDemo.manifest。

2. 创建数据集。

在PAI的数据集管理模块，通过本地上传方式将准备好的.manifest数据集上传，详情请参见创建数据集。

3. 创建打标任务并完成打标任务处理。

i. 通过智能标注（iTAG）管理员控制台，创建标注任务，详情请参见创建标注任务。

ii. 通过智能标注（iTAG）标注员控制台，进行数据标注，详情请参见标注图像（打标，质检及验收）。

4. 获取并处理打标结果。

i. 在**智能标注（iTAG）**页面的**任务中心**页签，单击标注完成的任务操作列下的**获取标注结果**，得到存放在指定OSS目录下的标注结果数据集。

ii. 将获得的标注结果数据集转换为模型构建需要的训练集和测试集的CSV格式数据表。

a. 文本NER模型的示例训练集请参见ner_train.csv，示例测试集请参见ner_dev.csv。

b. 文本细粒度标签分类模型的示例训练集请参见cate_train.csv，示例测试集请参见cate_dev.csv

> **说明**
>
> 您只需要标注二级类目即可。

5. 通过MaxCompute客户端的Tunnel命令将训练数据集和测试数据集上传至MaxCompute。

关于MaxCompute客户端的安装及配置请参见MaxCompute客户端（odpscmd），关于Tunnel命令详情请参见Tunnel命令。

本案例使用的示例命令如下所示。

```
# NER建表语句。
CREATE TABLE nlp_ner_train(content STRING, label STRING);
CREATE TABLE nlp_ner_dev(content STRING, label STRING);

# NER上传语句。
tunnel upload /Users/xxx/xxx/ner_train.csv nlp_ner_train -fd '\t';
tunnel upload /Users/xxx/xxx/ner_dev.csv nlp_ner_dev -fd '\t';

# 细粒度标签分类建表语句。
CREATE TABLE nlp_cate_train(content STRING, label STRING);
CREATE TABLE nlp_cate_dev(content STRING, label STRING);

# 细粒度标签分类上传语句。
tunnel upload /Users/xxx/xxx/cate_train.csv nlp_cate_train -fd '\t';
tunnel upload /Users/xxx/xxx/cate_dev.csv nlp_cate_dev -fd '\t';
```

6. 如果您需要训练文本细粒度标签分类模型，您还需要准备对应的文本标签体系sample_category.json，并存放在OSS目录下。

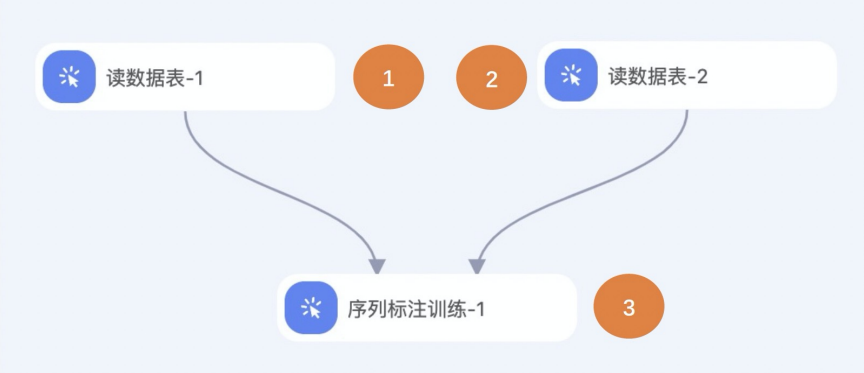
   文本标签体系由两层概念类目组成，分别为一级类目和二级类目，一级类目更加抽象，二级类目更加具体。这一文本标签体系组织成JSON格式，如下所示。

```
{
    "翻唱": "音乐",
    "手办": "二次元",
    "火锅": "美食"
}
```

构建文本NER模型

1. 登录PAI控制台。
2. 在Designer页面创建一个空白工作流，分别拖入以下组件，并根据下文的组件参数配置组件。

   创建工作流即组件配置的操作请参见创建及管理工作流。



区域	描述
①	配置实验的训练数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的训练表。例如nlp_ner_train。
②	配置实验的评估数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的评估表。例如nlp_ner_dev。
③	配置序列标注模型训练的参数。**序列标注训练**组件的配置详情请参见下文的表 1。

表1. 序列标注组件的配置

页签	参数	描述	本案例使用的示例值
	文本列选择	进行序列标注的内容在输入表中对应的列名。	content

页签	参数	描述	本案例使用的示例值
字段设置	标签列选择	标签在输入表中对应的列名。	label
	标签枚举值	标签列内容的枚举值，多个值之间使用半角逗号（,）分隔。	B-LOC,B-ORG,B-PER,I-LOC,I-ORG,I-PER,O
	样本权重列	特定列样本的数据增强，非必选。	无需填写
	模型存储路径	训练得到的模型的OSS存储路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/ner_model/
参数设置	优化器选择	优化器的类型，支持以下取值： ○ adam ○ adagrad ○ lamb	adam
	batchSize	训练过程中的批处理大小。如果使用多机多卡，则表示每个GPU上的批处理大小。	32
	sequenceLength	序列整体最大长度，取值范围为1~512。	64
	numEpochs	训练总Epoch的数量。	1
	学习率	模型构建过程中的学习率。	1e-5
	pretrainModelNameOrPath	选择使用的预训练语言模型。	pai-bert-base-zh
	模型额外参数	用户自定义参数，您无需填写。	
	指定Worker数	分布式服务器的数量，默认值表示1个Worker。	1

页签	参数	描述	本案例使用的示例值
执行调优			
	指定Worker的CPU卡数	每个Worker下的GPU卡数量。	1
	指定Worker的GPU卡数	每个Worker下的CPU卡数量。	1
	分布式策略	分布式策略，支持以下取值： ○ **MirroredStrategy**：单机多卡。 ○ **ExascaleStrategy**：多机多卡。	MirroredStrategy

构建文本细粒度标签分类模型

1. 登录PAI控制台。
2. 在Designer页面创建一个空白工作流，分别拖入以下组件，并根据下文的组件参数配置组件。

   创建工作流即组件配置的操作请参见创建及管理工作流。

区域	描述
④	配置实验的训练数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的训练表。例如nlp_cate_train。
⑤	配置实验的评估数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的评估表。例如nlp_cate_dev。
⑥	配置文本分类模型训练的参数。**文本分类训练**组件的配置详情请参见下文的表 2。

表2. 文本分类组件的配置

页签	参数	描述	本案例使用的示例值
字段设置	文本列选择	进行文本分类的内容在输入表中对应的列名。	content
	标签列选择	标签在输入表中对应的列名。	label
	标签枚举值	标签列内容的枚举值，多个值之间使用半角逗号（,）分隔。	翻唱,手办,火锅
	样本权重列	特定列样本的数据增强，非必选。	无需填写
	模型存储路径	训练得到的模型的OSS存储路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/cate_model/
参数设置	优化器选择	优化器的类型，支持以下取值： ○ adam ○ adagrad ○ lamb	adam
	batchSize	训练过程中的批处理大小。如果使用多机多卡，则表示每个GPU上的批处理大小。	32
	sequenceLength	序列整体最大长度，取值范围为1~512。	64
	numEpochs	训练总Epoch的数量。	1
	学习率	模型构建过程中的学习率。	1e-5
	pretrainModelNameOrPath	选择使用的预训练语言模型。	pai-bert-base-zh

	模型额外参数	用户自定义参数，可以配置是否训练多标签分类模型multi_label， 支持以下取值： ○ True ○ False 在本解决方案中，必须填写为True。	multi_label=True
执行调优	指定Worker数	分布式服务器的数量，默认值表示1个Worker。	1
	指定Worker的CPU卡数	每个Worker下的GPU卡数量。	1
	指定Worker的GPU卡数	每个Worker下的CPU卡数量。	1
	分布式策略	分布式策略，支持以下取值： ○ MirroredStrategy：单机多卡。 ○ ExascaleStrategy：多机多卡。	MirroredStrategy

离线批量预测

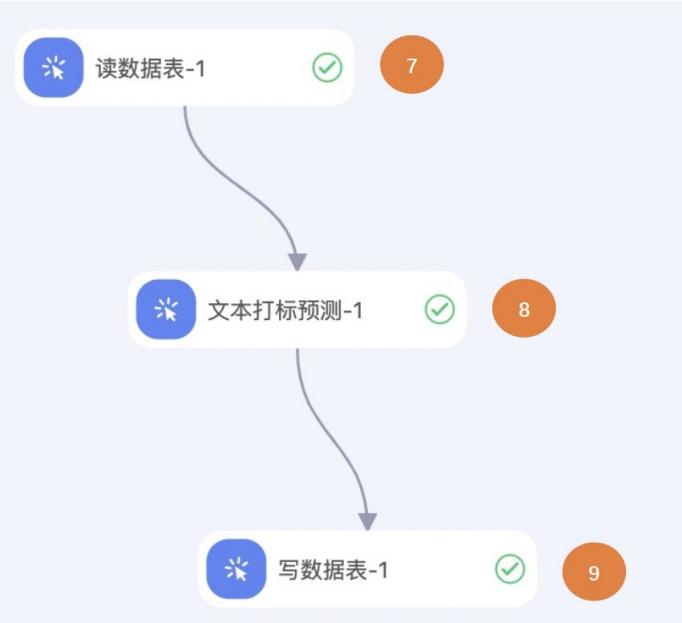
1. 准备并上传待打标文件。
   i. 将需要打标的文本转换为CSV格式数据表，示例测试集请参见test.csv。
   ii. 通过MaxCompute客户端的Tunnel命令将训练数据集和测试数据集上传至MaxCompute。

   关于MaxCompute客户端的安装及配置请参见MaxCompute客户端（odpscmd），关于Tunnel命令详情请参见Tunnel命令。本案例使用的示例命令如下所示。

```
# 建表语句。
CREATE TABLE nlp_text_test(content STRING);
# 上传语句。
tunnel upload /Users/xxx/xxx/test.csv nlp_text_test;
```

2. 在Designer页面创建一个空白工作流，分别拖入以下组件，并根据下文的组件参数配置组件。

创建工作流即组件配置的操作请参见创建及管理工作流。



区域	描述
⑦	配置实验的测试数据集，即配置**读数据表**组件的**表名**参数为MaxCompute的训练表。例如**nlp_text_test**。
⑧	配置文本打标预测的参数。**文本打标预测**组件的配置详情请参见下文的表 3。
⑨	配置实验的测试输出数据表，即配置**写数据表**组件的**表名**参数为MaxCompute的输出表。例如**nlp_text_test_out**。

表3. 文本打标预测组件的配置

页签	参数	描述	本案例使用的示例值
	Buckets	填写用户模型和其他配置文件所在的OSS Buckets。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/
	文本列序号	填写需要打标的文本在输入表的列序号（从0开始）。	0

参数设置	**是否执行默认实体识别**	是否使用PAI默认的模型进行实体识别，支持以下取值： ○ **True** ○ **False**	True
	是否执行关键词抽取	是否开启关键词抽取功能，支持以下取值： ○ **True** ○ **False**	True
	是否执行默认细粒度标签预测	是否使用PAI默认的模型进行细粒度标签预测，支持以下取值： ○ **True** ○ **False**	True
	用户自定义NER模型路径	用户自定义NER模型的OSS路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/ner_model/deployment/
	用户自定义关键词字典	用户自定义关键词字典的OSS路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/cate_model/deployment/user_dict.txt
	关键词抽取参数Alpha	关键词抽取参数Alpha取值在0-1之间，用于平衡TFIDF和TextRank这两个关键词抽取算法的权重。	0.5
	抽取关键词数	算法抽取关键词数量上限。	10

	用户自定义标签预测模型路径	用户自定义标签预测模型路径的OSS路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/cate_model/deployment/
	用户自定义标签配置	用户自定义标签配置的OSS路径。	oss://easynlp-sh.oss-cn-shanghai-internal.aliyuncs.com/text_mark/cate_model/deployment/sample_category.json
	标签预测阈值	标签预测阈值取值在0-1之间，当某标签预测概率大于该阈值则进行输出。	0.5
	是否输出根节点	是否输出预测的二级类目对应的一级类目， 支持以下取值： ○ **True** ○ **False**	False
执行调优	**Worker数量**	分布式服务器的数量，默认值表示1个Worker。	1
	Batch Size	每个Batch的大小。	16
	Sequence Length	文本序列长度。	128

3. 预测完成后查看输出表。输出表的Schema比输入表增加以下列。
   - keyword：如果选择执行关键词抽取，在此列中显示关键词抽取结果。
   - default_ner：如果选择执行默认实体识别，在此列中显示使用PAI默认的NER模型识别实体的结果。
   - user_ner：如果填写用户自定义NER模型路径，在此列中显示使用自定义NER模型识别实体的结果。
   - default_text_label：如果选择执行默认细粒度标签预测，在此列中显示使用PAI默认的细粒度标签预测模型的预测结果。

- user_text_label：如果填写用户自定义标签预测模型路径和用户自定义标签配置，在此列中显示使用自定义的细粒度标签预测模型的预测结果。

部署及调用模型服务

通过模型在线服务PAI-EAS，您可以将训练好的文本打标模型部署为在线服务，并在实际的生产环境中调用，从而进行推理实践。

- 首先将训练后的模型及其他相关配置文件打包，并配置在线服务参数。

  如果您希望使用PAI默认的模型进行部署，请跳过此步。以下介绍部署用户训练的模型的打包方式。

  首先将训练得到的模型，及其他配置文件打包为tar.gz格式，并且上传至用户的OSS Bucket。模型和配置的目录结构如下所示。

```
sentence_mark
|-- sentence_mark_main.py
|-- ner_model
| |-- variables
| | |-- variables.data-00000-of-00001
| | `-- variables.index
| |-- config.json
| |-- label_mapping.json
| |-- saved_model.pb
| `-- vocab.txt
|-- cate_model
| |-- variables
| | |-- variables.data-00000-of-00001
| | `-- variables.index
| |-- config.json
| |-- label_mapping.json
| |-- saved_model.pb
| |-- mlcls_mapping.json
| `-- vocab.txt
`-- keywords_dict.txt
```

  其中：

  - ner_model：为用户训练的自定义NER模型Saved Model的文件夹。
  - cate_model：为用户训练的自定义细粒度标签分类模型Saved Model的文件夹。
  - keywords_dict.txt：为用户自定义关键词词典。
  - mlcls_mapping.json：为用户自定义标签体系。
  - sentence_mark_main.py：为文本打标在线推理配置文件。

  以下给出sentence_mark_main.py文件的示例。用户需要修改的参数及其配置说明见下文的表4。

```
import media_ops as ops
from mediaflow import GraphBuilder
from mediaflow.data import MediaData
from mediaflow.config import window_policy


USE_KEYWORD = True
USE_NER = True
USE_TEXTLABEL = True
```

```
class MyGraphBuilder(GraphBuilder):
 def run_option(self):
  return {
   "enable_share_memory": False,
   "share_memory_size": 40000,
   "worker_threads": 128,
   "output_from_context": True,
   "rpc_keepalive": 1000000,
   "enable_numpy": True,
   "metric_print": True
  }

 def build_graph(self):
  keyword_config = {
   'user_dict':'/home/admin/docker_ml/workspace/model/sentence_mark/keywords_dict.txt',
   'text_source':'input',
            'top_k': '10',
            'alpha':'0.5'
  }
  ner_config ={
   'saved_model_path' : '/home/admin/docker_ml/workspace/model/sentence_mark/ner_model/',
   'seq_length' : '256',
   'text_source' : 'input'
  }
  text_label_config ={
   'two_level_mapping_path' : '/home/admin/docker_ml/workspace/model/sentence_mark/cate_m
odel/mlcls_mapping.json',
   'saved_model_path' : '/home/admin/docker_ml/workspace/model/sentence_mark/cate_model/'
,
   'seq_length': '128',
   'threshold': '0.5',
   'append_root': '0',
   'text_source':'input'
  }

  if USE_KEYWORD and USE_NER and USE_TEXTLABEL:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.common_op.keyword
_extract, keyword_config).\
    window(window_policy.select_all_window(), ops.nlp.text_classify.ner_predict, ner_conf
ig).\
     window(window_policy.select_all_window(), ops.nlp.text_classify.text_label_recogniti
on, text_label_config)
  elif USE_KEYWORD and USE_NER:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.common_op.keyword
_extract, keyword_config).\
    window(window_policy.select_all_window(), ops.nlp.text_classify.ner_predict, ner_conf
ig)
  elif USE_KEYWORD and USE_TEXTLABEL:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.common_op.keyword
_extract, keyword_config).\
    window(window_policy.select_all_window(), ops.nlp.text_classify.text_label_recognitio
n, text_label_config)
```

```
  elif USE_NER and USE_TEXTLABEL:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.text_classify.ner
_predict, ner_config).\
    window(window_policy.select_all_window(), ops.nlp.text_classify.text_label_recognitio
n, text_label_config)
  elif USE_KEYWORD:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.common_op.keyword
_extract, keyword_config)
  elif USE_NER:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.text_classify.ner
_predict, ner_config)
  elif USE_TEXTLABEL:
   MediaData('INPUT').window(window_policy.select_all_window(), ops.nlp.text_classify.tex
t_label_recognition, text_label_config)


GraphBuilder.register(MyGraphBuilder)
```

