

本作业分为两个难度，体验作业（低难度）和进阶作业（中等难度），推荐大家先完成体验作业，再尝试进阶作业

进阶作业相比于体验作业，需要额外完成前置的 finetune 步骤，即不再提供 tnews ckpt 文件，而是提供 bert\_base ckpt，通过完成 finetune，生成 tnews ckpt 文件，后续的 evaluation 过程与体验作业一致。创建华为云账号，申请公测、创建 OBS 桶与体验作业一致，如已完成，可以跳过

## 进阶作业：完成 finetune+Evaluation 流程

体验作业提供：bert\_base 文件、tnews 数据集、源代码

体验作业流程：

1. 申请华为云账号、申请 ModelArts 华为云昇腾集群服务公测
2. 从链接下载 bert\_base ckpt 文件、tnews 数据集和源代码
3. 创建 obs 桶
4. 上传 bert\_base ckpt 文件、tnews 数据集和源代码到 obs 桶
5. 提交 ModelArts 训练作业
6. 等待、查看结果，得到 tnews ckpt 文件
7. evaluation 流程，与体验作业一致

## 进阶作业详细步骤

1. 申请华为云账号、申请 ModelArts 华为云昇腾集群服务公测（如已完成，可跳过）



<https://www.huaweicloud.com/>

注册完毕之后，进入到 ModelArts 界面，点击昇腾集群公测，完成申请流程

# AI开发平台ModelArts

ModelArts是面向开发者的一站式AI开发平台，为机器学习与深度学习提供海量数据预处理及半自动化标注、大规模分布式Training、自动化模型生成，及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

按需/包周期付费可选，最低¥0.00/小时

购买套餐

进入控制台

ModelArts Pro

昇腾集群公测

AI开发者社区

ModelArts邀请好友 好礼1+1：AI算力 华为手表手环 周边纪念品随你拿

福利

<https://www.huaweicloud.com/product/modelarts.html>

2. 完成申请之后，开始下载所需文件，

包括数据集文件、bert\_base ckpt 文件和源代码文件（红色标出的文件需要重新下载，另外领个可以直接用体验作业中下载的文件）

**tnews 数据集：**

<https://e-share.obs-website.cn-north-1.myhuaweicloud.com?token=SvjC1tq+9EC0fILt0WI5w1SIJKMVsoAn6d3hiNI5bHpGQM MKuJiZhGfVhyPTgqUIFIVIYKtAmENUpN6ssH654op9P9KNXqX021meACXkf2EZ+O58B+FMoSmBJoaioi+97k52R89dhiOtO4HUMuIK6JYdZImjBOaZ1gxEB+gXCgHBKRNpzfcccU4d41LIA3BPBLqhJ24K0Q76EZdwPUFdf+oRgVqr9d4n01yNmZ9PIIYQILFSxfQGeIojXp5jif1EEISBUvsvSm880An8VEAqYCMdiLcyfVE9rv0YFodid0EPzGmngOMa5p2Q/P6q3jJsjaPsElkX4SMpQ80v4DkJ4J+un4gISVfwf5VWlxObY7CbnP2YIJ+K7F+m9OrRN8oDm4aUooUfi0u8xIKlij4NqFyWMjUuPCsAlHavtpB+WQqMFOZooFx E45VAIZSXQiWIL3UKYVSm/Eyvi+cyg8PON0I501QJkhONQH7inkdoUMinwQoWOx/lyBzKiw8f94rtRtCWWNwI9DZsz/tDSF+yeM o3hl7p7NILWwIFUbvFK2IUc=>

提取码：

123456

**bert 源代码：**

<https://21days-bert.obs.cn-north-1.myhuaweicloud.com:443/bert.zip?AccessKeyId=M7KX8KLMT0ZL1P8QWXZ5&Expires=1634522579&Signature=LY3ZdN9gtoxx8ULDQRkkPuOp1nk%3D>

