实验 2 下载开放图片数据集

实验时长：4 小时

实验难度：一般

实验摘要：计算机视觉应用结果非常依赖数据集质量，高质量的数据集即数据量庞大，标签正确的数据集，需要耗费大量的劳动力成本，在某些特殊领域例如医学图像，其标注工作需要具备专业知识人员才能完成，这导致这些领域的带有标签的数据量稀少。因此希望能够在数据丰富且标签准确的数据集中完成计算机视觉所做工作。可以充分利用现有的开放数据资源并且能够节省大量劳动力成本。本任务将完成常用于测试图像识别领域研究的数据集 CIFAR-10、MNIST 的下载。

实验建议：了解 MNIST 数据集相关知识。

实验目标：能够使用网络渠道完成开放图片数据集的采集和展示任务。

1、下载开放图片数据集

- 1.1、下载 PASCAL VOC 数据集
- 下载页面 http://host.robots.ox.ac.uk/pascal/VOC/databases.html#VOC2005_1
- 下载地址 http://host.robots.ox.ac.uk/pascal/VOC/download/voc2005_2.tar.gz
-
-

- 1.2、下载 MNIST 数据集

- 进入网址：http://yann.lecun.com/exdb/mnist/，在页面上可以找到一共 4 个文件名称，分别是训练集、训练集标签、测试集、测试集标签等 MNIST 数据集文件。

-

数据文件	大小	含义
train-images-idx3-ubyte.gz	9912422 bytes	训练数据集

数据文件	大小	含义
train-labels-idx 1-ubyte.gz	28881 bytes	训练数据集标
t10k-images-idx3-ubyte.gz	1648877 bytes	测试数据集
t10k-labels-idxl-ubyte.gz	4542 bytes	测试数据集标

- 下载解压这 4 个文件，并把解压后的文件放入新建的 MNIST_data 文件夹中。

- 

-

- 下一步

-

1.3、 认识 MNIST 数据集

MNIST 数据以一种非常简单的文件格式存储，可以看到直接下载的数据解压之后是以.idx3-ubyte、.idx1-ubyte 这种字节的形式存储，该格式是为存储向量和多维矩阵而设计。现代的计算机系统一般采用字节(Octet, 8 bit Byte)作为逻辑寻址单位。当物理单位的长度大于 1 个字节时，就要区分字节顺序(Byte Order, or Endianness)。常见的字节顺序有两种：Big Endian(High-byte first)和 Little Endian(Low-byte first)。Intel X86 平台采用 Little Endian，而 PowerPC 处理器则采用了 Big Endian。MNIST 数据集就是采用了 Big Endian 方式存储。各文件的存储文件说明如下：

训练集标签文件（(train-labels-idx1-ubyte)。

[offset]	[type]	[value]	[description]
0000	32 bit integer	0x00000801(2049)	magic number (MSB first)
0004	32 bit integer	60000	number of items
0008	unsigned byte	??	label
0009	unsigned byte	??	label
.....			
xxxx	unsigned byte	??	label

训练集图片文件(train-images-idx3-ubyte)。

[offset]	[type]	[value]	[description]
0000	32 bit integer	0x00000803(2051)	magic number
0004	32 bit integer	60000	number of images
0008	32 bit integer	28	number of rows
0012	32 bit integer	28	number of columns
0016	unsigned byte	??	pixel
0017	unsigned byte	??	pixel
.....			
xxxx	unsigned byte	??	pixel

测试集标签文件(t10k-labels-idx1-ubyte)。

[offset]	[type]	[value]	[description]
0000	32 bit integer	0x00000801(2049)	magic number (MSB first)
0004	32 bit integer	10000	number of items
0008	unsigned byte	??	label
0009	unsigned byte	??	label
.....			
xxxx	unsigned byte	??	label

测试集图像文件(t10k-images-idx3-ubyte)。

[offset]	[type]	[value]	[description]
0000	32 bit integer	0x00000803(2051)	magic number
0004	32 bit integer	10000	number of images
0008	32 bit integer	28	number of rows
0012	32 bit integer	28	number of columns
0016	unsigned byte	??	pixel
0017	unsigned byte	??	pixel
.....			
xxxx	unsigned byte	??	pixel

由于直接下载下来的数据文件不是任何标准的图像格式而是以字节的形式进行存储的，是无法通过解压或者应用程序打开的，可以通过编写程序来打开它。