



python 趣

dnn神经网络



手写平板电脑

python项



原油直播 间

python环

仿照CIFAR-10数据集格式，制作自己的数据集

2016-04-05 基陆伯 阅 12 分享： 微信 转藏到我的图书馆

前一篇博客：[C/C++ 图像二进制存储与读取](#)中，已经讲解了如何利用C/C++的方法存储与读取二进制图像文件，本文继续讲述如何根据[CIFAR-10](#)的格式制作自己的数据集。

所述博文与代码均已同步至GitHub：[yhleo/imageBinaryDataset](#)

主要代码文件有三个：

```
BinaryDataset.h
BinaryDataset.cpp
main.cpp
```

以 main.cpp 给出的一个小demo为例，首先指定一个原数据图片所在的文件夹：

```
1 std::string filefolder = "C:\\Samples\\train";
```

然后，自动获得该文件下的所有图片文件名：

```
1 std::vector<std::string> fileLists = binData.getFileLists(filefolder); // load file name
```

这里有一点需要说明一下，getFileLists() 是按照文件名升序顺序读取（大家都知道，文件名为字符串，comparable），文件命名最好不要以 1, 2, ..., 11, ... 这种方式存储，因为这么存，你就会发现 1 之后的文件可能不是你想的 2, 3, 4, ...，而是 11, 12, 13, ...。

如果你想按照顺序的某一堆数据是一种类别（我是这么做的，因为便于产生对应的 labels），建议使用等宽零位补齐的方式命名，例如：00001, 00002, ..., 00011, ...，那么文件读取的顺序就会如我们所设定。

总结一下实现方法（仅供参考）：

采集样本的时候可以先把类别存于不同的文件夹，命名就随意吧，如果是使用一些抠图软件，也不用纠结一个手工修改成自己想要的命名（这么做工作量很大，真的很蛋疼。。。）；

每一类数据整理好后，依次将每一类的数据，用程序读取并另存一份（读取使用getFileLists()，反正是一类的，也无所谓先后顺序）：

```
1 for ( int i=0; i<fileLists.size(); i++ )
2 {
3     char* curfile = new char[128];
4     sprintf(curfile, "C:\\Samples\\class-1\\%04d.jpg", i);
5     string fileName = filefolder + "\\";
6     fileName += fileLists[i];
7     cv::Mat image = cv::imread(fileName);
8     cv::imwrite(curfile, image);
9     delete[] curfile;
10 }
```

后面的其他类别也可以这样，为了按照顺序区分，依次进行其他类别的时候，只需要在改动文件夹后，将 sprintf(curfile, "C:\\Samples\\class-1\\%04d.jpg", i); 中的第三个参数 i 改为 i+k，这里 k 是前面一类或几类的样本总数。

最后，将重新命名的文件，存在一个文件夹里，记清楚类别对应的区间范围，以便生成 labels。



基陆伯 图书馆

★★★★★

11495 馆藏 33889

TA的推荐

TA的最新馆藏

永远成功的秘密，就是每天淘汰自己
我们将永生还是灭绝？人工智能很...
我们将永生还是灭绝？人工智能很...
[转] 赞美的大能
他们还不信我要到几时呢？
基督徒的委身 【】



推荐阅读

更多

BetaCat 的前生后世
揪出bug！解析调试神经网络的技巧
深度学习计算模型中“门函数（Ga...
简易的深度学习框架Keras代码解析...
国外公司开发新型移动无线网pCell...
enum的用法
再谈：义和团史实（转）
是还没有受洗，还没有正式参加某...
帧缓存



1 美亚保险官网	7 用英语介绍美国
2 美亚保险	8 led亮化照明
3 公司邮箱	9 企业邮箱申请
4 企业邮箱注册	10 中老年妈妈装
5 北京口腔医院	11 企业邮箱
6 钱爸管理财	12 英语学习

读取上述最终文件内的所有文件，接下来，生成 labels （ labels 一般用 [0, 9] 组成的整数）：

```
1 std::vector<int> image_labels(size_list, 0); // generate lables, here are all 0
```

当然，你也可以用 image_labels.push_back() 把所有的 labels 设置，但是熟悉 vector 的话，就会明白使用初始化长度，比那种做法更加高效（可以阅读本人的博客：[C++ 容器（一）：顺序容器简介](#)）。然后就相应地修改某些索引区间内的 label 值：

```
1 for ( int i=0; i<count_class_k; i++ )
2     image_labels[i] = 1;
```

都准备好后，就可以开始生成想要的二进制文件了：

```
1 std::string binfile = "C:\\Samples\\train.bin";
2 binData.images2BinaryFile( filefolder, fileLists, image_labels, binfile );
```