表4. 文本打标在线服务参数说明

参数	描述	本案例使用的示例值
功能控制参数		
USE_KEYWORD	控制是否输出文本关键词。	True
USE_NER	控制是否输出NER结果。	True
USE_TEXTLABEL	控制是否输出文本分类细粒度标签。	True
关键词抽取参数		
user_dict	用户自定义关键词字典的本地路径。如果不需要自定义关键词字典，请删除这一字段，否则无需修改。	/home/admin/docker_ml/workspace/model/sentence_mark/keywords_dict.txt
top_k	算法抽取关键词数量上限。	10
alpha	关键词抽取参数Alpha取值在0-1之间，用于平衡TFIDF和TextRank这两个关键词抽取算法的权重。	0.5
NER参数		

saved_model_path	用户自定义NER模型的本地路径。如果不需要自定义NER，**请将USE_NER设置为False，无需修改此字段**。	/home/admin/docker_ml/workspace/model/sentence_mark/ner_model/
seq_length	文本序列长度。	128
细粒度标签分类参数		
saved_model_path	用户自定义文本细粒度标签模型的本地路径。如果不需要自定义文本细粒度标签，**请将USE_TEXTLABEL设置为False，无需修改此字段**。	/home/admin/docker_ml/workspace/model/sentence_mark/cate_model/
two_level_mapping_path	用户自定义文本细粒度标签配置的本地路径。如果不需要自定义文本细粒度标签，**请将USE_TEXTLABEL设置为False，无需修改此字段。**	/home/admin/docker_ml/workspace/model/sentence_mark/cate_model/mlcls_mapping.json
threshold	标签预测阈值取值在0-1之间，当某标签预测概率大于该阈值则进行输出。	0.5
append_root	是否输出预测的二级类目对应的一级类目。	0
seq_length	文本序列长度。	128

2. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏，选择**模型部署 > 模型在线服务（EAS）**。
3. 部署模型服务。
   i. 在**PAI EAS模型在线服务**页面，单击**模型上传部署**。
   ii. 在**资源和模型**面板，填写以下参数，并单击**下一步**。

iii. 如果您希望部署自行训练的模型，在model_path处填写第1步打包的文件在OSS Bucket的路径。

```
{
    "cloud": {
        "computing": {
            "instance_type": "ecs.gn6i-c4g1.xlarge"
        }
    },
    "data_image": "registry.cn-shanghai.aliyuncs.com/eas/mediaflow:py36-0.3.3",
    "generate_token": "true",
    "metadata": {
        "cpu": 4,
        "cuda": "9.0",
        "gpu": 1,
        "instance": 1,
        "memory": 15000,
        "name": "sentence_mark",
        "rpc": {
            "enable_jemalloc": true,
            "enable_service_hang_detect": true,
            "keepalive": 500000,
            "max_batch_size": 64,
            "max_batch_timeout": 500,
            "service_hang_detect_period": 5,
            "worker_threads": 2
        }
    },
    "model_entry": "sentence_mark/sentence_mark_main.py",
    "model_path": "oss://xxx/xxx/sentence_mark.tar.gz",
    "name": "sentence_mark",
    "processor": "mediaflow_py3",
    "requirements": "http://pai-vision-data-hz.oss-cn-zhangjiakou.aliyuncs.com/VIP/
VideoOp/release/202201_offline/requirements.txt"
}
```

iv. 单击**部署**，等待一段时间即可完成模型部署。

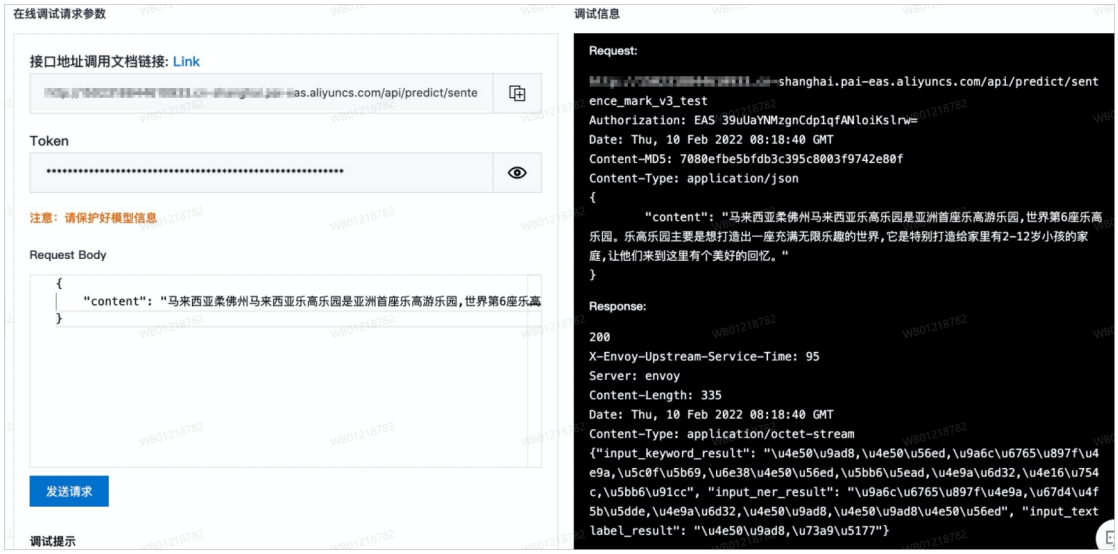
4. 调试模型服务。

i. 在**PAI EAS模型在线服务**页面，单击目标服务**操作**列下的**在线调试**。

ii. 在调试页面的**在线调试请求参数**区域的Request Body处填写如下内容。

```
{
          "content": "马来西亚柔佛州马来西亚乐高乐园是亚洲首座乐高游乐园，世界第6座乐高乐园。
乐高乐园主要是想打造出一座充满无限乐趣的世界，它是特别打造给家里有2-12岁小孩的家庭，让他们来到这
里有个美好的回忆。"
}
```

iii. 单击**发送请求**，即可在**调试信息**区域查看预测结果，如下图所示。



5. 查看模型服务的公网地址和访问Token。
   i. 在**PAI EAS模型在线服务**页面，单击目标服务**服务方式**列下的**调用信息**。
   ii. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的**访问地址**和**Token**。
6. 使用脚本进行批量调用。
   i. 创建调用模型服务的Python脚本**eas_sentence_mark.py**。

```
#!/usr/bin/env python
#encoding=utf-8
from eas_prediction import PredictClient
from eas_prediction import StringRequest
if __name__ == '__main__':
    #下面的client = PredictClient()入参需要替换为实际的访问地址。
    client = PredictClient('http://1664xxxxxxxxxx.cn-hangzhou.pai-eas.aliyuncs.com', 'sentence_mark')
    #下面的Token需要替换为实际值。
    client.set_token('MTgxNjE1NGVmMDdjNDRkY2Q5NWE4xxxxxxxxxxxxxxxxxxxxxxxx')
    client.init()
    #输入请求需要根据模型进行构造，此处仅以字符串作为输入输出的程序为例。
    request = StringRequest('{"content": "马来西亚柔佛州马来西亚乐高乐园是亚洲首座乐高游乐园，世界第6座乐高乐园。乐高乐园主要是想打造出一座充满无限乐趣的世界，它是特别打造给家里有2-12岁小孩的家庭，让他们来到这里有个美好的回忆。"}')
    for x in range(0, 50000):
        resp = client.predict(request)
        # print(str(resp.response_data, 'utf8'))
print("test ending")
```

   ii. 将**eas_sentence_mark.py**的Python脚本上传至您的任意环境，并在脚本上传后的当前目录执行如下调用命令。

```
python3 <eas_sentence_mark.py>
```

其中**<eas_sentence_mark.py>**需要替换为实际的Python脚本名称。

7. 监控服务指标。调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

   i. 在**PAI EAS模型在线服务**页面，单击已调用服务**服务监控**列下的图标。

   ii. 在**服务监控**页签，即可查看模型调用的指标水位。从下面的服务监控水位图中可以看到，本案例部署的模型时延在100ms左右。您自己的模型时延以实际为准。

6.相似图像匹配与图像检索解决方案

针对图像检索业务场景，PAI提供了端到端的相似图像匹配和图像检索解决方案。本文介绍如何基于未标注的数据构建图像自监督模型，助力您快速搭建相似图像匹配和图像检索业务系统，进而实现以图搜图。

背景信息

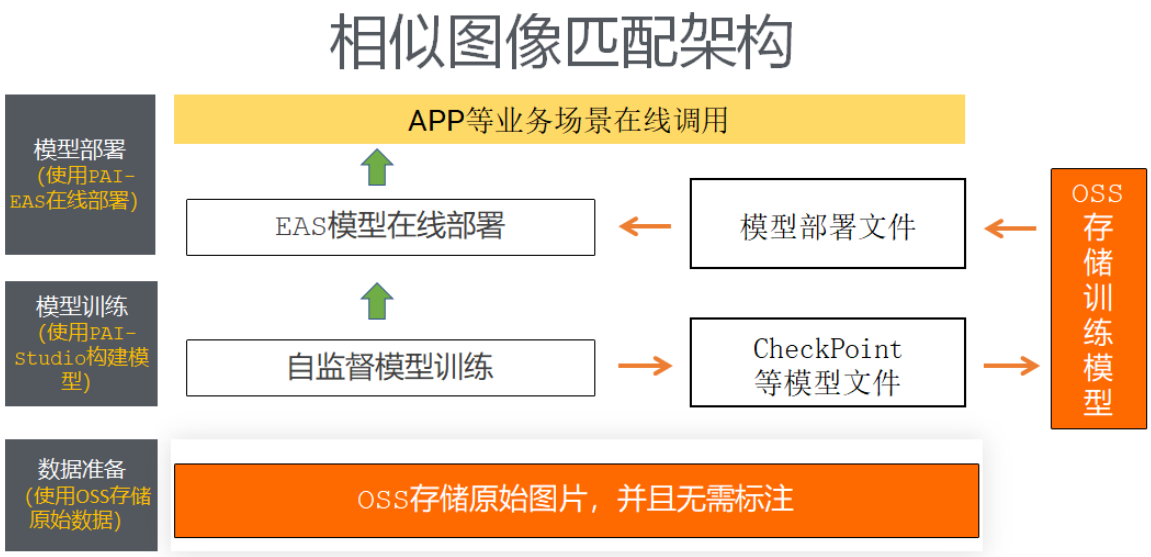
针对图像检索的业务场景（例如电商业务中需要根据已有图像搜索目标商品），PAI提供了如下解决方案，帮助您快速搭建相似图像匹配和图像检索业务系统：

- 解决方案

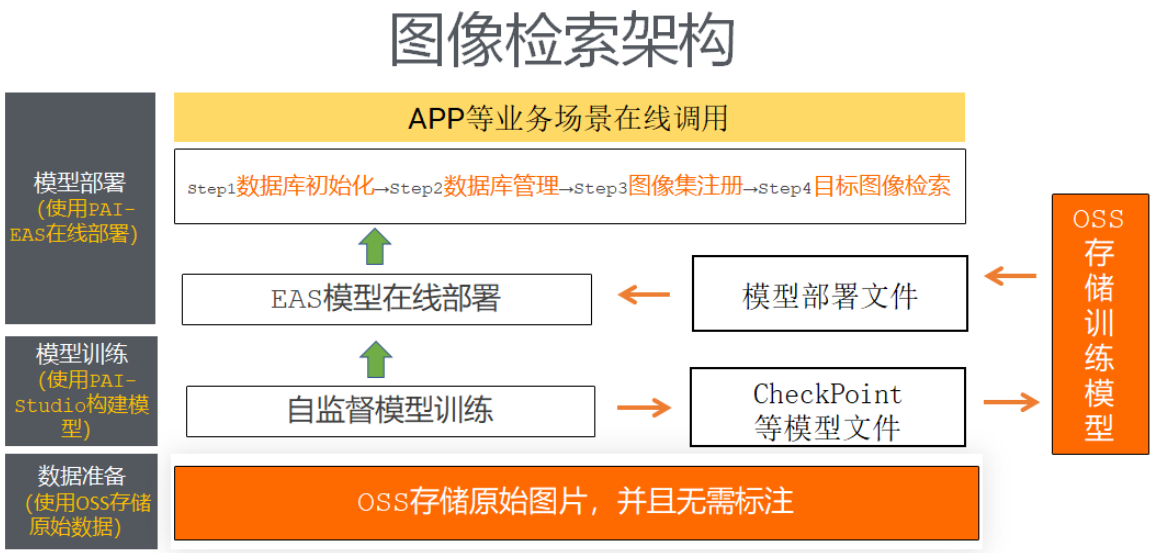
  PAI在相似图像匹配和图像检索领域提供了端到端、轻量化的纯白盒解决方案。您只需要准备原始的图像数据，无需标注就能够构建模型。然后利用PAI的可视化建模平台快速自定义构建图像自监督模型。最后将模型在PAI上进行部署推理，形成完整的端到端流程，从而实现相似图像匹配和图像检索的业务系统。

- 方案架构

  相似图像匹配解决方案的架构图如下所示。



图像检索解决方案的架构图如下所示。



- 方案优势
  - 无需标注：图像无需标注，使用原始数据即可建模，从而节省大量人力成本。
  - 纯白盒：可以根据自己具体的业务场景，纯自定义构建模型。
  - 端到端：从最初的数据准备到最后的模型推理，提供全链路系统构建流程。
  - 轻量级：无论是工程师还是初级用户，都能快速搭建图像匹配和检索系统。

前提条件

- 已开通**PAI（Studio、DSW、EAS）后付费**，详情请参见开通。
- 已开通AI工作空间，并添加MaxCompute计算资源和DLC计算资源，详情请参见AI工作空间（旧版）。
- 已创建OSS存储空间（Bucket），用于存储原始图像和训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创建专属资源组。
- 已完成DLC使用权限授权，授权方法详情请参见云产品依赖与授权：DLC。

操作流程

基于阿里云机器学习PAI平台，构建相似图像匹配与图像检索解决方案的流程如下：

1. 准备数据

   本文使用的算法为图像自监督算法，训练过程中无需标注数据，您将原始数据存入OSS即可直接进行模型训练。此外，如果使用图像检索解决方案，则需要在本地准备检索图的图像数据库。从而在图像检索模型的部署过程中，将准备好的本地数据注册到数据库中。

   PAI提供了原始数据集，您可以直接使用它进行数据准备。关于数据集的下载方式，请参见Deepfashion2图像数据集。

2. 构建图像自监督模型

   利用可视化建模PAI-Studio平台，基于自己特定的业务场景，采用自监督的图像深度学习训练组件，将原始的尚未标注的图像直接进行训练。对于相似图像匹配场景和图像检索场景，您都可以使用该自监督组件进行模型训练，两种场景在模型训练部分无差别。

3. 部署及调用模型服务

通过模型在线服务PAI-EAS，您可以将训练好的图像自监督模型部署为在线服务，并在实际的生产环境中的相似图像匹配和图像检索两个场景下进行推理实践，详情请参见部署及调用相似图像匹配的模型服务和部署及调用图像检索的模型服务。

准备数据

PAI提供了原始数据集，您可以直接使用它进行数据准备。关于数据集的下载方式，请参见Deepfashion2图像数据集。

1. 准备模型训练所需的数据。
   i. 将原始图片上传至OSS的某一目录中。关于如何将文件上传至OSS，请参见上传文件。
   ii. 根据图像上传的OSS文件目录，生成OSS图像目标索引的TXT文件。索引文件的示例请参见demo_img_list.txt。
   iii. 将生成好的TXT索引文件存储到OSS的某一目录下。
2. 如果您使用图像检索解决方案，则需要准备图像检索数据库。

