Paddlepaddle OCR安装文档（Mac版）

1. 安装方法

首先根据GitHub中的PaddlePaddleOCR官方指导文件中的官方文档教程中进行安装：

<https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.3/doc/doc_ch/quickstart.md>

1. **环境准备：**

<https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.3/doc/doc_ch/environment.md>

我是已经创建并安装好Anaconda环境

1. **创建好环境后，安装paddlepaddle**

根据情况，这里我选择安装CPU版本：

python3 -m pip install paddlepaddle -I https://mirror.baidu.com/pypi/simple

1. **之后根据快速上手教程：**

<https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.3/doc/doc_ch/quickstart.md>

进行快速开始：

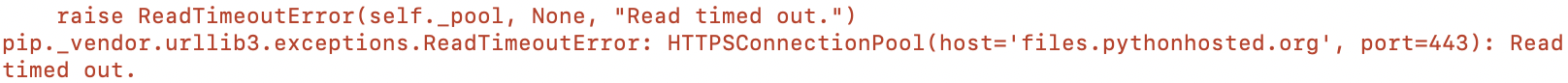
* 首先：安装PaddleOCR whl包，输入命令：

pip install "paddleocr>=2.0.1" # 推荐使用2.0.1+版本

这里因为下载时间过长，报错：

图片包含 文本

描述已自动生成…



分析是因为下载时间过长不响应导致的问题，处理方法：



将原命令的源换为：

pip install "paddleocr>=2.0.1" -i <https://mirror.baidu.com/pypi/simple>

之后安装顺利且速度得到保障。

* 其次，可以使用简单的命令行尝试效果，例如（注意，如果不使用gpu在之后的所有使用中都要进行声明）：

paddleocr --image\_dir ./imgs/11.jpg --use\_angle\_cls true --use\_gpu false

* 之后，根据需要安装拌面分析模块：

paddleocr --image\_dir ./imgs/11.jpg --use\_angle\_cls true --use\_gpu false

之后就可以设置好命令后，查看使用效果了。

1. 本地使用样例

在本地使用paddlepaddleOCR有两种选择，

1. 命令行输入
2. .py脚本使用

以下，为我使用脚本文件（方法二）对于中文文本的效果检测

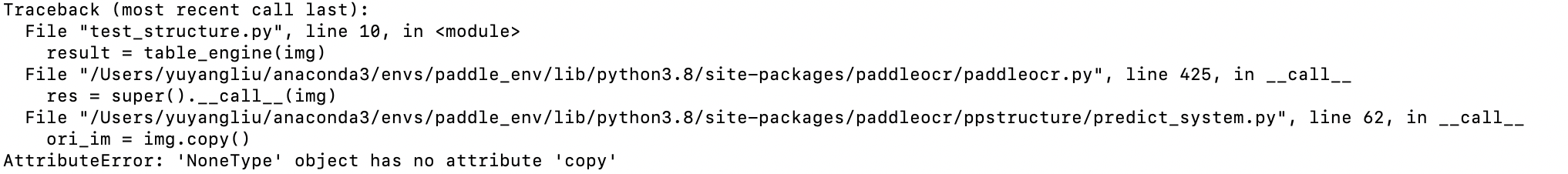
根据下链接中的样例：

<https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.3/doc/doc_ch/quickstart.md>

遇到问题：

使用命令行或者脚本都可以使用paddlepaddleocr中的功能

在使用pp-structure的时候存在一些例如下图的问题：



处理方式：

最好的方式就是把相对路径改为绝对路径，原代码即可正常运行



一定注意引号中不要有多余空格！

* **对于含有表格的图片：**

报纸上的文字

描述已自动生成

当图片是上图样式时，使用PP-structure脚本，得到的

**使用版面分析得到的文本效果图，如图：**

报纸上的表格

中度可信度描述已自动生成

**效果分析：**

识别率一般，文章的标题部分暂时无法识别到。在图片中文字信息的对应效果良好，但是在其配套生成的txt文件中文字部分不能有序排列，具体可以在文件夹中的res.txt中查看

**生成的excel如图：**

图形用户界面, 应用程序, 表格, Excel

描述已自动生成

**效果分析：**

生成表格效果相对较好（相对少数会有表格中内容被合并的情况出现）。

* **对于纯文本内容的图片**

原图：



**使用全分析脚本后得到效果：**

报纸上的文字

描述已自动生成

**效果分析：**

得到的结果识别率较好，但是识别后读取生成的txt文件中的文本信息与原始文本信息的对应程度一般，主要原因是因为不同字号的标题导致文本内容分块处理的时候顺序变乱。

具体信息可以在res.txt文件中查看。

1. 结论

综上所述，PaddlePaddleOCR的

优点：

1. 文档健全，便于本地部署，在有无gpu情况下都可以运行
2. 相关代码齐全，可以进行train和finetune
3. 可以将原始图片中的图片、表格、文字，分别进行提取，并且单独形成文件存储

缺点：

1. 目前尚未解决双栏读取的问题（可以尝试攻克）
2. 识别率一般，位于角落部分的内容有时无法识别