

# 目录

1	deeplearning	1
1.1	降维	1
1.1.1	自编码	1
1.1.2	自动降噪编码	1
1.1.3	手写体数据自编码	2
1.2	稀疏编码	7
1.2.1	稀疏编码的概率表示	8
1.3	PCA	10
1.3.1	数学定义	10
1.4	kl 散度	11
1.4.1	相对熵	12
2	Tensorflow 基础	13
2.1	Tensorflow 基础函数	13
2.1.1	Variable	13
2.1.2	placeholder	13
2.1.3	batch normalization	14
2.2	常见的激活函数	14
2.2.1	relu	14
2.2.2	relu6	15
2.2.3	sigmoid	17
2.2.4	relu 和 softplus	18
2.2.5	dropout	20
2.3	CNN 常用函数	20
2.3.1	卷积函数	20
2.3.2	常见的分类函数	21
2.4	优化方法	22

2.4.1	BGD . . . . .	22
2.4.2	SGD . . . . .	22
2.4.3	momentum . . . . .	23
2.4.4	Nesterov Momentum . . . . .	23
2.4.5	Adagrad . . . . .	23
2.4.6	RMSprop . . . . .	23
2.4.7	Adam . . . . .	23
2.4.8	构造简单的神经网络拟合数据 . . . . .	25
2.5	TensorBoard . . . . .	27
2.5.1	TensorBoard Histogram Dashboard . . . . .	29
2.5.2	一个简单的例子 . . . . .	29
2.5.3	Overlay Mode . . . . .	32
2.5.4	多个分布 . . . . .	33
2.5.5	更多分布 . . . . .	35
2.5.6	poisson 分布 . . . . .	38
2.5.7	结合所有的数据到一张图向上 . . . . .	39
2.6	CNN 手写体数据识别 . . . . .	40
2.6.1	mnist 数据集 . . . . .	40
2.7	RNN . . . . .	44
2.7.1	The Problem Long-Term Dependencies . . . . .	45
2.7.2	LSTM 网络 . . . . .	45
2.7.3	LSTMs 想法的核心 . . . . .	46
2.7.4	一步步的设置 . . . . .	46
2.7.5	LSTM 的多种变体 . . . . .	47
2.7.6	向量字表示 . . . . .	49
2.7.7	RNN . . . . .	57
3	Tensorflow 进阶 . . . . .	61
3.1	模型存储和加载 . . . . .	61
3.2	用 GPU . . . . .	62
3.2.1	手工配置设备 . . . . .	62
3.2.2	允许 GPU 的内存增长 . . . . .	63
3.3	如何利用 Inception 的最后一层重新训练新的分类 . . . . .	65
3.3.1	训练花 . . . . .	65
3.3.2	瓶颈 . . . . .	66
3.3.3	训练 . . . . .	66

3.3.4	用 TensorBoard 可视化 . . . . .	67
3.3.5	用重新训练的模型 . . . . .	67
3.3.6	在你自己的分类上训练 . . . . .	67
3.3.7	创建一个训练图像集合 . . . . .	68
3.3.8	训练步骤 . . . . .	68
3.3.9	扭曲 . . . . .	68
3.3.10	超参数 . . . . .	69
3.3.11	训练, 验证, 测试集 . . . . .	69
3.3.12	更对模型架构 . . . . .	70
3.4	TF layer 向导: 建立一个卷积神经网络 . . . . .	70
3.4.1	开始 . . . . .	70
3.4.2	介绍卷积神经网络 . . . . .	71
3.4.3	建立 CNN MNIST 分类器 . . . . .	71
3.4.4	输入层 . . . . .	72
3.4.5	第一层卷积层 . . . . .	72
3.4.6	池化层 1 . . . . .	73
3.4.7	二层卷积和池化 . . . . .	73
3.4.8	Dense layer . . . . .	74
3.4.9	Logits Layers . . . . .	74
3.4.10	常见的预测 . . . . .	75
3.4.11	计算 Loss . . . . .	75
3.4.12	配置训练操作 . . . . .	76
3.4.13	增加评估度量 . . . . .	76
3.5	训练评估 CNN MNIST 分类器 . . . . .	76
3.5.1	载入训练和测试数据 . . . . .	77
3.5.2	创建 Estimator . . . . .	77
3.5.3	建立 Logging Hook . . . . .	77
3.5.4	选练模型 . . . . .	78
3.5.5	评估模型 . . . . .	78
3.5.6	运行模型 . . . . .	78
4	扩展 . . . . .	81
4.1	TensorFlow 架构 . . . . .	81
4.2	概述 . . . . .	81
4.2.1	Client . . . . .	83
4.2.2	Distributed master . . . . .	83

4.2.3 Worker Service . . . . .	84
4.3 内核实现 . . . . .	85
5 Performance	87
5.1 最好的实践 . . . . .	87
5.2 从源代码创建安装 . . . . .	87
5.2.1 利用队列读取数据 . . . . .	88
5.2.2 在 CPU 上的预处理 . . . . .	88
5.2.3 用大文件 . . . . .	89
5.2.4 用 NCHW 图像数据格式 . . . . .	89
5.2.5 用融批规范 . . . . .	89
5.3 性能向导 . . . . .	90
5.4 好性能模型 . . . . .	90
5.5 Benchmark . . . . .	90
5.6 如何用 TensorFlow 量化神经网络 . . . . .	90
5.6.1 为什么做量化工作 . . . . .	90
5.6.2 为什么量化 . . . . .	90
5.6.3 为什么不直接训练低精度 . . . . .	91
5.6.4 你能如何量化你的模型 . . . . .	91
5.6.5 如何量化处理工作 . . . . .	92
5.6.6 量化 Tensor 将呈现什么 . . . . .	94
5.6.7 下一步 . . . . .	95
6 常用的 python 模块	97
6.1 Argparse . . . . .	97
6.1.1 ArgumentParser 对象 . . . . .	98
6.1.2 prog . . . . .	98
6.1.3 add_argument() 方法 . . . . .	103
6.2 path . . . . .	131
6.2.1 函数说明 . . . . .	131
6.2.2 例子 . . . . .	133
6.2.3 常见问题 . . . . .	134
6.3 正则表达式介绍 . . . . .	144
6.4 RE 库的主要功能函数 . . . . .	147
6.4.1 re 表达式中的 flags . . . . .	149
6.5 常用的 sys 函数 . . . . .	154

6.6 collections . . . . .	161
6.7 base64 . . . . .	162
6.8 struct . . . . .	163
6.9 hashlib . . . . .	164
6.10 itertools . . . . .	165
6.11 contextlib . . . . .	166
6.12 XML . . . . .	167
6.13 HTMLParser . . . . .	168
6.14 ZipFile . . . . .	169
6.15 url . . . . .	170
6.15.1 urllib.request . . . . .	170
6.16 requests . . . . .	171
6.16.1 发送请求 . . . . .	171
6.16.2 requests 库的 7 个主要方法 . . . . .	171
6.16.3 request 对象的属性 . . . . .	171
6.16.4 理解 encoding 和 apparent_encoding . . . . .	172
6.16.5 理解 Requests 库的异常 . . . . .	172
6.16.6 HTTP 协议 . . . . .	172
7 Bazel . . . . .	175
7.1 Bazel start . . . . .	175
7.1.1 用工作空间 . . . . .	175
8 Tensorflow 技巧 . . . . .	177
8.1 文件读取 . . . . .	177
9 Tensorflow API . . . . .	179
9.1 tf.app.flags . . . . .	179
9.1.1 DEFINE_boolean . . . . .	179
9.1.2 DEFINE_boolean . . . . .	179
9.1.3 DEFINE_float . . . . .	179
9.1.4 DEFINE_integer . . . . .	180
9.1.5 DEFINE_string . . . . .	180
9.1.6 tf.squeeze . . . . .	180
9.1.7 tf.metrics . . . . .	181
9.1.8 tf.stack . . . . .	181

9.1.9	tf.reshape . . . . .	182
9.1.10	tf.random_crop . . . . .	182
9.1.11	tf.random_gamma . . . . .	183
9.1.12	tf.random_normal . . . . .	184
9.1.13	tf.random_normal_initializer . . . . .	185
9.1.14	tf.random_possion . . . . .	186
9.1.15	random_shuffle . . . . .	187
9.1.16	tf.random_uniform . . . . .	187
9.1.17	tf.random_uniform_initializer . . . . .	188
9.1.18	tf.one_hot . . . . .	189
9.1.19	tf.unstack . . . . .	191
9.2	tf.image . . . . .	193
9.2.1	tf.image.decode_gif . . . . .	193
9.2.2	tf.image.decode_jpeg . . . . .	193
9.2.3	tf.image.encode_jpeg . . . . .	194
9.2.4	tf.image.decode_png . . . . .	194
9.2.5	tf.image.encode_png . . . . .	195
9.2.6	tf.image.decode_image . . . . .	195
9.2.7	tf.image.resize_images . . . . .	195
9.3	layer . . . . .	197
9.3.1	tf.layers.average_pooling1d . . . . .	197
9.3.2	tf.layers.average_pooling2d . . . . .	197
9.3.3	tf.layers.average_pooling3d . . . . .	198
9.3.4	tf.layers.batch_normalization . . . . .	199
9.3.5	conv1d . . . . .	201
9.3.6	conv2d . . . . .	202
9.3.7	conv2d_transpose . . . . .	204
9.3.8	conv3d . . . . .	205
9.3.9	conv3d_transpose . . . . .	207
9.3.10	dense . . . . .	208
9.3.11	dropout . . . . .	209
9.3.12	max_pool1d . . . . .	210
9.3.13	max_pool2d . . . . .	210
9.3.14	max_pool3d . . . . .	211
9.3.15	separable_conv2d . . . . .	212

目录	vii
9.4 tf.train . . . . .	214
9.4.1 优化器 . . . . .	214



# Chapter 1

## deeplearning

### 1.1 降维

#### 1.1.1 自编码

人工神经网络（ANN）本身就是具有层次结构的系统，如果给定一个神经网络，我们假设其输出与输入是相同的，然后训练调整其参数，得到每一层中的权重。自然地，我们就得到了输入  $I$  的几种不同表示（每一层代表一种表示），这些表示就是特征。在研究中可以发现，如果在原有的特征中加入这些自动学习得到的特征可以大大提高精确度，甚至在分类问题中比目前最好的分类算法效果还要好！这种方法称为 AutoEncoder（自动编码器）。自动编码器就是一种尽可能复现输入信号的神经网络。为了实现这种复现，自动编码器就必须捕捉可以代表输入数据的最重要的因素，就像 PCA 那样，找到可以代表原信息的主要成分。我们将 input 输入一个 encoder 编码器，就会得到一个 code，这个 code 也就是输入的一个表示，那么我们怎么知道这个 code 表示的就是 input 呢？我们加一个 decoder 解码器，这时候 decoder 就会输出一个信息，那么如果输出的这个信息和一开始的输入信号 input 是很像的（理想情况下就是一样的），那很明显，我们就有理由相信这个 code 是靠谱的。所以，我们就通过调整 encoder 和 decoder 的参数，使得重构误差最小，这时候我们就得到了输入 input 信号的第一个表示了，也就是编码 code 了。因为是无标签数据，所以误差的来源就是直接重构后与原输入相比得到。

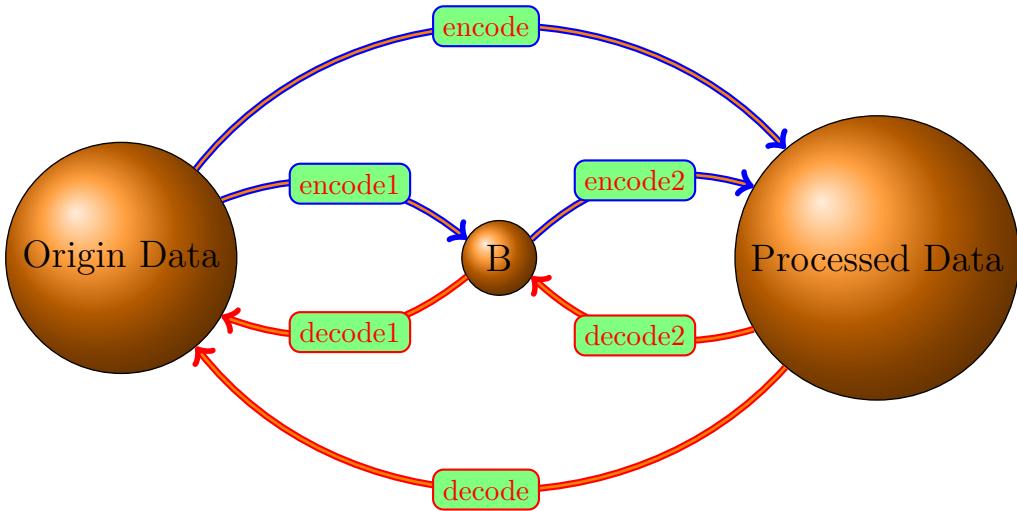
#### 1.1.2 自动降噪编码

以一定的概率分布擦出原始数据（将数据置为 0），这样操作后的数据称为破損数据，这样的数据有两个作用：

1. 通过破損数据和非破損数据相比，破損数据训练出来的权重噪声小（可能不小心删除了噪声）。

2. 破损数据一定程度上减轻了训练数据和测试数据之间的代沟。由于数据部分被擦除，因而训练出来的权重的健壮性就提高了。

### 1.1.3 手写体数据自编码



```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 data_path = '/home/hpc/文档/mnist_tutorial/mnist'
4
5 from tensorflow.examples.tutorials.mnist import input_data
6 mnist = input_data.read_data_sets(data_path, one_hot=False)
7
8
9 # Visualize decoder setting
10 learning_rate = 0.01
11 training_epochs = 5
12 batch_size = 256
13 display_step = 1
14 examples_to_show = 10
15
16 n_input = 784 # MNIST data input (img shape: 28*28)
17
18 x = tf.placeholder(tf.float32, [None, n_input])
19
20 n_hidden_1 = 256
21 n_hidden_2 = 128
22 weights = {
23     'encode_h1': tf.Variable(tf.random_normal([n_input, n_hidden_1])),
  
```

```

24     'encode_h2': tf.Variable(tf.random_normal([n_hidden_1, n_hidden_2])),  

25     'decode_h2': tf.Variable(tf.random_normal([n_hidden_2, n_hidden_1])),  

26     'decode_h1': tf.Variable(tf.random_normal([n_hidden_1, n_input]))  

27 }  

28 bias = { 'encode_h1': tf.Variable(tf.random_normal([n_hidden_1])),  

29     'encode_h2': tf.Variable(tf.random_normal([n_hidden_2])),  

30     'decode_h2': tf.Variable(tf.random_normal([n_hidden_1])),  

31     'decode_h1': tf.Variable(tf.random_normal([n_input]))  

32  

33 }  

34 def encode(x):  

35     layer_1 = tf.nn.sigmoid(tf.add(tf.matmul(x, weights['encode_h1']), bias[  

36                                     'encode_h1']))  

37     layer_2 = tf.nn.sigmoid(tf.add(tf.matmul(layer_1, weights['encode_h2']), bias[  

38                                     'encode_h2']))  

39     return layer_2  

40  

41 def decode(x):  

42     layer_1 = tf.nn.sigmoid(tf.add(tf.matmul(x, weights['decode_h2']), bias[  

43                                     'decode_h2']))  

44     layer_2 = tf.nn.sigmoid(tf.add(tf.matmul(layer_1, weights['decode_h1']), bias[  

45                                     'decode_h1']))  

46     return layer_2  

47  

48 encode_op = encode(x)  

49 decode_op = decode(encode_op)  

50 y_pred = decode_op  

51 y_true = x  

52 cost = tf.reduce_mean(tf.square(y_pred - y_true))  

53 optimizer = tf.train.AdamOptimizer(learning_rate).minimize(cost)  

54  

55 with tf.Session() as sess:  

56     init = tf.global_variables_initializer()  

57     sess.run(init)  

58     total_batch = int(mnist.train.num_examples/batch_size)  

59     for epoch in range(training_epochs):  

60         for i in range(total_batch):  

61             batch_xs, batch_ys = mnist.train.next_batch(batch_size)  

62             _, c = sess.run([optimizer, cost], feed_dict={x: batch_xs})  

63             if epoch%display_step==0:  

64                 print("Epoch:", '%04d' % (epoch+1), 'cost=', '{:.9f}'.format(c))

```

```

63     print('Optimize finish')
64     encode_decode = sess.run(y_pred, feed_dict={x: mnist.test.images[:, examples_to_show]})
65     f, a = plt.subplots(2, 10, figsize=(10, 2))
66     for i in range(examples_to_show):
67         a[0][i].imshow(sess.run(tf.reshape(mnist.test.images[i], [28, 28])))
68         a[1][i].imshow(sess.run(tf.reshape(encode_decode[i], [28, 28])))
69     plt.savefig('auto_encode.png', dpi=800)

