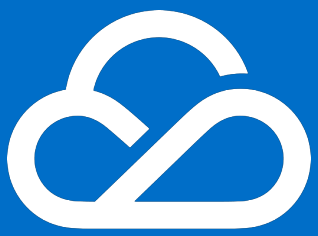
机智机器学习

参数说明

产品文档



腾讯云

【版权声明】

【商标声明】

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

文档声明 2
数据格式 4
参数设置 11

数据格式

1 LDA

1.1 输入数据

一行一个文本，utf-8或gbk编码。例如：

习近平在英国议会讲话：中英关系创下多个第一

国家旅游局接入内地游客香港被殴案调查

美国中考中国区成绩被取消 因高分太多

1.2 输出数据

含义如下：

*.docDistOnTopic: 主题到文档聚类结果(文档主题分布取top1)；

*.model.others: 训练中所使用的参数；

*.model.tassign: 训练集tf-idf值；

*.model.wordmap.txt: 词频统计；

*.topicDistOnDoc: 主题到文档的分布；

*.topic_words.txt: 主题到词的分布。

2 CNN

2.1 模型训练

2.1.1 输入数据

输入数据是一个压缩包，内容是所有图片和对应的索引文件，包括训练集索引文件train.txt，验证集的索引文件valid.txt，标签索引文件label.txt。

压缩包的内容结构：

```
dataset/train.txt
dataset/valid.txt
dataset/label.txt
dataset/image/…
```

图片存放在dataset/image/，子目录结构由用户自定义。

train.txt和valid.txt的格式一样，第一列是图片路径（相对路径，例如image/1.jpg），第二列是图片分类的编号（从0开始递增）。

label.txt每一行是一个分类名字，对应的分类编号是行号减1，即从0开始标号。

2.2.2 输出数据

输出一个二进制的模型文件task_id.model，三个json格式的训练信息文件info.txt，stat.txt，accuracy.txt

info.txt格式：

```
{
"start_time": "2015-10-19 21:40:13", // 开始时间
"end_time": "2015-10-19 21:40:13", // 结束时间
"train_image_num": 200000, // 训练集图片数
"valid_image_num": 200000, // 验证集图片数
"category_num": 1000, // 分类数
"total_iteration": 100000, // 总迭代次数
"cur_iteration": 56000, // 当前的迭代次数
"progress": 0.56, // 训练进度，浮点数0-1
"top1_accuracy": 0.65, // 当前的top1准确率，浮点数0-1
"top5_accuracy": 0.83, // 当前的top5准确率，浮点数0-1
"stat_file": "stat.txt", // 统计信息，json格式
"detail_accuracy": "accuracy.txt", // 按照分类统计准确率信息，json格式
}
```

stat.txt格式：

```
{
"stat": [
{
"iter": 100, // 迭代次数
"top1_accuracy": 0.55, // 当前的top1准确率, 浮点数0-1
"top5_accuracy": 0.73, // 当前的top5准确率, 浮点数0-1
"loss": 3.25 // 损失函数的输出值
},
{
"iter": 200, // 迭代次数
"top1_accuracy": 0.65, // 当前的top1准确率, 浮点数0-1
"top5_accuracy": 0.83, // 当前的top5准确率, 浮点数0-1
"loss": 2.15 // 损失函数的输出值
}
]
}
```

accuracy.txt格式:

```
{
"accuracy": [
{
"index": 0, // 分类编号
"name": "宝马", // 分类名
"top1_accuracy": 0.55, // 当前的top1准确率, 浮点数0-1
"top5_accuracy": 0.73, // 当前的top5准确率, 浮点数0-1
},
{
"index": 1, // 分类编号
"name": "奔驰", // 分类编号
"top1_accuracy": 0.65, // 当前的top1准确率, 浮点数0-1
```

```
"top5_accuracy": 0.83, // 当前的top5准确率, 浮点数0-1
}
]
}
```