   部署图像检索模型时，您需要准备检索的图像数据库，并对注册到数据库中的图像进行特征提取，从而在目标图像的推理过程中实现在已存在图像数据库中对上传的图像进行相似的快速检索。因此，您需要在本地准备图像检索数据库中的图像数据，为后续模型部署环节的注册数据库做准备。

构建图像自监督模型

1. 登录PAI控制台。
2. 在左侧导航栏，选择**模型开发和训练 > 可视化建模（Studio2.0）**。
3. 基于实验模板，创建相似图像匹配与图像检索实验。
   i. 在**可视化建模（Studio）**页面，单击**实验模板**页签。
   ii. 单击CV页签。
   iii. 在**相似图像匹配与图像检索**区域，单击**创建**。

   iv. 在**新建模板**对话框，先输入实验**名称**和**描述**、选择绑定的**AI工作空间**和实验存放的**位置**，再单击**确定**即可创建实验。
4. 进入实验，并配置组件参数。
   i. 在**可视化建模（Studio）**页面，单击**实验列表**页签。
   ii. 在**实验列表**页面，选中刚才创建好的模板实验，并单击**进入实验**。
   iii. 系统根据预置的模板，自动构建实验，如下图所示。

区域	描述
①	配置实验的数据集，即配置**读OSS数据**组件的**OSS数据路径**参数为准备的图像目标索引TXT文件的OSS目录。例如 `oss://demo-zhoulou.oss-cn-hangzhou.aliyuncs.com/demo_image_match/df2_data/meta/train_crop_label_lt_10k_nolabel.txt`，该数据集是PAI在华东1（杭州）准备好的数据集，您可以直接使用。
②	将初始数据集转换为图像自监督训练组件所需的训练集。**数据转tfrecord**组件的配置详情请参见下文的数据转tfrecord组件配置。
③	配置图像自监督模型训练的参数。**图像自监督训练**组件的配置详情请参见下文的图像自监督训练组件配置。

数据转tfrecord组件配置

页签	参数	描述	本案例的示例值
字段设置	**转换配置文件路径**	转化配置文件的OSS路径。在PAI-Studio中无需使用该配置文件。	无需填写
	输出tfrecord路径	组件运行成功后，系统会自动在该路径下输出训练集和测试集径。	`oss://pai-online-hangzhou.oss-cn-hangzhou-internal.aliyuncs.com/demo_image_match/df2_data/tfrecord/test_210910/`
	输出tfrecord前缀	自定义输出TFRecord文件名称的前缀。	df2_val

页签	参数	描述	本案例的示例值
参数设置	**转换数据用于何种模型训练**	**数据转tfrecord**组件的输出数据可以用于以下类型的模型训练： ■ CLASSIFICATION：图像分类或多标签分类 ■ DETECTION：物体检测 ■ SEGMENTATION：语义分割 ■ POLYGON_SEGMENTATION：多边形语义分割 ■ INSTANCE_SEGMENTATION：实体分割 ■ TEXT_END2END：端到端OCR ■ TEXT_RECOGNITION：单行文字识别 ■ TEXT_DETECTION：文字检测 ■ VIDEO_CLASSIFICATION：视频分类	CLASSIFICATION
	类别列表文件路径	类别列表文件的OSS路径。由于自监督训练组件无需标签组，因此，您上传空文件即可。例如使用示例test.config。	`oss://pai-online-hangzhou.oss-cn-hangzhou-internal.aliyuncs.com/demo_image_match/df2_data/meta/test.config`
	测试数据分割比例	测试数据分割比例。如果设置为0，则所有数据转换为训练数据。设置为0.1表示10%的数据作为验证集。	0
	图片最大边限制	如果配置了该参数，则大图片会被Resize后存入TFRecord，从而节省存储、提高数据读取速度。	无需填写
	测试图片最大边限制	同**图片最大边限制**，用于配置测试数据。	无需填写
	默认类别名称	默认类别名称。对于在类别列表文件中未找到的类别，系统将其映射到默认类别。	无需填写
	错误类别名称	含有该类别的物体和Box会被过滤，不参与训练。	无需填写

页签	参数	描述	本案例的示例值
	忽略类别名称	仅用于检测模型，含有该类别的Box会在训练中被忽略。	无需填写
	转换类名称	标注数据的来源类型，支持以下取值： ■ **PAI标注格式** ■ **亲测标注格式** ■ **自监督标注格式**	PAI标注格式
	分隔符	用于标记内容的分隔符。	无需填写
	图片编码方式	TFRecord中图片的编码方式。	jpg
执行调优	**读取并发数**	训练过程读取并发数。	10
	写tfrecord并发数	训练过程写TFRecord并发数。	1
	每个tfrecord保存图片数	训练过程每个TFRecord保存的图片数量。	1000
	worker个数	训练过程中的Worker数量。	5
	CPU Core个数	训练过程中的CPU Core数量。	800
	memory大小	训练过程中的内存大小，单位为MB。	20000

图像自监督训练组件配置

页签	参数	描述	本案例的示例值
	训练模型类型	模型训练的类型，支持以下取值： ■ **MOCO_R50** ■ **MOBY_TIMM** ■ **MOCO_TIMM** ■ **SWAV_R50** 在后续的模型推理中，**MOBY_TIMM**的特征维度为384，其余类型的特征维度皆为2048。	MOBY_TIMM

参数设置	参数	描述	本案例的示例值
	训练所用oss目录	存储训练模型的OSS目录。	oss://pai-online-hangzhou.oss-cn-hangzhou-internal.aliyuncs.com/demo_image_match/train/moby_0902/
	训练数据oss路径	训练数据集的OSS路径。如果通过上游组件传递训练数据，则无需指定该参数。	无需填写
	是否使用预训练的模型	建议使用预训练模型，以提高训练模型的精度。	否
	预训练模型oss路径	如果有自己的预训练模型，则将该参数配置为自己预训练模型的OSS路径。如果没有配置该参数，则使用PAI提供的默认预训练模型。	无需填写
参数设置	自监督模型使用的backbone	识别模型的网络名称，系统支持主流的识别模型resnet_50。	resnet_50
	优化方法	模型训练的优化方法，仅支持AdamW。	AdamW
	初始学习率	网络训练初始的学习率。	0.001
	训练batch_size	训练的批大小，即单次模型迭代或训练过程中使用的样本数量。	64
	总的训练迭代epoch轮数	总的训练迭代轮数。	80
	保存checkpoint的频率	保存模型文件的频率。取值1表示对所有训练数据都进行一次迭代。	10
执行调优	读取训练数据线程数	读取训练数据的线程数。	4
	evtorch model开启半精度	开启半精度会使模型的推理速度显著提升，同时准确率略有降低。	否
	单机或分布式(maxCompute/DLC)	模型训练使用的计算资源，支持以下取值： ■ 单机DLC ■ 分布式DLC	分布式DLC

页签	参数	描述	本案例的示例值
	worker个数	使用**分布式DLC**计算时，您需要配置用于计算的Worker数量。	1
	gpu机型选择	计算资源的GPU机型。 **说明** 自监督模型对资源的消耗较大，建议选择单机4卡或单机8卡的机器进行训练。	48vCPU+368GB Mem+4xv100-ecs.gn6e-c12g1.12xlarge

部署及调用相似图像匹配的模型服务

1. 进入**PAI EAS模型在线服务**页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。
2. 部署模型服务。
   i. 在**PAI EAS模型在线服务**页面，单击**模型上传部署**。
   ii. 在**资源和模型**面板，配置参数（此处仅介绍与本案例相关的核心参数配置方法，其他参数的解释请参见控制台上传部署），并单击**下一步**。

参数	描述
自定义模型名称	模型的名称，建议结合实际业务进行命名，以区分不同的模型服务。
资源组种类	建议使用专属资源组部署模型服务，从而避免公共资源组资源有限时导致的服务排队。关于如何创建专属资源组，请参见专属资源组。
资源种类	如果您选择**资源组种类**为**公共资源组**，则需要选择**资源种类**为GPU。
Processor种类	选择EasyVision。
模型类型	选择**通用图像比对**。
模型文件	本案例中训练好的模型均存储在OSS中，因此选择**OSS文件导入**。 您选择**训练所用oss目录**下的.pt模型文件即可。例如选择epoch_50_export.pt。

iii. 在**部署详情及配置确认**面板，配置模型服务占用资源的相关参数。

参数	描述
实例数	本案例配置为1。实际应用时建议配置多个服务实例，以避免单点部署带来的风险。
卡数	本案例配置为1。实际应用时根据情况配置。
核数	本案例配置为1。实际应用时根据情况配置。
内存数（M）	本案例配置为16384 MB。实际应用时根据情况配置。

iv. 单击**部署**，等待一段时间即可完成模型部署。

3. 查看模型服务的公网地址和访问Token。
   i. 在**PAI EAS模型在线服务**页面，单击目标服务**服务方式**列下的**调用信息**。
   ii. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的访问地址和**Token**。
4. 使用脚本进行批量调用。
   i. 创建相似图像匹配的Python脚本img_match.py。

```
import requests
import base64
import sys
import json
hosts = 'http://16640818xxxxxxx.cn-hangzhou.pai-eas.aliyuncs.com/api/predict/img_ma
tch_001'
head = {
            "Authorization":"ZDgzYzdlODg5NTA3ODZiOTY0OWRxxxxxxxxxxxxxxxxxxxxx"
        }
imagea_path =  "./001.jpg"   # 对比图像1的本地路经。
imageb_path =  "./002.jpg"   # 对比图像2的本地路经。
ENCODING = 'utf-8'
datas = json.dumps( {
            "imagea": base64.b64encode(open(imagea_path, 'rb').read()).decode(ENCOD
ING),
            "imageb": base64.b64encode(open(imageb_path, 'rb').read()).decode(ENCOD
ING),
        })
for x in range(0,1):
    resp = requests.post(hosts, data=datas, headers=head)
    print(resp.content)
print("test endding")
```

   ii. 将Python脚本上传至您的任意环境，并在脚本上传后的当前目录执行如下调用命令。

```
python3 <img_match.py>
```

其中<img_match.py>需要替换为实际的Python脚本名称。

得到类似如下的预测结果。

```
b'{"request_id": "5ba88f4b-0104-4861-a548-6c85b15d8d40", "success": true, "similarity": [0.8721832633018494],
"l2_distance": [20.392292022705078]}'
test endding
```

其中similarity表示图像的相似度，取值0~1之间，数值越大表示相似度越高。

5. 监控服务指标。

   调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

   i. 在**PAI EAS模型在线服务**页面，单击已调用服务**服务监控**列下的图标。

   ii. 在**服务监控**页签，即可查看模型调用的指标水位。

   从下面的服务监控水位图中可以看到，本案例部署的模型时延在80 ms左右。您自己的模型时延以实际为准。

部署及调用图像检索的模型服务

1. 进入**PAI EAS模型在线服务**页面。

   i. 登录PAI控制台。

   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。

   iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入**EAS-模型在线服务**页面。

2. 部署模型服务。

   i. 在**PAI EAS模型在线服务**页面，单击**模型上传部署**。

ii. 在**资源和模型**面板，配置参数（此处仅介绍与本案例相关的核心参数配置方法，其他参数的解释请参见控制台上传部署），并单击**下一步**。

参数	描述
自定义模型名称	模型的名称，建议结合实际业务进行命名，以区分不同的模型服务。
资源组种类	建议使用专属资源组部署模型服务，从而避免公共资源组资源有限时导致的服务排队。关于如何创建专属资源组，请参见专属资源组。
资源种类	如果您选择**资源组种类**为**公共资源组**，则需要选择**资源种类**为GPU。
Processor种类	选择EasyVision。
模型类型	选择**商品检索**。
模型文件	本案例中训练好的模型均存储在OSS中，因此选择**OSS文件导入**。 您选择**训练所用oss目录**下的.pt模型文件即可。例如选择epoch_50_export.pt。

iii. 在**部署详情及配置确认**面板，配置模型服务占用资源的相关参数。

参数	描述
实例数	本案例配置为1。实际应用时建议配置多个服务实例，以避免单点部署带来的风险。
卡数	本案例配置为1。实际应用时根据情况配置。
核数	本案例配置为1。实际应用时根据情况配置。
内存数（M）	本案例配置为16384 MB。实际应用时根据情况配置。

iv. 单击**部署**，等待一段时间即可完成模型部署。

3. 查看模型服务的公网地址和访问Token。

i. 在**PAI EAS模型在线服务**页面，单击目标服务**服务方式**列下的**调用信息**。

ii. 在**调用信息**对话框的**公网地址调用**页签，查看公网调用的访问地址和Token。

4. 调用图像检索服务进行推理

图像检索服务需要首先建立图像检索数据库，然后将注册到数据库中的图像进行特征提取，最后从图像数据库的数据中，对上传的图像进行相似快速检索。整个过程需要使用的接口包括数据库初始化、增加数据库、增加数据及检索数据等接口，如下表所示，所有接口说明请参见概述。下文提供一个简单的示例供您参考。

功能	描述	接口文档
初始化	首先基于某一指定的OSS路径进行数据库的初始化。	初始化接口
数据库管理	支持对图像数据库进行增加、查询、删除及存储操作。	数据库管理层接口

功能	描述	接口文档
图像数据注册	指定已存在的图像数据库，对图像数据进行增加、查询、删除及修改操作。	数据库层接口
图像数据检索	指定已存在的图像数据库，对上传的图片进行图像相似的快速检索。	检索数据（db_search）

> **说明** 该方案主要针对轻量级数据。如果您的检索库数据量预期大于1000万，请单独提交工单。

i. 创建数据库初始化的Python脚本retrieval_init.py。

```
import requests
import base64
import sys
import json
ENCODING = 'utf-8'
our_oss_io_config = dict(ak_id='LTAIcsxxxxxxxxxx',     # 阿里云账号的AccessKey。
                ak_secret='gsiu1HjLGDxxxxxxxxxxxxx',  # 阿里云账号的AccessKey Secret
。
                hosts='oss-cn-hangzhou-internal.aliyuncs.com',    # 内网Endpoint。
                buckets=['tongxin-lly'])     # OSS Bucket。
datas = json.dumps({
            "function_name": "init",
            "function_params": {
                "backend": "oss",
                "root_path": "oss://tongxin-lly/img_retrieval/db0914/",  # OSS具体路
经。
                "oss_io_config" : our_oss_io_config,
            },
        })
hosts = 'http://16640818xxxxxxx.cn-hangzhou.pai-eas.aliyuncs.com/api/predict/img_re
trieval_001'
head = {
            "Authorization":"NTE3Y2M2ZmEzZGQ3ZGRkOGM4ZDxxxxxxxxx"
        }
r = requests.post(hosts, data=datas, headers=head)
print(r.content)
print("test endding")
```

ii. 创建增加数据库的Python脚本retrieval_add.py。

```
import requests
import base64
import sys
import json
ENCODING = 'utf-8'
datas = json.dumps({
            "function_name": "add",
            "function_params": {
                "database_name" : "tongxin_demo",
                "feature_dim_dict": {'feature' : 384},    #模型训练的特征维度。模型训练时，如果选择MOBY_TIMM，则取值为384，否则取值均为2048。
                "feature_distance_dict":{'feature':'Cosine'},    # 计算特征距离的衡量方法。
            },
        })
hosts = 'http://16640818xxxxxx.cn-hangzhou.pai-eas.aliyuncs.com/api/predict/img_retrieval_001'
head = {
            "Authorization":"NTE3Y2M2ZmEzZGQ3ZGRkOGM4ZDE1MDxxxxxxxxxxxxxx"
        }
r = requests.post(hosts, data=datas, headers=head)
print(r.content)
print("test endding")
```

iii. 创建增加数据的Python脚本retrieval_db_set.py。

```
import requests
import base64
import sys
import json
ENCODING = 'utf-8'
image_path =  "./010.jpg"
datas = json.dumps({
            "function_name": "db_set",
            "function_params": {
                "database_name":"tongxin_demo",
                "image" : base64.b64encode(open(image_path, 'rb').read()).decode(ENCODING),
                "group_id" : "888",     # 存入该数据的群组名。
                "intra_id" : 10         # 该图像群组名下的具体ID。
            },
        })
hosts = 'http://16640818xxxxxx.cn-hangzhou.pai-eas.aliyuncs.com/api/predict/img_retrieval_001'
head = {
            "Authorization":"NTE3Y2M2ZmEzZGQ3ZGRkOGM4Zxxxxxxxxxxxx"
        }
r = requests.post(hosts, data=datas, headers=head)
print(r.content)
print("test endding")
```

iv. 创建检索数据的Python脚本retrieval_db_search.py。

```
import requests
import base64
import sys
import json
ENCODING = 'utf-8'
image_path =  "./test.jpg"
datas = json.dumps({
            "function_name": "db_search",
            "function_params": {
                "database_name":"tongxin_demo",
                "image" : base64.b64encode(open(image_path, 'rb').read()).decode(EN
CODING),
                'search_topk' : 3,     # 返回的最相似的k张图像.
            },
        })
hosts = 'http://1664081855xxxxxx.cn-hangzhou.pai-eas.aliyuncs.com/api/predict/img_r
etrieval_001'
head = {
            "Authorization":"NTE3Y2M2ZmEzZGQ3ZGRkOGM4ZDE1MDVhNzxxxxxxxxx"
        }
r = requests.post(hosts, data=datas, headers=head)
print(r.content)
print("test endding")
```

v. 将上述创建的Python脚本全部上传至您的任意环境，并在脚本上传后的当前目录依次执行调用命令。

```
python3 <retrieval_xxx.py>
```