源代码下载完成后，在本地解压缩

### Bert\_base ckpt:

[https://21days-bert.obs.cn-north-4.myhuaweicloud.com:443/bert\\_base.ckpt?AccessKeyId=M7KX8KLMT0ZL1P8QWXZ5&Expires=1634522231&Signature=1CSNQY%2BK%2BmQxRBkqBwBHMbf0%2BE4%3D](https://21days-bert.obs.cn-north-4.myhuaweicloud.com:443/bert_base.ckpt?AccessKeyId=M7KX8KLMT0ZL1P8QWXZ5&Expires=1634522231&Signature=1CSNQY%2BK%2BmQxRBkqBwBHMbf0%2BE4%3D)

这个文件比较大，请大家耐心下载

### 3. 创建 OBS 桶(如已完成，可跳过，直接使用之前创建的桶)

进入官网界面 <https://www.huaweicloud.com/product/obs.html>，点击右上角的控制台按钮



在控制台界面中，在左上角选择北京四节点，同时在左边的弹出式菜单的搜索框中，搜索 OBS，并选择“对象存储服务 OBS”



此时我们已经进入了 OBS 对象存储服务的界面，点击右上角“创建桶”按钮



按照以下模板创建桶

数据冗余存储策略 ?

**多AZ存储** **单AZ存储**

多AZ存储能提高您的数据可用性，同时会采用相对较高的计费标准。 [价格详情](#)

多AZ存储属性一旦启用，后续无法修改。

---

桶名称

**bert-21days**

! 请求的桶名已经存在，或被其他用户占用，请重新输入。

命名规则:

- 需全局唯一，不能与已有的任何桶名称重复。
- 删除桶或并行文件系统后，需要等待30分钟才能创建同名桶或并行文件系统。
- 长度范围为3到63个字符，支持小写字母、数字、中划线 (-)、英文句号 (.)。
- 禁止两个英文句号 (.) 或英文句号 (.) 和中划线 (-) 相邻，禁止以英文句号 (.) 和中划线 (-) 开头或结尾。
- 禁止使用IP地址。
- 如果名称中包含英文句号 (.)，访问桶或对象时可能会进行安全证书校验。

---

存储类别

**标准存储** **低频访问存储** **归档存储**

适用于有大量热点文件或小文件，且需要频繁访问（平均一个月多次）并快速获取数据的业务场景。

上传对象时，对象默认与桶的存储类别相同，也可以根据适用场景修改。 [了解更多](#)

桶策略

**私有** **公共读** **公共读写**

任何用户都可以对桶内对象进行读操作，仅桶拥有者可以进行写操作。

默认加密 ?

**开启** **关闭**

**推荐** 密钥管理全免费，核心数据更安全。

归档数据直读

**开启** **关闭**

数据冗余存储策略选择“单AZ 存储”