2.2 模型预测

2.2.1 输入数据

输入数据是一个压缩包，内容是待预测的图片和对应的索引文件test.txt。

压缩包的内容结构：

```
dataset/test.txt
dataset/image/…
```

图片存放在dataset/image/，子目录结构由用户自定义。

test.txt每一行是一个图片路径（相对路径，例如image/1.jpg）。

2.2.2 输出数据

输出明文的预测结果result.txt，训练信息info.txt。

result.txt第1列是图片路径，第2列和第3列是概率最高的图片分类名字和概率，依次类推，展示概率最高的5种结果，一共11列。

info.txt格式：

```
{
"start_time": "2015-10-19 21:40:13", // 开始时间
"end_time": "2015-10-19 21:40:13", // 结束时间
```

```
"total_image_num": 100000, // 需要预测的图片数
"finished_image_num": 50000, // 完成预测的图片数
"progress": 0.5, // 训练进度，浮点数0-1
"result": "result.txt" // 预测结果文件
}
```

2.3 特征提取

2.3.1 输入数据

输入数据是一个压缩包，内容是待抽取特征的图片和对应的索引文件test.txt。

压缩包的内容结构：

```
dataset/test.txt
dataset/image/...
```

图片存放在dataset/image/，子目录结构由用户自定义。

test.txt每一行是一个图片路径（相对路径，例如dataset/image/1.jpg）。

2.3.2 输出数据

输出明文的特征文件result.txt，训练信息info.txt。

result.txt第一列是图片路径，第二列是参数列表，参数之间用逗号隔开。

info.txt格式：

```
{
"start_time": "2015-10-19 21:40:13", // 开始时间
"end_time": "2015-10-19 21:40:13", // 结束时间
```

```
"total_image_num": 100000, // 需要预测的图片数
"finished_image_num": 50000, // 完成预测的图片数
"progress": 0.5, // 训练进度，浮点数0-1
"result": "result.txt" // 预测结果文件
}
```

3 LR

3.1 输入数据

每行一条记录，字段由空格分隔。记录第1个字段为点击数，第2个字段为展示数，后面跟着由维数号和特征值组成的特征向量稀疏表示。
例如：
0 3 1 1.0 3 4.0 6 1.0
含义：点击数为0，展示数为3，x[1]=1.0, x[3]=4.0, x[6]=1.0，向量x的其余维度都为0值

3.2 输出数据

【训练输出模型文件】

只有一行，由冒号':'连接的<下标，值>对组成，空格分隔。
例如：
0:0 1:0.205509 2:0.0848729 3:0.0226618 4:-0.0138736 5:-0.0479451 6:-0.0174602
含义：权重向量的各维值分别为
w[0]=0
w[1]=0.205509
w[2]=0.0848729
w[3]=0.0226618
w[4]=-0.0138736
w[5]=-0.0479451
w[6]=-0.0174602

【预测输出】

对应输入数据中的每一行特征，输出一个[0,1]之间的浮点数

例如：

0.145311

0.00615487

0.00381483

0.00600852

0.034186

4 Word2Vec

4.1 输入数据

一行一个文本，utf-8或gbk编码。例如：

习近平在英国议会讲话：中英关系创下多个第一

国家旅游局接入内地游客香港被殴案调查

美国中考中国区成绩被取消 因高分太多

4.2 输出数据：

`*.bin：` 词向量的二进制文件；

`*.txt：` 词向量的文本可读文件；

欢迎加入腾讯机智官方交流qq群：252119476

参数设置

1 LDA

参数字段	参数名称	参数意义	默认值
Topic_size	主题数	主题聚类个数，需小于训练文档数	100
Iterate_size	迭代次数	小于500。一般500次即可收敛。	500

2 CNN

参数字段	参数名称	参数意义	默认值
Model_type	模型类型	GoogleNet或AlexNet	GoogleNet
Epoch	训练轮次	数据训练多少轮	10
Finetune	参数初始化	随机或finetune。finetune是用训练好的模型参数进行初始化，模型是用imageNet 1000种图片数据训练的。	随机

3 LR

参数字段	参数名称	参数意义	默认值
Dimension	维数	实际用于训练的特征向量维数，下标不得大于等于该值。	1000
Max_iteration_num	最大迭代次数	目标函数优化求解的最大迭代次数.	80

4 Word2Vec

参数字段	参数名称	参数意义	默认值
Layer_size	维数	词向量的表示维数	50
Model_type	模型	使用模型：Cbow/Skip-Gram	Cbow
Enable_hs	使用hs	是否使用分层softmax	1
Iterate_site	迭代次数		5
Window	前后窗口		5
Sample	采样词频阈值		0.001
Negative	负采样数		5
Min_count	最小词频	低于最小词频的词将丢弃	5
Alpha	学习率	初始学习率	0.025

欢迎加入腾讯机智官方交流qq群：252119476