Cifar-10 与 MNIST 训练集

Dai Jialun

July 12, 2015

1. Cifar10

CIFAR-10 数据集由 10 类总共 60000 张 32×32 的彩色图像组成，其中每一类都有 6000 张图像。在 60000 张图像中，有 50000 张的训练图像与 10000 张的测试图像。

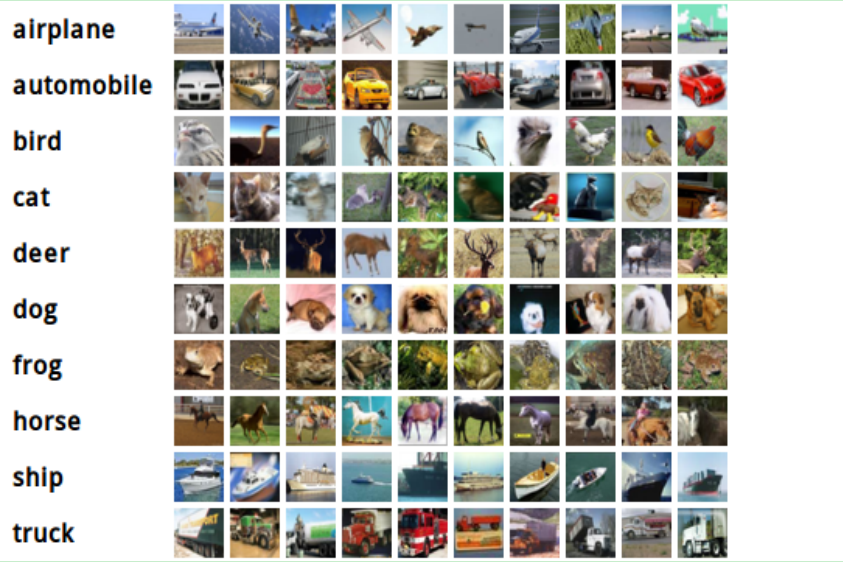


Figure 1: cifar10

数据集被分成 5 个训练组和 1 个测试组，每一组都有 10000 张图像。其中，测试组从 10 类中的每一类随机挑选出的 1000 张图像，即测试组中每一类图像数量相同。训练组从所剩下的图像中随机挑选，因此在某些训练组中，存在某一类中的图像可能比其他类图像多，即某个训练组中每一类图像数量可能不同。但是从整体上看，在 5 组训练组的合计中，每一类有 5000 张图像。

另外在数据集中，每一类的图像都是独立存在，不存在有一张图像在两个类别中出现的情况。

Cifar10 数据集的版本：

- CIFAR-10 python version
- CIFAR-10 Matlab version

- CIFAR-10 binary version (suitable for C programs)

Python/Matlab 版（cifar-10-batches-py / cifar-10-batches-mat）：在这里介绍了数据集的 Matlab 版本。Python 版本的结构与其是相同的。

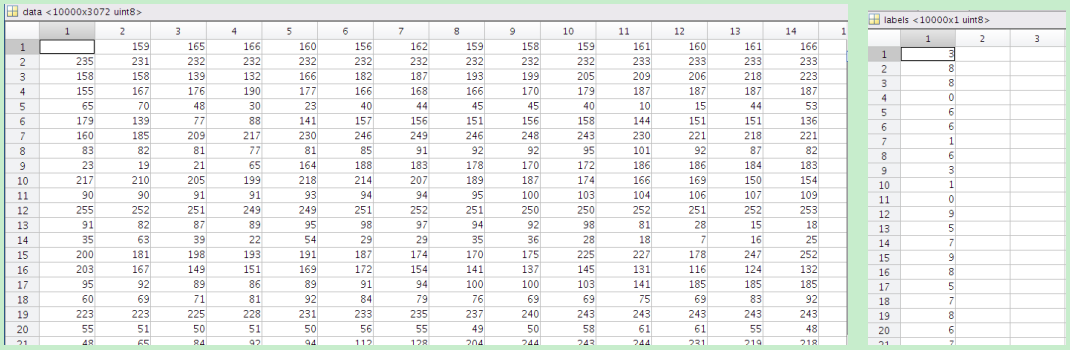
在下载的文档中，包含了 data_batch_1，data_batch_2，data_batch_3，data_batch_4，data_batch_5，以及 test_batch。通过以上方式加载后，每一个组文件将会生成一个包含以下文件的文件夹：

data 一个类型为 unit8 的 10000×3072 的 numpy 矩阵。已知每一张图像都是 3 通道，大小为 32×32。这个矩阵的每一行存储一个 32×32 图像的 3 个不同通道的像素值。每一行表示一张图像，刚开始的 1024 列（32×32）表示红色通道像素值，接下来的 1024 列（32×32）表示绿色通道像素值，最后的 1024 列（32×32）表示蓝色通道像素值。图像是行优先存储的。

labels 一个列表，一列包含了 10000 个数字，范围为 0~9。在列表中，第 i 个数字表示在数组数据中，第 i 张图像的标注。

数据集包含了另外一个 batches.meta 文件。它也包含了一个对象，有以下内容：

label_names 一个含有 10 个目标的列表，其描述了标注数字对应具体的类别名字。例如，label_names[0]=“airplane”



(a) (b)

