cspdataset

一、工具说明

二、数据集格式说明

voc格式

coco格式

三、使用说明

1. pip 安装

2. 命令行式使用

3. 参数说明

4. 示例

5. 补充说明

一、工具说明

- 处理 COCO, VOC 格式的数据集
- 包含数据集互转、切分、检查

二、数据集格式说明

voc格式

VOC
|
|---------Annotations
|-------xxx.xml
|
|---------ImageSets
|-------Main
|-------trainval.txt
|---------JPEGImages
|-------xxx.jpg

|

|---------labels.txt

注：图片文件名与对应标注数据xml文件文件名相同

coco格式

COCO

|

|---------Annotations

|-------train.json

|

|---------Images

|-------train

|-------xxx.jpg

三、使用说明

1. pip 安装

```
pip install cspdataset
```

2. 命令行式使用

```
datatools COMMAND
使用帮助
datatools --help
```

```
脚本调用
import datatools
```

3. 参数说明

```
usage: datatools.py [-h]
                    --form {coco,COCO,VOC,voc}
                    --data_dir DATA_DIR
                    [--split]
```

```
                [--division_ratio DIVISION_RATIO [DIVISION_RATIO ...]]
                [--transform]
                [--output_transform OUTPUT_TRANSFORM]
                [--check]
                [--output_check OUTPUT_CHECK]
                [--labelfile LABELFILE] [--eva EVA]
```

```
tools for voc/coco dataset
optional arguments:
-h, --help                                  查看帮助信息
--form {coco,COCO,VOC,voc}                  指定待转换数据集格式
--data_dir DATA_DIR                         数据集地址目录
--split                                     执行数据集切分
--division_ratio DIVISION_RATIO [DIVISION_RATIO ...]  数据集切分比例
--transform                                 执行数据集转换
--output_transform OUTPUT_TRANSFORM         转化后数据集地址目录
--check                                     执行数据集检查
--output_check OUTPUT_CHECK                 数据集检查结果保存目录
--labelfile LABELFILE                       指定voc格式标签检查文件，如labels.txt
```

4. 示例

- 数据集转化

```
datatools --transform --form voc --data_dir /your/voc/folder --output_transform
/the/folder/tosave/coco
```

- 数据集切分

```
datatools --split --form voc --data_dir /your/voc/folder --division_ratio 0.9 0.8
#0.9 训练集占总数据集比例，0.1用作训练后的模型评估
#0.8 训练集中用于训练的比例，0.2用于训练过程中的评估
```

- 数据集检查

```
datatools --check --form voc --data_dir /your/voc/folder --labelfile /the/label.txt --
output_check /the/folder/tosave/check/result
```

5. 补充说明

数据集检查

包括对原始图片、标注数据（xml或json文件）的检查

图片：图像是否损坏，无法加载

标注文件：没有图片尺寸；图片尺寸为0；类别错误；坐标越界；没有标注坐标；坐标信息为负数；
标注文件无坐标信息（缺bbox）

输出：包含错误文件文件名的 .txt 文件

数据集切分

voc格式：在VOC/ImageSets/Main 生成 trainval.txt、train.txt、val.txt、test.txt，分别为训练数据集、训练数据集中用作训练部分、训练数据集中用作评估部分、测试数据集。内容为图片文件名

cooc格式：在coco/Annotations/生成 trainval.json、train.json、val.json、test.json，在coco/Images目录下生成存放应图片文件的trainval、train、val、test文件夹

labels.txt 格式

hat 安全帽

head 人头