桶名称自己取一个不重复的名称即可

桶策略采用公共读

其他单策略使用默认单即可

点击创建按钮之后，OBS 桶就创建完毕了

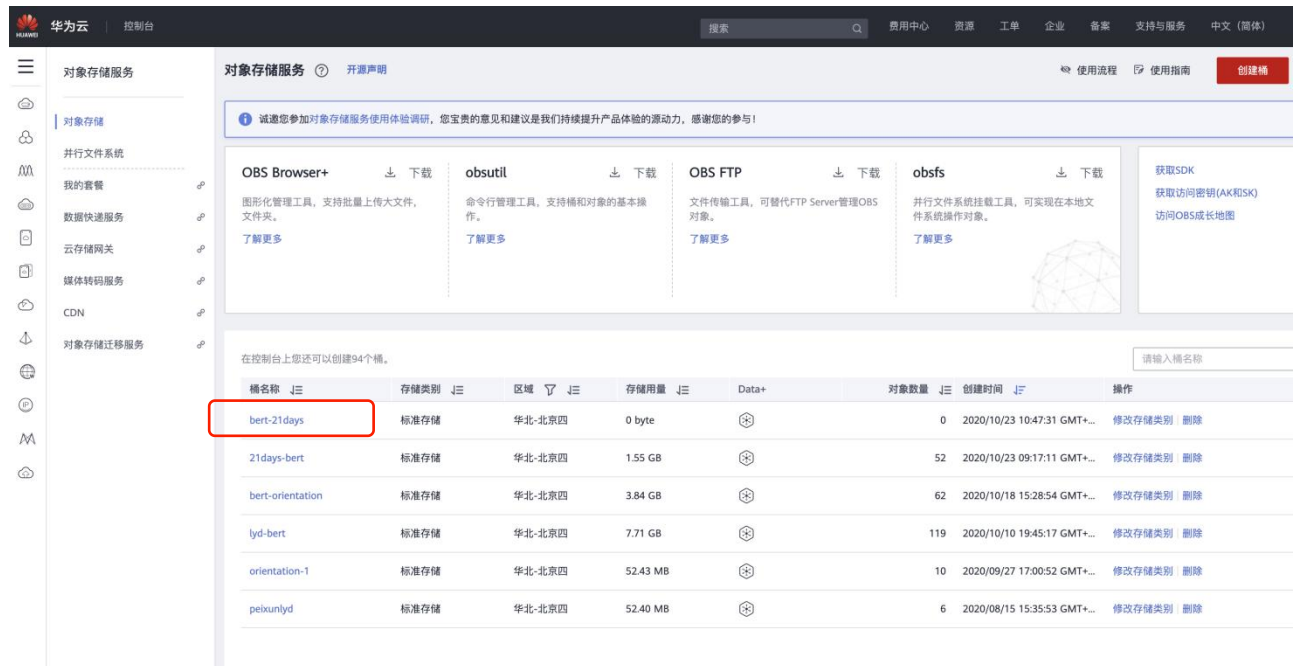
#### 4. 上传文件

这一步与体验作业的区别仅在于进阶作业不上传 **tnews ckpt**，而是上传 **bert\_base ckpt** 到代码根目录下

如果之前已经完成了体验任务，我们需要把之前上传的 **tnews ckpt** 文件删除

这一步我们需要在刚刚创建的 OBS 桶内新建一些文件夹，并且上传我们刚刚下载好的文件

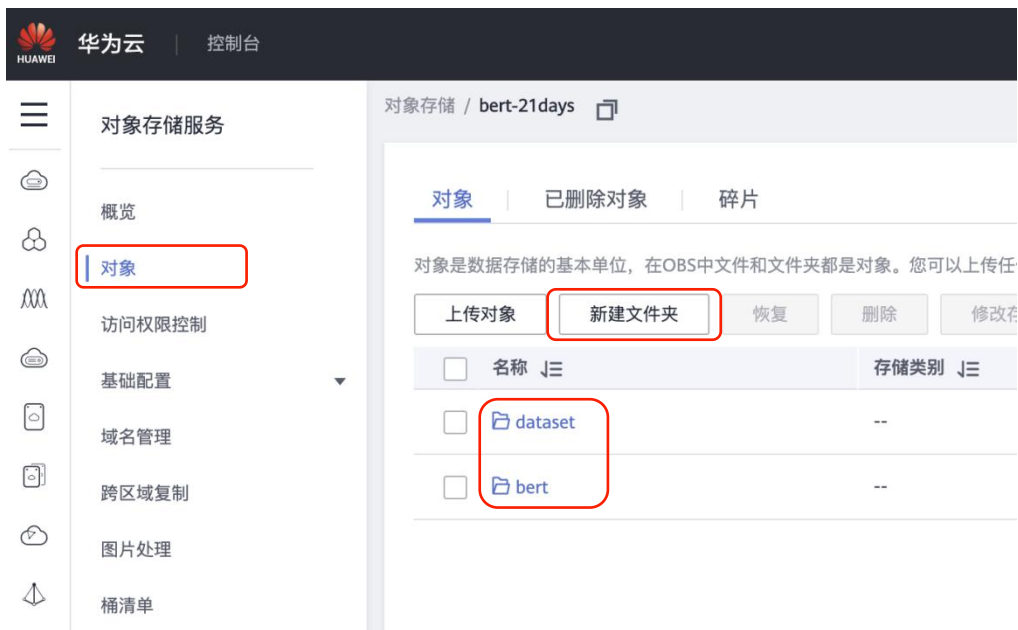
在 OBS 主页面点击进入刚才创建出来的桶



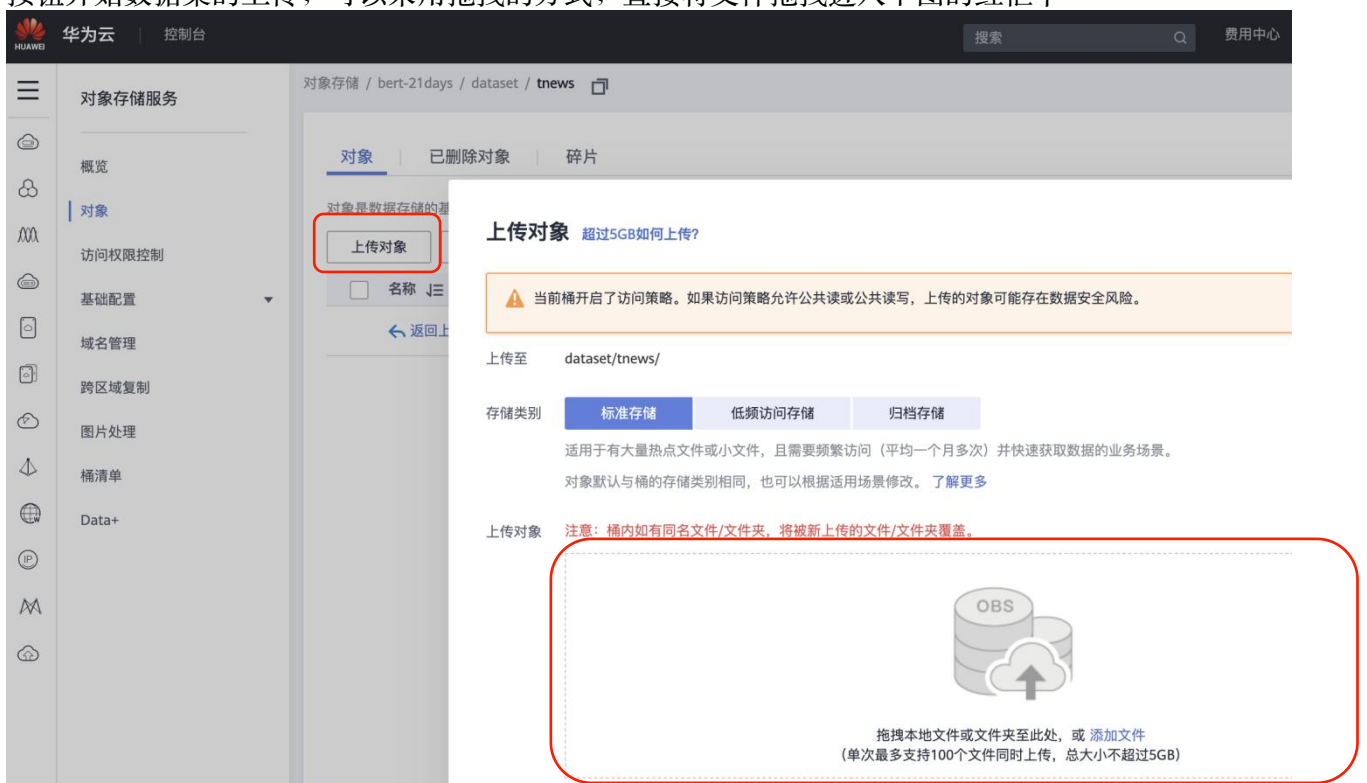
The screenshot shows the Huawei Cloud OBS console. The left sidebar contains navigation options like '对象存储' (Object Storage) and '并行文件系统' (Parallel File System). The main area displays the '对象存储服务' (Object Storage Service) page. At the top, there are links for 'OBS Browser+', 'obsutil', 'OBS FTP', and 'obsfs'. Below this, a table lists existing buckets. The bucket 'bert-21days' is highlighted with a red box.