其中<retrieval_xxx.py>需要依次替换为上述创建的初始化脚本、增加数据库脚本及增加数据脚本。

vi. 针对部署好的数据库，进行目标图像检索，即调用检索数据脚本。

```
python3 retrieval_db_search.py
```

得到的推理预测结果如下图所示。

```
b'{"request_id": "e0f7231c-b42e-4723-842d-70e00d669805", "success": true, "status": true, "info": {"distance":
 [0.7133828401565552, 0.6841099262237549, 0.6735824346542358], "group_id": ["999", "999", "888"], "intra_id":
[7, 8, 4]}}'
test endding
```

从上述模型推理的返回结果可以看出，模型服务返回了该数据库中与目标图像相似度最高的三张图像的group_id和对应intra_id。

7.智能客服对话系统解决方案

针对问题咨询场景中出现大量相关领域的问题，PAI提供了智能客服对话系统解决方案，以降低客户等待时间和人工客服成本。本文以汽车售前咨询业务领域为例，介绍如何基于人工智能算法，快速构建智能客服对话系统。

您可前往阿里云AI体验馆中的智能客服页面，体验“汽车售前咨询”的服务示例。

背景信息

在企业服务用户的过程中，每天都会出现大量相关领域内的问题。传统的人工客服作息服务解答效率低，且成本高。针对该问题，PAI提出了如下解决方案，搭建智能客服对话系统，从而帮助企业实现在线服务的智能化人机协作、提高坐席服务的效率、降低人力开销及新人业务学习成本：

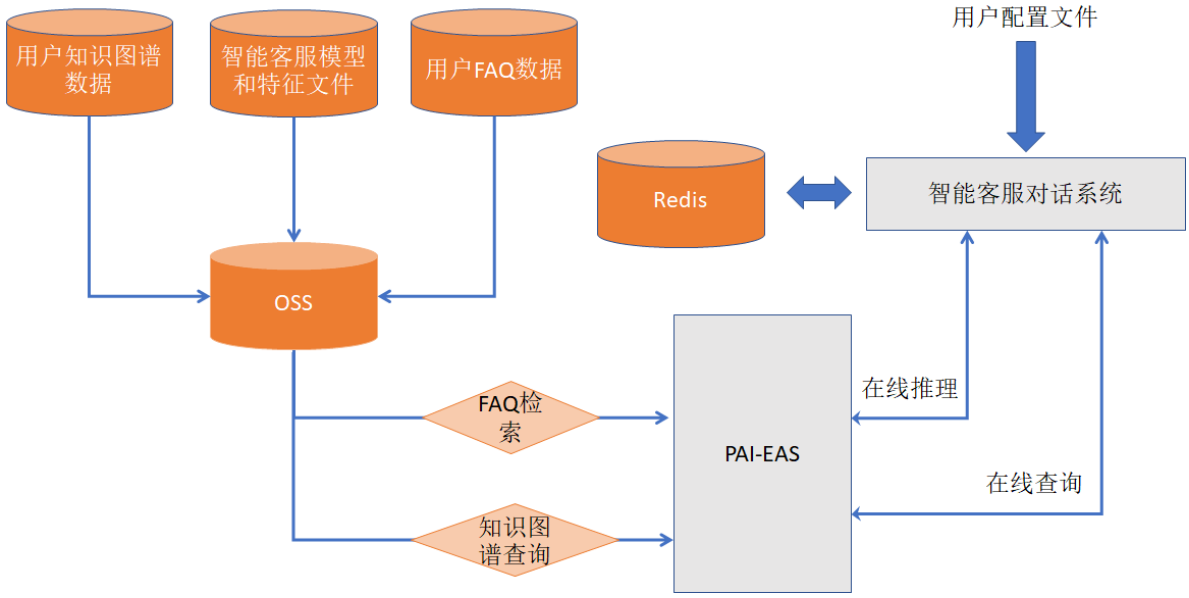
- 解决方案

  PAI在智能客服领域提供了端到端的纯白盒解决方案。您只需要准备好自己相关领域的常见问题FAQ（Frequently Asked Questions）和知识图谱数据，就可以利用PAI搭建自定义的人工智能流程，形成从算法构建到模型部署的端到端解决方案，从而实现对应领域的智能客服业务系统。

  本文以汽车售前咨询业务领域为例，为您展示如何快速搭建智能客服机器人，从而实现机器人自动回答关于汽车售前业务的相关问题。该方式不仅能够节省客户的咨询等待和信息检索时间，而且能够节省企业运营人力成本，使精力可以集中在高价值客户上。

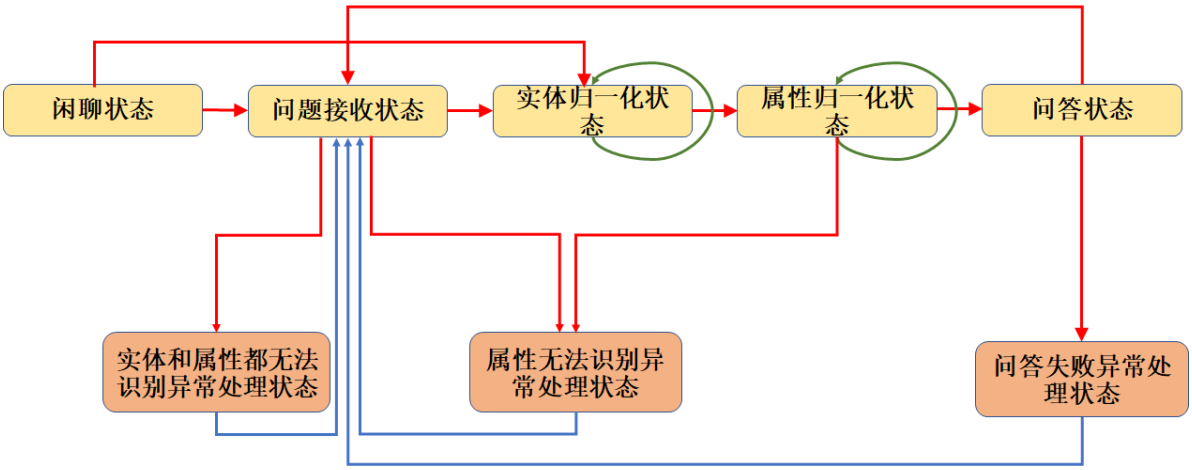
- 方案架构

  智能客服对话系统工程部署架构图如下所示。



  您先将知识图谱数据、模型和特征文件及FAQ数据存储在OSS上，然后将智能客服对话系统服务部署在PAI-EAS上，就可以使用FAQ检索和知识图谱查询提供的功能进行智能客服问答。

智能客服对话系统内部状态跳转逻辑如下图所示。



系统初始化后进入闲聊状态，对用户输入“您好”等话语进行响应。响应完成后系统进入问题接收状态，此刻等待用户提问。接入通过实体归一化、属性归一化明确用户提问中的实体和属性，接着进入问答状态进行FAQ检索或知识图谱查询。当出现异常状况时，进入异常处理模式，在异常处理中引导用户进入正确的查询。

- 方案优势
  - 纯白盒：您可以根据自己具体的业务场景自定义智能客服业务系统。
  - 端到端：从最初的数据准备到最后的模型部署推理，提供全链路的系统构建流程。
  - 有据可依：对话系统内所有的回答都是有依据的，避免纯深度学习方案的不可解释性。
  - 鲁棒可控：系统任何地方出了问题，都有相应的异常处理与Bad Case分析应对机制。

前提条件

- 已开通PAI（Studio、DSW、EAS）后付费，详情请参见开通。
- 已开通AI工作空间，并添加MaxCompute计算资源和DLC计算资源，详情请参见创建工作空间。
- 已创建OSS存储空间（Bucket），用于存储标签文件和训练获得的模型文件。关于如何创建存储空间，请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，请参见创建专属资源组。
- 已下载并配置PAI-EAS的客户端工具easycmd，本文使用该工具进行模型部署相关操作，详情请参见下载并认证客户端。
- 已开通ECS并创建ECS安全组，用于配置PAI-EAS在线服务。开通请参见通过控制台使用ECS实例（快捷版）云服务器ECS快速入门，安全组创建请参见创建安全组。
- 已创建专有网络VPC和专有网络交换机，用于配置PAI-EAS在线服务。关于如何创建专有网络VPC和交换机，详情请参见创建和管理专有网络和使用交换机。
- 已创建数据库Redis实例，用于配置PAI-EAS在线服务，详情请参见步骤1：创建实例。

操作流程

基于阿里云机器学习PAI平台，构建智能客服对话系统的流程如下：

1. 步骤一：基于FAQ构建检索问答系统

   检索问答系统在对您提出的查询进行预处理的基础上，从FAQ数据集中找到该查询的同义问句，从而匹配并回答预置好的答案。

2. 步骤二：基于KBQA构建知识图谱问答系统

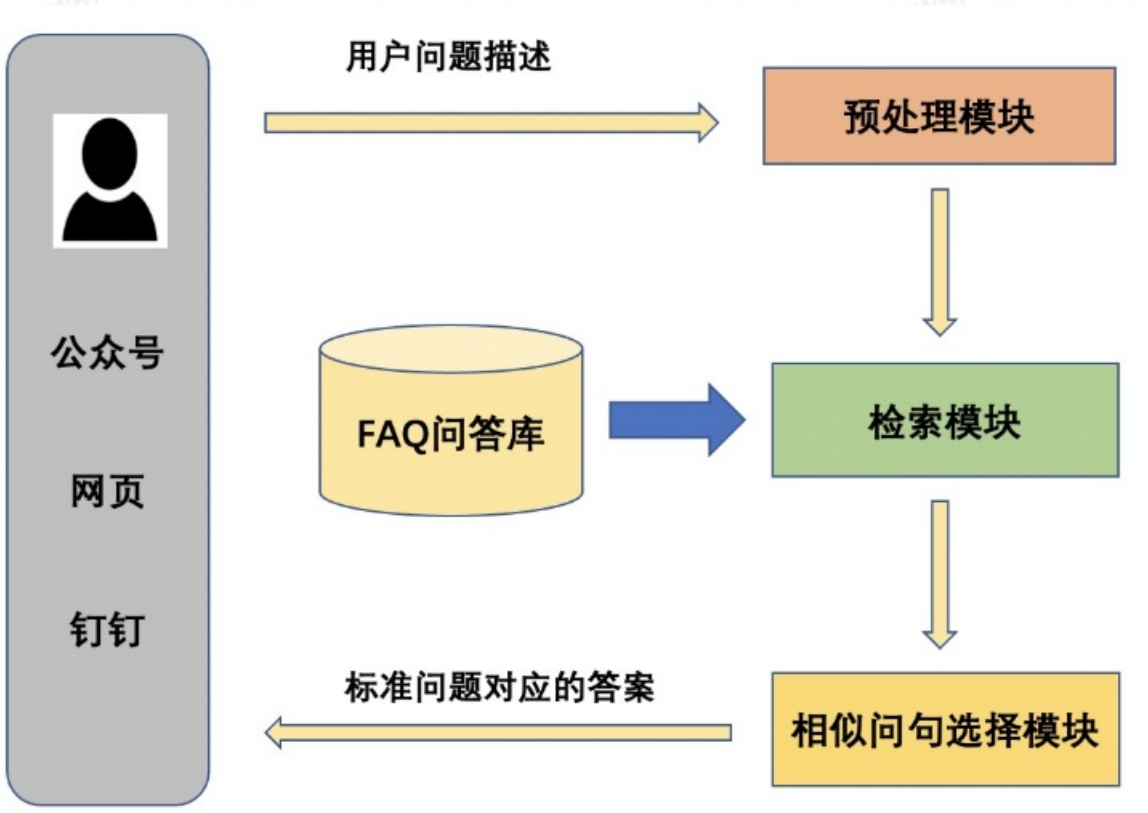
对问题进行语义理解和解析，进而基于知识图谱构建的知识库进行查询和推理，从而得到答案。

3. 步骤三：配置智能客服在线服务

   部署智能客服在线服务，并进行在线调试。

步骤一：基于FAQ构建检索问答系统

基于FAQ的问答系统总体结构如下图所示。



该系统包括以下关键模块：

- 频繁问答对数据集：FAQ问答库。
- 预处理模块：负责对查询进行分词等操作。
- 检索模块：负责从FAQ中检索可能与您查询相似的若干标准问句。
- 相似问句选择模块：负责从候选相似问句中，选择与您查询最相似的标准问句。

上述的检索模块在对您提出的查询进行预处理的基础上，从FAQ数据集中检索K个与查询最相似的问句。然后相似问句选择模块会从这K个候选相似问句中，选出与查询相似度最高的，并判断是否具有足够的置信度。如果相似度最高的候选相似问句的置信度达标，则相似问句选择模块会判定该问句是用户查询的同义问句，从而返回提前预置的答案。

1. 准备FAQ频繁问答对的数据集。

   FAQ是业务场景中最常问的问题，也称为标准问题。对于这类问题，您可以提前编制答案，构成问答对，即QA对（Question Answer Pair）。问答系统的知识库中存储的FAQ数据集，实际上是标准问题与其答案构成的QA对数据集。如果您提出的问题和FAQ数据集中的某个标准问题相似，那么您问题的答案就是该标准问题的答案。

构建频繁问答对数据集需要根据场景特点，确定问题涉及的范围。然后根据开放数据源及自有数据集，采集问答对内容或编写问答对内容。此外，本解决方案支持业务方持续对FAQ数据集进行优化，业务方可以增加一个标准问题、增加同一标准问题的同义问句、增加问题的答案配置集等。

本文以汽车售前的智能客服场景为例，展示FAQ所需的数据，文件列表及层级关系如下所示。

```
car_faqs
├── query_label_idx_map.txt          # 标注问题ID编号文件。
├── questions                        # 标准问题同义配置文件。
│   ├── 保养灯的使用说明
│   ├── 保养一次多少钱
│   ├── 变速箱故障灯亮
│   ├── ......
├── answers                          # 标准答案配置文件。
│   ├── 某汽车品牌_A型号_2021款_标准续航升级版
│   ├── 某汽车品牌_A型号_2021款_长续航版
│   ├── ......
```

以下为构建上述FAQ所需的数据文件的具体步骤。

i. 构建标注问题ID编号文件。

为提升检索模块的召回率，通常会为每一个标准问句配置若干同义问句。因此您需要对FAQ频繁问答库中的标准问题进行ID化编号。例如针对上面FAQ中的标准问题，ID化后如下所示（仅为示例文件query_label_idx_map.txt的部分数据样例）。

```
有几个排气管  0
是双排气管吗  1
问某款车是什么悬挂 2
有换挡拨片吗 3
问发动机产地 4
变速箱是哪里产的 5
问上市时间 6
问某款车上市了吗 7
可以加装换挡拨片吗 8
一保多少钱 9
```