```

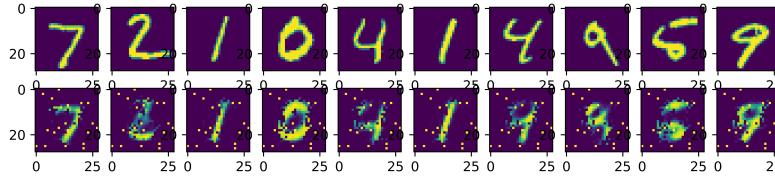


图 1.1: 原图和自编码解码后的图像

编码器输出可视化:

```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3
4 from tensorflow.examples.tutorials.mnist import input_data
5 path = '/home/hpc/文档/mnist_tutorial/mnist'
6 mnist = input_data.read_data_sets(path, one_hot=False)
7
8 learning_rate = 0.01
9 training_epochs = 5
10 batch_size = 256
11 display_step = 1
12 examples_to_show = 10
13
14 n_input = 784 # MNIST data input (img shape: 28*28)
15
16 X = tf.placeholder("float", [None, n_input])
17
18 n_hidden_1 = 256 # 1st layer num features
19 n_hidden_2 = 128 # 2nd layer num features
20
21 learning_rate = 0.01 # 0.01 this learning rate will be better! Tested
22 training_epochs = 10

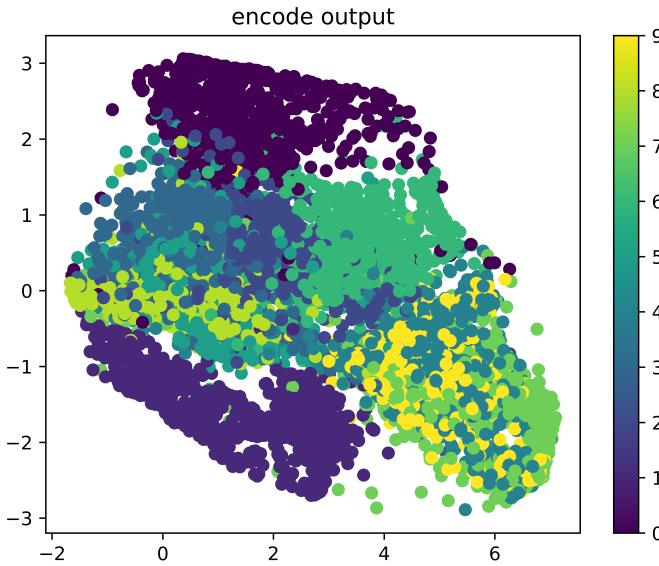
```



```

66 layer_3 = tf.nn.sigmoid(tf.add(tf.matmul(layer_2, weights[ 'decoder_h3' ]),
67                         biases[ 'decoder_b3' ]))
68 layer_4 = tf.nn.sigmoid(tf.add(tf.matmul(layer_3, weights[ 'decoder_h4' ]),
69                         biases[ 'decoder_b4' ]))
70 return layer_4
71
72 encoder_op = encoder(X)
73 decoder_op = decoder(encoder_op)
74
75 y_pred = decoder_op
76 y_true = X
77
78 cost = tf.reduce_mean(tf.pow(y_true - y_pred, 2))
79 optimizer = tf.train.AdamOptimizer(learning_rate).minimize(cost)
80
81
82 with tf.Session() as sess:
83     init = tf.global_variables_initializer()
84     sess.run(init)
85     total_batch = int(mnist.train.num_examples/batch_size)
86     for epoch in range(training_epochs):
87         for i in range(total_batch):
88             batch_xs, batch_ys = mnist.train.next_batch(batch_size) # max(x) =
89                                     # min(x) = 0
90             _, c = sess.run([optimizer, cost], feed_dict={X: batch_xs})
91             if epoch % display_step == 0:
92                 print("Epoch:", "%04d" % (epoch+1),
93                       "cost=", "{:.9f}".format(c))
94     print("Optimization Finished!")
95 encode_decode = sess.run(
96     y_pred, feed_dict={X: mnist.test.images[:examples_to_show] })
97 encoder_result = sess.run(encoder_op, feed_dict={X: mnist.test.images})
98 plt.scatter(encoder_result[:, 0], encoder_result[:, 1], c=mnist.test.labels)
99 plt.title('encode output')
100 plt.colorbar()
101 plt.savefig('auto_encode_v.png', dpi=800)

```



## 1.2 稀疏编码

稀疏编码算法是一种无监督学习方法，它用来寻找一组“超完备”基向量来更高效地表示样本数据。稀疏编码算法的目的就是找到一组基向量  $\phi_i$ ，使得我们能将输入向量  $\mathbf{x}$  表示为这些基向量的线性组合：

$$\mathbf{x} = \sum_{i=1}^k a_i \phi_i \quad (1.1)$$

虽然形如主成分分析技术 (PCA) 能使我们方便地找到一组“完备”基向量，但是这里我们想要做的是找到一组“超完备”基向量来表示输入向量  $\mathbf{x} \in \mathbb{R}^n$ （也就是说， $k > n$ ）。超完备基的好处是它们能更有效地找出隐含在输入数据内部的结构与模式。然而，对于超完备基来说，系数  $a_i$  不再由输入向量  $\mathbf{x}$  唯一确定。因此，在稀疏编码算法中，我们另加了一个评判标准“稀疏性”来解决因超完备而导致的退化 (degeneracy) 问题。

这里，我们把“稀疏性”定义为：只有很少的几个非零元素或只有很少的几个远大于零的元素。要求系数  $a_i$  是稀疏的意思就是说：对于一组输入向量，我们只想有尽可能少的几个系数远大于零。选择使用具有稀疏性的分量来表示我们的输入数据是有原因的，因为绝大多数的感官数据，比如自然图像，可以被表示成少量基本元素的叠加，在图像中这些基本元素可以是面或者线。同时，比如与初级视觉皮层的类比过程也因此得到了提升。

我们把有  $m$  个输入向量的稀疏编码代价函数定义为：

$$\underset{a_i^{(j)}, \phi_i}{\text{minimize}} \sum_{j=1}^m \left\| \mathbf{x}^{(j)} - \sum_{i=1}^k a_i^{(j)} \phi_i \right\|^2 + \lambda \sum_{i=1}^k S(a_i^{(j)}) \quad (1.2)$$

此处  $S(\cdot)$  是一个稀疏代价函数，由它来对远大于零的  $a_i$  进行“惩罚”。我们可以把稀疏编码目标函式的第一项解释为一个重构项，这一项迫使稀疏编码算法能为输入向量  $\mathbf{x}$  提供一个高拟合度的线性表达式，而公式第二项即“稀疏惩罚”项，它使  $\mathbf{x}$  的表达式变得“稀疏”。常量  $\lambda$  是一个变换量，由它来控制这两项式子的相对重要性。

虽然“稀疏性”的最直接测度标准是“L0”范式 ( $S(a_i) = \mathbf{1}(|a_i| > 0)$ )，但这是不可微分的，而且通常很难进行优化。在实际中，稀疏代价函数  $S(\cdot)$  的普遍选择是 L1 范式代价函数  $S(a_i) = |a_i|_1$  及对数代价函数  $S(a_i) = \log(1 + a_i^2)$ 。

此外，很有可能因为减小  $a_i$  或增加  $\phi_i$  至很大的常量，使得稀疏惩罚变得非常小。为防止此类事件发生，我们将限制  $\|\phi\|^2$  要小于某常量  $C$ 。包含了限制条件的稀疏编码代价函数的完整形式如下：

$$\underset{a_i^{(j)}, \phi_i}{\text{minimize}} \sum_{j=1}^m \|x^{(j)} - \sum_{i=1}^k a_i^{(j)} \phi_i\|^2 + \lambda \sum_{i=1}^k S(s_i^{(j)}) \quad \|\phi_i\|^2 \leq C, \forall i = 1, \dots, k \quad (1.3)$$

### 1.2.1 稀疏编码的概率表示

到目前为止，我们所考虑的稀疏编码，是为了寻找到一个稀疏的、超完备基向量集，来覆盖我们的输入数据空间。现在换一种方式，我们可以从概率的角度出发，将稀疏编码算法当作一种“生成模型”。

我们将自然图像建模问题看成是一种线性叠加，叠加元素包括  $k$  个独立的源特征  $\phi_i$  以及加性噪声：

$$\mathbf{x} = \sum_{i=1}^k a_i \phi_i + v(\mathbf{x}) \quad (1.4)$$

我们的目标是找到一组特征基向量  $\phi$ ，它使得图像的分布函数  $P(\mathbf{x} | \phi)$  尽可能地近似于输入数据的经验分布函数  $P^*(\mathbf{x})$ 。一种实现方式是，最小化  $P^*(\mathbf{x})$  与  $P(\mathbf{x} | \phi)$  之间的 KL 散度，此 KL 散度表示如下：

$$D(P^*(\mathbf{x}) || P(\mathbf{x} | \phi)) = \int P^*(\mathbf{x}) \log \left( \frac{P^*(\mathbf{x})}{P(\mathbf{x} | \phi)} \right) d\mathbf{x} \quad (1.5)$$

因为无论我们如何选择  $\phi$ ，经验分布函数  $P^*(\mathbf{x})$  都是常量，也就是说我们只需要最大化对数似然函数  $P(\mathbf{x} | \phi)$ 。假设  $v$  是具有方差  $\sigma^2$  的高斯白噪声，则有下式：

$$P(\mathbf{x} | \mathbf{a}, \phi) = \frac{1}{Z} \exp \left( -\frac{(\mathbf{x} - \sum_{i=1}^k a_i \phi_i)^2}{2\sigma^2} \right) \quad (1.6)$$

为了确定分布  $P(\mathbf{x} | \phi)$ ，我们需要指定先验分布  $P(\mathbf{a})$ 。假定我们的特征变量是独立的，我们就可以将先验概率分解为：

$$P(\mathbf{a}) = \prod_{i=1}^k P(a_i) \quad (1.7)$$

此时，我们将“稀疏”假设加入进来——假设任何一幅图像都是由相对较少的一些源特征组合起来的。因此，我们希望  $a_i$  的概率分布在零值附近是凸起的，而且峰值很高。一个方便的参数化先验分布就是：

$$P(a_i) = \frac{1}{Z} \exp(-\beta S(a_i)) \quad (1.8)$$

这里  $S(a_i)$  是决定先验分布的形状的函数。

当定义了  $P(\mathbf{x} | \mathbf{a}, \phi)$  和  $P(\mathbf{a})$  后，我们就可以写出在由  $\phi$  定义的模型之下的数据  $\mathbf{x}$  的概率分布：

$$P(\mathbf{x} | \phi) = \int P(\mathbf{x} | \mathbf{a}, \phi) P(\mathbf{a}) d\mathbf{a} \quad (1.9)$$

那么，我们的问题就简化为寻找：

$$\phi^* = \operatorname{argmax}_\phi \langle \log(P(\mathbf{x} | \phi)) \rangle \quad (1.10)$$

这里  $\langle \cdot \rangle$  表示的是输入数据的期望值。

不幸的是，通过对  $\mathbf{a}$  的积分计算  $P(\mathbf{x} | \phi)$  通常是难以实现的。虽然如此，我们注意到如果  $P(\mathbf{x} | \phi)$  的分布（对于相应的  $\mathbf{a}$ ）足够陡峭的话，我们就可以用  $P(\mathbf{x} | \phi)$  的最大值来估算以上积分。估算方法如下：

$$\phi^{*'} = \operatorname{argmax}_{\mathbf{a}} \langle \max_{\mathbf{a}} \log(P(\mathbf{x} | \phi)) \rangle \quad (1.11)$$

跟之前一样，我们可以通过减小  $a_i$  或增大  $\phi$  来增加概率的估算值（因为  $P(a_i)$  在零值附近陡升）。因此我们要对特征向量  $\phi$  加一个限制以防止这种情况发生。最后，我们可以定义一种线性生成模型的能量函数，从而将原先的代价函数重新表述为：

$$E(x, a | \phi) := -\log(P(x | \phi, \mathbf{a}) P(\mathbf{a})) \quad (1.12)$$

$$= \sum_{j=1}^m \|x^{(j)} - \sum_{i=1}^k a_i^{(j)} \phi_i\|^2 + \lambda \sum_{i=1}^k S(a_i^{(j)}) \quad (1.13)$$

其中  $\lambda = 2\sigma 2\beta$ ，并且关系不大的常量已被隐藏起来。因为最大化对数似然函数等同于最小化能量函数，我们就可以将原先的优化问题重新表述为：

$$\phi^*, \mathbf{a}^* = \operatorname{argmin}_{\phi, \mathbf{a}} \sum_{j=1}^m \left\| \mathbf{x}^{(j)} - \sum_{i=1}^k a_i^{(j)} \phi_i \right\|^2 + \lambda \sum_{i=1}^k S(a_i^{(j)}) \quad (1.14)$$

使用概率理论来分析，我们可以发现，选择 L1 惩罚和  $\log(1 + a_i^2)$  惩罚作为函数  $S(\cdot)$ ，分别对应于使用了拉普拉斯概率  $P(a_i) \propto \exp(-\beta|a_i|)$  和柯西先验概率  $P(a_i) \propto \frac{\beta}{1+a_i^2}$ 。

### 1.3 PCA

在多元统计分析中，主成分分析（英语：Principal components analysis, PCA）是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。由于主成分分析依赖所给数据，所以数据的准确性对分析结果影响很大。主成分分析由卡尔·皮尔逊于 1901 年发明，用于分析数据及建立数理模型。其方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值）。PCA 是最简单的以特征量分析多元统计分布的方法。其结果可以理解为对原数据中的方差做出解释：哪一个方向上的数据值对方差的影响最大？换而言之，PCA 提供了一种降低数据维度的有效办法；如果分析者在原数据中除掉最小的特征值所对应的成分，那么所得的低维度数据必定是最优化的（也即，这样降低维度必定是失去讯息最少的方法）。主成分分析在分析复杂数据时尤为有用，比如人脸识别。PCA 是最简单的以特征量分析多元统计分布的方法。通常情况下，这种运算可以被看作是揭露数据的内部结构，从而更好的解释数据的变量的方法。如果一个多元数据集能够在一个高维数据空间坐标系中被显现出来，那么 PCA 就能够提供一幅比较低维度的图像，这幅图像即为在讯息最多的点上原对象的一个‘投影’。这样就可以利用少量的主成分使得数据的维度降低了。PCA 跟因子分析密切相关，并且已经有很多混合这两种分析的统计包。而真实要素分析则是假定底层结构，求得微小差异矩阵的特征向量。

#### 1.3.1 数学定义

PCA 的数学定义是：一个正交化线性变换，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依次类推。定义一个  $n \times m$  的矩阵， $X^T$  为去平均值（以平均值为中心移动至原点）的数据，其行为数据样本，列为数据类别（注意，这里定义的是  $X^T$  而不是  $X$ ）。则  $X$  的奇异值分解为  $X = W\Sigma V^T$ ，其中  $m \times m$  矩阵  $W$  是  $XX^T$  的本征矢量矩阵， $\Sigma$  是  $m \times n$  的非负矩形对角矩阵， $V$  是  $m \times n$  的  $X^T X$  的本征矢量矩阵。据此，

$$\begin{aligned} Y^T &= X^T W \\ &= V \Sigma^T W^T W \\ &= V \Sigma^T \end{aligned} \tag{1.15}$$

当  $m < n$  时， $V$  在通常情况下不是唯一定义的，而  $Y$  则是唯一定义的。 $W$  是一个正交矩阵， $Y^T$  是  $X^T$  的转置，且  $Y^T$  的第一列由第一主成分组成，第二列由第二主成分组成，依次类推。为了得到一种降低数据维度的有效办法，我们可以利用  $W_L$  把  $X$  映射到一个只应

用前面 L 个向量的低维空间中去：

$$\mathbf{Y} = \mathbf{W}_{\mathbf{L}}^T \mathbf{X} = \Sigma_L \mathbf{V}^T \quad (1.16)$$

其中  $\Sigma_L = \mathbf{I}_{L \times m} \Sigma$  且  $\mathbf{I}_{L \times m}$  为  $L \times mL \times m$  的单位矩阵。X 的单向量矩阵 W 相当于协方差矩阵的本征矢量  $C = XX^T$ ,

$$XX^T = W\Sigma\Sigma^TW^T \quad (1.17)$$

在欧几里得空间给定一组点数，第一主成分对应于通过多维空间平均点的一条线，同时保证各个点到这条直线距离的平方和最小。去除掉第一主成分后，用同样的方法得到第二主成分。依此类推。在  $\Sigma$  中的奇异值均为矩阵  $XX^T$  的本征值的平方根。每一个本征值都与跟它们相关的方差是成正比的，而且所有本征值的总和等于所有点到它们的多维空间平均点距离的平方和。PCA 提供了一种降低维度的有效办法，本质上，它利用正交变换将围绕平均点的点集中尽可能多的变量投影到第一维中去，因此，降低维度必定是失去讯息最少的方法。PCA 具有保持子空间拥有最大方差的最优正交变换的特性。然而，当与离散余弦变换相比时，它需要更大的计算需求代价。非线性降维技术相对于 PCA 来说则需要更高的计算要求。PCA 对变量的缩放很敏感。如果我们只有两个变量，而且它们具有相同的样本方差，并且成正相关，那么 PCA 将涉及两个变量的主成分的旋转。但是，如果把第一个变量的所有值都乘以 100，那么第一主成分就几乎和这个变量一样，另一个变量只提供了很小的贡献，第二主成分也将和第二个原始变量几乎一致。这就意味着当不同的变量代表不同的单位（如温度和质量）时，PCA 是一种比较武断的分析方法。但是在 Pearson 的题为”On Lines and Planes of Closest Fit to Systems of Points in Space”的原始文件里，是假设在欧几里得空间里不考虑这些。一种使 PCA 不那么武断的方法是使用变量缩放以得到单位方差。

## 1.4 kl 散度

相对熵 (relative entropy) 又称为 KL 散度 (Kullback-Leibler divergence, 简称 KLD)，信息散度 (information divergence)，信息增益 (information gain) KL 散度是两个概率分布 P 和 Q 差别的非对称性度量。KL 散度是用来度量基于 Q 的编码来编码来自 P 的样本平均所需的额外的位元数。典型情况下，P 表示数据的真实分布，Q 表示数据的理论分布，模型分布或 P 的近似分布。

对于离散随机变量，其概率分布 P 和 Q 的 KL 散度可以按下面定义为

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (1.18)$$

即按概率 P 求得的 P 和 Q 的对数差的平均值。KL 散度仅当 P 和 Q 各自总和均为 1，且对任何 t 皆满足对于  $Q(i) > 0$  及  $P(i) > 0$  时才有定义。式子出现  $0 \ln 0$  其值按 0 处理，对于

连续随机变量，其概率分布  $P$  和  $Q$  可按计分方式定义为：

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \quad (1.19)$$

其中  $p$  和  $q$  分别表示分布  $P$  和  $Q$  的概率密度。

### 1.4.1 相对熵

由 Gibbs 不等式可知，当且仅当  $P = Q$  时  $D_{KL}(P||Q)$  为 0。尽管从直觉上 KL 散度是个度量或距离函数，但是它实际上不是一个真正的度量或距离。因为 KL 散度不具有对称性：从分布  $P$  到  $Q$  的距离（或度量）通常并不等于从  $Q$  到  $P$  的距离（或度量）。