桶名称	存储类别	区域	存储用量	Data+	对象数量	创建时间	操作
bert-21days	标准存储	华北-北京四	0 byte		0	2020/10/23 10:47:31 GMT+...	修改存储类别 删除
21days-bert	标准存储	华北-北京四	1.55 GB		52	2020/10/23 09:17:11 GMT+...	修改存储类别 删除
bert-orientation	标准存储	华北-北京四	3.84 GB		62	2020/10/18 15:28:54 GMT+...	修改存储类别 删除
lyd-bert	标准存储	华北-北京四	7.71 GB		119	2020/10/10 19:45:17 GMT+...	修改存储类别 删除
orientation-1	标准存储	华北-北京四	52.43 MB		10	2020/09/27 17:00:52 GMT+...	修改存储类别 删除
peixunlyd	标准存储	华北-北京四	52.40 MB		6	2020/08/15 15:35:53 GMT+...	修改存储类别 删除

在左侧的弹出菜单中，点击“对象”，然后在右边点击新建文件夹,在根目录下新建两个文件夹：**bert**和**dataset**

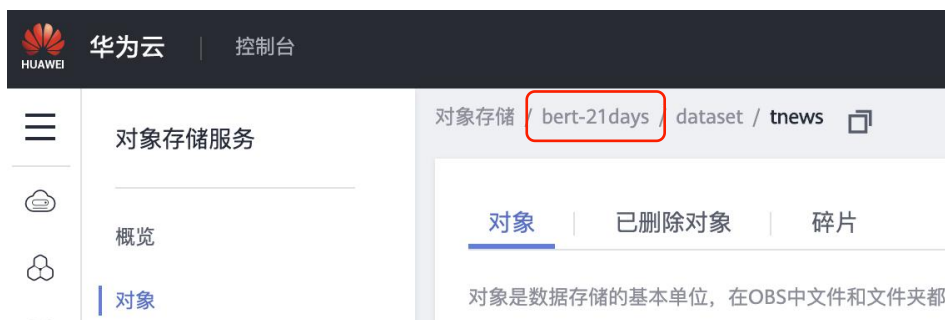


点击进入 **dataset** 文件夹，再创建一个二级文件夹 **tnews**，并且进入 **tnews** 文件夹，点击“上传对象”按钮开始数据集的上传，可以采用拖拽的方式，直接将文件拖拽进入下图的红框中



在完成了数据集的上传之后，回到桶的根目录，并且点击进入 **bert** 文件夹，同样的方法，开始上传解压缩之后的源代码文件，这里推荐使用拖拽的方式，把解压之后的文件和文件夹全选，然后拖拽进入上图的框中





最后上传 bert\_base ckpt 文件，这里我们直接把 bert\_base ckpt 文件上传到 bert 目录即代码根目录下，由于 bert\_base ckpt 这个文件比较大，所以上传时间会略微长一些

如果之前已经完成了体验任务，我们需要把之前上传的 tnews ckpt 文件删除

<input type="checkbox"/>	tnews-3_3335.ckpt	标准存储	1.14 GB	未加密	--	2020/10/23 13:58:30 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	bert_base.ckpt	标准存储	392.48 MB	未加密	--	2020/10/23 13:38:39 GMT+08:00	复制路径   删除   配置对象策略   下载为   修改存储类别
<input type="checkbox"/>	vocab.txt	标准存储	106.97 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	tokenization.py	标准存储	11.93 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	run_pretrain.py	标准存储	11.79 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾

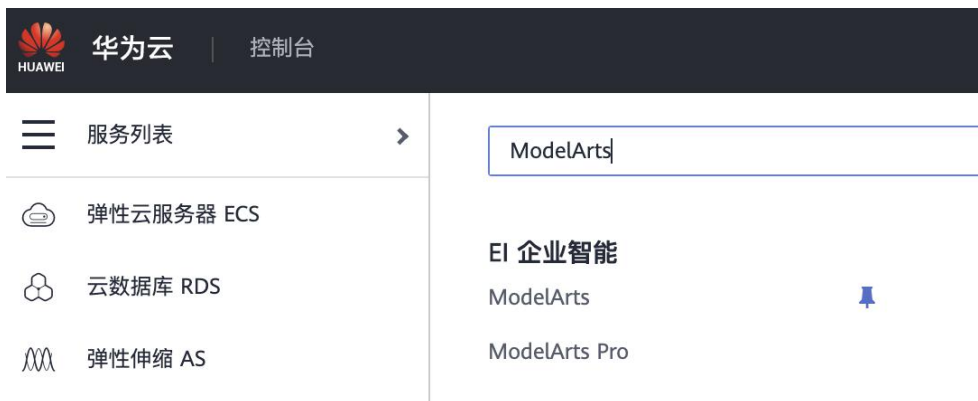
上传全部完成之后，截图如下：

对象是数据存储的基本单位。在OBS中文件和文件夹都是对象。您可以上传任何类型（文本、图片、视频等）的文件，并在桶中对这些文件进行管理。 <a href="#">了解更多</a>							
上传对象		新建文件夹		恢复	删除	修改存储类别	
<input type="checkbox"/>	名称	存储类别	大小	加密状态	恢复状态	最后修改时间	操作
← 返回上一级							
<input type="checkbox"/>	src	--	--	--	--	--	分享   复制路径   更多 ▾
<input type="checkbox"/>	scripts	--	--	--	--	--	分享   复制路径   更多 ▾
<input type="checkbox"/>	bert_base.ckpt	标准存储	392.48 MB	未加密	--	2020/10/23 13:38:39 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	vocab.txt	标准存储	106.97 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	tokenization.py	标准存储	11.93 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	run_pretrain.py	标准存储	11.79 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	README.md	标准存储	9.31 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	squadeval.py	标准存储	4.20 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	pretrain_eval.py	标准存储	5.71 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	finetune.py	标准存储	9.33 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	evaluation.py	标准存储	9.67 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾
<input type="checkbox"/>	convert_example.py	标准存储	4.10 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	下载   分享   更多 ▾

至此，上传部分结束

## 5. 这一步与体验作业类似，在参数输入方面有小部分区别

提交 ModelArts 训练作业，类似我们进入 OBS 的方式，在左侧服务列表的弹出菜单中，在搜索框中输入 modelarts，并且选择下面的“ModelArts”



进入 ModelArts 界面之后，点击左侧“训练管理”，在下拉菜单中点击“训练作业”，进入训练作业界面后，点击中上方的“创建”按钮



在创建训练任务的界面中，名称同样的，自己取一个不重复的名称即可

★ 名称  ✓

下面的参数配置中，以下几个参数需要配置：

“算法来源”选择“常用框架”，

“AI 引擎”选择“Ascend-Powered-Engine”和“MindSpore-0.5-python3.7”

“代码目录”选择之前在 obs 建立的 bert 目录即代码根目录

“启动文件”选择代码根目录下的 finetune.py

“数据来源”选择“数据存储位置”

在“数据存储位置”中之前在 obs 中建立的 dataset 文件夹下的 tnews 文件夹



“训练输出位置”这里我们直接把输出位置选择为 bert 目录下即代码根目录下，直接把训练好的 tnews 模型输出到代码根目录，方便我们做 evaluation，这样做 evaluation 时就不再需要把 tnews ckpt 文件从 output 文件夹拷贝过来了

“作业日志路径”可以在 obs 根目录下创建一个“log”文件夹用来存放日志输出

“资源池”选择“公共资源池”，“类型”选择“Ascend”，“规格”选择“Ascend: 1 \* Ascend 910 CPU: 24 核 96GiB”

其他参数选择默认即可

最终参数设置结果如下图所示：

一键式参数配置

如您已保存过参数配置，可单击 [这里](#)

★ 算法来源

算法管理 预置算法 **常用框架** 自定义 [MoXing手册](#)

选择常用引擎创建训练作业。

★ AI引擎 Ascend-Powered-Engine MindSpore-0.5-python3.7-aarch...

