# 负面评论识别

提到负面评论，我们首先要聊下水军。水军顾名思义，是指在论坛大量灌水的人员 ，他们受雇于[网络公关公司](http://baike.baidu.com/item/%E7%BD%91%E7%BB%9C%E5%85%AC%E5%85%B3%E5%85%AC%E5%8F%B8" \t "_blank)、以发帖回帖为主要手段、为雇主进行网络造势的网络人员，有专职和兼职之分。一般来讲，发帖回帖造势常常需要成百上千个人共同完成，那些临时在网上征集来的发帖的人在行内被叫做“网络水军”。2009年12月央视报道网络水军这一新兴现象之后，受到社会广泛关注，不少长期在线的网虫们纷纷加入网络水军一族。水军的大行其道，严重影响了社会舆论，尤其是有组织的针对公众人物、企业的诽谤、攻击行为，造成严重的社会影响。大型的社交媒体以及知名论坛，早期都需要雇佣大量的运营支撑人员来人工鉴别处理。是否可以使用机器学习的技术来达到一定程度的自动化识别负面评论呢？我们利用垃圾邮件识别的技术积累，尝试来完成这一任务。



1. 网络水军

本章主要以MNIST数据集为例子介绍识别数字型验证码。

本章1.1介绍验证码识别使用的数据集。

本章1.2介绍使用的特征提取方法，包括一维向量和二维向量。

本章1.3介绍使用的模型以及对应的验证结果，包括K近邻、支持向量机和深度学习。

本章演示代码请参考本书Github上的review.py文件。

## 数据集

验证码识别使用的数据集为MNIST数据集，MNIST是一个入门级的计算机视觉数据集，它包含各种手写数字图片如图1-4 所示：



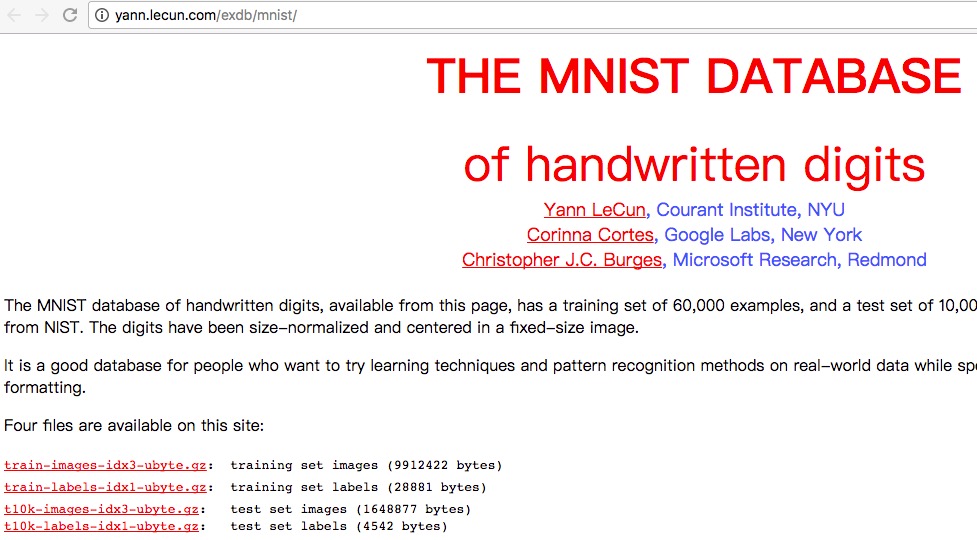
1. MNIST图片示例

它也包含每一张图片对应的标签，告诉我们这个是数字几。比如图1-4 这四张图片的标签分别是5，0，4，1。数据集包括60000行的训练数据集和10000行的测试数据集。每一个MNIST数据单元有两部分组成：一张包含手写数字的图片和一个对应的标签。每一张图片包含28X28个像素点，可以把这个数组展开成一个向量，长度是 28x28 = 784。

1. MNIST数据集合详解

|  |  |
| --- | --- |
| **文件名称** | **文件用途** |
| [train-images-idx3-ubyte.gz](http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz) | 60000个图片训练样本 |
| [train-labels-idx1-ubyte.gz](http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz) | 60000个图片训练样本的标注 |
| [t10k-images-idx3-ubyte.gz](http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz) | 10000个图片测试样本 |
| [t10k-labels-idx1-ubyte.gz](http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz) | 10000个图片测试样本的标注 |