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

自信息和散度的关系： $I(m) = D_{KL}(\delta_{im}||p_i)$ 。互信息和散度：

$$\begin{aligned} I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\ &= E_x D_{KL}(P(Y|X)||P(Y)) \\ &= E_y D_{KL}(P(X|Y)||P(X)) \end{aligned}$$

信息熵和散度：

$$\begin{aligned} H(X) &= (i) E_x I(x) \\ &= (ii) \log N - D_{KL}(P(X)||P_U(X)) \end{aligned}$$

条件熵和散度：

$$\begin{aligned} H(X|Y) &= \log N - D_{KL}(P(X,Y)||P_U(X)P(Y)) \\ &= (i) \log N - D_{KL}(P(X,Y)||P(X)P(Y)) - D_{KL}(P(X)||P_U(X)) \\ &= H(x) - I(X;Y) \\ &= (ii) \log N - E_Y D_{KL}(P(X|Y)||P_U(X)) \end{aligned}$$

交叉熵与散度： $H(p,q) = E_p[-\log q] = H(p) + D_{KL}(p||q)$

# Chapter 2

## Tensorflow 基础

### 2.1 Tensorflow 基础函数

#### 2.1.1 Variable

```
1 #tensorflow 1.2.1
2 import tensorflow as tf
3 var = tf.Variable(0)
4 add_operation = tf.add(var,1)
5 update_operation = tf.assign(var,add_operation)
6 with tf.Session() as sess:
7     sess.run(tf.global_variables_initializer())
8     for _ in range(3):
9         sess.run(update_operation)
10        print(sess.run(var))
```

#### 2.1.2 placeholder

```
1 #tensorflow 1.2
2 import tensorflow as tf
3 x1 = tf.placeholder(dtype=tf.float32,shape=None)
4 y1 = tf.placeholder(dtype=tf.float32,shape=None)
5 z1 = x1+y1
6 x2 = tf.placeholder(dtype=tf.float32,shape=[2,1])
7 y2 = tf.placeholder(dtype=tf.float32,shape=[1,2])
8 z2 = tf.matmul(x2,y2)
9 with tf.Session() as sess:
10    z1_value = sess.run(z1,feed_dict={x1:1,y1:2})
```

```

11 z1_value, z2_value = sess.run([z1, z2], feed_dict={x1: 1, y1: 2, x2: [[2], [2]], y2: [[
12     print(z1_value)
13     print(z2_value)

```

### 2.1.3 batch normalization

§ 数据  $x$  为 Tensor。

- mean: 为  $x$  的均值, 也是一个 Tensor。
- var: 为  $x$  的方差, 也为一个 Tensor。
- offset: 一个偏移, 也是一个 Tensor。
- scale: 缩放倍数, 也是一个 Tensor。
- variable\_epsilon, 一个不为 0 的浮点数。
- name: 操作的名字, 可选。

batch normalization 计算方式是:

$$x = (x - \bar{x}) / \sqrt{Var(x) + variable\_epsilon} \quad (2.1)$$

$$x = x \times scale + offset \quad (2.2)$$

$$(2.3)$$

$$\text{均值: } \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i \quad (2.4)$$

$$\text{方差: } \sigma^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \bar{x})^2 \quad (2.5)$$

## 2.2 常见的激活函数

### 2.2.1 relu

relu 函数在自变量  $x$  小于 0 时值全为 0, 在  $x$  大于 0 时, 值和自变量相等。

```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 x = tf.linspace(-10., 10., 100)
4 y = tf.nn.relu(x)
5 with tf.Session() as sess:

```

```

6 [x,y] = sess.run([x,y])
7 plt.plot(x,y,'r',6,6,'bo')
8 plt.title('relu')
9 ax = plt.gca()
10 ax.annotate("", 
11             xy=(6, 6), xycoords='data',
12             xytext=(6, 4.5), textcoords='data',
13             arrowprops=dict(arrowstyle="->",
14                             connectionstyle="arc3"),
15             )
16 ax.annotate("",xy=(6,6),xycoords='data',
17             xytext=(10, 6), textcoords='data',
18             arrowprops=dict(arrowstyle="->",
19                             connectionstyle="arc3"),
20             )
21 )
22 ax.grid(True)
23 plt.xlabel('x')
24 plt.ylabel('relu(x)')
25 plt.savefig('relu.png',dpi = 600)

```

### 2.2.2 relu6

relu6 函数和 relu 不同之处在于在  $x$  大于等于 6 的部分值保持为 6。

```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 x = tf.linspace(-10.,10.,100)
4 y = tf.nn.relu6(x)
5 with tf.Session() as sess:
6     [x,y] = sess.run([x,y])
7     plt.plot(x,y,'r',6,6,'bo')
8     plt.title('relu6')
9     ax = plt.gca()
10    ax.annotate("", 
11                xy=(6, 6), xycoords='data',
12                xytext=(6, 4.5), textcoords='data',
13                arrowprops=dict(arrowstyle="->",
14                                connectionstyle="arc3"),
15                )
16    ax.grid(True)
17    plt.xlabel('x')
18    plt.ylabel('relu6(x)')

```

```
19 plt.savefig('relu6.png', dpi = 600)
```

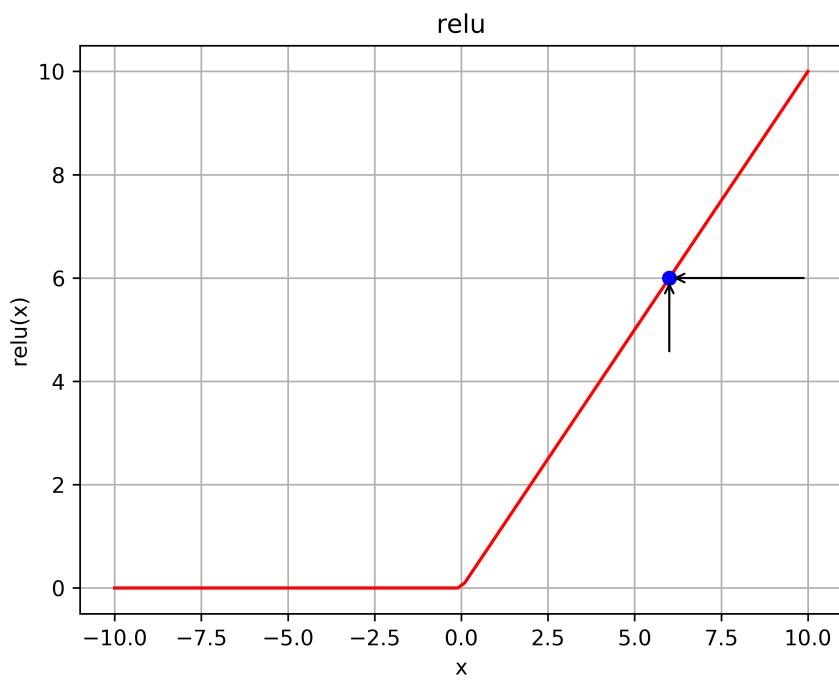


图 2.1: relu

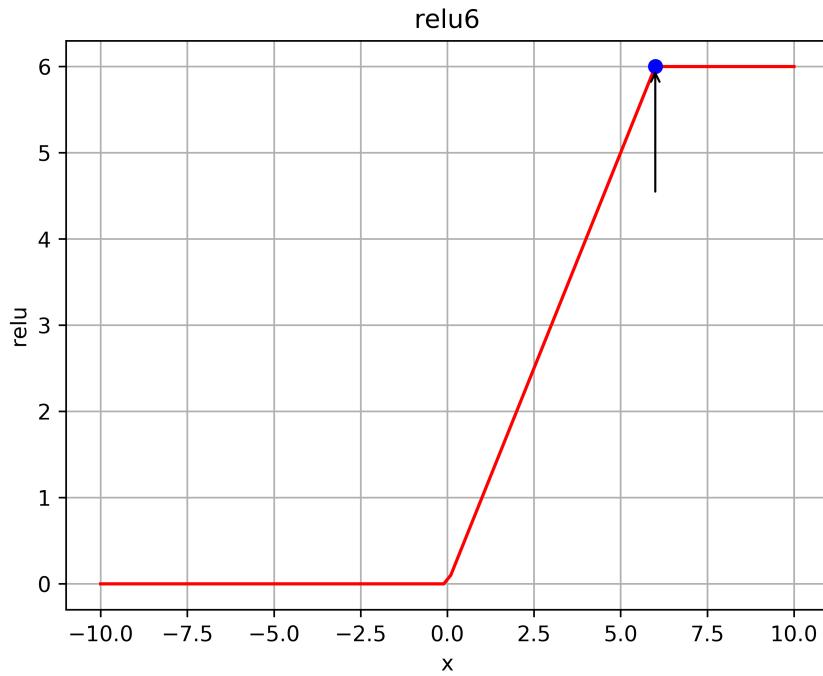


图 2.2: relu6

### 2.2.3 sigmoid

```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 import matplotlib.patches as mpatches
4 x = tf.linspace(-10.,10.,100)
5 y1 = tf.nn.sigmoid(x)
6 y2 = tf.nn.tanh(x)
7 red_patch = mpatches.Patch(color = 'red',label = 'sigmoid')
8 blue_patch = mpatches.Patch(color = 'blue',label = 'tanh')
9 with tf.Session() as sess:
10     [x,y1,y2] = sess.run([x,y1,y2])
11     plt.plot(x,y1,'r',x,y2,'b')
12     ax = plt.gca()
13     ax.annotate(r"\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}",
14                 xy=(0,0),xycoords="data",
15                 xytext=(1,0),textcoords="data",
16                 arrowprops=dict(arrowstyle="->",
17                                 connectionstyle="arc3"),
18 )
19     ax.annotate(r"\text{sigmoid}(x) = \frac{1}{1+e^{-x}}",

```

```

20     xy=(0,0.5),xycoords="data",
21     xytext=(1,0.5),textcoords="data",
22     arrowprops=dict (arrowstyle="->",
23     connectionstyle="arc3"),
24 )
25 plt.xlabel('x')
26 plt.grid(True)
27 plt.legend(handles = [red_patch,blue_patch])
28 plt.savefig('activate.png',dpi=600)

```

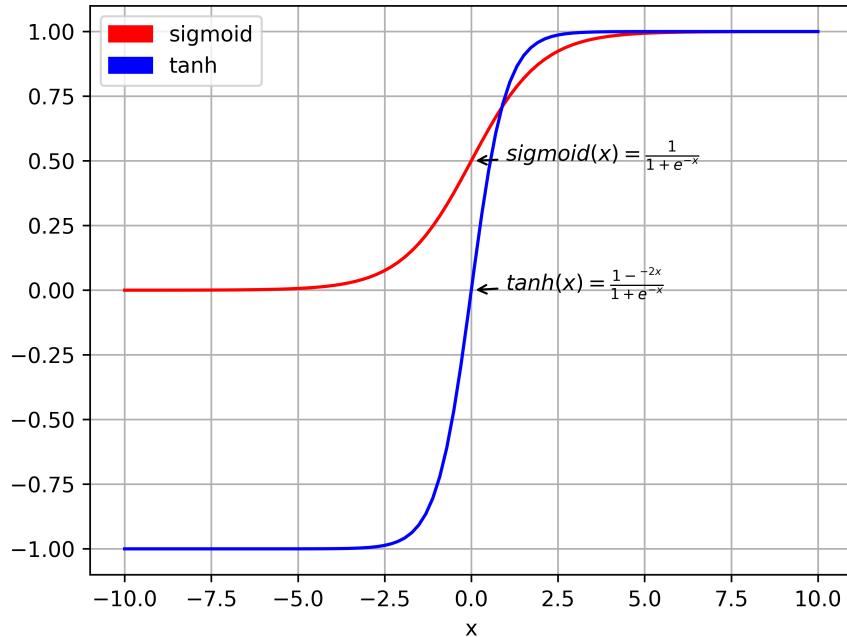


图 2.3: activate\_fun

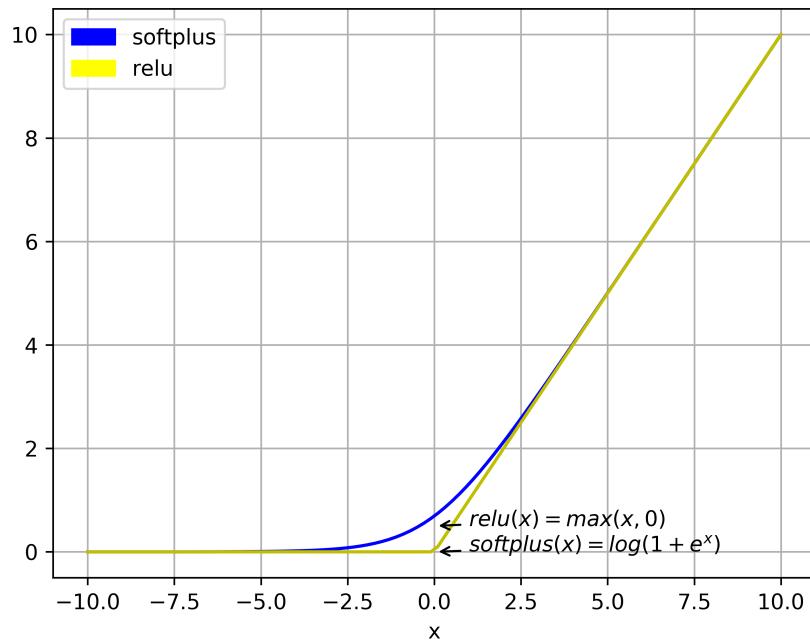
#### 2.2.4 relu 和 softplus

```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 import matplotlib.patches as mpatches
4 x = tf.linspace(-10.,10.,100)
5 y2 = tf.nn.softplus(x)
6 y3 = tf.nn.relu(x)
7 blue_patch = mpatches.Patch(color = 'blue',label = 'softplus')
8 yellow_patch = mpatches.Patch(color = 'yellow',label = 'relu')

```

```
9 with tf.Session() as sess:  
10     [x,y2,y3] = sess.run([x,y2,y3])  
11     plt.plot(x,y2,'b',x,y3,'y')  
12     ax = plt.gca()  
13     plt.xlabel('x')  
14     ax.annotate(r"$softplus(x)=\log(1+e^x)$",  
15                 xy=(0, 0), xycoords="data",  
16                 xytext=(1, 0), textcoords="data",  
17                 arrowprops=dict(arrowstyle="->",  
18                                   connectionstyle="arc3"),  
19             )  
20     ax.annotate(r"$relu(x)=\max(x, 0)$",  
21                 xy=(0, 0.5), xycoords="data",  
22                 xytext=(1, 0.5), textcoords="data",  
23                 arrowprops=dict(arrowstyle="->",  
24                                   connectionstyle="arc3"),  
25             )  
26  
27     plt.grid(True)  
28     plt.legend(handles = [blue_patch,yellow_patch])  
29     plt.savefig('relu_softplus.png',dpi=600)
```



### 2.2.5 dropout

将神经元以概率 keep\_prob 绝对是否被抑制。如果被抑制该神经元的输出为 0 如果不被抑制,该神经元的输出将被放大到原来的  $1/\text{keep\_prop}$ 。默认情况下,每个神经元是否被抑制是相互独立的。但是是否被抑制也可以通过 noise\_shape 来调节。当 noise\_shape[i]=shape(x)[i] 时,x 中的元素相互独立。如果 shape(x)=[k,1,1,n], 那么每个批通道都是相互独立的, 但是每行每列的数据都是关联的, 也就是说要么都为 0, 要么还是原来的值。

```

1 import tensorflow as tf
2 a = tf.constant([[-1., 2., 3., 4.]])
3 with tf.Session() as sess:
4     b = tf.nn.dropout(a, 0.5, noise_shape=[1, 4])
5     print(sess.run(b))
6     c = tf.nn.dropout(a, 0.5, noise_shape=[1, 1])
7     print(sess.run(c))

```

`[[ -2. 0. 0. 8.]]`

`[[ -0. 0. 0. 0.]]`

当输入数据特征相差明显时, 用 tanh 效果会很好, 但在循环过程中会不断扩大特征效果并显示出来。当特征相差不明显时, sigmoid 效果比较好。同时, 用 sigmoid 和 tanh 作为激活函数时, 需要对输入进行规范化, 否则激活厚的值全部进入平坦区, 隐藏层的输出会趋同, 丧失原来的特征表达, 而 relu 会好很多, 优势可以不需要输入规范化来避免上述情况。因此, 现在大部分卷积神经网络都采用 relu 作为激活函数。

## 2.3 CNN 常用函数

### 2.3.1 卷积函数

`tf.nn.conv2d(input,filter,padding,stride=None,diation_rate=None, name = None,data_format=None)`

- input: 一个 tensor, 数据类型必须是 float32, 或者是 float64
- filter: 一个 tensor, 数据类型必须和 input 相同。
- strides: 一个长度为 4 的一组证书类型数组, 每一维对应 input 中每一维对应移动的步数, strides[1] 对应 input[1] 移动的步数。
- padding: 有两个可选参数'VALID' (输入数据维度和输出数据维度不同) 和'SAME' (输入数据维度和输出数据维度相同)
- use\_cudnn\_on\_gpu: 一个可选的布尔值, 默认情况下时 True。

- name: 可选，操作的一个名字。

```

1 import tensorflow as tf
2 input_data = tf.Variable(tf.random_normal(shape = [10,9,9,3],mean=0,stddev=1),
3                         dtype = tf.float32)
4 kernel = tf.Variable(tf.random_normal(shape = [2,2,3,2],mean = 0,stddev=1,dtype=
5                         tf.float32))
6
7 y = tf.nn.conv2d(input_data,kernel,strides=[1,1,1,1],padding='SAME')
8 init = tf.global_variables_initializer()
9 with tf.Session() as sess:
10     sess.run(init)
11     print(sess.run(y).shape)

```

输出形状为 [10,9,9,2]。

### 2.3.2 常见的分类函数

`tf.nn.sigmoid_cross_entropy_with_logits(logits,targets,name=None)`

- logits:[batch\_size,num\_classes]
- targets:[batch\_size,size]
- 输出： loss[batch\_size,num\_classes]