在上述标准问题ID编号文件中，问题和数字之间使用Tab分隔。

ii. 构建标准问题同义配置文件，并将所有标准问题的同义配置文件放到questions文件夹中。

为每个标准问题配置若干个同义问题。以“保养灯的使用说明”为例，文件的每一行内容对应一个和“保养灯的使用说明”同义的问题。以下为示例文件保养灯的使用说明的部分数据样例。

```
新款保养灯怎么设定
保养灯是什么样子的图标
保养灯什么消出
保养灯归是什么标志?
保养灯长什么样子
保养灯如何设置。因为过保不想去4S店了,有谁知道的说一下.
保养灯归零，是怎么，弄得，请问各位兄弟，谢谢!
保养灯和燃油指示灯同时闪烁是怎么回事?
有保养灯的没有?
汽车保养灯归零
```

iii. 构建标准答案配置文件，并将所有的标准答案配置文件放到文件夹answers中。

以某汽车品牌_A型号_2021款_标准续航升级版这款车为例，答案配置文件的示例文件某汽车品牌_A型号_2021款_标准续航升级版的部分样例数据如下所示。

```
[有几个排气管]  0个
[是双排气管吗]  不是
[问某款车是什么悬挂] 前麦弗逊式独立悬架，后多连杆独立悬架
[有换挡拨片吗]  没有
[问发动机产地]  美国
[变速箱是哪里产的]  上海
[问上市时间] 2021年1月
[问某款车上市了吗]  2021年1月已上市
[可以加装换挡拨片吗] 可以加装，但需要加装模块
[一保多少钱] 5000公里做首保，首保免费
```

标准答案配置文件中的[]表示问题类型，后面展示问题的答案。其中答案和问题类型之间使用Tab分隔。

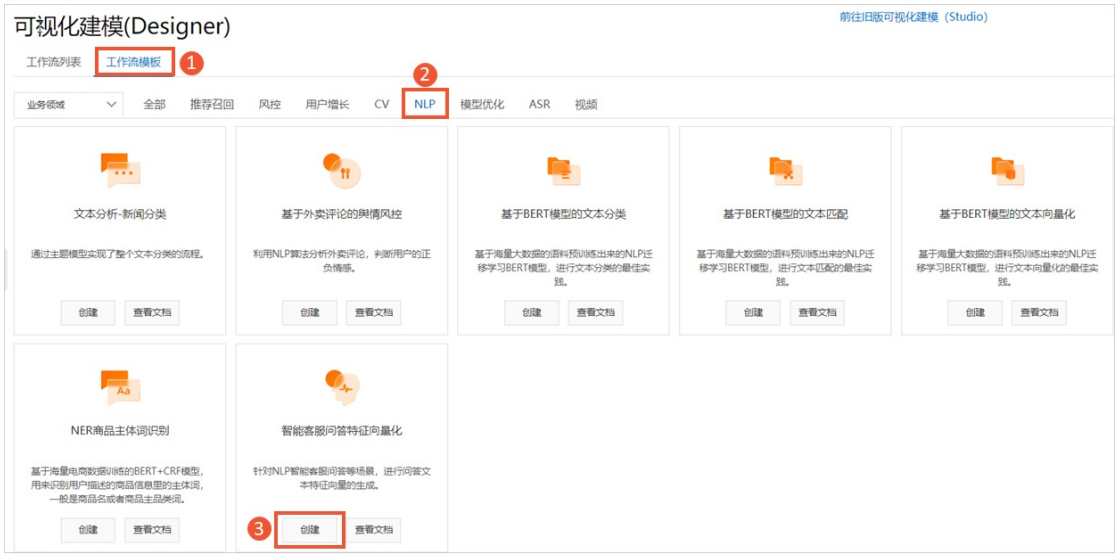
2. 对问答进行特征向量化。

利用可视化建模PAI-Designer提供的基于海量大数据语料预训练获得的NLP迁移学习模型，构建查询向量表征模型。示例使用的训练集为train.csv、评估集为dev.csv、测试集为test.csv。

i. 进入AI工作空间，详情请参见工作空间管理。

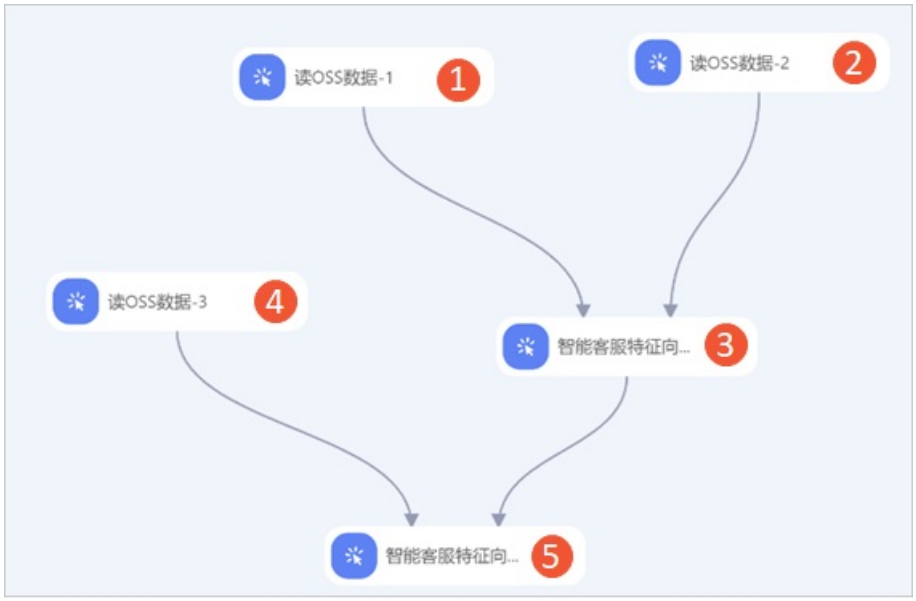
ii. 在左侧导航栏，选择**模型开发和训练 > 可视化建模（Designer）**。

iii. 基于工作流模板，创建智能客服问答特征向量化工作流。



iv. 在**工作流列表**页面，选中刚才创建好的模板工作流，并单击**进入工作流**。

v. 系统根据预置的模板，自动构建工作流，如下图所示。运行过程中，您右键单击工作流中的节点，可以查看运行日志。



区域	描述
①	配置实验的训练集，即配置**读OSS数据**组件的**OSS数据路径**参数为准备的训练集文件的OSS目录。例如 `oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/chatbot_demo/FAQ/train.csv` ，该数据集是PAI在华东2（上海）提供的公开数据集，您可以直接使用。
②	配置实验的评估集，即配置**读OSS数据**组件的**OSS数据路径**参数为准备的评估集文件的OSS目录。例如 `oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/chatbot_demo/FAQ/dev.csv` ，该数据集是PAI在华东2（上海）提供的公开数据集，您可以直接使用。
③	配置智能客服特征向量训练的参数。智能客服特征向量训练组件的配置详情请参见下文的特征向量训练组件配置。
④	配置实验的测试集，即配置**读OSS数据**组件的**OSS数据路径**参数为准备的测试集文件的OSS目录。例如 `oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/chatbot_demo/FAQ/test.csv` ，该数据集是PAI在华东2（上海）提供的公开数据集，您可以直接使用。
⑤	配置智能客服特征向量预测的参数。智能客服特征向量预测组件的配置详情请参见下文的特征向量预测组件设置。

特征向量训练组件配置

页签	参数	描述	本案例的示例值
	输入Schema数据	训练向量召回内容的schema。	`sent:str:1,label:str:1`
	文本列选择	训练向量召回的内容在输入表中对应的列名。	sent
	标签列选择	标签在输入格式中对应的列名。	label

字段设置 页签	参数	描述	本案例的示例值
	模型存储路径	存储训练模型的OSS路径。	`oss://pai-online-shanghai.oss-cn-shanghai-internal.aliyuncs.com/chatbot_demo/FAQ/saved_model_dir/`
参数设置	**每个GPU上的每步训练的样本数 Batch Size**	每个GPU上的每步训练的样本数目。	32
	序列长度	模型的Sequence Length。	128
	训练轮数	训练模型时，使用整遍数据集的次数。	3
	模型保存间隔步数	训练模型时，每隔多少步保存一次模型参数。	100
	学习率	模型构建过程中的学习率。	1e-5
	双塔模型用户自定义参数	用户自定义参数。	`pretrain_model_name_or_path=hfl/chinese-roberta-wwm-exttwo_tower=Truesiamese=Trueloss_type=hinge_lossmargin=0.45gamma=32embedding_size=128`
执行调优	**训练需要的机器数**	分布式服务器的数量，默认值表示1个Worker。	1
	训练需要的每台机器上的CPU核数	每个Worker下的CPU卡数量。	1
	训练需要的每台机器上的GPU卡数	每个Worker下的GPU卡数量。	1
	GPU机器类型	选择工作流运行的计算资源。	GPU计算型8核60G

特征向量预测组件设置

页签	参数	描述	本案例的示例值
	输入Schema数据	训练向量召回内容的schema。	`sent:str:1,label:str:1`

页签	参数	描述	本案例的示例值
字段设置	**文本列选择**	训练向量召回的内容在输入表中对应的列名。	sent
	预测输出文件	预测结果的输出文件。	`oss://pai-algo-tongrun-test.oss-cn-shanghai-internal.aliyuncs.com/etm_text_match_two_tower/features.out` t
参数设置	**每个GPU上的每步预测的样本数 Batch Size**	每个GPU上的每步预测的样本数目。	32
	序列长度	模型的Sequence Length。	128
	双塔模型用户自定义参数	用户自定义参数。	`two_tower=True siamese=True`
执行调优	**训练需要的机器数**	分布式服务器的数量，默认值表示1个Worker。	1
	训练需要的每台机器上的CPU核数	每个Worker下的GPU卡数量。	1
	训练需要的每台机器上的GPU卡数	每个Worker下的CPU卡数量。	1
	GPU机器类型	选择工作流运行的计算资源。	GPU计算型8核60G

步骤二：基于KBQA构建知识图谱问答系统

基于知识库问答KBQA（Knowledge Base Question Answering）是指给定自然语言问题，通过对问题进行语义理解和解析，进而基于知识图谱构建的知识库进行查询和推理，从而得出答案。基于KBQA的知识图谱问答系统的架构图如下所示。

该系统包括以下关键模块：

- 实体链接：将问句中提到的实体关联到唯一的实体表达。
- 语义角色标注：对属性的类型进行自动打标。
- 意图分类：判断问句属于什么问题类型。
- 查询链路生成：生成查询知识库的查询语句。
- 答案生成：负责按照问题意图的类型，在答案模版里生成答案。

上述的实体链接会根据规则对问句进行模式匹配，将问句中的实体映射到系统中的实体规范表达。系统自动标注出属性归属的类别树，并且根据是否为叶子属性进行属性引导，直至定位到叶子属性上，再对问句进行意图分类。最后系统生成查询路径对树形知识图谱进行查询，返回叶子结点的属性值，将生成的答案返回给用户。

以下为您介绍如何准备**基于实体和属性描述的知识图谱数据**。

针对实体、属性名称和属性值，基于KBQA的知识图谱问答系统采用可读性强的JSON构建数据文件。属性名可以包括子属性，以“某汽车品牌_A型号_2021款_标准续航升级版”为例，车轮制动属性包括前制动器类型、后制动器类型、驻车制动类型这些子属性。您需要将业务中的数据转换为JSON格式。下面为示例文件某汽车品牌_A型号_2021款_标准续航升级版.json的部分样例数据。

```
{
    "实体": {
        "名称": "某汽车品牌_A型号_2021款_标准续航升级版",
        "关键词": "某汽车品牌|A型号|2021款|标准续航升级版"
    },
    "属性": {
        "厂商": "某汽车品牌中国",
        "品牌": "某汽车品牌",
        "经销商报价": "27.6万",
        "级别": "中型SUV",
        "上市时间": "2021.07",
        "纯电续航里程": "525km",
        "充电时间": "快冲一小时，慢充10小时",
        "百公里加速时间": "5.6s",
        "整车保修期限": "4年或8万公里"
    },
    "变速箱": {
        "变速箱档位个数": "6个",
        "变速箱类型": "手动变速箱"
    },
    "配置": {
        "内部配置": {
            "空调": {
                "空调控制方式": "自动",
                "温度分区控制": "有",
                "后座出风口": "无"
            },
            "行车电脑显示屏": "有",
            "倒车视频影像": "有"
        }
    }
}
```

步骤三：配置智能客服在线服务

1. 创建安全组。

   创建安全组时，您需要注意以下参数的配置。

区域	参数	描述
基本信息	**网络**	选择已经创建的专有网络VPC。如果没有专有网络VPC，单击**创建专有网络**创建VPC，并完成专有网络中的交换机创建。
	协议类型	选择自定义TCP。
	端口范围	配置为 `6379/6379` 。

访问规则的入方向 区域	参数	描述
	授权对象	填写已创建的PAI-EAS专属资源组对应的服务器IP。 **说明** 此处不能填写0.0.0.0/0，该策略会导致使用到该安全组的ECS、RDS等资源被外部入侵。

2. 创建Redis实例，详情请参见步骤1：创建实例。

   创建Redis实例时，您需要注意以下参数配置，其他参数使用默认值即可。

参数	描述
网络类型	选择专有网络。
专有网络	选择已创建的专有网络。
虚拟交换机	选择已创建专有网络下的虚拟交换机（vSwitch）。
密码设置	选择立即设置，并填写自定义的密码。

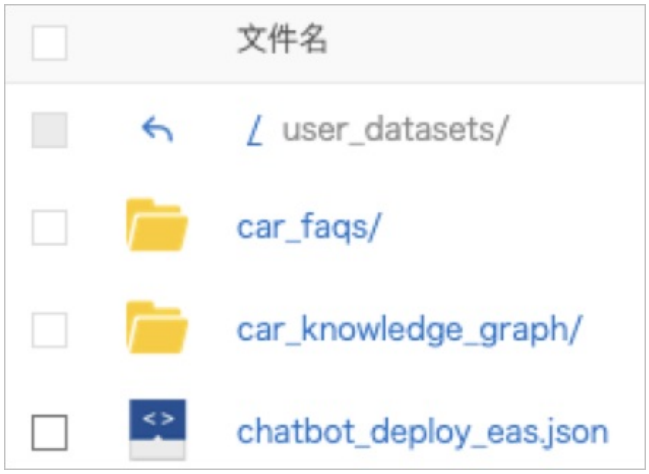
3. 获取Redis的连接地址和端口号。
   i. 单击已创建的Redis实例，进入实例详情页面。
   ii. 在实例信息的连接信息中，查看专有网络的**连接地址**和**端口号**，在后续的部署中会使用到这两个信息。
4. 准备数据。

i. 创建检索引擎配置文件chatbot_service_desc.json。

检索引擎配置文件中需要配置如下参数，您也可以直接下载检索引擎配置的示例文件chatbot_deploy_eas.json。

参数	描述	示例
faq_modules	计算文本相似度的算法。例如基于余弦相似度的特征向量匹配、基于模糊匹配的字符串匹配。	`"faq_modules": ["feature_vector_faq_match", "fuzzy_closest_faq_match"]`
pretrained_feature_vectors_path	离线批处理预测输出的特征向量库。	`/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datasets/car_faqs/query_features.out`
feature_generator_model_dir	特征向量生成器的路径。	`/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datasets/car_faqs/text_match_two_tower_model_dir`
faq_data_dir	faq数据路径。	`/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datasets/car_faqs/answers`
faq_query_label_idx_map	query label的映射表。	`/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datasets/car_faqs/query_label_idx_map.txt`
faq_confidence_threshold	faq的置信度。	0.4
knowledge_graph_data_dir	存储基于实体和属性描述的知识图谱数据的文件夹。	`/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datasets/car_knowledge_graph`
redis_host	Redis实例Host。	`r-uf62y0k7ho7pku54z5.redis.rds.aliyuncs.com`
redis_port	Redis端口。	6379
redis_password	Redis密码。	您自己定义。

ii. 将步骤一：基于FAQ构建检索问答系统和步骤二：基于KBQA构建知识图谱问答系统中准备好的数据文件存储到自己的OSS Bucket中，如下图所示。关于如何将文件上传至OSS，请参见上传文件。



目录中的文件清单如下所示。

```
user_datasets
├── car_faqs
│   ├── query_label_idx_map.txt          # 步骤一中标注问题ID编号文件。
│   ├── questions                        # 步骤一中标准问题同义配置文件夹。
│   ├── answers                          # 步骤一中标准答案配置文件夹。
│   ├── ext_match_two_tower_model_dir    # 步骤一中模型训练得到的模型文件。
│   ├── query_features.out               # 步骤一中模型预测得到的特征文件。
├── car_knowledge_graph                  # 步骤二中准备的知识图谱数据。
│   ├── 某汽车品牌_A型号_2021款_标准续航升级版.json
│   ├── 某汽车品牌_A型号_2021款_长续航版.json
├── chatbot_service_desc.json            # 上一步中创建的检索引擎配置文件。
```