MNIST的网址为http://ai.stanford.edu/~amaas/data/sentiment/



1. MNIST官网

也可以使用离线版的MNIST文件，下载链接为：

http://www.iro.umontreal.ca/~lisa/deep/data/mnist/mnist.pkl.gz

文件读取方式为：

import pickle

import gzip

def load\_data():

with gzip.open('./mnist.pkl.gz') as fp:

training\_data, valid\_data, test\_data = pickle.load(fp)

return training\_data, valid\_data, test\_data

## 特征提取

### 方法一：一维向量

这也是MNIST默认的特征提供方式，将28乘以28的图片转换成了维度为784的一维向量。

import tflearn.datasets.mnist as mnist  
X, Y, testX, testY = mnist.load\_data(one\_hot=True)

### 方法二：二维向量

为了适应特定模型，需要将原本的一维向量转换回原有的28乘以28的二维向量，具体方法使用数组的reshape函数即可。

X = X.reshape([-1, 28, 28, 1])  
testX = testX.reshape([-1, 28, 28, 1])

## 模型训练与验证

### 方法一：K近邻算法

特征提取方式使用一维向量，其中邻居数量n\_neighbors设置为15。

clf = neighbors.KNeighborsClassifier(n\_neighbors=15)

实例化K近邻，对训练数据集进行训练。

y\_pred = clf.predict(x\_test)  
print metrics.accuracy\_score(y\_test, y\_pred)

运行程序，经过训练，在测试数据集上预测准确率为96.28%。

KNeighborsClassifier(algorithm='auto', leaf\_size=30, metric='minkowski',

metric\_params=None, n\_jobs=1, n\_neighbors=15, p=2,

weights='uniform')

0.9628

### 方法二：支持向量机算法

特征提取方式使用一维向量，使用支持多分类的支持向量机。

clf = svm.SVC(decision\_function\_shape='ovo')

实例化支持向量机，对训练数据集进行训练。

y\_pred = clf.predict(x\_test)  
print metrics.accuracy\_score(y\_test, y\_pred)

运行程序，经过训练，在测试数据集上预测准确率为94.39%。

SVC(C=1.0, cache\_size=200, class\_weight=None, coef0=0.0,

decision\_function\_shape='ovo', degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, probability=False, random\_state=None, shrinking=True,

tol=0.001, verbose=False)