最后已成不需要进行 sigmoid 操作。

`tf.nn.softmax(logits,dim=-1,name=None)`: 计算 Softmax

$$\text{softmax} = \frac{x^{\logits}}{\text{reduce\_sum}(e^{\logits}, dim)}$$

`tf.nn.log_softmax(logits,dim=-1,name = None)` 计算 log softmax

$$\text{logsoftmax} = \log(\text{reduce\_softmax}(\exp(\logits), dim))$$

`tf.nn.softmax_cross_entropy_with_logits(_sentinel=None,labels=None,logits=None,dim=-1,name=None)` 输出 loss:[batch\_size] 保存的时 batch 中每个样本的交叉熵。`tf.nn.sparse_softmax_cross_entropy`

- logits: 神经网络最后一层的结果。
- 输入 logits:[batch\_size,num\_classes],labels:[batch\_size], 必须在 [0,num\_classes]
- loss[batch], 保存的是 batch 每个样本的交叉熵。

## 2.4 优化方法

- `tf.train.GradientDescentOptimizer`
- `tf.train.AdadeltaOptimizer`
- `tf.train.AdagradDAOptimizer`
- `tf.train.AdagradOptimizer`
- `tf.train.MomentumOptimizer`
- `tf.train.AdamOptimizer`
- `tf.train.FtrlOptimizer`
- `tf.train.RMSPropOptimizer`

### 2.4.1 BGD

BGD(batch gradient descent) 批量梯度下降。这种方法是利用现有的参数对训练集中的每一个输入生成一个估计输出  $y_i$ , 然后跟实际的输出  $y_i$  比较, 统计所有的误差, 求平均后的到平均误差作为更新参数的依据。啊他的迭代过程是:

1. 提取训练集集中所有内容  $\{x_1, \dots, x_n\}$ , 以及相关的输出  $y_i$ ;
2. 计算梯度和误差并更新参数。

这种方法的优点是: 使用所有数据计算, 都保证收敛, 并且不需要减少学习率。缺点是每一步需要使用所有的训练数据, 随着训练的进行, 速度会变慢。那么如果将训练数据拆分成一个个 batch, 每次抽取一个 batch 数据更新参数, 是不是能加速训练? 这就是 SGD。

### 2.4.2 SGD

SGD(stochastic gradient descent): 随机梯度下降。这种方法的主要思想是将数据集拆分成一个个的 batch, 随机抽取一个 batch 计算并更新参数, 所以也称为 MBGD(minibatch gradient descent) SGD 在每次迭代计算 mini-batch 的梯度, 然后对参数进行更新。和 BGD 相比, SGD 在训练数据集很大时也能以较快的速度收敛, 但是它有两个缺点:

1. 需要手动调整学习率, 此外选择合适的学习率比较困难。尤其在训练时, 我们常常想对常出现的特征更快速的更新, 对不常出现的特征更新速度慢些, 而 SGD 更新参数时对所有参数采用一样的学习率, 因此无法满足要求。
2. SGD: 容易收敛到局部最优。

### 2.4.3 momentum

Momentum 是模拟物理学中的动量概念，更新时在一定程度上保留之前的更新方向，利用当前批次再次微调本次更新参数，因此引入了一个新的变量  $v$ ，作为前几次梯度的累加。因此，momentum 能够更新学习率，在下降初期，前后梯度方向一致时能加速学习；在下降的中后期，在局部最小值附近来回振荡，能够抑制振荡加快收敛。

### 2.4.4 Nesterov Momentum

标准的 Monentum 法首先计算一个梯度，然后在加速更新梯度的方向进行一个大的跳跃 Nesterov 首先在原来加速的梯度方向进行一个大的跳跃，然后在改为值设置计算梯度值，然后用这个梯度值修正最终的更新方向。

### 2.4.5 Adagrad

Adagrade 能够自适应的为各个参数分配不同的学习率，能够控制每个维度的梯度方向，这种方法的优点是能实现学习率的自动更改，如果本次更新时梯度大，学习率就衰减得快，如果这次更新时梯度小，学习率衰减得就慢些。

### 2.4.6 RMSprop

和 Momentum 类似，通过引入衰减系数使得每个回合都衰减一定比例。在实践中，对循环神经网络效果很好。

### 2.4.7 Adam

名称来自自适应矩阵 (adaptive moment estimation).Adam 根据损失函数针对每个参数的一阶矩，二阶矩估计动态调整每个参数的学习率。

```

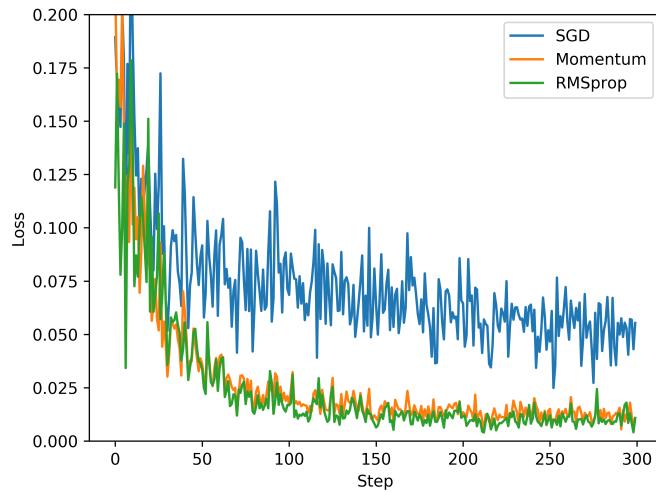
1 import numpy as np
2 import tensorflow as tf
3 import matplotlib.pyplot as plt
4 tf.set_random_seed(0)
5 np.random.seed(0)
6 LR = 0.01
7 BATCH_SIZE = 32
8 x = np.linspace(-1, 1, 100).reshape(-1, 1)
9 noise = np.random.normal(0, 0.1, size=x.shape)
10 y = np.power(x, 2)+noise
11 class Net:
12     def __init__(self, opt, **kwargs):
13         self.x = tf.placeholder(tf.float32, [None, 1])

```

```

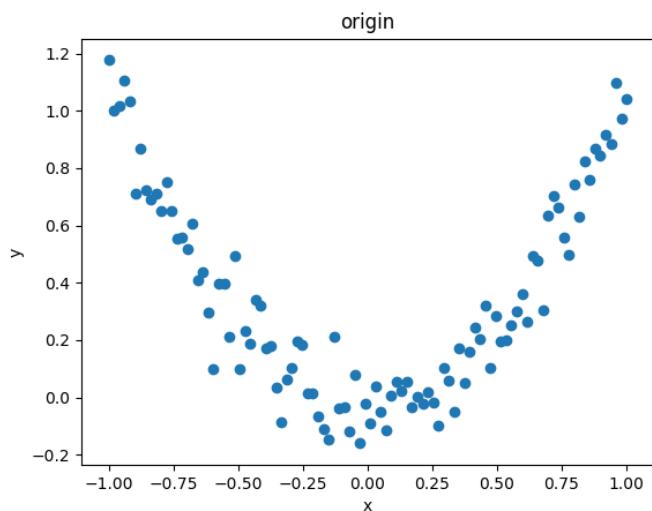
14     self.y = tf.placeholder(tf.float32, [None,1])
15     l = tf.layers.dense(self.x,20,tf.nn.relu)
16     out = tf.layers.dense(l,1)
17     self.loss = tf.losses.mean_squared_error(self.y,out)
18     self.train = opt(LR,**kwargs).minimize(self.loss)
19 net_SGD = Net(tf.train.GradientDescentOptimizer)
20 net_momentum = Net(tf.train.MomentumOptimizer,momentum=0.9)
21 net_RMSprop = Net(tf.train.RMSPropOptimizer)
22 net_Adam = Net(tf.train.AdamOptimizer)
23 nets = [net_SGD,net_momentum,net_RMSprop,net_Adam]
24 sess = tf.Session()
25 sess.run(tf.global_variables_initializer())
26 losses_his = [[],[],[]]
27 for step in range(300):
28     index = np.random.randint(0,x.shape[0],BATCH_SIZE)
29     b_x = x[index]
30     b_y = y[index]
31     for net,l_his in zip(nets,losses_his):
32         _,l = sess.run([net.train,net.loss],{net.x:b_x,net.y:b_y})
33         l_his.append(l)
34 labels = ['SGD','Momentum','RMSprop','Adam']
35 for i,l_his in enumerate(losses_his):
36     plt.plot(l_his,label=labels[i])
37 plt.legend(loc='best')
38 plt.xlabel('Step')
39 plt.ylabel('Loss')
40 plt.ylim(0,0.2)
41 plt.savefig('Opt.png',dpi=600)

```



#### 2.4.8 构造简单的神经网络拟合数据

原始数据为  $y = x^2$  的基础上添加随机噪声。原始数据的散点图如下



```

1 #tensorflow 1.2.1
2 import tensorflow as tf
3 import matplotlib.pyplot as plt
4 import numpy as np
5 tf.set_random_seed(0)
6 np.random.seed(0)

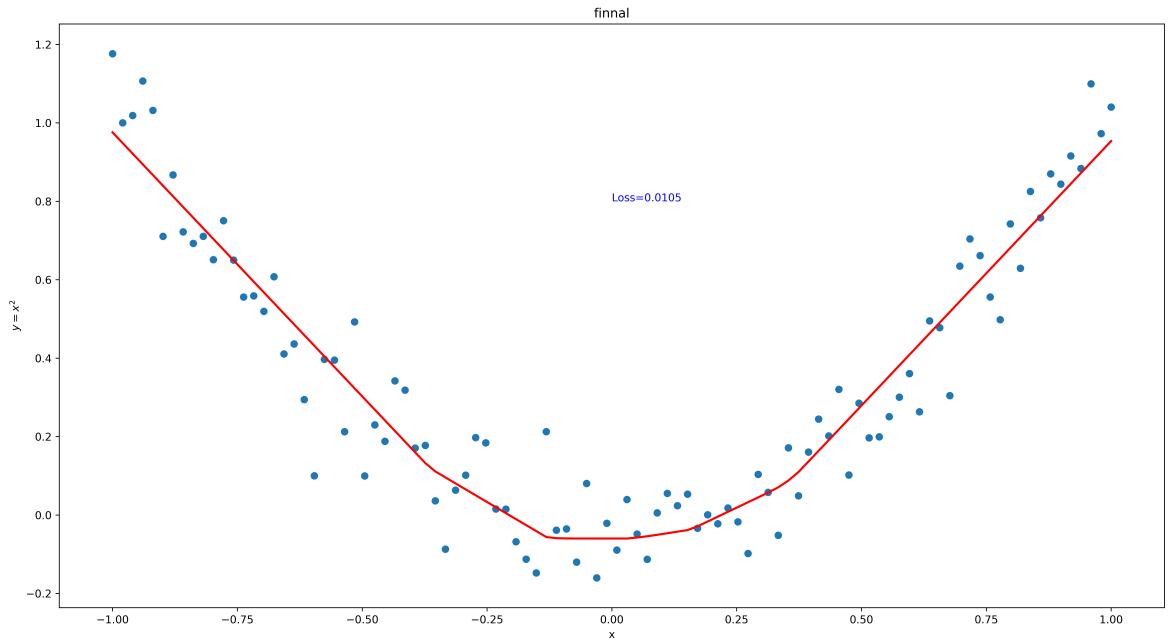
```

```

7 #生成数据
8 step = 100
9 x = np.linspace(-1,1,step).reshape(-1,1)
10 noise = np.random.normal(0,0.1,size=x.shape)
11 y = np.power(x,2)+noise
12
13 tf_x = tf.placeholder(tf.float32,x.shape)
14 tf_y = tf.placeholder(tf.float32,x.shape)
15 l1 = tf.layers.dense(tf_x,10,tf.nn.relu)
16 output = tf.layers.dense(l1,1)
17
18 loss = tf.losses.mean_squared_error(tf_y,output)
19 optimizer = tf.train.GradientDescentOptimizer(learning_rate=0.5)
20 train_op = optimizer.minimize(loss)
21
22 sess = tf.Session()
23 sess.run(tf.global_variables_initializer())
24 plt.ion()
25 for step in range(100):
26     _,l,pred = sess.run([train_op,loss,output],{tf_x:x,tf_y:y})
27     if step%5==0:
28         plt.cla()
29         plt.scatter(x,y)
30         plt.title(r'$y=x^2+noise$')
31         plt.plot(x,pred,'r-',lw=2)
32         plt.text(0,0.8,'Loss=%f' % l,fontdict={'size':10,'color':'blue'})
33         plt.xlabel("x")
34         plt.ylabel(r"$y=x^2$")
35         plt.pause(0.1)
36 plt.ioff()
37 plt.show()

```

最终拟合数据:



## 2.5 TensorBoard

```

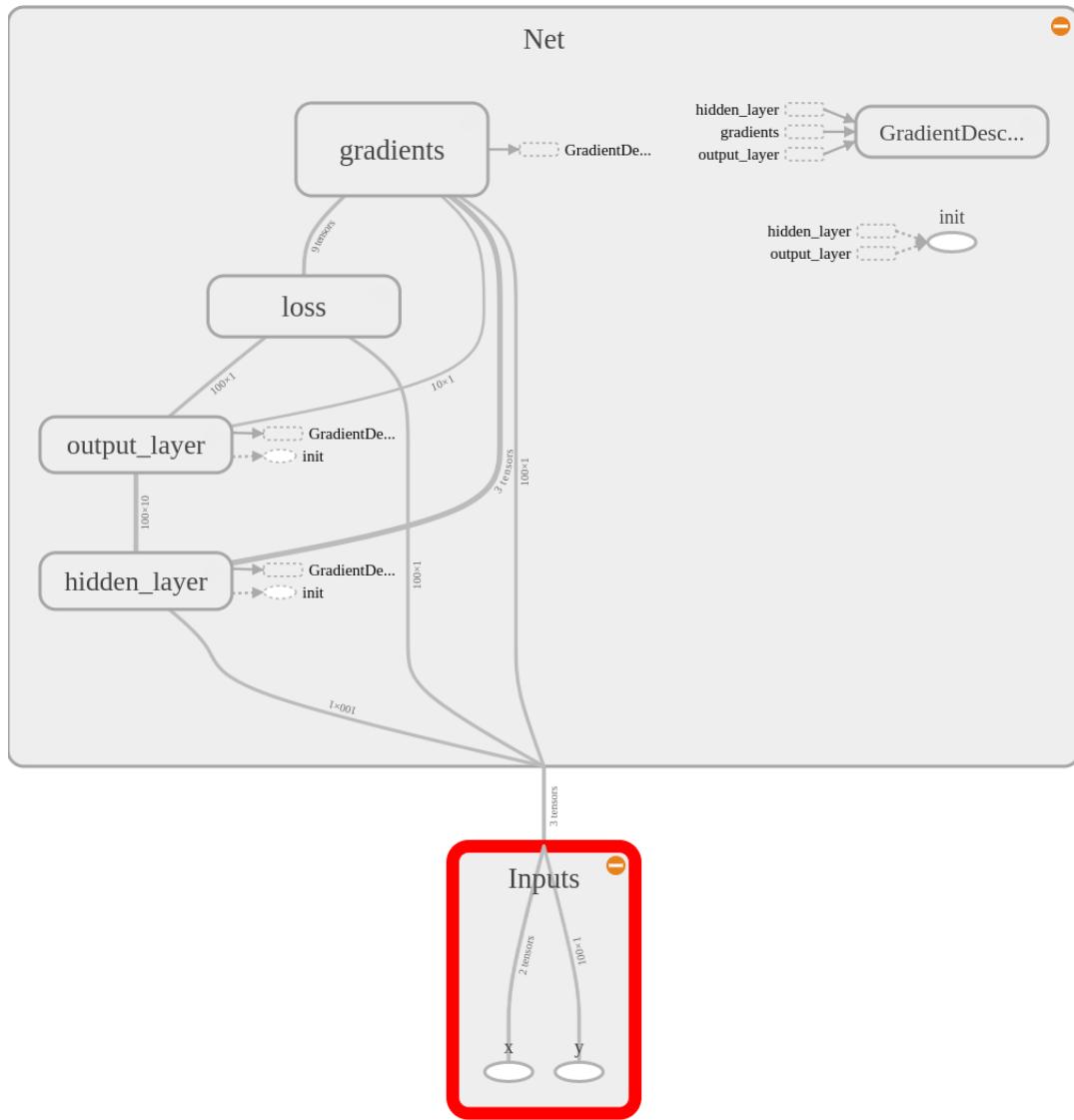
1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3
4 tf.set_random_seed(1)
5 x0 = tf.random_normal((100,2),2,2,tf.float32,0)
6 y0 = tf.zeros(100)
7 x1 = tf.random_normal((100,2),-2,2,tf.float32,0)
8 y1 = tf.ones(100)
9 x = tf.reshape(tf.stack((x0,x1),axis=1),(200,2))
10 y = tf.reshape(tf.stack((y0,y1),axis=1),(200,1))
11 with tf.Session() as sess:
12     x = sess.run(x)
13     y = sess.run(y)
14
15 tf_x = tf.placeholder(tf.float32, x.shape)      # input x
16 tf_y = tf.placeholder(tf.int32, y.shape)        # input y
17
18 # neural network layers
19 l1 = tf.layers.dense(tf_x, 10, tf.nn.relu)       # hidden layer

```

```

20 output = tf.layers.dense(11, 2)                      # output layer
21
22 loss = tf.losses.sparse_softmax_cross_entropy(labels=tf_y, logits=output)
23                                         # compute cost
23 accuracy = tf.metrics.accuracy(                  # return (acc, update_op), and create 2
24                                         labels=tf.squeeze(tf_y), predictions=tf.argmax(output, axis=1),)[1]
25 optimizer = tf.train.GradientDescentOptimizer(learning_rate=0.05)
26 train_op = optimizer.minimize(loss)
27
28 sess = tf.Session()
29
30                                         # control training and others
31 init_op = tf.group(tf.global_variables_initializer(), tf.
32                                         local_variables_initializer())
33 sess.run(init_op)      # initialize var in graph
34
35 plt.ion()    # something about plotting
36 for step in range(100):
37     _, acc, pred = sess.run([train_op, accuracy, output], {tf_x: x, tf_y: y})
38     if step % 2 == 0:
39         plt.cla()
40         plt.scatter(x[:, 0], x[:, 1], c=pred.argmax(1), s=100, lw=0, cmap='RdYlGn')
41         plt.text(1.5, -4, 'Accuracy=% .2f' % acc, fontdict={'size': 20, 'color': 'red'})
42         plt.pause(0.1)
43 plt.ioff()
44 plt.show()