到这里，已经制作好了二进制数据集，我很懒，想直接基于 tensorflow/models/image/cifar10 模块的源码跑我定义的数据集，想想只要跟 cifar10 数据集类似，那肯定没什么问题，下面是官网下载的 cifar-10-binary.tar 解压后内容：

学习 (E) > tensorflow > cifar-10-batches-bin

名称	修改日期	类型	大小
batches.meta	2009/6/5 3:44	TXT 文件	1 KB
data_batch_1.bin	2009/6/5 3:36	BIN 文件	30,010 KB
data_batch_2.bin	2009/6/5 3:37	BIN 文件	30,010 KB
data_batch_3.bin	2009/6/5 3:38	BIN 文件	30,010 KB
data_batch_4.bin	2009/6/5 3:37	BIN 文件	30,010 KB
data_batch_5.bin	2009/6/5 3:35	BIN 文件	30,010 KB
readme	2009/6/5 4:46	Chrome HTML D...	1 KB
test_batch.bin	2009/6/5 3:36	BIN 文件	30,010 KB

这份数据集比较大，训练样本有 50000，测试样本 10000（我的数据集并没有这么大，但是又有什么关系呢！）。

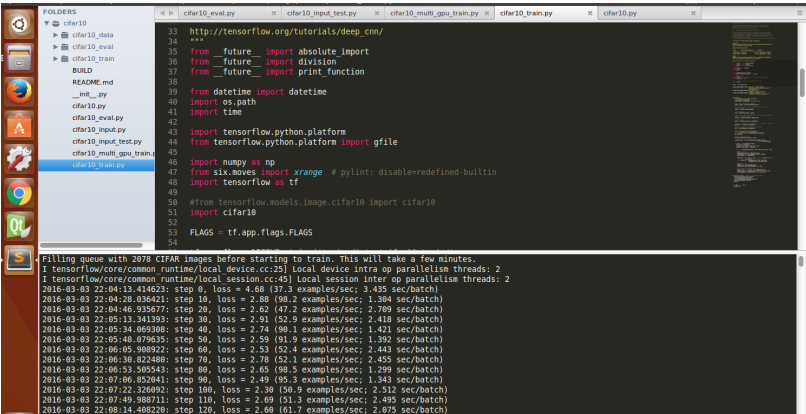
看，这是我的数据集：

名称	修改日期	类型	大小
batches.meta	2016/3/3 21:28	TXT 文件	1 KB
data_batch_1.bin	2016/3/3 21:14	BIN 文件	15,594 KB
readme	2016/3/3 21:16	TXT 文件	1 KB
test_batch.bin	2016/3/3 20:51	BIN 文件	997 KB

是不是很迷你~

然后，将 tensorflow/models/image/cifar10 模块的拷贝中的部分参数修改成为适合自己数据集的，就OK了~

献上运行截图（训练测试集有 5196 张样本，所以 $5196 \times 0.4 = 2078$ ）：



训练了两天，跑完后，评估精度为：0.896。

转藏到我的图书馆 献花（0） 分享： 微信

来自：基陆伯 > 《Tensorflow》 以文找文 | 举报

上一篇：TensorFlow CNN 测试CIFAR-10数据集
下一篇：Tensorflow MNIST 数据集测试代码入门

猜你喜欢



霸业传奇



六房间直播



角色扮演页游



秀色直播间



现货白银



传奇私



原油分析师



文档管理系统



现货贵金属




原油交易点差

类似文章


- 不看这篇日志也许会节省你十分钟，但是却...
- 浓汤老豆腐——不放调料也足够鲜咸的秘诀....
- 《三教图》赏析
- 神秘的海底世界
- 专题集邮 >> 毛泽东诗词专题
- 成人版西游记(强烈推荐)
- 文化杂谈视频集(1000多部)
- 浏览人数最多的照片

精选文章


- 百年老卤店绝密配方 各种卤味齐全！
- 短靴这样配，显高显腿长！
- QQ号被盗的解决办法
- 50碗形形色色的陕西面
- 激发大脑之音
- 大道理，值得一看
- 一万年来，中国人征服了哪些食物
- 视频教程《跟我学英语》 80集全




自我护理能力



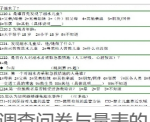
六房间直播




人有三个错误不能




《算命字典》举例



调查问卷与量表的



lol竞猜的首页



中国居民膳食营养



霍兰德职业兴趣测



苏果lol的首页



lol职业联赛2f的首

- 1 股市明日必涨停的3只牛股!

2 让你20天成为中医脉诊高手

3 一台电脑在家月入3万元
- 1 美亚保险官网

2 美亚保险

3 公司邮箱

4 北京口腔医院

5 企业邮箱注册

6 英语学习

发表评论：
请 登录 或者 注册 后再进行评论 社交帐号登录：



太阳公元二手房



拓展项目



金赐贵金属

关闭