0.9439

### 方法三：深度学习算法之DNN

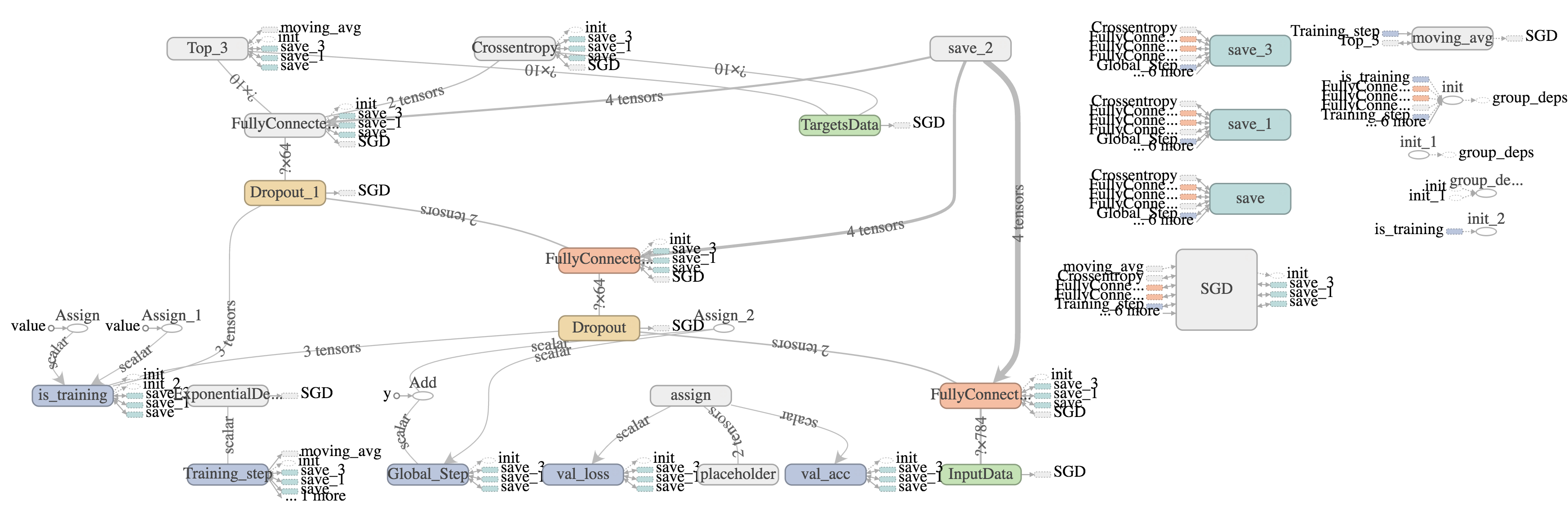
特征提取方式使用一维向量，定义输入层为维度为784的一维向量。

input\_layer = tflearn.input\_data(shape=[None, 784])

构造DNN模型，使用3个全连接层。

dense1 = tflearn.fully\_connected(input\_layer, 64, activation='tanh',  
 regularizer='L2', weight\_decay=0.001)  
dropout1 = tflearn.dropout(dense1, 0.8)  
dense2 = tflearn.fully\_connected(dropout1, 64, activation='tanh',  
 regularizer='L2', weight\_decay=0.001)  
dropout2 = tflearn.dropout(dense2, 0.8)  
softmax = tflearn.fully\_connected(dropout2, 10, activation='softmax')  
# Regression using SGD with learning rate decay and Top-3 accuracy  
sgd = tflearn.SGD(learning\_rate=0.1, lr\_decay=0.96, decay\_step=1000)  
top\_k = tflearn.metrics.Top\_k(3)  
net = tflearn.regression(softmax, optimizer=sgd, metric=top\_k,  
 loss='categorical\_crossentropy')

完整的DNN结构如图1-6 所示。



1. 用于识别MNIST的DNN结构图

实例化DNN算法并训练10轮。

# Training  
model = tflearn.DNN(net, tensorboard\_verbose=0)  
model.fit(X, Y, n\_epoch=10, validation\_set=(testX, testY),  
 show\_metric=True, run\_id="mnist")

运行程序，经过10轮训练，准确率达到了99.49%。

Training Step: 8600 | total loss: 0.28404 | time: 5.892s

| SGD | epoch: 010 | loss: 0.28404 - top3: 0.9842 | val\_loss: 0.12718 - val\_acc: 0.9949 -- iter: 55000/55000