```



### 2.5.1 TensorBoard Histogram Dashboard

TensorBoard Histogram Dashboard 显示 TensorFlow 图中的 Tensor 如何随着时间变化。

### 2.5.2 一个简单的例子

正态分布变量，均值随着和时间移动。TensorFlow 有一个操作 `tf.random_normal` 可以完美的达到这个目的。正如通常情况下 TensorBoard，我们将用 `summary op` 融合数据

据。在这种情况下'tf.summary.histogram'。这里有一个代码段将生成一些包含正态分布直方图数据的总结，这里均值随着时间增大。

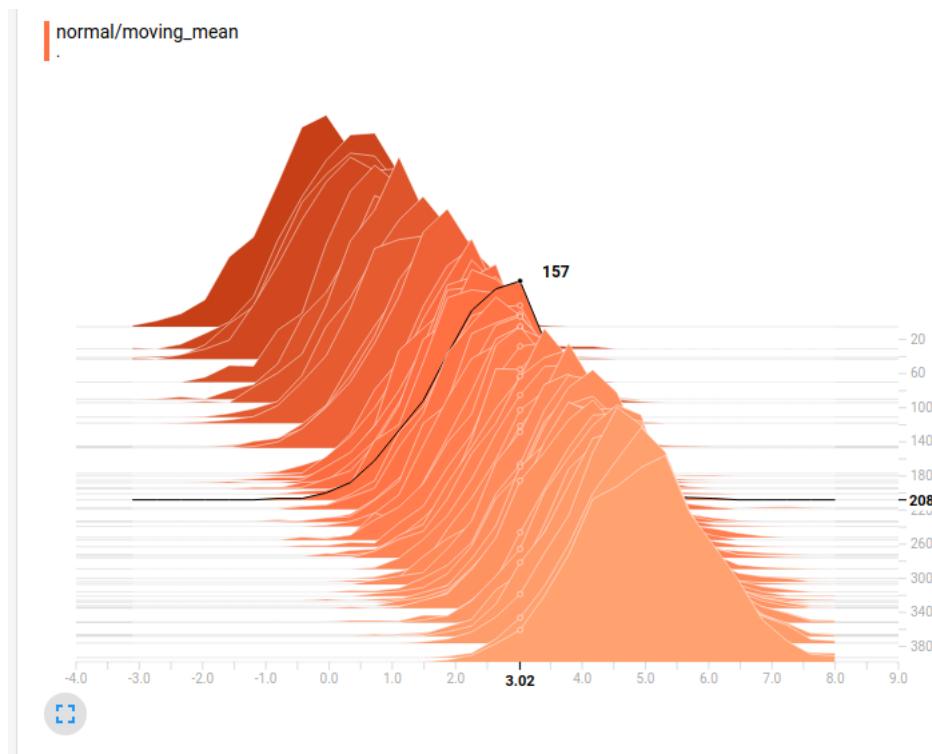
```

1 import tensorflow as tf
2 k = tf.placeholder(tf.float32)
3 mean_moving_normal = tf.random_normal(shape=[1000], mean=(5*k), stddev=1)
4 summaries = tf.summary.histogram('normal/moving_mean', mean_moving_normal)
5 sess = tf.Session()
6 writer = tf.summary.FileWriter('./histogram_example')
7 N = 400
8 for step in range(N):
9     k_val = step/float(N)
10    summ = sess.run(summaries, feed_dict={k:k_val})
11    writer.add_summary(summ, global_step=step)

```

在当前代码中运行下边的代码启动 TensorFlow 载入数据

```
1 tensorboard --logdir=./histogram_example
```



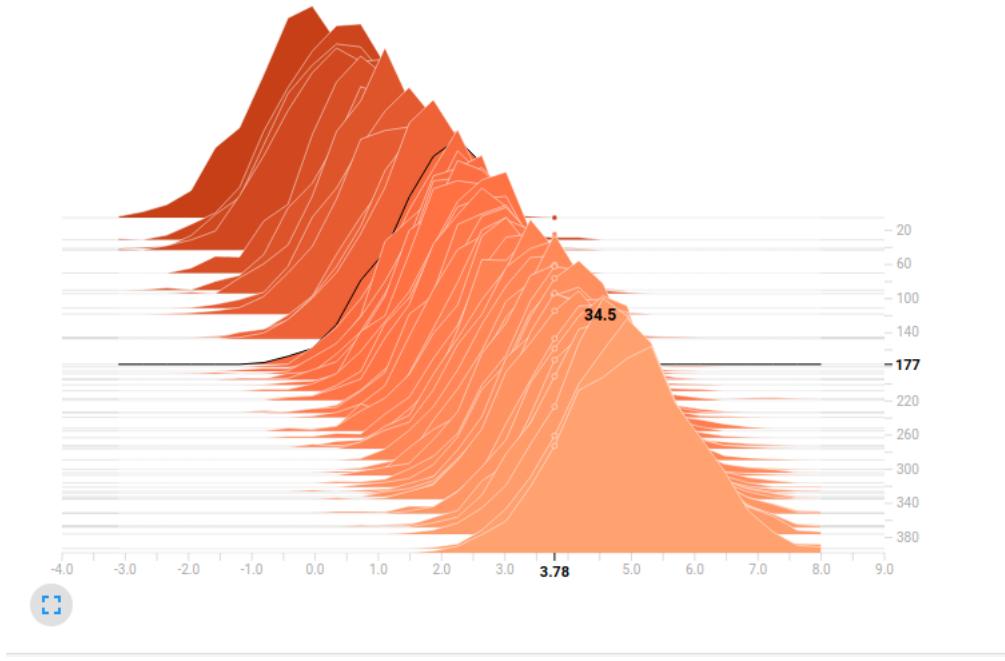
tf.summary.histogram 接受任一尺寸和大小的 Tensor，压缩他们进入直方数据结构组成一些小的数据宽度和数量组层的 bin 将该够，例如我们像组成数 [0.5,1.1,1.3,2.2,2.9,2.99] 成 3

个 bin，我们可以创建三个 bin：一个包含 0 到 1 之间的一切 (0.5)，一个包含 1-2(1.1,1.3) 之间，一个包含 2-3(2.2,2.9,2.99)

TensorFlow 用类是的方法创建 bins，但是不想我们上面的例子，它不创建整数读额 bins，瑞与大型数据，稀疏数据，这样的也许导致上千个 bin，bins 时指数分布时，一些 bins 相比于一些非常大数的 bin 接近于 0。然而，可视化指数分布 bin 时一个技巧，如果高被编码为数量，bin 宽度更大的空间，甚至他们有相同的元素，相比较之下统计数量使得豪赌比较变得可能，直方图采集数据仅均匀的 bins，这可能导致不幸的人工操作。

在直方图可视化器的每一个切片显示为一个单个的直方图。切片安装步数组组织。例老的切片 (e.g. step 0) 比较靠后变为更深，然而新的 slices 接近于前景色，颜色更轻，右边的 y 轴显示了步数。

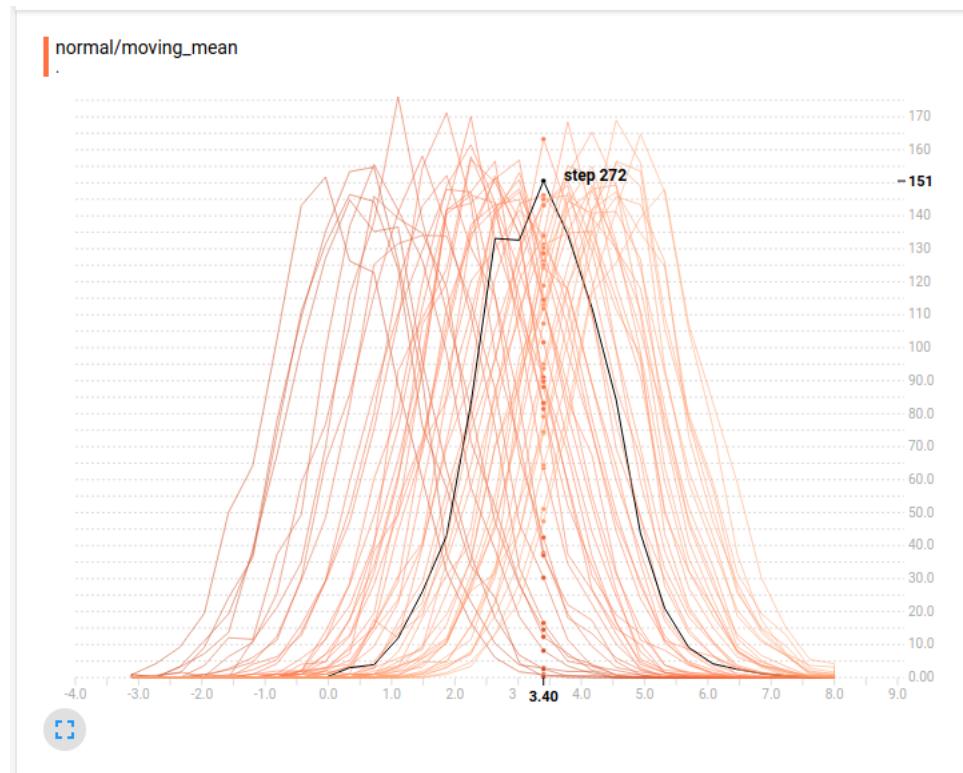
你可以在直方图上滑动鼠标看到更多的详细星系。你如下面的图你可以看到直方图的时间不为 177 有一个 bin 中心在 3.78 有 bin 中有 34.5 个元素。



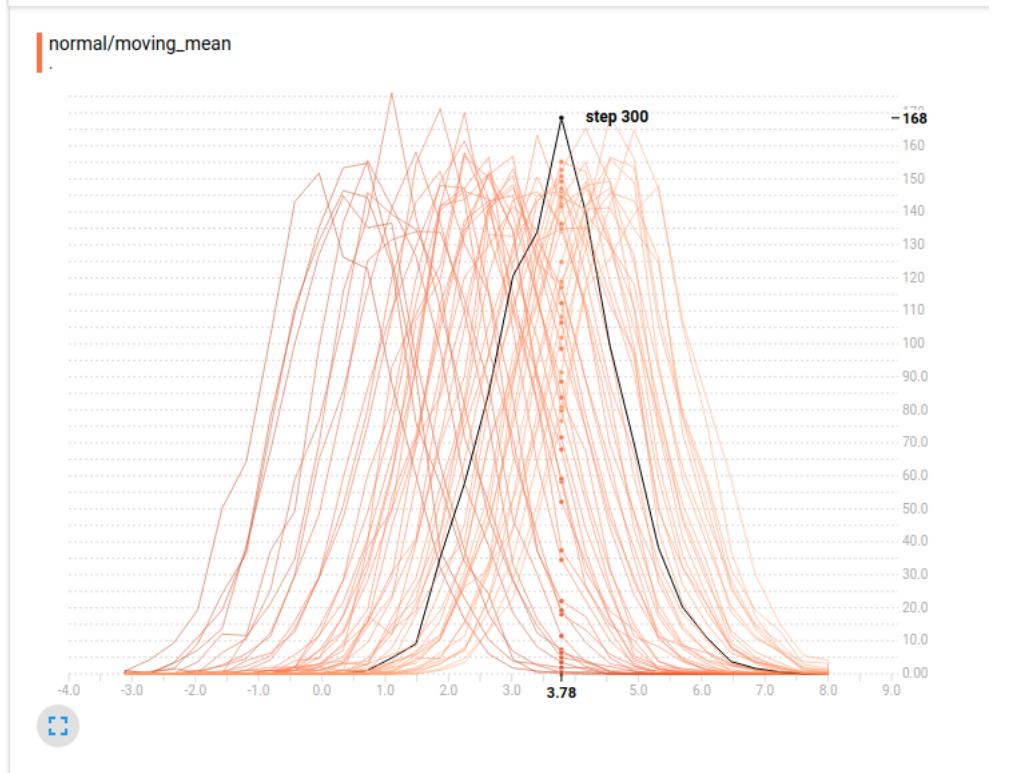
你也许注意到注意到直方切片在统计步数和时间上不总是偶数，这是因为 TensorBoard 用[reservoir sampling](#)保持直方图的子集，为了节约内存，Reservior sampling 保证每个采样有一个相等的可能性被包含进去，但是因为它时一个随机算法，采样并不在每个偶数步发生。

### 2.5.3 Overlay Mode

控制面板上允许你打开直方图模式为 offset 为 overlay。在 offset 模式下，可视化转动 45 度，因此单个的直方图切片不再展开，而是所有的图共享一个相同的 y 轴上。



现在表上的每个切片被线分开，y 轴显示每个 bucket 项目数量，深色线时老的，早期的时间不，浅色线时最近的新时间不，你可以用鼠标在表上查看更多的信息。



overlay 可视化在你想直接比较不同直方图的数量。

#### 2.5.4 多个分布

直方图控制面板对多分布下的可视化很有用，当我们通过链接两个不同的正态分布构造一个简单的二两分布，代码如下：

```

1 import tensorflow as tf
2 k = tf.placeholder(tf.float32)
3 mean_moving_normal = tf.random_normal(shape=[1000], mean=(3*5), stddev=1)
4 tf.summary.histogram('normal/moving_mean', mean_moving_normal)
5 variance_shrinking_normal = tf.random_normal(shape=[100], mean=0, stddev=1-(k))
6 tf.summary.histogram('normal/shrinking_varance', variance_shrinking_normal)
7 normal_combined = tf.concat([mean_moving_normal, variance_shrinking_normal], 0)
8 tf.summary.histogram('normal/bimodal', normal_combined)
9 summaris = tf.summary.merge_all()
10 sess = tf.Session()
11 writer = tf.summary.FileWriter('./ histgram_example1')
12 N = 400
13 for step in range(N):

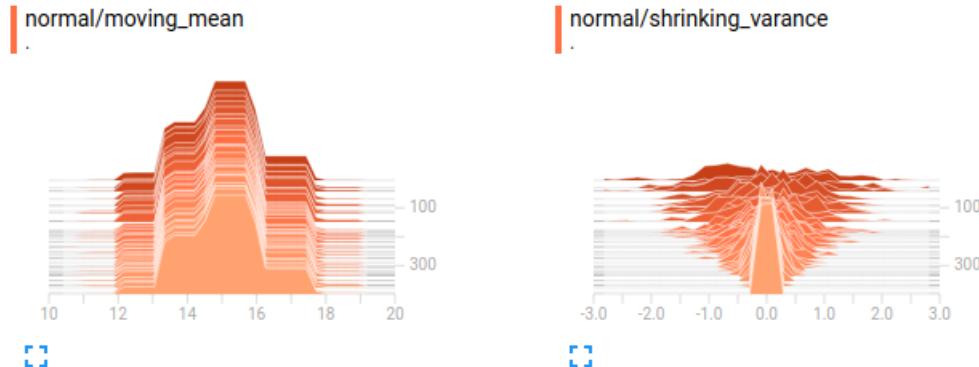
```

```

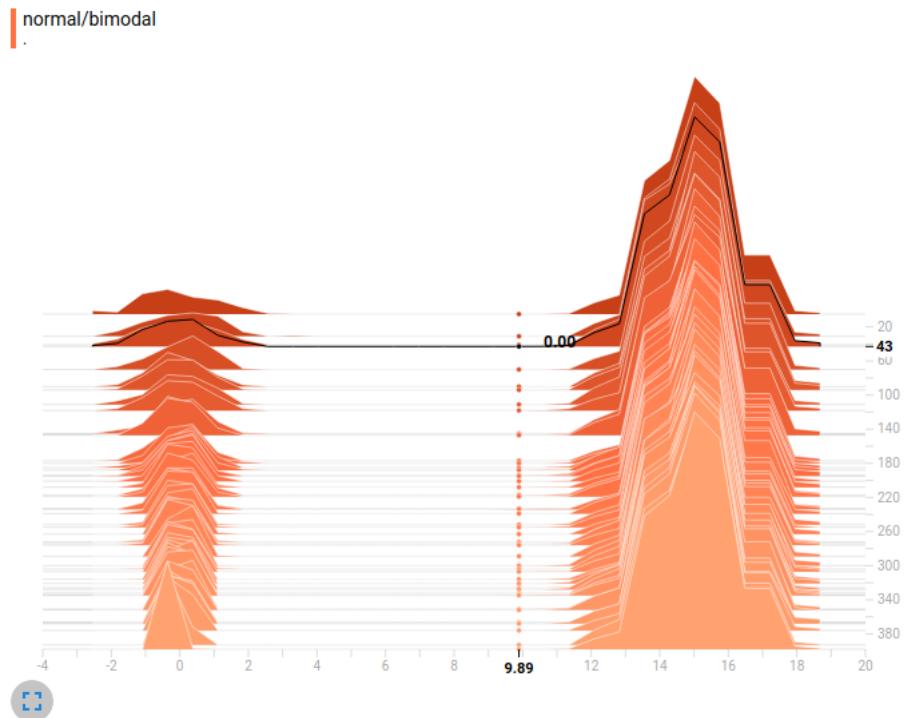
14 k_val = step/float(N)
15 summ = sess.run(summaris, feed_dict={k:k_val})
16 writer.add_summary(summ, global_step=step)