★ 代码目录 ? /bert-21days/bert/ 选择

★ 启动文件 ? /bert-21days/bert/finetune.py 选择

★ 数据来源 ?

数据集 **数据存储位置**

★ 数据存储位置 /bert-21days/dataset/tnews/ 选择

★ 训练输出位置 ? /bert-21days/bert/ 选择

请尽量选择空目录来作为训练输出路径。

运行参数 ?

train\_url = /bert-21days/bert/

data\_url = /bert-21days/dataset/tnews/

⊕ 增加运行参数

作业日志路径 ? /bert-21days/log/ 选择 清除

日志默认保存在服务，会不定期清除，请选择相应路径用来上传日志。

然后就可以提交任务了

## 6. 等待、查看结果

finetune 需要的时间比 evaluation 更长，约 10 分钟左右，同时等待的时间也主要取决于当前提交训练作业的人数，同样的，大家在提交作业之后定时回来查看状态即可

如果显示“运行成功”，即代表 finetune 运行作业已经完成，我们可以，可以点击作业名称，进入到作业内部详情，再点击“日志”，查看输出



在日志里上下拖动滑动条，找到 epoch 3 和 end exit 这两项，代表 finetune 完成

epoch: 3, step: 10005, outputs are (1.8981895, False)

[Modelarts Service Log]2020-10-23 13:58:53,337 - INFO - Begin destroy FMK processes

[Modelarts Service Log]2020-10-23 13:58:53,338 - INFO - FMK of device2 (pid: [152]) has exited

[Modelarts Service Log]2020-10-23 13:58:53,338 - INFO - End destroy FMK processes

[ModelArts Service Log]modelarts-pipe: total length: 88438

=== begin proc exit ===

=== begin stop slogd ===

=== end pro exit ===

同时我们回到 OBS，bert 目录即代码根目录下面，检查是否生成了 tnews ckpt 文件

<input type="checkbox"/>	名称	存储类别	大小	加密状态	恢复状态	最后修改时间	操作
<a href="#">返回上一级</a>							
<input type="checkbox"/>	src	--	--	--	--	--	<a href="#">分享</a> <a href="#">复制路径</a> <a href="#">更多</a>
<input type="checkbox"/>	scripts	--	--	--	--	--	<a href="#">分享</a> <a href="#">复制路径</a> <a href="#">更多</a>
<input type="checkbox"/>	tnews-graph.meta	标准存储	1.01 MB	未加密	--	2020/10/23 14:04:17 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	tnews-3_3335.ckpt	标准存储	1.14 GB	未加密	--	2020/10/23 14:04:17 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	bert_base.ckpt	标准存储	392.48 MB	未加密	--	2020/10/23 13:38:39 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	vocab.txt	标准存储	106.97 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	tokenization.py	标准存储	11.93 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	run_pretrain.py	标准存储	11.79 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	README.md	标准存储	9.31 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	squadeval.py	标准存储	4.20 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	pretrain_eval.py	标准存储	5.71 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	finetune.py	标准存储	9.33 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	evaluation.py	标准存储	9.67 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>
<input type="checkbox"/>	convert_example.py	标准存储	4.10 KB	未加密	--	2020/10/23 11:30:05 GMT+08:00	<a href="#">下载</a> <a href="#">分享</a> <a href="#">更多</a>

可以看到，finetune 生成的 tnews ckpt 文件已经存储到了代码根目录下面，接下来的 evaluation 任务与体验作业是一样的，我们直接可以使用这个生成的 tnews ckpt 文件去创建 evaluation 任务并提交即可，参数也与之之前一致。

学员可以截图 finetune 日志和 obs 桶里生成的 tnews ckpt 文件并上传完成作业