5. 部署在线服务。

i. 创建服务描述文件chatbot_service_desc.json（通过CPU部署）。

chatbot_service_desc.json的内容如下所示。

```
{
    "containers": [
        {
            "image": "registry-vpc.cn-shanghai.aliyuncs.com/eas/smart_chatbot:v2021
1129_pre",
            "env": [
                {
                    "name": "chatbot_config",
                    "value": "/data/eas/EasyTexMiner/examples/solutions/chatbot/use
r_datasets/chatbot_deploy_eas.json"
                },
                {
                    "name": "LC_ALL",
                    "value": "zh_CN.utf8"
                }
            ],
            "command": "/data/eas/ENV/bin/python /data/eas/EasyTexMiner/examples/so
lutions/chatbot/easybot/chatbot_app.py",
            "port": 8079
        }
    ],
    "metadata": {
        "cpu": 1,
        "instance": 1,
        "memory": 10000,
        "resource": "eas-r-ia5cabi7v7fd86****"
    },
    "name": "chatbot_test",
    "oss_endpoint": "oss-cn-shanghai-internal.aliyuncs.com",
    "oss_mount_path": "/data/eas/EasyTexMiner/examples/solutions/chatbot/user_datas
ets",
    "oss_path": "oss://chatbot-test/user_datasets",
    "cloud": {
        "networking": {
            "security_group_id": "sg-uf6ec8tbwnldltrsa1iq",
            "vswitch_id": "vsw-uf624jjhqb7bgcqeg221r"
        }
    }
}
```

上述配置文件的字段详情如下表所示。

字段	描述	是否需要修改
image	智能客服专属镜像的路径	无需修改
env	传递给镜像的环境变量	无需修改
command	启动脚本	无需修改
name	部署在线服务的名称	可以自定义

字段	描述	是否需要修改
resource	部署在PAI-EAS的专属资源组的资源组ID	根据实际情况修改
oss_endpoint	OSS地址的Endpoint	根据实际情况修改
oss_mount_path	挂载到镜像中的本地路径	无需修改
oss_path	OSS上的存放数据的路径	根据实际情况修改
security_group_id	创建的安全组ID	无需修改
vswitch_id	创建的交换机ID	无需修改

ii. 通过eascmd如下命令部署服务。关于eascmd的命令详情，请参见命令使用说明。

```
eascmd create chatbot_service_desc.json
```

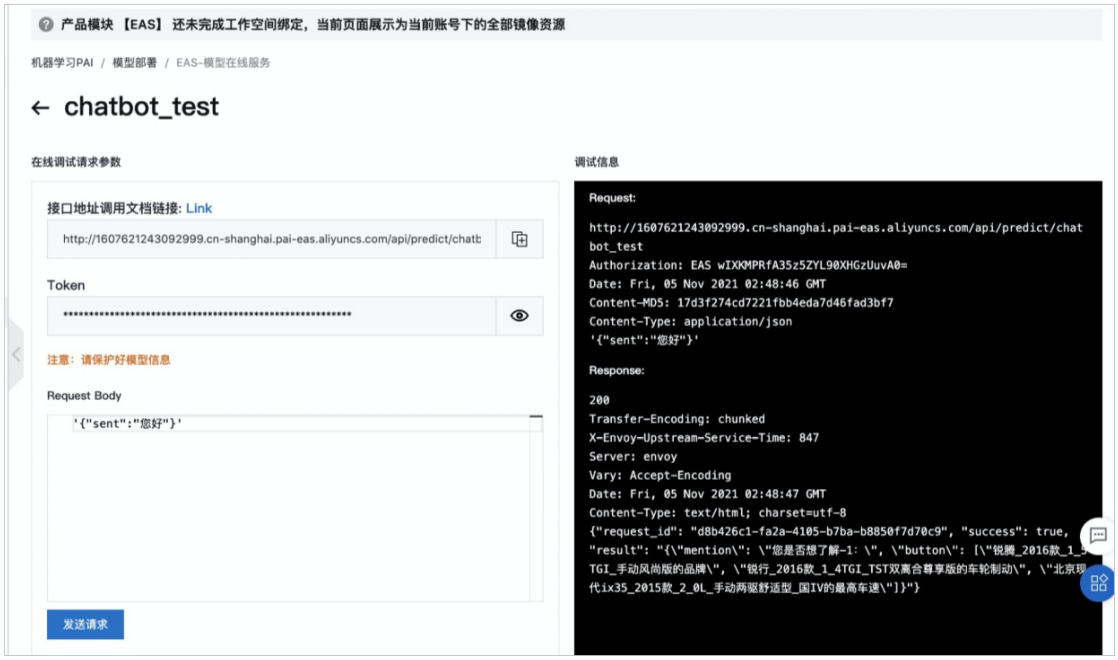
iii. 开启PAI-EAS专属资源组的VPC直连，详情请参见VPC高速直连。

iv. 调试在线服务。

在**PAI EAS 模型在线服务**页面的服务列表，找到目标服务，单击**操作**列下的**在线调试**，即可进入调试页面。

在调试页面，输入**Request Body**，再单击**发送请求**，即可在右侧的**调试信息**区域查看返回结果。

例如，**Request Body**输入 `'{"sent":"您好"}'`，返回的结果如下图所示。



从返回的调试结果可以看出，创建好的智能客服系统针对上述问题进行了引导式的知识回复。

8.通用视频打标解决方案

为了帮助您更好的理解视频、检索视频、查找视频，PAI提供了通用视频打标解决方案。该解决方案通过构建并训练视频打标模型、将模型部署为EAS服务、调用服务来实现视频分类。本文为您介绍通用视频打标的解决方案。

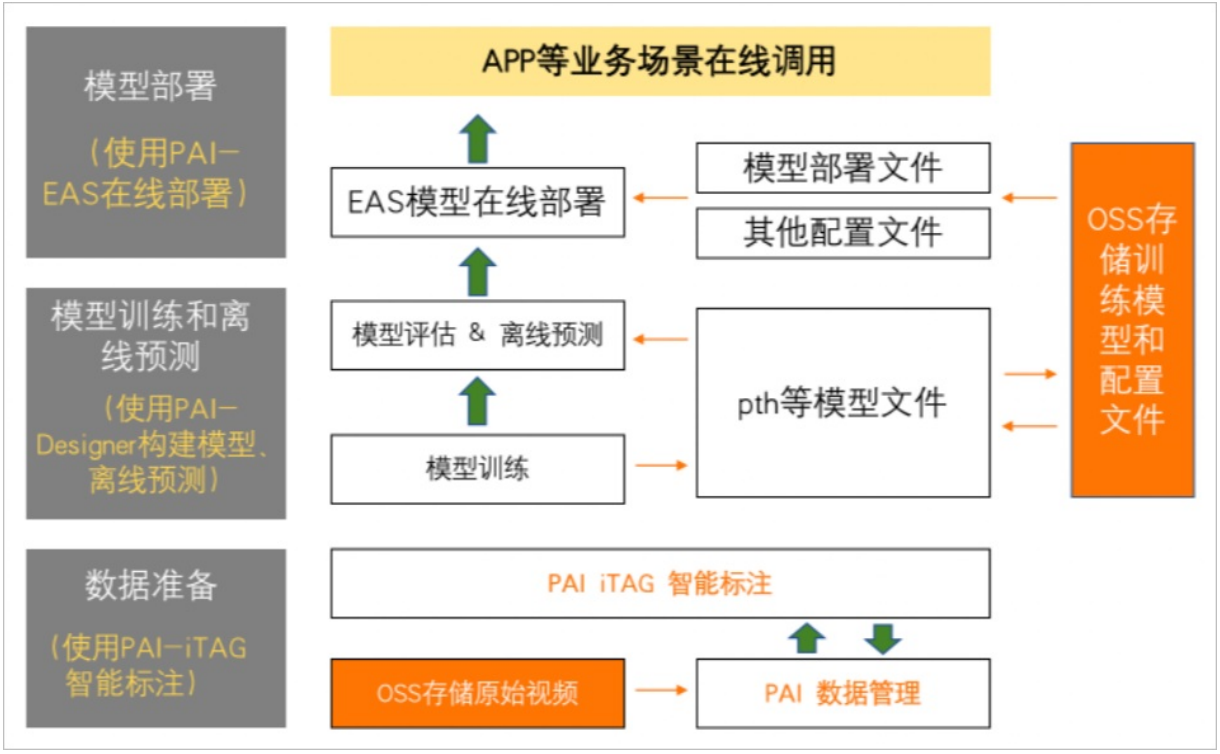
前提条件

在开始执行操作之前，请确认您已经完成以下准备工作：

- 已开通PAI（Designer、DSW、EAS）后付费，详情请参见开通。
- 已开通并创建工作空间，且添加了DLC计算资源，详情请参见开通并创建默认工作空间。
- 已创建OSS存储空间（Bucket），用于存储原始数据、标签列表文件及训练获得的模型文件。关于如何创建存储空间，详情请参见创建存储空间。
- 已创建PAI-EAS专属资源组，本文训练好的模型会部署至该资源组。关于如何创建专属资源组，详情请参见创建专属资源组。

使用流程

基于阿里云机器学习PAI平台，视频打标解决方案的流程如下。



1. 步骤一：准备数据

   将视频数据及标签文件上传到OSS存储空间，用于后续的模型训练。若数据未标注可基于智能标注（iTAG）进行原始数据标注，详情请参见智能标注（iTAG）。

2. 步骤二、构建视频打标模型及离线批量预测

   在可视化建模平台PAI-Designer上，通过视频分类训练组件构建模型。该组件支持视频帧输入和视频帧组合文本的多模态输入两种形式；输出支持单标签和多标签两种形式。

   完成模型训练后，您可以通过通用视频预测组件，对视频数据进行离线批量打标。

3. 步骤三：模型部署及调用模型服务

   您可以将训练好的视频打标模型部署为PAI-EAS在线服务，并在实际的生产环境中调用模型服务，从而进行在线推理。

步骤一：准备数据

首先您需要根据不同的任务类型准备标签文件和视频数据，并将标签文件和视频数据上传到OSS。注意事项如下所示：

- 在模型训练和推理过程中，会将OSS Bucket挂载为/root/data。您需要将标签文件中的oss://Bucket替换为/root/data。

  ```
  /root/data/path/to/video/Z_qZNBiWdlw_000005_000015.mp4 5
  ```

- 对于多标签预测任务，您需要将标签文件内容转换为如下形式。

  ```
  /root/data/path/to/video/-fjW8olkBjI_000034_000044.mp4 0 1 0 1 1 1
  ```

- 对于视频帧输入的任务元素之间使用空格作为分隔符，对于包含视频帧和文本输入的任务元素之间使用Tab（\t）作为分隔符。

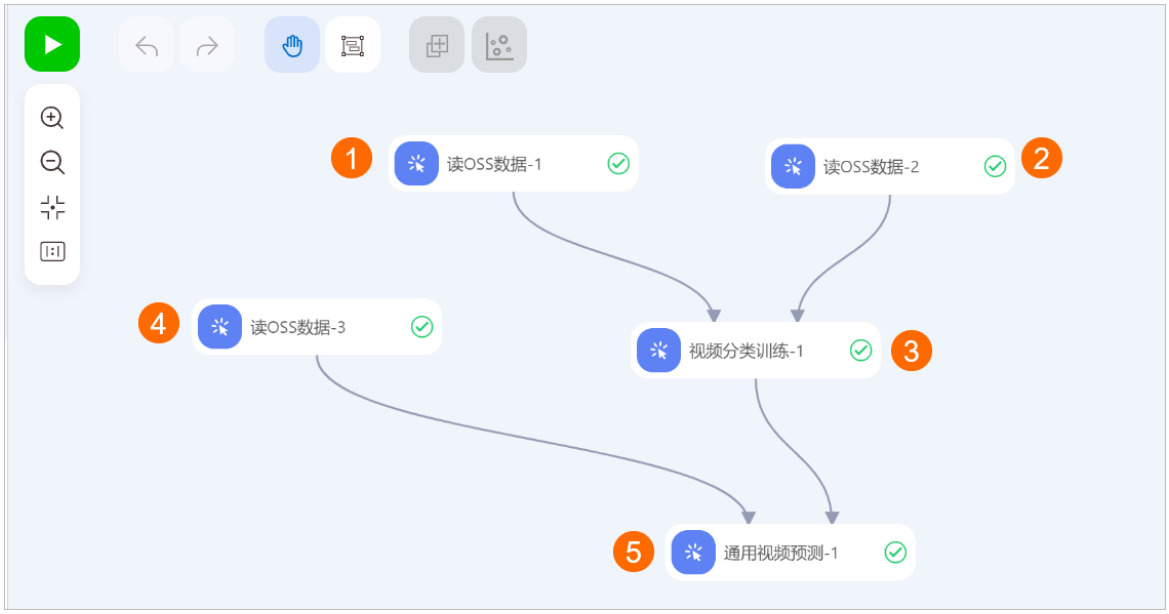
  > **说明** 文本输入是视频对应的描述文本。通过引入额外的文本描述作为输入，可以大幅提升模型的预测精度。

标签文件的示例如下表所示。

任务类型	示例
视频帧输入和单标签预测组合形式	videoinput_multiclass.txt
视频帧输入和多标签预测组合形式	videoinput_multilabel.txt
视频帧输入、文本输入和单标签预测组合形式	videotextinput_multiclass.txt
视频帧输入、文本输入和多标签预测组合形式	videotextinput_multilabel.txt

步骤二、构建视频打标模型及离线批量预测

1. 进入PAI-Designer页面，并创建空白工作流，具体操作请参见创建工作流：空白工作流。
2. 在工作流列表，选择已创建的空白工作流，单击**进入工作流**。
3. 在工作流页面，分别拖入以下组件，并配置组件参数。



区域	描述
①	配置实验的训练数据集，即配置**读OSS数据**组件的**OSS数据路径**为存放训练数据标签文件的OSS路径。
②	配置实验的验证数据集，即配置**读OSS数据**组件的**OSS数据路径**为存放验证数据标签文件的OSS路径。
③	配置视频分类训练的参数。**视频分类训练**组件的配置详情请参见下文的视频分类训练参数配置。
④	配置实验的测试数据集，即配置**读OSS数据**组件的**OSS数据路径**为存放测试数据标签文件的OSS路径。
⑤	配置视频分类预测的参数。**通用视频预测**组件的配置详情请参见下文的通用视频预测参数配置。

视频分类训练参数配置

页签	参数	描述	本案例使用的示例值
	训练所用oss目录	训练过程中模型和Log保存的OSS路径。	`oss://em-dlc-sh.oss-cn-shanghai-internal.aliyuncs.com/EasyMM/DataSet/videotext/experiments/swin_t_bert/` **说明** 需要修改为您使用的OSS路径。

字段设置 页签	参数	描述	本案例使用的示例值
	训练数据文件oss路径	存放训练数据标签文件的OSS路径，如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	无需设置
	评估数据文件oss路径	存放验证数据标签文件的OSS路径，如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	无需设置
	预训练模型oss路径	存放预训练模型的OSS路径，导入预训练模型的权重文件。非必填。	无需设置
参数设置	**视频分类模型网络**	视频分类模型类别，其中swin_t_bert为视频帧组合文本的多模态输入，其他类别为视频帧输入。	swin_t_bert
	是否多标签	是否使用多标签预测形式。	True
	分类类别数目	分类类别数目。	1206
	初始学习率	模型训练的初始学习率，默认值为0.1。	5e-5
	训练迭代轮数	训练迭代的总轮数，默认值为10。	1
	warmup初始学习率	当视频分类模型网络为x3d_xs、x3d_l、x3d_m时，该参数支持设置。 采用warmup学习策略的初始学习率，默认值为0.01。	无需设置
	warmup迭代轮数	当视频分类模型网络为x3d_xs、x3d_l、x3d_m和swin_t_bert时，该参数支持设置。 采用warmup学习策略的迭代轮数，默认值为35。	35
	训练batch_size	训练时使用的批数据量。	4
	模型保存频率	保存模型的迭代轮数。	1

页签	参数	描述	本案例使用的示例值
执行调优	单机或分布式	支持单机或分布式训练模式： ○ single_dlc：单机模式。 ○ distribute_dlc：分布式训练模式。	single_dlc
	是否使用GPU	输入数值每100代表1张GPU卡。例如，800表示8张GPU卡。	100
	gpu类型选择	GPU机型选择	64vCPU+256GB Mem+8xv100-ecs.gn6v-c8g1.16xlarge
	worker个数	采用分布式训练模式时的机器数量。	无需设置