```

上面的例子是滑动平均，现在我们已有一个收缩的变量分布。



当我们链接她们在一起，我们得到一个清晰解释分歧，二进制结构的表格：



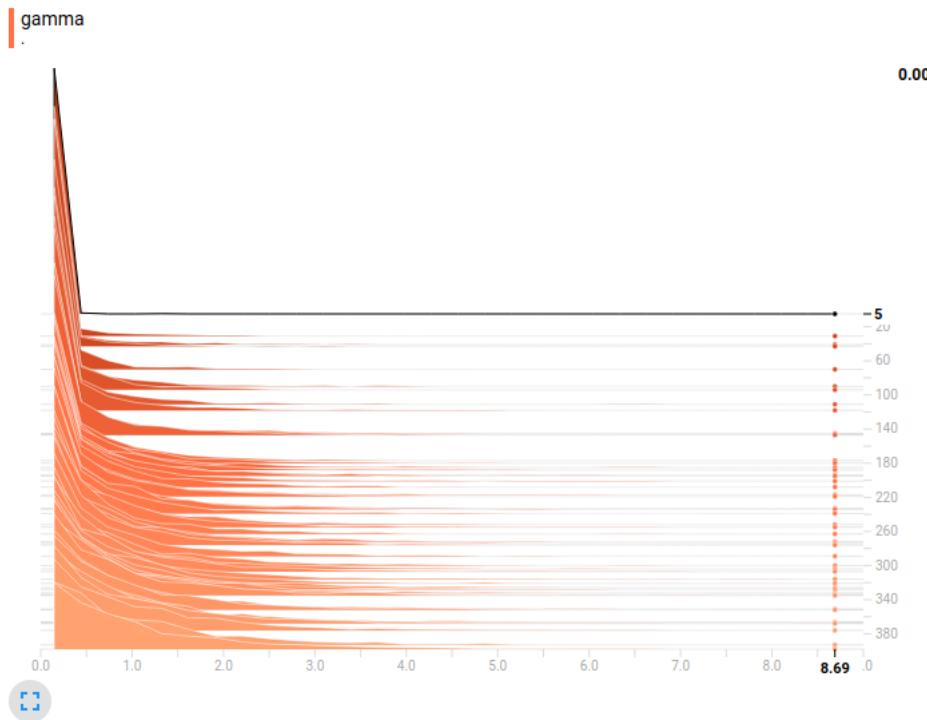
### 2.5.5 更多分布

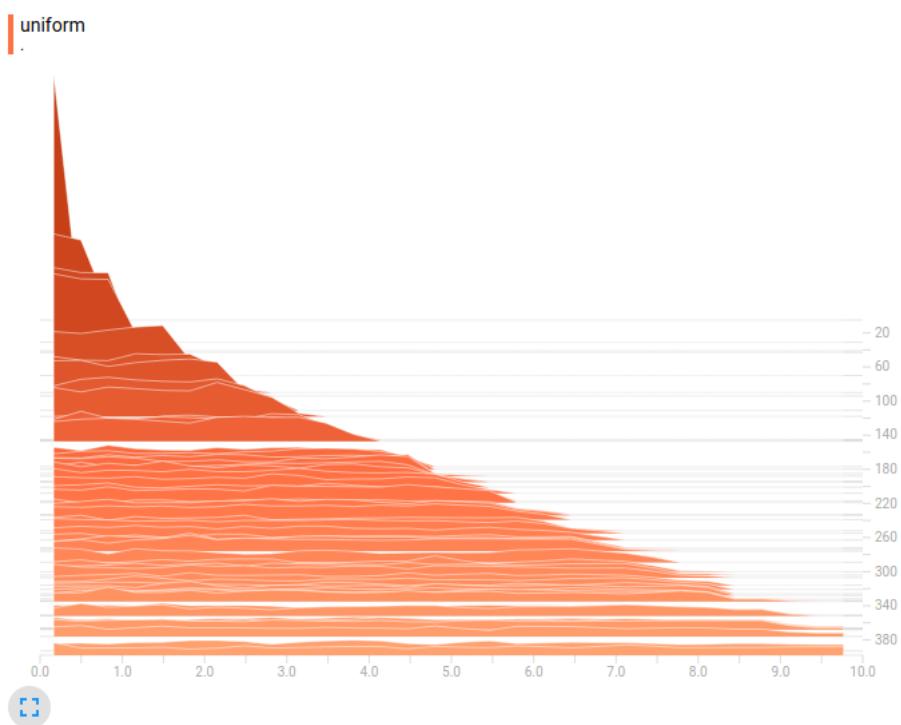
生成可视化更多分布，结合他们到表中：

```

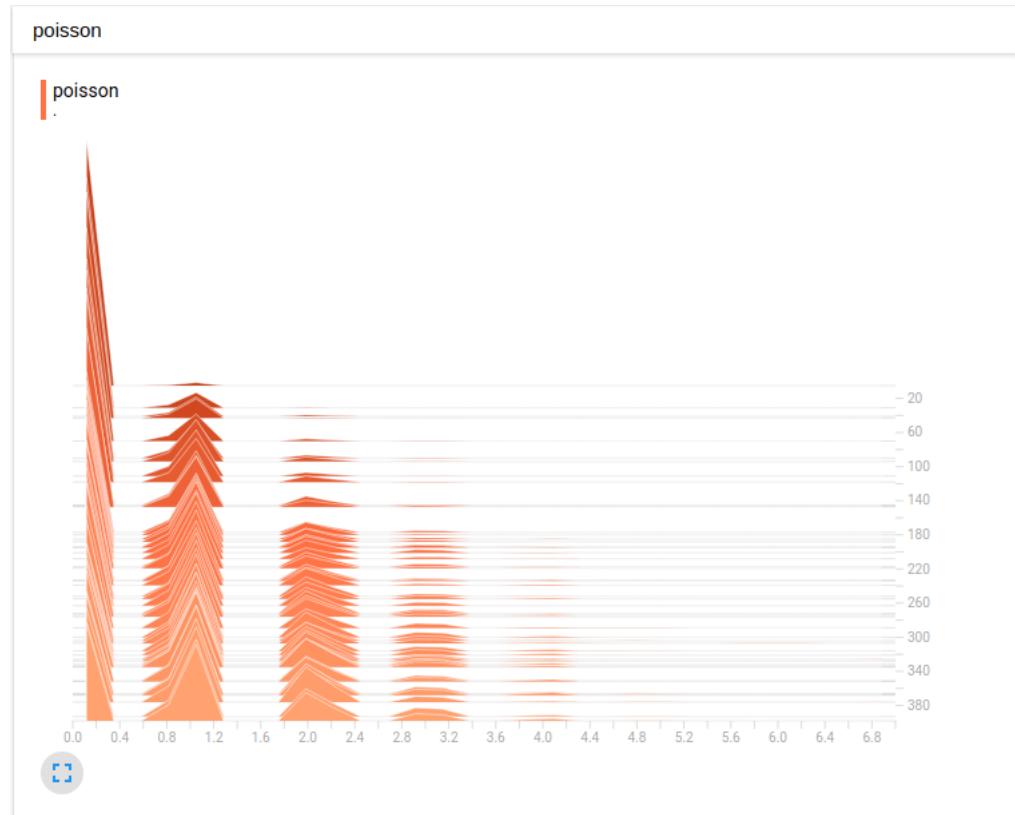
1 import tensorflow as tf
2 k = tf.placeholder(tf.float32)
3 # Make a normal distribution ,with a shift mean
4 mean_moving_normal = tf.random_normal(shape=[1000],mean=(5*k),stddev=1)
5 tf.summary.histogram('normal/moving_mean',mean_moving_normal)
6 variance_shrinking_normal = tf.random_normal(shape=[1000],mean=0,stddev=1-(k))
7 tf.summary.histogram('normal/shrinking_variance',variance_shrinking_normal)
8 normal_combined = tf.concat([mean_moving_normal,variance_shrinking_normal],0)
9 tf.summary.histogram("normal/bimodal",normal_combined)
10 #add gamma distribution
11 gamma = tf.random_gamma(shape=[1000],alpha=k)
12 tf.summary.histogram('gamma',gamma)
13 poisson = tf.random_poisson(shape=[1000],lam=k)
14 tf.summary.histogram('poisson',poisson)
15 #add a uniform distribution
16 uniform = tf.random_uniform(shape=[1000],maxval=k*10)
17 tf.summary.histogram('uniform',uniform)
18 #finnally combine everything together
19
20 all_distributions = [mean_moving_normal,variance_shrinking_normal,gamma,poisson,
21                      uniform]
22 all_combined = tf.concat(all_distributions,0)
23 tf.summary.histogram('all_combined',all_combined)
24 summaries = tf.summary.merge_all()
25 sess = tf.Session()
26 writer = tf.summary.FileWriter('./histogram_example2')
27 N = 400
28 for step in range(N):
29     k_val = step/float(N)
30     summ = sess.run(summaries,feed_dict={k:k_val})
31     writer.add_summary(summ,global_step=step)

```



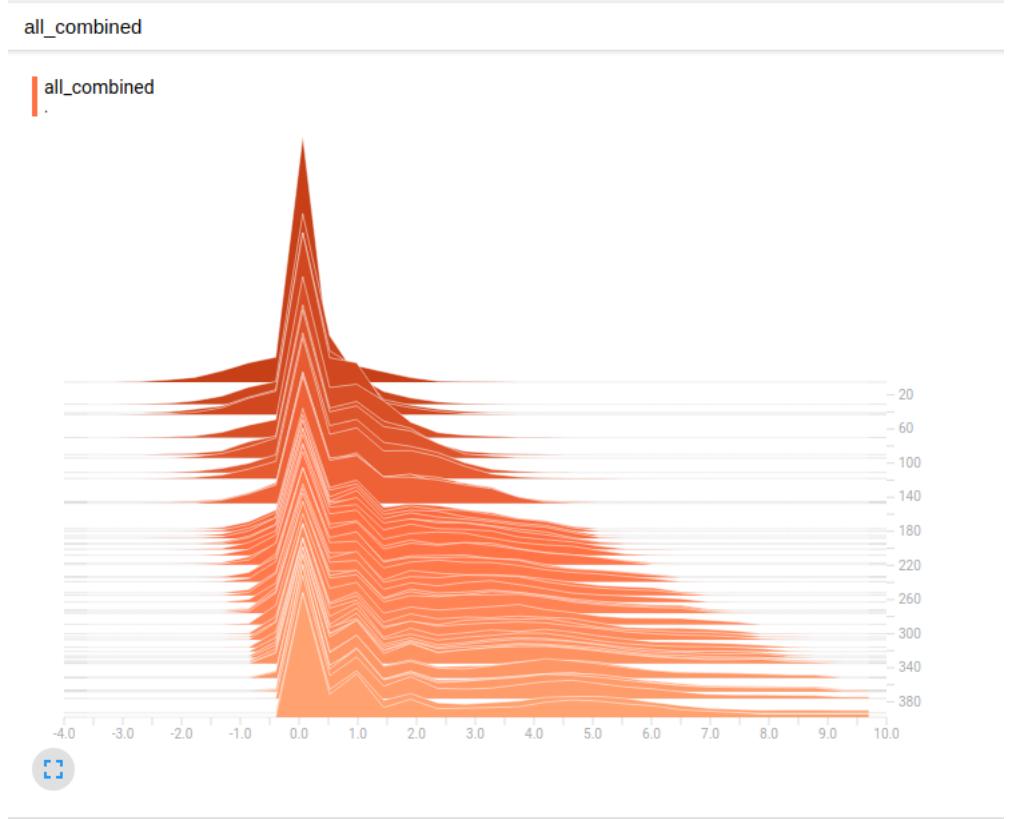


### 2.5.6 poisson 分布



poisson 分布定义在整数上，因此所有被生成的值都是整数，直方图压缩移动数据到浮点 bins，导致可视化在整数值上显示一点点突起。

### 2.5.7 结合所有的数据到一张图向上



```
1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 import numpy as np
4 tf.set_random_seed(0)
5 np.random.seed(0)
6 x = np.linspace(-1, 1, 100).reshape(-1, 1)
7 noise = np.random.normal(0, 0.1, size=x.shape)
8 y = np.power(x, 2) + noise
9 def gendata():
10     t = np.linspace(-1, 1, 100).reshape(-1, 1)
11 def save():
12     print('This is save')
13     tf_x = tf.placeholder(tf.float32, x.shape)
14     tf_y = tf.placeholder(tf.float32, y.shape)
15     l = tf.layers.dense(tf_x, 10, tf.nn.relu)
16     o = tf.layers.dense(l, 1)
```

```

17 loss = tf.losses.mean_squared_error(tf_y,o)
18 train_op = tf.train.GradientDescentOptimizer(learning_rate=0.5).minimize(
19             loss)
20 sess = tf.Session()
21 sess.run(tf.global_variables_initializer())
22 saver = tf.train.Saver()
23 for step in range(100):
24     sess.run(train_op,{tf_x:x,tf_y:y})
25 saver.save(sess,'params',write_meta_graph=False)
26 pred,l = sess.run([o,loss],{tf_x:x,tf_y:y})
27 plt.figure(1,figsize=(10,5))
28 plt.subplot(121)
29 plt.scatter(x,y)
30 plt.plot(x,pred,'r-',lw=5)
31 plt.text(-1,1.2,'save loss=%f'%l,fontdict={'size':15,'color':'red'})
32 def reload():
33     print('This is reload')
34     tf_x = tf.placeholder(tf.float32,x.shape)
35     tf_y = tf.placeholder(tf.float32,y.shape)
36     l_ = tf.layers.dense(tf_x,10,tf.nn.relu)
37     o_ = tf.layers.dense(l_,1)
38     loss_ = tf.losses.mean_squared_error(tf_y,o_)
39     sess = tf.Session()
40     saver = tf.train.Saver()
41     saver.restore(sess,'params')
42     pred,l = sess.run([o_,loss_],{tf_x:x,tf_y:y})
43     plt.subplot(122)
44     plt.scatter(x,y)
45     plt.plot(x,pred,'r-',lw=5)
46     plt.text(-1,1.2,'Reload Loss=%f'%l,fontdict={'size':15,'color':'red'})
47     plt.show()
48 save()
49 tf.reset_default_graph()
50 reload()

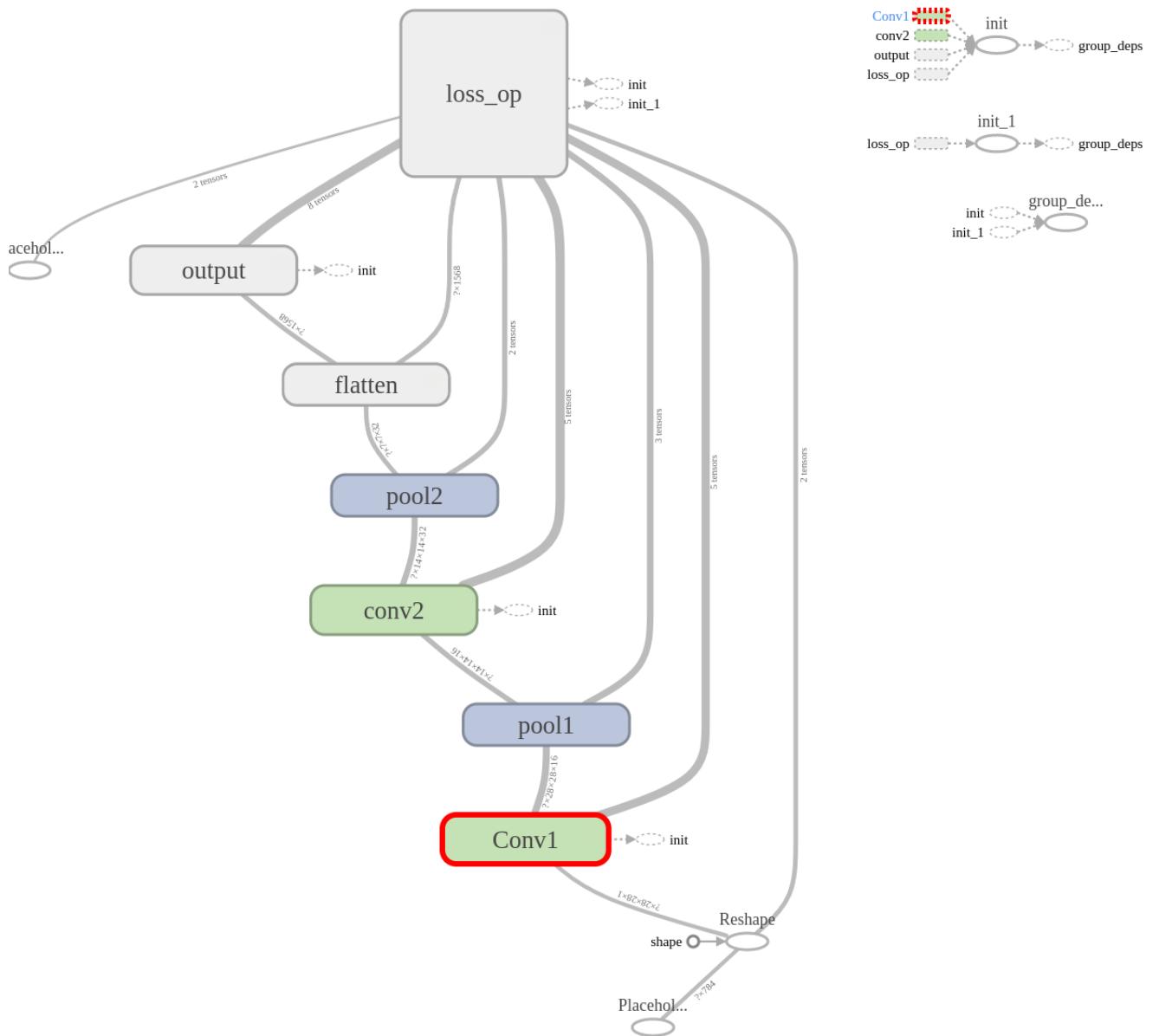
```

## 2.6 CNN 手写体数据识别

### 2.6.1 mnist 数据集

手写体数据训练集有 55000 张手写体数据图片。测试集有 10000 张图片。每张图片是大小为 32\*32 的灰度图片。卷积神经网络结构：

- 第一层卷积层：卷积核 16 个，卷积核大小为  $5 \times 5$ ,strides=1,padding 为 SAME，激活函数为 relu(输出大小为  $28 \times 28 \times 16$ )。
- 第一层池化层：池化层大小为 2,strides 为 2( $14 \times 14 \times 16$ )。第二层卷积层：卷积核 32，大小为  $5 \times 5$ ,strides=1,padding 为 SAME，激活函数为 relu。 $(14 \times 14 \times 32)$
- 第二层池化层：池化层大小为 2,strides 为 2( $7 \times 7 \times 32$ )。
- flatten:1568。



```

1 import tensorflow as tf
2 import matplotlib.pyplot as plt
3 import numpy as np
4 from tensorflow.examples.tutorials.mnist import input_data
5
6 tf.set_random_seed(0)
7 np.random.seed(0)
8

