Tensorflow深度学习之二十：CIFAR-10数据集介绍

原创 子为空 2017-12-13 16:55:16 16154 收藏 16 版权

分类专栏： 深度学习 Tensorflow 文章标签： 深度学习 数据 Tensorflow CIFAR-10

一、CIFAR-10

CIFAR-10数据集由10类32x32的彩色图片组成，一共包含60000张图片，每一类包含6000图片。其中50000张图片作为训练集，10000张图片作为测试集。

CIFAR-10数据集被划分成了5个训练的batch和1个测试的batch，每个batch均包含10000张图片。测试集batch的图片是从每个类别中随机挑选的1000张图片组成的,训练集batch以随机的顺序包含剩下的50000张图片。不过一些训练集batch可能出现包含某一类图片比其他类的图片数量多的情况。训练集batch包含来自每一类的5000张图片，一共50000张训练图片。

下图显示的是数据集的类，以及每一类中随机挑选的10张图片：

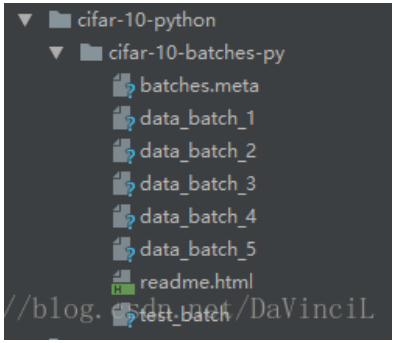
二、CIFAR-10数据集解析

官方给出了多个CIFAR-10数据集的版本，以下是链接：

Version	Size	md5sum
CIFAR-10 python version	163 MB	c58f30108f718f92721af3b95e74349a
CIFAR-10 Matlab version	175 MB	70270af85842c9e89bb428ec9976c926
CIFAR-10 binary version (suitable for C programs)	162 MB	c32a1d4ab5d03f1284b67883e8d87530

此处我们下载python版本。

下载完成后，解压，得到如下目录结构的文件夹：



其中：

名称	作用
batches.meta	程序中不需要使用该文件
data_batch_1	训练集的第一个batch，含有10000张图片
data_batch_2	训练集的第二个batch，含有10000张图片
data_batch_3	训练集的第三个batch，含有10000张图片
data_batch_4	训练集的第四个batch，含有10000张图片
data_batch_5	训练集的第五个batch，含有10000张图片
readme.html	网页文件，程序中不需要使用该文件
test_batch	测试集的batch，含有10000张图片

上述文件结构中，每一个batch文件包含一个python的字典（dict）结构，结构如下：

名称	作用
b’data’	是一个10000x3072的array，每一行的元素组成了一个32x32的3通道图片，共10000张
b’labels’	一个长度为10000的list，对应包含data中每一张图片的label
b’batch_label’	这一份batch的名称
b’filenames’	一个长度为10000的list，对应包含data中每一张图片的名称