Figure 2: ILSVRC 2013

二进制版本的数据集（cifar-10-batches-bin）：

二进制版本包含了 data_batch_1.bin，data_batch_2.bin，data_batch_3.bin，data_batch_4.bin，data_batch_5.bin，以及 test_batch_bin。每个文件的格式如下：

```
<1 × label> <3072 × pixel>
...
<1 × label> <3072 × pixel>
```

label_names <10x1 cell>

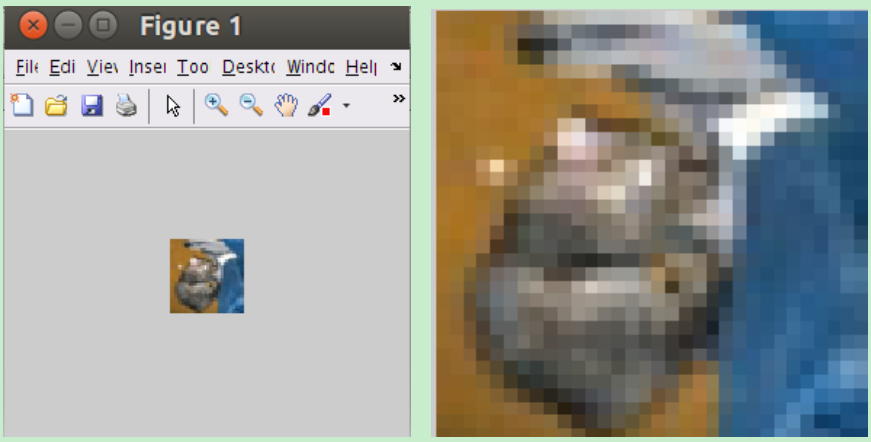
	1	2	3
1	airplane		
2	automobile		
3	bird		
4	cat		
5	deer		
6	dog		
7	frog		
8	horse		
9	ship		
10	truck		
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Figure 3: cifar10

也就是说，刚开始的第一列是训练或测试图像的标注，是一个范围在 0~9 之间的数字。接下来的 3072 列是图像 3 通道的像素值。最前面的 1024 列（32×32）是红色通道的像素值，接下来的 1024 个数（32×32）是绿色通道的像素值，最后的 1024 个数（32×32）是蓝色通道的像素值。这些值是按行优先排列的。

每一个文件包含 10000 个这样像 3073 个列的图像行，即为 10000×3073。

另外，还有一个 batches.meta.txt 的文件。这是一个 ASCII 文件，表示 0~9 之间的数字对应的各个类别名字。这仅仅是一个含有 10 个类别名字的列表，每一个一行。



(a) (b)

Figure 4: ILSVRC 2013

2. MNIST

MNIST 数据集是手写数字图像，其训练集有 60000 张图像，测试集有 10000 张图像。这只是一个更大的数据集 NIST 的一个子集。这些数字图像已经经过尺寸的规范化与在图像的置中心处理。

这个数据集适合一些人想用学习技术以及模式识别方法来处理真实世界的数据，而且不需要花太多功夫在预处理和格式化方面的工作上。

这个数据集总共有 4 个文件：

- **train-images-idx3-ubyte.gz** training set images (9912422 bytes)
- **train-labels-idx1-ubyte.gz** training set labels (28881 bytes)
- **t10k-images-idx3-ubyte.gz** test set images (1648877 bytes)
- **t10k-labels-idx1-ubyte.gz** test set labels (4542 bytes)

在 NIST 中的手写数字图像，都为灰度保真图，而且经过规范化处理，图像大小为 28×28。

MNIST 数据集是由 NIST 的 Special Database 3 （SD3）和 Special Database 1（SD1）所组成，二者包含了手写数字的二值图像。NIST 中最初将 SD-3 作为训练集，SD-1 作为测试集。然而，SD-3 比 SD-1 更清晰，更容易识别。主要是因为 SD-3 是从 Census Bureau 的员工处收集的，而 SD-1 是从高中生处收集的。从学习实验中得出可靠的结论要求结果与训练集的选择无关，并且测试样本是完整数据集。因此，必须通过混合 NIST 数据集来建立一个全新的数据集。

MNIST 训练集是由 SD-3 的 30000 张图像和 SD-1 的 300000 张图像组成。测试集是由 SD-3 的 5000 张图像和 SD-1 的 5000 张图像组成。训练集中所包含的图像是从大约 250 个人中收集到的。基本可以判定训练集与测试集的数字是不同的人所写的。

SD-1 是由 500 个不同的人所写的 58527 张数字图像。在 SD-1 中，数据是杂乱无章地排列的。这个排列方式与 SD-3 不同，在 SD-3 中每个人所写的数字图像单独存放。我们可用人工方法来分辨 SD-1 的数字图像由哪个人缩写。我们将 SD-1 分为两部分：由前 250 个人所写的数字图像放在训练集中，剩下的 250 个人所写数字放在测试集中。因此，我们有了两个子集，每个子集有大约 30000 张图像。新的训练集从 SD-3 中的 #0 开始抽取足够多的图像，形成一个 60000 张的训练图像。与其相似，新的测试集是由 SD-3 中的 #35000 开始，形成一个 60000 的测试图像。只有测试集中的 10000 张测试图像（从 SD-1 中的 5000 张图像和从 SD-3 中的 5000 张图像）在这里是可用。当然，60000 张的训练图像也是可用的。

MNIST 数据库的文件格式：

数据存储在一个非常简单的用来存放向量和多维矩阵的文件格式中。在文件中所有的整数都