```

```

9 BATCH_SIZE = 50
10 LR = 0.001
11 mnist = input_data.read_data_sets('/home/hpc/文档/mnist_tutorial/mnist', one_hot
12 = True)
13 test_x = mnist.test.images[:2000]
14 test_y = mnist.test.labels[:2000]
15
16 tf_x = tf.placeholder(tf.float32, [None, 28*28])
17 images = tf.reshape(tf_x, [-1, 28, 28, 1])
18 tf_y = tf.placeholder(tf.int32, [None, 10])
19 with tf.variable_scope('Conv1'):
20     conv1 = tf.layers.conv2d(
21         inputs = images,
22         filters = 16,
23         kernel_size = 5,
24         strides = 1,
25         padding = 'same',
26         activation = tf.nn.relu
27     )
28     tf.summary.histogram('conv1', conv1)
29 with tf.variable_scope('pool1'):
30     pool1 = tf.layers.max_pooling2d(
31         conv1,
32         pool_size=2,
33         strides =2
34     )
35     tf.summary.histogram('max_pool1', pool1)
36 with tf.variable_scope('conv2'):
37     conv2 = tf.layers.conv2d(pool1, 32, 5, 1, 'SAME', activation=tf.nn.relu)
38     tf.summary.histogram('conv2', conv2)
39 with tf.variable_scope('pool2'):
40     pool2 = tf.layers.max_pooling2d(conv2, 2, 2)
41     tf.summary.histogram('max_pool', pool2)
42 with tf.variable_scope('flatten'):
43     flat = tf.reshape(pool2, [-1, 7*7*32])
44 with tf.variable_scope('output'):
45     output = tf.layers.dense(flat, 10)
46 with tf.variable_scope('loss_op'):
47     loss = tf.losses.softmax_cross_entropy(onehot_labels=tf_y, logits=output)
48     train_op = tf.train.AdamOptimizer(LR).minimize(loss)
49     accuracy = tf.metrics.accuracy(labels = tf.argmax(tf_y, axis=1), predictions =
50                                     tf.argmax(output, axis=1),)[1]
51     tf.summary.scalar('loss', loss)

```

```

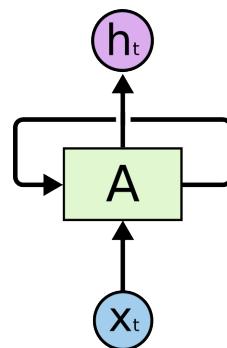
50     tf.summary.scalar('accuracy', accuracy)
51 sess = tf.Session()
52 merge_op = tf.summary.merge_all()
53 init_op = tf.group(tf.global_variables_initializer(), tf.
54                     local_variables_initializer())
55 sess.run(init_op)
56 writer = tf.summary.FileWriter('./log', sess.graph)
57 for step in range(600):
58     b_x, b_y = mnist.train.next_batch(BATCH_SIZE)
59     _, loss_, result = sess.run([train_op, loss, merge_op], {tf_x:b_x, tf_y:b_y})
60     writer.add_summary(result, step)
61     if step%50 == 0:
62         accuracy_, flat_representation = sess.run([accuracy, flat], {tf_x:test_x,
63                                         tf_y:test_y})
64         print('Step:', step, '| train loss: %.4f | loss_: %.2f | test accuracy: %.2f' %
65               accuracy_)
66 test_output = sess.run(output, {tf_x:test_x[:10]})
```

## 2.7 RNN

人不能抓住每一秒的思考，当你读这篇文章的时候，你能基于你之前的对单词的理解明白文章的每一个单词的意思，你思考的时候不需要丢掉所有的东西，你的思想有持续性。

传统的神经网络很难做到这点，这也是传统神经网络的主要缺点。例如你想分类电影中的不同时间点的事件，传统神经网络用不清楚如何用之前的事件了解新的事件。

RNN 通过循环处理这个问题，允许信息保留。



上面的图表示一个 RNN 单元， $A$  得到输入  $x_t$  和输出  $h_t$ ， $A$  允许信息被循环从一步到下一步，一个循环神经网络可以看成是多个相同单元的复制。铺开 RNN 可以得到这个链式结构揭示了循环神经网络和序列或者列表密切相关，它适用于这种数据。

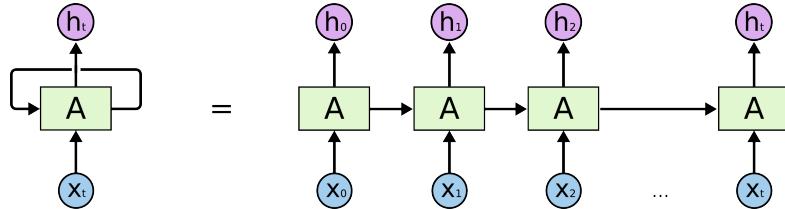
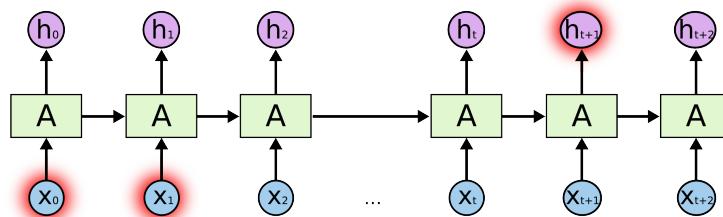


图 2.4: unrolled RNN

### 2.7.1 The Problem Long-Term Dependencies

语言模型中常用先前的一个词预测下一个词，如果我们尝试预测“the clouds are in the sky”我们不需要很多上下文信息 RNN 通过之前的信息就能学到。但是我们尝试预测这样一个句子“I grew up in France... I speak fluent **France**”，之前的信息暗示下一个单词可能是语言的名字，如果我们想去缩小语言的范围，我们需要上下文**France**，可相关信息和这个需要点的间隔很大。理论上 RNN 有能力处理“long-term dependencies”，人能小心的挑



选参数解决这个烦人的问题，然而不幸的是 RNN 似乎不能做到，原因由 [Hochreiter \(1991\)](#) [[German](#)] and [Bengio, et al. \(1994\)](#) 提出。

### 2.7.2 LSTM 网络

Long Short Term Memory networks 通常简称为 LSTMs 是一个特殊的 RNN，能学习 learning long-term dependencies，他被 [Hochreiter Schmidhuber \(1997\)](#) 引入，然后被提炼，在大型文体处理上效果很好因而被广泛的使用。

LSTMs 明确的设计去解决 long-term dependency problem。

所有的循环神经网络都有重复的链式形式。在标准的 RNNs，重复的模块有一个非常简单的结构，像 tanh Layer。LSTMs 也有这样类似的结构，但是 congruent 模块有点不同，有一个神经网络层有四个相互作用部分，在上面的图上，每一根线上携带的都是一个向量，从一个输出节点到其他输入，粉色圆圈代表按点操作，黄色盒子是学习好的神经网络层，线融合表示串联，copy 表示将一条线复制一份。

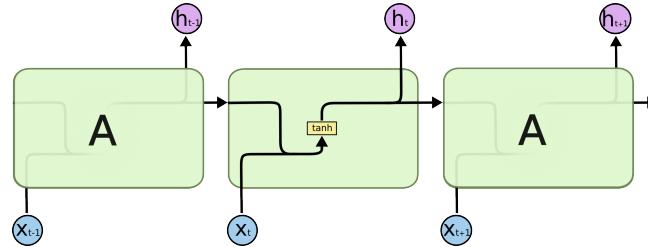


图 2.5: The repeating module in a standard RNN contains a single layer

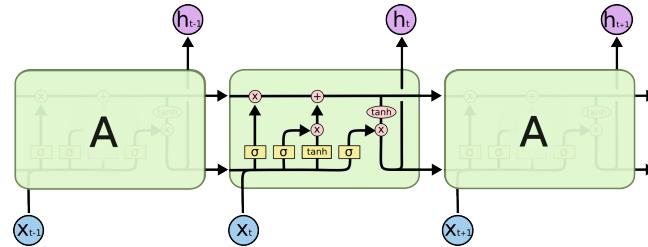


图 2.6: The repeating module in an LSTM contains four interacting layers

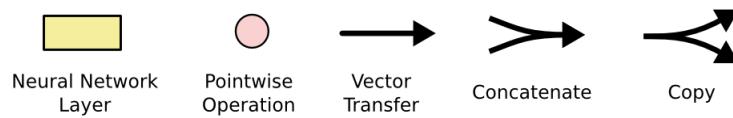
### 2.7.3 LSTM 想法的核心

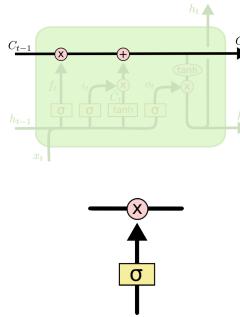
LSTMs 的核心是图像顶部的水平流过的 cell state, cell state 像一个传送带, 它笔直的沿着整条链跑, 和一些次要的线性交互, 很容易实现信息不改变的流动。LSTM 能删除或者增加信息到 cell state, 被控制的结构称为门。门是一种让信息通过的手段, 由一个 sigmoid 神经网络层和 pointwise 惩罚操作组成。sigmod Layer 输出 0 到 1 之间的数, 描述多少组件应该被通过, 0 表示不允许通过 1 表示让一切通过, LSTMs 有三个门, 保护和控制 cell state。

### 2.7.4 一步步的设置

第一步是 LSTMs 决定什么信息应该被传送, 这个决定每一个称为忘记门的 sigmoid layer 组成, 通过  $h_{t-1}$  和  $x_t$  输出 0 到 1 之间的数给当前的  $C_{t-1,1}$  表示完全保持, 0 表示丢弃。

对于上面的语言模型, cell state 也许包含 the gender of the present subjects, 以至于正确的带名字能被使用, 当我们看一个新的 subject, 我们想图忘记 the gender of the old subject。下一步是决定什么新的信息将被存储在 cell state 中, 这分为两部分





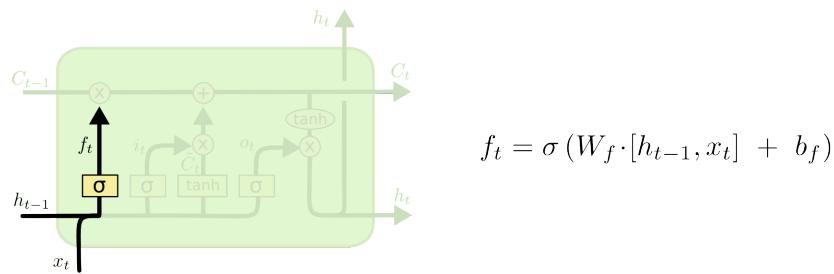
1. Sigmoid layer 调用 input gate layer 决定更新哪个值。
2. tanh layer 创建一个可能被添加到 state 新的候选向量。 $\tilde{C}_t$

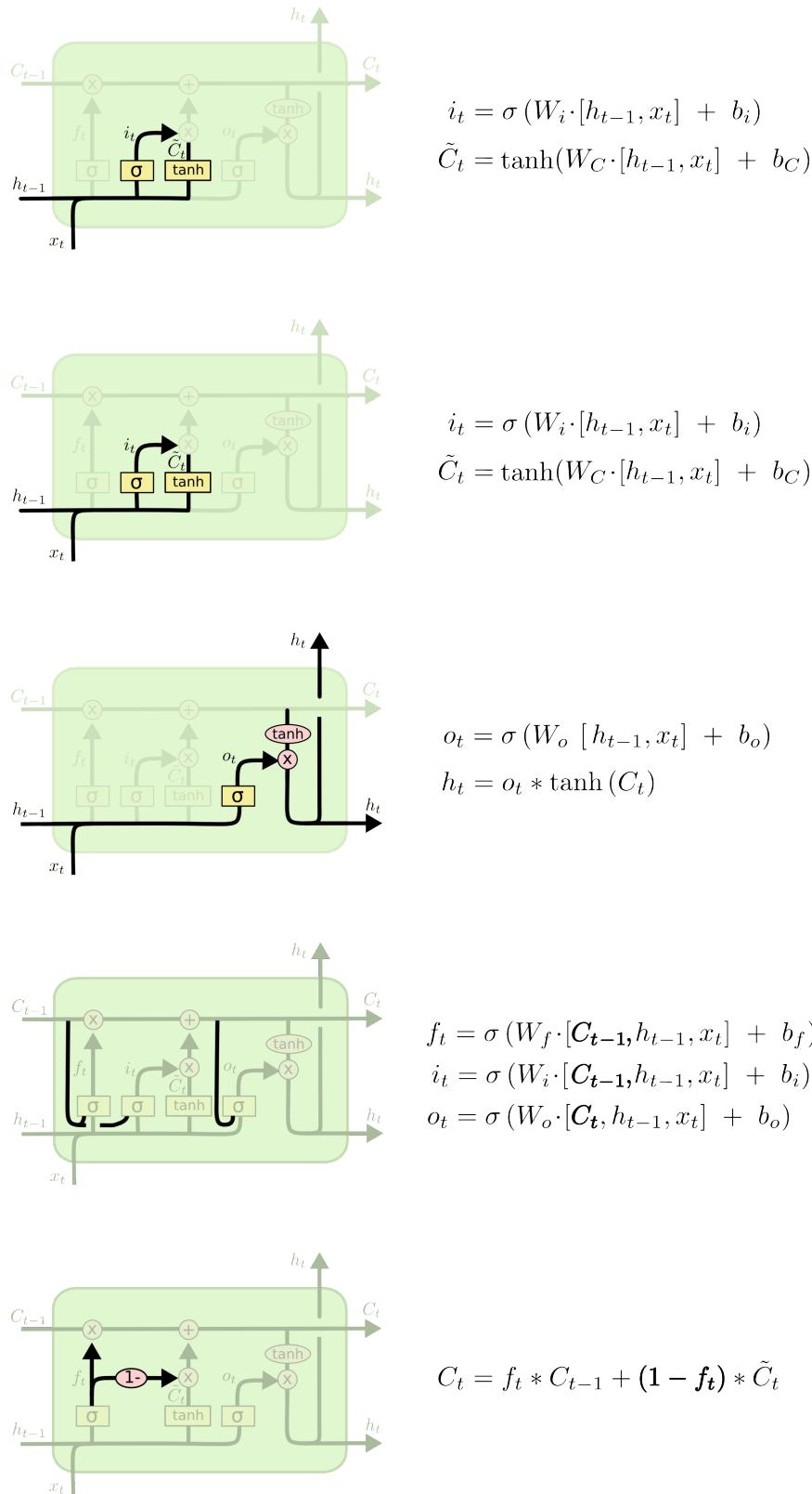
下一步我们结合两个不走创建一个更新状态。, 在我们的语言模型例子中, 我们想要增加 gender of the new subject 到 cell state 取代我们将要忘记的数据 现在更新老的 cell state  $C_{t-1}$  到新的 cell state  $C_t$ , 我们用老的  $c_{t-1}$  乘上  $f_t$  忘记我们之前决定忘记的事, 然后我们增加  $i_t * \tilde{C}_t$ . 这是新的候选值, 表示我们更新每个状态值的规模。在例子中的语言模型, 我们删掉了一个老的 subject's gender 增加新的信息。最后我们需要决定我们输出什么, 输出取决于我们的 cell state, 但是将被过滤, 所限我们运行 sigmoid layer 决定我们将输出那一部分。然后我们放通过 tanh 将 cell state 映射到-1,1, 然后乘上 sigmoid 门的输出, 以至于我们仅仅输出我们决定输出的部分。

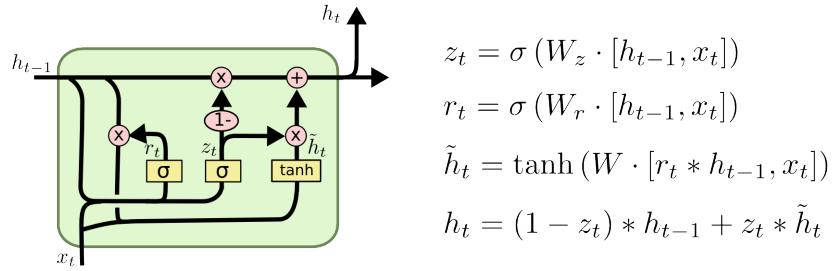
对于语言模型的例子, 因为它仅仅看 subject, 它也许想输出关于动词的信息, 例子中的下一个, 例如, 它也许输出是否 subject 是单数或者复数, 以至于从一个动词应该能知道接下来应该是动词的什么形式。

### 2.7.5 LSTM 的多种变体

[Gers Schmidhuber \(2000\)](#), 它增加了 peephole connections, 这一位置我们让 gate layer 通过 cell state 上面的图增加了 peepholes 到所有的门, 但是一些论文给出一些 peepholes 和 not others。另一个变体用两个 forget 和输入门。而不是分别决定忘记或者添加信息, 我们一起决定, 我们需要输入一些值是忘记, 我们仅仅忘记老的值输入新值到 state 一个更







引人注目的变体是 Gate Recurrent Unit 或者称为 (GRU), 由 Cho, et al. (2014) 引入, 它结合忘记和输入门为一个单独的更新们, 它也融合 cell state 和 hidden state 做了些改变, 这结果模型比标准的 LSTM 模型简单, 现在也越来越流行。这些仅仅是流行的 LSTM 变体, 有一些其它的像 Yao, et al. (2015) 的 Depth Gated RNNs, 用完全不同的方法处理 long-term dependencies, 像 Koutnik, et al. (2014) 的 Clockwork RNNs。

那个算法是最好的? 他们的差别大吗? Greff, et al. (2015) 做了一些比较了一些流行的变体, 发现他们基本相同。Jozefowicz, et al. (2015) 比较了超过 1 万中架构, 找到了一些在确定问题上比 LSTMs 好的架构。