通用视频预测参数配置

页签	参数	描述	本案例使用的示例值
字段设置	数据来源	仅支持输入OSS数据。	OSS
	模型oss路径	存放预测模型的OSS路径。如果您使用视频分类训练作为上游节点，则该参数无需设置。	无需设置
	输入oss文件路径	存放测试数据标签文件的OSS路径。如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	无需设置
	输出oss文件路径	存放预测结果文件的OSS路径。	oss://em-dlc-sh.oss-cn-shanghai-internal.aliyuncs.com/EasyMM/DataSet/videotext/experiments/swin_t_bert/out.txt 说明 out.txt文件存放的路径需要修改为您使用的OSS路径。
	输出oss资源目录	预测过程产生其他文件存放的OSS目录，对于视频分类任务，该参数无需设置。	无需设置
	模型类型	预测模型的类型。	swin_t_bert

页签	参数	描述	本案例使用的示例值
参数设置	**分类类别数目**	分类任务的分类类别数目。	1206
	测试batch_size	测试时使用的批数据量。	1
执行调优	**gpu机型选择**	GPU的机型选择。	64vCPU+256GB Mem+8xv100-ecs.gn6v-c8g1.16xlarge

4. 通过连线将上图各组件节点组织构建为模型工作流。单击画布上方的**运行**。
   实验运行成功之后，即可查看预测结果文件。

预测完成后的输出文件out.txt包含三列，各列描述信息如下所示：

- 视频数据名称。
- 预测标签。
- 每个标签类别的置信度。

步骤三：模型部署及调用模型服务

您可以将训练好的视频打标模型部署为PAI-EAS在线服务，并在实际生产环境中调用模型服务，从而进行在线推理。

1. 首先将训练得到的模型及其他相关配置文件打包为tar.gz格式，并上传至您的OSS Bucket。模型和配置文件的目录结构如下所示。

```
video_text_label
├── video_text_label_model
│   ├── chinese-roberta-wmm-ext
│   │   ├── config.json
│   │   ├── pytorch_model.bin
│   │   ├── special_tokens_map.json
│   │   ├── tokenizer_config.json
│   │   └── vocab.txt
│   ├── video_item.py
│   └── video_text_multiclass.pth
└── video_text_label_ext.py
```

其中：

- video_text_label_model：放置您训练的视频打标模型和模型配置文件。
- video_text_label_ext.py：视频打标在线推理的配置文件。

  video_text_label_ext.py的内容示例如下所示。

```
import mediaflow
import media_ops as ops
import mediaflow.deractor as deractor
from mediaflow import GraphBuilder
from mediaflow.data import Video
from mediaflow.function_data import Context, DataFrame, ImageFrame, AudioFrame, Video
Meta
from mediaflow.config import ImageDecodeConfig, AudioDecodeConfig, EncodeConfig, wind
ow_policy
import math
import cv2
@deractor.window
def short_side_scale(image_frames, ctx):
    new_frames = []
    for i in range(len(image_frames)):
        im = image_frames[i].get_numpy()
        height, width, _ = im.shape
        new_width, new_height = 600, 600
        if width < height:
            new_height = int(math.floor((float(height) / width) * 600))
        else:
            new_width = int(math.floor((float(width) / height) * 600))
        im = cv2.resize(im, (new_width, new_height))
        image_frame = image_frames[i]
        image_frame.set_numpy(im)
        new_frames.append(image_frame)
    return new_frames
class MyGraphBuilder(GraphBuilder):
    def run_option(self):
        return {
            "enable_share_memory": False,
            "share_memory_size": 40000,
            "worker_threads": 128,
            "output_from_context": True,
            "rpc_keepalive": 1000000,
            "enable numpy": True,
```

```
            "metric_print": True
        }
    def build_graph(self):
        image_decode_config = {"use_gpu": "False",'fps':'6'}
        video_item_recognition_config ={
            'model_path' : '/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/video_text_multiclass.pth',
            'config_path' : '/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/video_item.py',
            'token_path' : '/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/chinese-roberta-wmm-ext/',
            'prob_threshold' : '0.5'
        }
        image_frames = Video('video').decode_image(ImageDecodeConfig(image_decode_config))
        image_frames = image_frames.window(window_policy.select_all_window(),short_side_scale)
        image_frames_item = image_frames.window(window_policy.count_window(window_count=64), ops.video.apps.video_item_recognition,video_item_recognition_config)
GraphBuilder.register(MyGraphBuilder)
```

您需要按照以下参数说明，修改以上代码的配置参数。

参数类型	参数	描述	本案例使用的示例值
视频解码参数	use_gpu	是否使用GPU进行视频解码。	False
	fps	视频解码时的采样频率。	6
视频标签分类参数	model_path	自定义视频分类模型的本地路径。	/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/video_text_multiclass.pth
	config_path	模型结构定义文件路径。	/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/video_item.py
	token_path	文本token配置文件路径。	/home/admin/docker_ml/workspace/model/video_text_label/video_text_label/chinese-roberta-wmm-ext/
	prob_threshold	标签预测阈值取值在0-1之间，当某标签预测概率大于该阈值则进行输出。	0.5

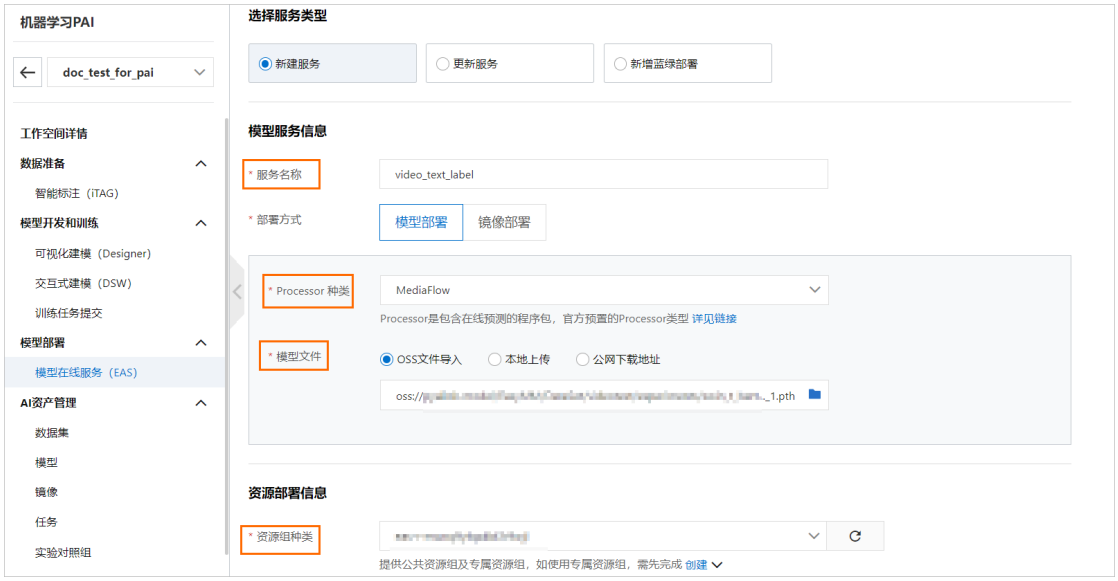
2. 进入PAI EAS模型在线服务页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。

iii. 在工作空间页面的左侧导航栏选择**模型部署 > 模型在线服务（EAS）**，进入EAS-模型在线服务页面。

3. 部署模型服务。

i. 在PAI EAS模型在线服务页面，单击**部署服务**。

ii. 在部署服务页面，配置**服务名称**、**Processor种类**、**模型文件**、**资源组种类**参数，其中**资源组种类**选择已创建的专属资源组，参数配置详情请参见控制台上传部署。



iii. 在对应配置编辑面板，配置部署信息，具体配置信息如下所示。

```
{
      "cloud": {
        "computing": {
            "instance_type": "ecs.gn6i-c4g1.xlarge"
        }
    },
    "model_entry": "video_text_label/video_text_label_ext.py",
    "name": "video_text_label",
    "data_image": "registry.cn-shanghai.aliyuncs.com/eas/mediaflow:py36-0.3.3",
    "model_path": "http://pai-vision-data-hz.cn-hangzhou.oss.aliyun-inc.com/VIP/mod
els_ext/video_text_label.tar.gz",
    "generate_token": "true",
    "processor": "mediaflow_py3",
    "model_config" : {
        "graph_pool_size":8,
        "wait_resource":false,
        "worker_threads":16
    },
    "metadata": {
      "eas.handlers.disable_failure_handler" :true,
      "rpc.batching": false,
      "rpc.worker_threads": 2,
      "rpc.enable_jemalloc": true,
      "rpc.keepalive": 500000,
      "rpc.enable_service_hang_detect": true,
      "rpc.service_hang_detect_period": 5,
      "rpc.max_batch_timeout": 500,
      "cpu": 4,
      "instance": 1,
      "cuda": "9.0",
      "rpc.max_batch_size": 64,
      "memory": 20000,
      "gpu": 1
    },
    "requirements":"http://pai-vision-data-hz.oss-cn-zhangjiakou.aliyuncs.com/VIP/V
ideoOp/release/202202_offline/requirements.txt"
  }
```

如果您希望部署自行训练的模型，model_path需要替换为第一步打包的tar.gz文件的OSS Bucket路径。

iv. 单击**部署**，等待一段时间即可完成模型部署。

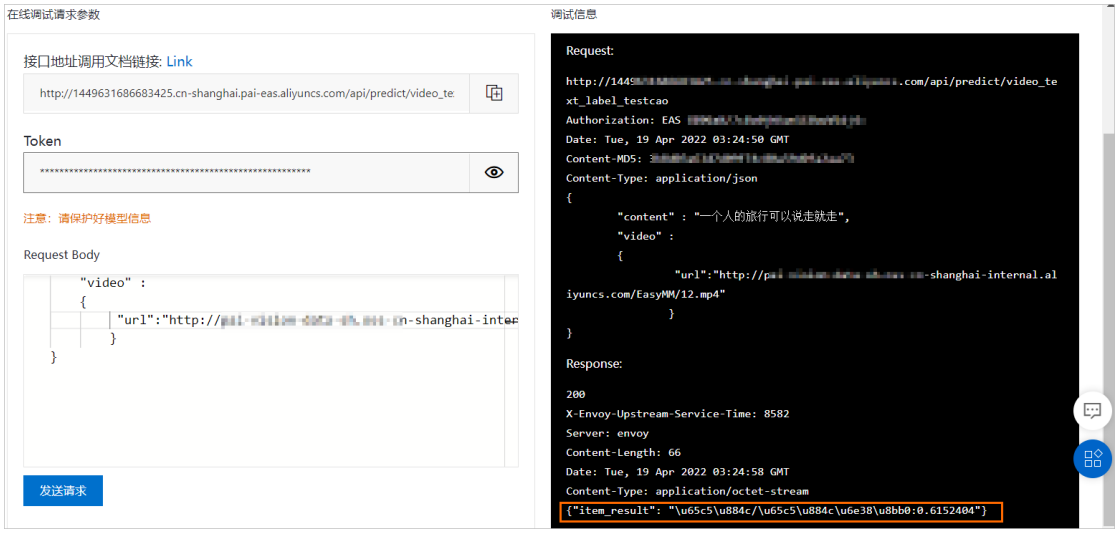
4. 模型服务在线调试。

i. 在PAI EAS模型在线服务页面，单击目标服务**操作**列下的**在线调试**。

ii. 在调试页面**在线调试请求参数**区域的**Request Body**处，输入以下内容。

```
{
    "content" : "一个人的旅行可以说走就走",
    "video" :
      {
        "url":"http://pai-vision-data-hz.oss-cn-zhangjiakou.aliyuncs.com/EasyMM/vid
eo_test/12.mp4"
        }
}
```

iii. 单击**发送请求**，即可在**调试信息**区域查看预测结果。预测结果中会给出item_result字段，包含模型预测的标签信息和对应的置信度。



5. 查看模型服务的公网地址和访问Token。
   i. 在PAI EAS模型在线服务页面，单击目标服务**服务方式**列下的**调用信息**。
   ii. 在调用信息对话框的**公网地址调用**页签，查看公网调用的**访问地址调用文档链接**和**Token**。
6. 使用脚本批量调用模型服务。
   i. 创建调用模型服务的Python脚本eas_video_text_label.py。

```
import requests
import json
#下面的Token需要替换为实际值。
headers = {
    "Authorization":"YTllNjxxxxxxxxxxVlMDliNjdhODExZWIzM2UwYTY5Yg=="
}
data = '{"content" : "一个人的旅行可以说走就走","video" :{"url":"http://pai-vision-dat
a-hz.oss-cn-zhangjiakou.aliyuncs.com/EasyMM/video_test/12.mp4"}}'
data = data.encode('utf-8')
for i in range(20):
      #下面的requests入参需要替换为实际的访问地址。
    result = requests.get('http://1502xxxxxxxxxx.cn-shanghai.pai-eas.aliyuncs.com/
api/predict/video_text_label_test',headers=headers,data=data).text
    print(result)
```

ii. 将eas_video_text_label.py上传至您的环境，并在脚本上传的当前目录执行如下调用命令。

```
python <eas_video_text_label.py>
```

其中，<eas_video_text_label.py>需要替换为实际的Python脚本名称。

7. 监控服务指标。

调用模型服务后，您可以查看模型调用的相关指标水位，包括QPS、RT、CPU、GPU及Memory。

i. 在PAI EAS模型在线服务页面，单击目标服务**服务监控**列下的图标。

ii. 在**服务监控**，查看模型服务调用的指标水位。

9.通用端视觉解决方案

为了解决端侧模型开发和部署的问题，PAI提供了通用端视觉解决方案。该解决方案提供端侧模型训练和部署套件，经过模型训练、模型剪枝和模型量化，最终导出可供部署到端侧产品的模型，并将模型部署到端侧，服务于下游推荐或其他应用场景。本文为您介绍通用端视觉的解决方案。

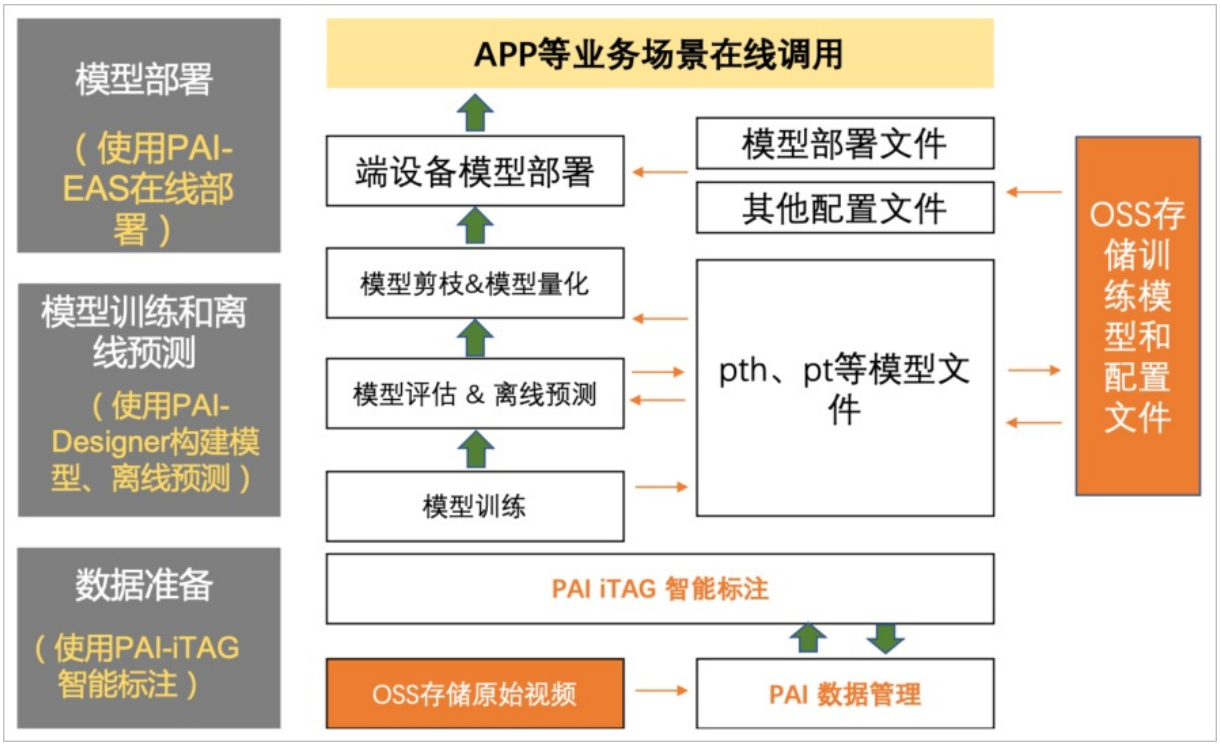
前提条件

在开始执行操作之前，请确认您已经完成以下准备工作：

- 已开通PAI（Designer、DSW、EAS）后付费，详情请参见开通。
- 已开通并创建工作空间，且添加了DLC计算资源，详情请参见开通并创建默认工作空间。
- 已创建OSS存储空间（Bucket），用于存储原始数据、标签列表文件及训练获得的模型文件。关于如何创建存储空间，详情请参见创建存储空间。