### 方法四：深度学习算法之CNN

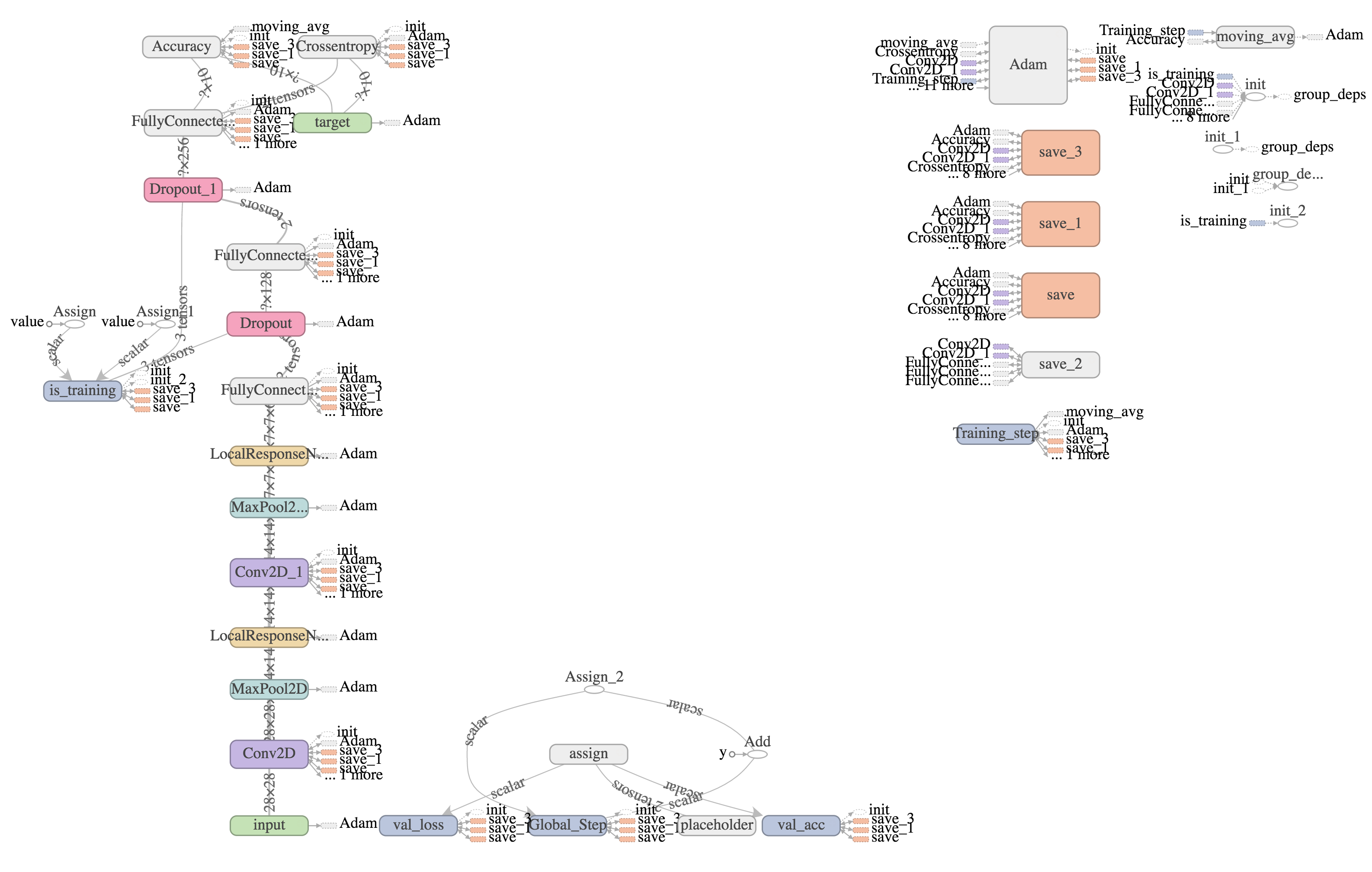
特征提取方式使用二维向量，定义输入层为28乘以28的二维向量。

network = input\_data(shape=[None, 28, 28, 1], name='input')

构造CNN模型，使用两层卷积两层全连接结构。

network = conv\_2d(network, 32, 3, activation='relu', regularizer="L2")  
network = max\_pool\_2d(network, 2)  
network = local\_response\_normalization(network)  
network = conv\_2d(network, 64, 3, activation='relu', regularizer="L2")  
network = max\_pool\_2d(network, 2)  
network = local\_response\_normalization(network)  
network = fully\_connected(network, 128, activation='tanh')  
network = dropout(network, 0.8)  
network = fully\_connected(network, 256, activation='tanh')  
network = dropout(network, 0.8)  
network = fully\_connected(network, 10, activation='softmax')  
network = regression(network, optimizer='adam', learning\_rate=0.01,  
 loss='categorical\_crossentropy', name='target')

完整的CNN结构如图1-7 所示。



1. 用于识别MNIST的CNN结构图

实例化CNN算法并训练10轮。

model = tflearn.DNN(network, tensorboard\_verbose=0)  
model.fit({'input': X}, {'target': Y}, n\_epoch=10,  
 validation\_set=({'input': testX}, {'target': testY}),  
 snapshot\_step=100, show\_metric=True, run\_id='convnet\_mnist')

运行程序，经过10轮训练，准确率达到了96.98%。

Training Step: 8600 | total loss: 0.39358 | time: 139.542s

| Adam | epoch: 010 | loss: 0.39358 - acc: 0.9366 | val\_loss: 0.12829 - val\_acc: 0.9698 -- iter: 55000/55000

## 本章小结

本章基于MNIST数据集介绍了验证码的识别方法。针对MNIST数据集，特征提取方法有一维向量和二维向量两种方法。训练模型介绍了K近邻、支持向量机以及深度学习的DNN和CNN算法，其中DNN算法表现优异，识别率达到了99%以上。

### 参考文献

* <http://china.huanqiu.com/hot/2013-09/4320971.html>
* <http://dl.sohu.com/20130911/n386353688.shtml>
* Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). [Learning Word Vectors for Sentiment Analysis.](http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf) The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).