使用流程

基于阿里云机器学习PAI平台，端视觉解决方案的流程如下。



1. 步骤一：准备数据

   将训练集、测试集、数据类别列表等数据上传到OSS存储空间，用于后续的模型训练。若数据未标注可基于智能标注（iTAG）进行原始数据标注，详情请参见智能标注（iTAG）。

2. 步骤二：构建端侧视觉检测模型、压缩模型和导出模型

   在可视化建模平台PAI-Designer上，根据上传的图像数据，构建ev-torch框架下的YOLOX_EDGE检测模型。通过PAI-Designer平台提供的模型剪枝和模型量化组件，对ev-torch框架模型进行模型压缩，并导出pt模型。

3. 步骤三：部署模型服务

   您可以将输出的pt模型部署到端侧，用于下游推荐或其他应用场景。

步骤一：准备数据

首先您需要根据选择的模型准备对应类型的数据。以检测模型yolovx为例，您可以准备COCO类型的数据，并将数据上传到OSS。需要上传的数据包括：训练数据、训练数据标签、验证数据、验证数据标签、数据类别列表。

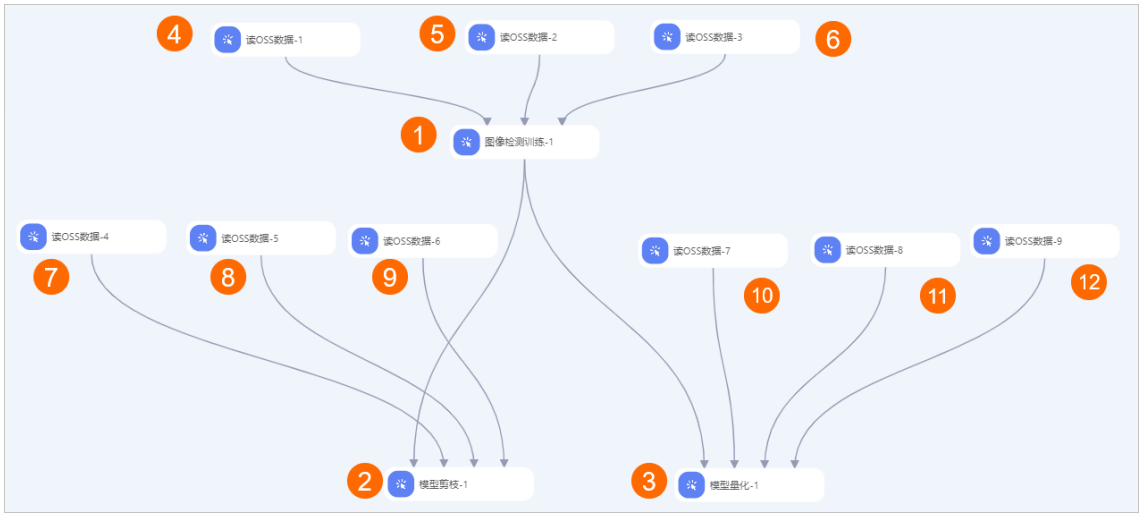
您也可以基于智能标注（ITAG）进行原始数据标注，并将获得的训练集和测试集数据上传到OSS，用于后续的模型训练，详情请参见智能标注（ITAG）。

步骤二：构建端侧视觉检测模型、压缩模型和导出模型

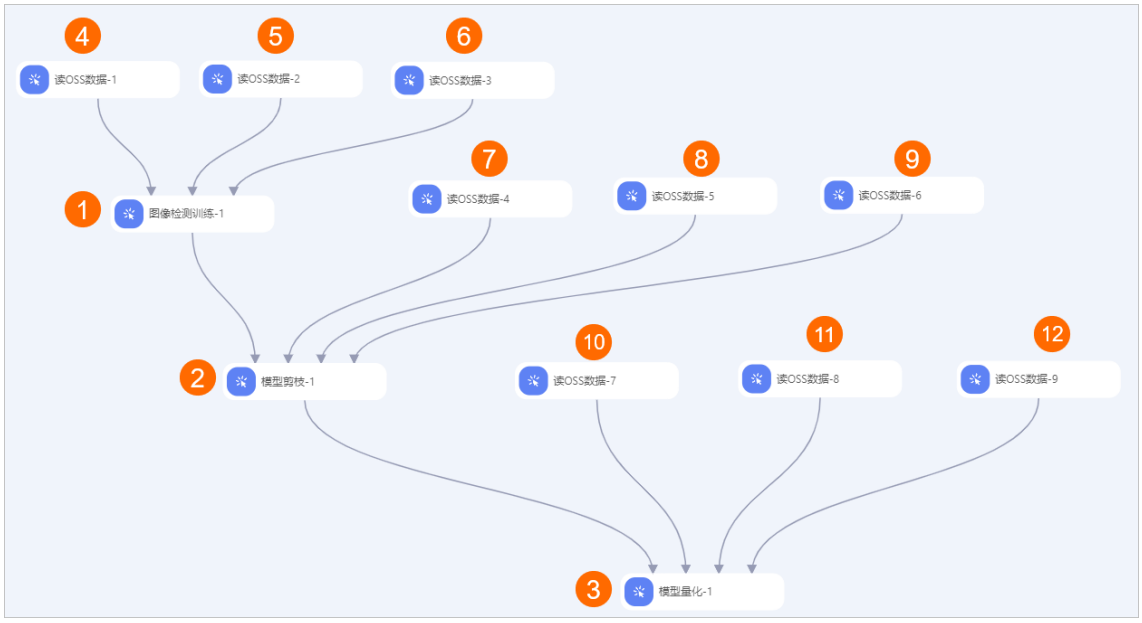
1. 进入PAI-Designer页面，并创建空白工作流，具体操作请参见创建工作流：空白工作流。
2. 在工作流列表，选择已创建的空白工作流，单击**进入工作流**。
3. 在工作流页面，分别拖入以下组件，并配置组件参数。

   您可以按照以下任意一种方式，通过连线将节点组织构建为模型工作流。

   - 使用图像检测训练组件获得的模型文件，分别进行模型剪枝和模型量化。



   - 使用图像检测训练组件获得的模型文件，先经过模型剪枝，再进行模型量化。



区域	描述
①	配置图像检测训练的参数。**图像检测训练**组件的配置详情请参见配置组件参数。
②	配置模型剪枝组件参数。**模型剪枝**组件的配置详情请参见配置组件参数。
③	配置模型量化组件参数。**模型量化**组件的配置详情请参见配置组件参数。
④	配置实验的训练数据集，即配置**读OSS数据-1**组件的**OSS数据路径**为存放训练数据文件的OSS路径。
⑤	配置实验的测试数据集，即配置**读OSS数据-2**组件的**OSS数据路径**为存放测试数据文件的OSS路径。
⑥	配置实验的数据类别列表文件，即配置**读OSS数据-3**组件的**OSS数据路径**为存放数据类别列表文件的OSS路径。
⑦	配置实验的训练数据集，即配置**读OSS数据-4**组件的**OSS数据路径**为存放训练数据文件的OSS路径。
⑧	配置实验的测试数据集，即配置**读OSS数据-5**组件的**OSS数据路径**为存放测试数据文件的OSS路径。
⑨	配置实验的数据类别列表文件，即配置**读OSS数据-6**组件的**OSS数据路径**为存放数据类别列表文件的OSS路径。
⑩	配置实验的训练数据集，即配置**读OSS数据-7**组件的**OSS数据路径**为存放训练数据文件的OSS路径。
⑪	配置实验的测试数据集，即配置**读OSS数据-8**组件的**OSS数据路径**为存放测试数据文件的OSS路径。
⑫	配置实验的数据类别列表文件，即配置**读OSS数据-9**组件的**OSS数据路径**为存放数据类别列表文件的OSS路径。

配置组件参数

页签	参数	描述	本案例使用的示例值

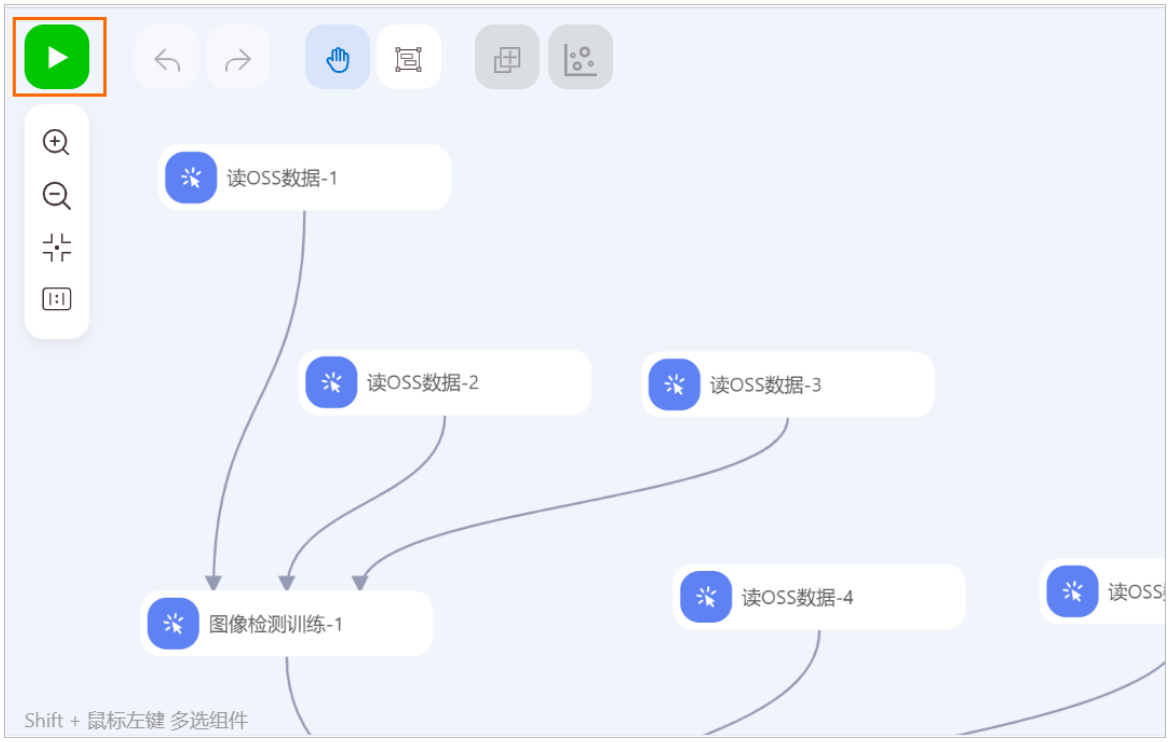
页签	参数	描述	本案例使用的示例值
字段设置	**训练模型类型**	训练模型的类型。 **图像检测训练**支持以下模型类型： ○ SSD ○ FasterRCNN ○ RFCN ○ YOLOV5 ○ YOLOX ○ YOLOX_EDGE **模型剪枝**和**模型量化**支持以下模型类型： ○ YOLOX ○ YOLOX_EDGE	YOLOX_EDGE
	训练所用oss目录	训练过程中模型和Log保存的OSS路径。	oss://examplebucket-cn-shanghai-internal.aliyuncs.com/test/test_model_compression_3-8/ **说明** 需要修改为您使用的OSS路径，且**图像检测训练**、**模型剪枝**和**模型量化**三个组件配置的路径不同。
	训练数据oss路径	存放训练数据文件的OSS路径，如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	无需填写
	训练集标注结果文件路径	存放训练集标注结果文件的OSS路径。	oss://examplebucket-cn-shanghai.aliyuncs.com/data/model_compression/annotations/instances_train2017_sub.json **说明** 需要修改为您使用的OSS路径。
	评估数据oss路径	存放测试数据文件的OSS路径，如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	无需填写

页签	参数	描述	本案例使用的示例值
	测试集标注结果文件路径	存放测试集标注结果文件的OSS路径。	oss://examplebucket-cn-shanghai.aliyuncs.com/data/model_compression/annotations/instances_train2017_sub.json **说明** 需要修改为您使用的OSS路径。
	类别列表文件oss路径	存放数据类别列表文件的OSS路径，如果您使用读OSS数据组件作为上游节点，则该参数无需设置。	oss://examplebucket-cn-shanghai.aliyuncs.com/data/model_compression/label.list **说明** 需要修改为您使用的OSS路径。
	YOLOX数据源格式	数据的处理格式，取值如下： ○ COCO格式 ○ PAI标注格式	COCO格式
	预训练模型oss路径	存放预训练模型的OSS路径，导入预训练模型的权重文件。非必填。	无需填写
	YOLOX端检测模型类型	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 YOLOX端检测的模型类别，预置不同的模型结构。	yolox-customized
	剪枝算法类型	仅**模型剪枝**组件需要配置该参数。 剪枝算法类型，默认为AGP。	AGP
	剪枝算法	仅**模型剪枝**组件需要配置该参数。 剪枝算法，默认为taylorfo。	taylorfo
	yolox端模型depth	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 yolox端模型depth，浮点型。范围为[0.01,1.0]。	1.0

页签	参数	描述	本案例使用的示例值
参数设置	**yolox端模型width**	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 yolox端模型width，浮点型。范围为[0.01,1.0]。	1.0
	yolox端模型parameters size	仅**图像检测训练**组件**训练模型类型**选择YOLOX_EDGE时，才需要配置该参数。 yolox端模型parameters size，限制最大模型参数，浮点型。-1表示不做限制。	-1
	yolox端模型Gflops	仅**图像检测训练**组件**训练模型类型**选择YOLOX_EDGE时，才需要配置该参数。 yolox端模型Gflops，限制最大模型Gflops，浮点型。-1表示不做限制。	-1
	YOLOX端检测模型激活函数类型	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 YOLOX端检测模型激活函数类型，取值如下： ○ relu ○ lrelu ○ silu ○ hsilu	relu
	测试置信度:[0.01,1.0]	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 测试置信度，浮点型。范围为[0.01,1.0]。	0.01
	nms阈值:[0.01,1.0]	仅**字段设置**中**训练模型类型**选择YOLOX_EDGE时，需要配置该参数。 nms阈值，浮点型。范围为[0.01,1.0]。	0.65
	检测类别数目	检测类别数目。	1
	图像尺度	用于yolox，图片缩放后的尺度，分别表示高和宽。	256 256
	初始学习率	模型训练的初始学习率，浮点型。	0.01

页签	参数	描述	本案例使用的示例值
	训练batch_size	训练batch_size。	8
	评估batch_size	评估batch_size。	8
	总的训练迭代epoch轮数	总的训练迭代epoch轮数。	20
	warmup epochs	采用warmup学习策略时的迭代轮数。	5
	最后稳定lr的epochs	采用warmup学习策略lr稳定后继续训练的迭代轮数。	5
	保存checkpoint的频率	保存模型的迭代轮数。	5
执行调优	**读取训练数据线程数**	读取训练数据的线程数。	4
	evtorch model 开启半精度	仅**图像检测训练**组件需要配置该参数。 evtorch model是否开启半精度。	不选中
	单机或分布式(MaxCompute/DLC)	组件运行的引擎。系统会根据您设置的训练模型类型，自动匹配。	分布式DLC
	worker个数	采用分布式训练时的机器数量。	1
	cpu机型选择	仅**图像检测训练**组件的**训练模型类型**选择YOLOV5、YOLOX、YOLOX_EDGE时，需要配置该参数。 选择运行的CPU规格。	32vCPU+128GB Mem-ecs.g6.8xlarge
	gpu机型选择	选择运行的GPU规格。	28vCPU+112GB Mem+1xp100-ecs.gn5-c28g1.7xlarge

4. 单击画布上方的**运行**。



5. 实验运行成功后，您可以在**训练所用OSS目录**配置的OSS路径，下载压缩后的pt模型文件。

   后续您可以使用下载的模型文件部署模型服务。

步骤三：部署模型服务

实验运行成功后，您可以在**训练所用OSS目录**配置的OSS路径下载prune_model.pt和quantize_model.pt的TorchScript模型文件，并将模型文件部署到CPU机器上，服务于下游推荐或其他应用场景。