### 2.7.6 向量字表示

#### Vector Representation of Words

通常图像或音频系统处理的是由图片中所有单个原始像素点强度值或者音频中功率谱密度的强度值, 把它们编码成丰富、高维度的向量数据集。对于物体或语音识别这一类的任务, 我们所需的全部信息已经都存储在原始数据中 (显然人类本身就是依赖原始数据进行日常的物体或语音识别的)。然后, 自然语言处理系统通常将词汇作为离散的单一符号, 例如”cat”一词或可表示为 Id537, 而”dog”一词或可表示为 Id143。这些符号编码毫无规律, 无法提供不同词汇之间可能存在的关联信息。换句话说, 在处理关于”dogs”一词的信息时, 模型将无法利用已知的关于”cats”的信息 (例如, 它们都是动物, 有四条腿, 可作为宠物等等)。可见, 将词汇表达为上述的独立离散符号将进一步导致数据稀疏, 使我们在训练统计模型时不得不寻求更多的数据。而词汇的向量表示将克服上述的难题。向量空间模型 (VSMs) 将词汇表达 (嵌套) 于一个连续的向量空间中, 语义近似的词汇被映射为相邻的数据点。向量空间模型在自然语言处理领域中有着漫长且丰富的历史, 不过几乎所有利用这一模型的方法都依赖于 分布式假设, 其核心思想为出现于上下文情景中的词汇都有相类似的语义。采用这一假设的研究方法大致分为以下两类: 基于计数的方法 (e.g. 潜在语义分析), 和 预测方法 (e.g. 神经概率化语言模型)。

其中它们的区别在如下论文中又详细阐述 Baroni :et al, 不过简而言之: 基于计数的方法计算某词汇与其邻近词汇在一个大型语料库中共同出现的频率及其它统计量, 然后将这些统计量映射到一个小型且稠密的向量中。预测方法则试图直接从某词汇的邻近词汇对其

进行预测，在此过程中利用已经学习到的小型且稠密的嵌套向量。

Word2vec 是一种可以进行高效率词嵌套学习的预测模型。其两种变体分别为：连续词袋模型（CBOW）及 Skip-Gram 模型。从算法角度看，这两种方法非常相似，其区别为 CBOW 根据源词上下文词汇（'the cat sits on the'）来预测目标词汇（例如，'mat'），而 Skip-Gram 模型做法相反，它通过目标词汇来预测源词汇。Skip-Gram 模型采取 CBOW 的逆过程的动机在于：CBOW 算法对于很多分布式信息进行了平滑处理（例如将一整段上下文信息视为一个单一观察量）。很多情况下，对于小型的数据集，这一处理是有帮助的。相形之下，Skip-Gram 模型将每个“上下文-目标词汇”的组合视为一个新观察量，这种做法在大型数据集中会更为有效。本教程余下部分将着重讲解 Skip-Gram 模型。

## 处理噪声的对比训练

神经概率化语言模型通常使用极大似然法 (ML) 进行训练，其中通过 softmax function 来最大化当提供前一个单词  $h$  (代表"history")，后一个单词的概率  $w_t$ (目标词概率)

$$P(w_t|h) = \text{softmax(score}(w_t, h)) = \frac{\exp\{\text{score}(w_t, h)\}}{\sum_{\text{Word } w' \text{ in Vocab}} \exp\{\text{score}(w', h)\}}$$

当  $\text{score}(w_t, h)$  计算了文字  $w_t$  和 上下文  $h$  的相容性（通常使用向量积）。我们使用对数似然函数来训练训练集的最大值，比如通过：

$$J_{ML} = \log P(w_t|h) = \text{score}(w_t, h) - \log(\sum_{\text{Word } w' \text{ in Vocab}} \exp\{\text{score}(w', h)\})$$

这里提出了一个解决语言概率模型的合适的通用方法。然而这个方法实际执行起来开销非常大，因为我们需要去计算并正则化当前上下文环境  $h$  中所有其它  $V$  单词  $w'$  的概率得分，在每一步训练迭代中。

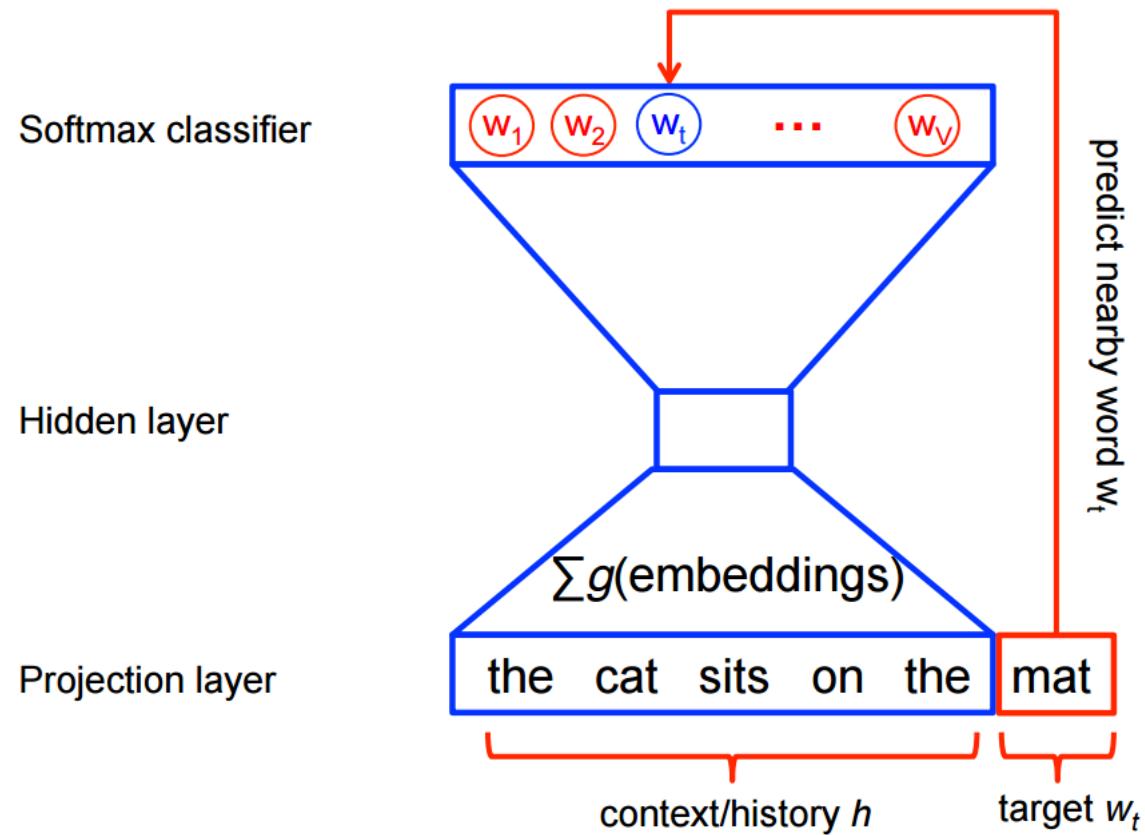


图 2.7: CBOW 方法

从另一个角度来说，当使用 word2vec 模型时，我们并不需要对概率模型中的所有特征进行学习。而 CBOW 模型和 Skip-Gram 模型为了避免这种情况发生，使用一个二分类器（逻辑回归）在同一个上下文环境里从  $k$  虚构的（噪声）单词  $\hat{w}$  区分真正的目标单词  $w_t$ ，下面详细参数 CBOW 模型，对于 Skip-Gram 模型只要简单的反向操作即可。

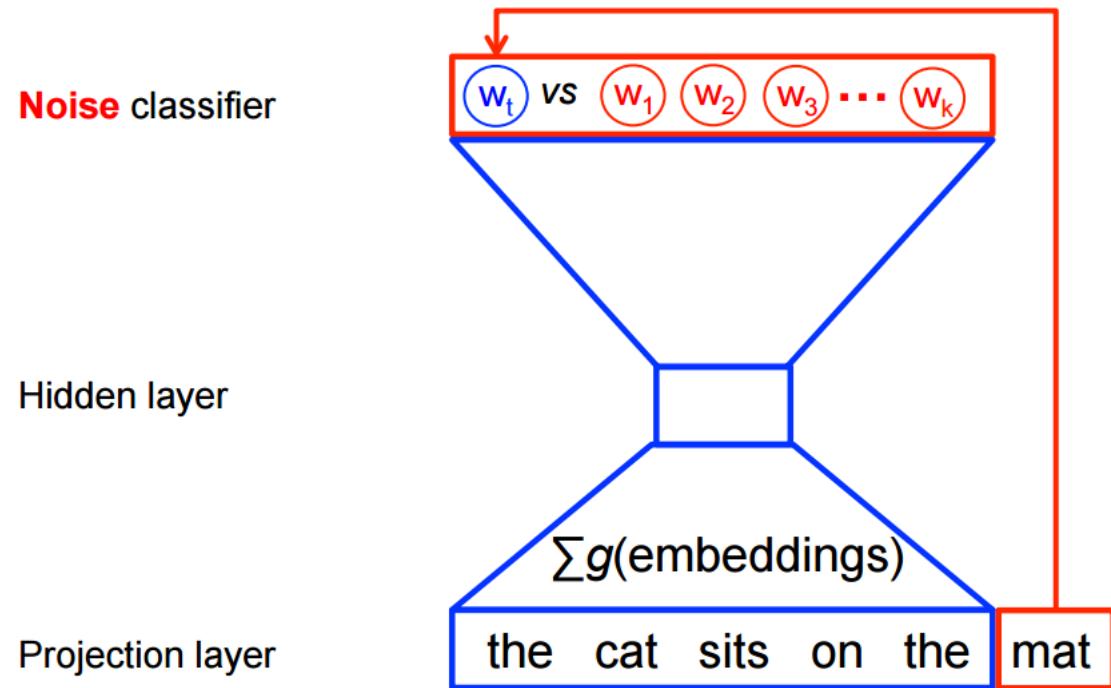


图 2.8: Skip-Gram

从数学的角度来说，我们的目标是对每个样本最大化:

$$J_{NEG} = \log Q_\theta(D = 1|w_t, h) + k \mathbb{E}_{\hat{w} \sim P_{noise}} [\log Q_\theta(D = 0|\hat{w}, h)]$$

其中  $Q_\theta(D = 1|w, h)$  代表的是当前上下文  $h$ ，根据所学得嵌套向量  $\theta$  目标单词  $w$  使用二分类逻辑回归计算得出的概率。在实践中，我们通过在噪声分布中绘制比对文字来获得近似的期望值（通过计算蒙特卡洛平均值）。

当真实地目标单词被分配到较高的概率，同时噪声单词的概率很低时，目标函数也就达到最大值了。从技术层面来说，这种方法叫做**负抽样**，而且使用这个损失函数在数学层面上也有很好的解释：这个更新过程也近似于 softmax 函数的更新。这在计算上将会有很大的优势，因为当计算这个损失函数时，只是有我们挑选出来的  $k$  个 噪声单词，而没有使用整个语料库  $V$ 。这使得训练变得非常快。我们实际上使用了与**noise-contrastive estimation (NCE)**介绍的非常相似的方法，这在 TensorFlow 中已经封装了一个很便捷的函数 `tf.nn.nce_loss()`。

### Skip-gram 模型

下面来看一下这个数据集

the quick brown fox jumped over the lazy dog

我们首先对一些单词以及它们的上下文环境建立一个数据集。我们可以以任何合理的方式定义‘上下文’，而通常上这个方式是根据文字的句法语境的（使用语法原理的方式处理当前目标单词可以看一下这篇文献 [Levy et al.](#)，比如说把目标单词左边的内容当做一个‘上下文’，或者以目标单词右边的内容，等等。现在我们把目标单词的左右单词视作一个上下文，使用大小为 1 的窗口，这样就得到这样一个由(上下文, 目标单词)组成的数据集：

([the, brown], quick), ([quick, fox], brown), ([brown, jumped], fox), ...

前文提到 Skip-Gram 模型是把目标单词和上下文颠倒过来，所以在这个问题中，举个例子，就是用'quick' 来预测'the' 和'brown'，用'brown' 预测'quick' 和'fox'。因此这个数据集就变成由(输入, 输出)组成的：

(quick, the), (quick, brown), (brown, quick), (brown, fox), ...

目标函数通常是对整个数据集建立的，但是本问题中要对每一个样本（或者是一个 batch\_size 很小的样本集，通常设置为  $16 \leq \text{batch\_size} \leq 512$ ）在同一时间执行特别的操作，称之为[随机梯度下降 \(SGD\)](#)。我们来看一下训练过程中每一步的执行。

假设用 t 表示上面这个例子中 quick 来预测 the 的训练的单个循环。用 num\_noise 定义从噪声分布中挑选出来的噪声（相反的）单词的个数，通常使用一元分布， $P(w)$ 。为了简单起见，我们就定 num\_noise=1，用 sheep 选作噪声词。接下来就可以计算每一对观察值和噪声值的损失函数了，每一个执行步骤就可表示为：

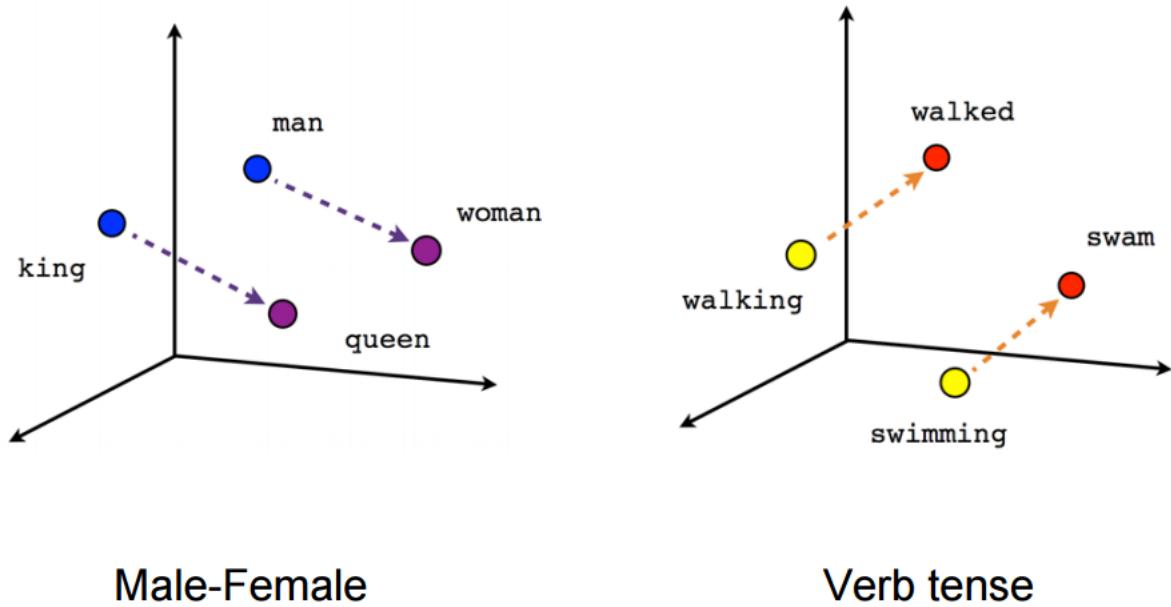
$$J_{NEG}^{(t)} = \log Q_\theta(D = 1 | the, quick) + \log(Q_\theta(D = 0 | sleep, quick))$$

整个计算过程的目标是通过更新嵌套参数  $\theta$  来逼近目标函数（这个例子中就是使目标函数最大化）。为此我们要计算损失函数中嵌套参数  $\theta$  的梯度，比如

$$\frac{\partial}{\partial} J_{NEG}$$

（幸好 TensorFlow 封装了工具函数可以简单调用！）。对于整个数据集，当梯度下降的过程中不断地更新参数，对应产生的效果就是不断地移动每个单词的嵌套向量，直到可以把真实单词和噪声单词很好得区分开。

我们可以把学习向量映射到 2 维中以便我们观察，其中用到的技术可以参考[t-SNE 降维技术](#)。当我们用可视化的方式来观察这些向量，就可以很明显的获取单词之间语义信息的关系，这实际上是非常有用的。当我们第一次发现这样的诱导向量空间中，展示了一些特定的语义关系，这是非常有趣的，比如文字中 male-female, gender 甚至还有 country-capital 的关系，如下方的图所示（也可以参考 [Mikolov et al., 2013](#) 论文中的例子）。



这也解释了为什么这些向量在传统的 NLP 问题中可作为特性使用，比如用在对一个演讲章节打个标签，或者对一个专有名词的识别（看看如下这个例子 [Collobert et al.](#) 或者 [Turian et al.](#)）。

不过现在让我们用它们来画漂亮的图表吧！

这里谈得都是嵌套，那么先来定义一个嵌套参数矩阵。我们用唯一的随机值来初始化这个大矩阵。

```
1 embeddings = tf.Variable(
2     tf.random_uniform([vocabulary_size, embedding_size], -1.0, 1.0))
```

对噪声-比对的损失计算就使用一个逻辑回归模型。对此，我们需要对语料库中的每个单词定义一个权重值和偏差值。（也可称之为输出权重与之对应的输入嵌套值）。定义如下：

```
1 nce_weights = tf.Variable(
2     tf.truncated_normal([vocabulary_size, embedding_size],
3                         stddev=1.0 / math.sqrt(embedding_size)))
4 nce_biases = tf.Variable(tf.zeros([vocabulary_size]))
```

我们有了这些参数之后，就可以定义 Skip-Gram 模型了。简单起见，假设我们已经把语料库中的文字整型化了，这样每个整型代表一个单词（细节请查看 `_basic.py`）。Skip-Gram 模型有两个输入。一个是一组用整型表示的上下文单词，另一个是目标单词。给这些输入建立占位符节点，之后就可以填入数据了。

```
1 train_inputs = tf.placeholder(tf.int32, shape=[batch_size])
2 train_labels = tf.placeholder(tf.int32, shape=[batch_size, 1])
```

然后我们需要对批数据中的单词建立嵌套向量，TensorFlow 提供了方便的工具函数。

```
1 embed = tf.nn.embedding_lookup(embeddings, train_inputs)
```

好了，现在我们有了每个单词的嵌套向量，接下来就是使用噪声-比对的训练方式来预测目标单词。

```
1 loss = tf.reduce_mean(
2     tf.nn.nce_loss(nce_weights, nce_biases, embed, train_labels,
3                     num_sampled, vocabulary_size))
```

我们对损失函数建立了图形节点，然后我们需要计算相应梯度和更新参数的节点，比如说在这里我们会使用随机梯度下降法，TensorFlow 也已经封装好了该过程。

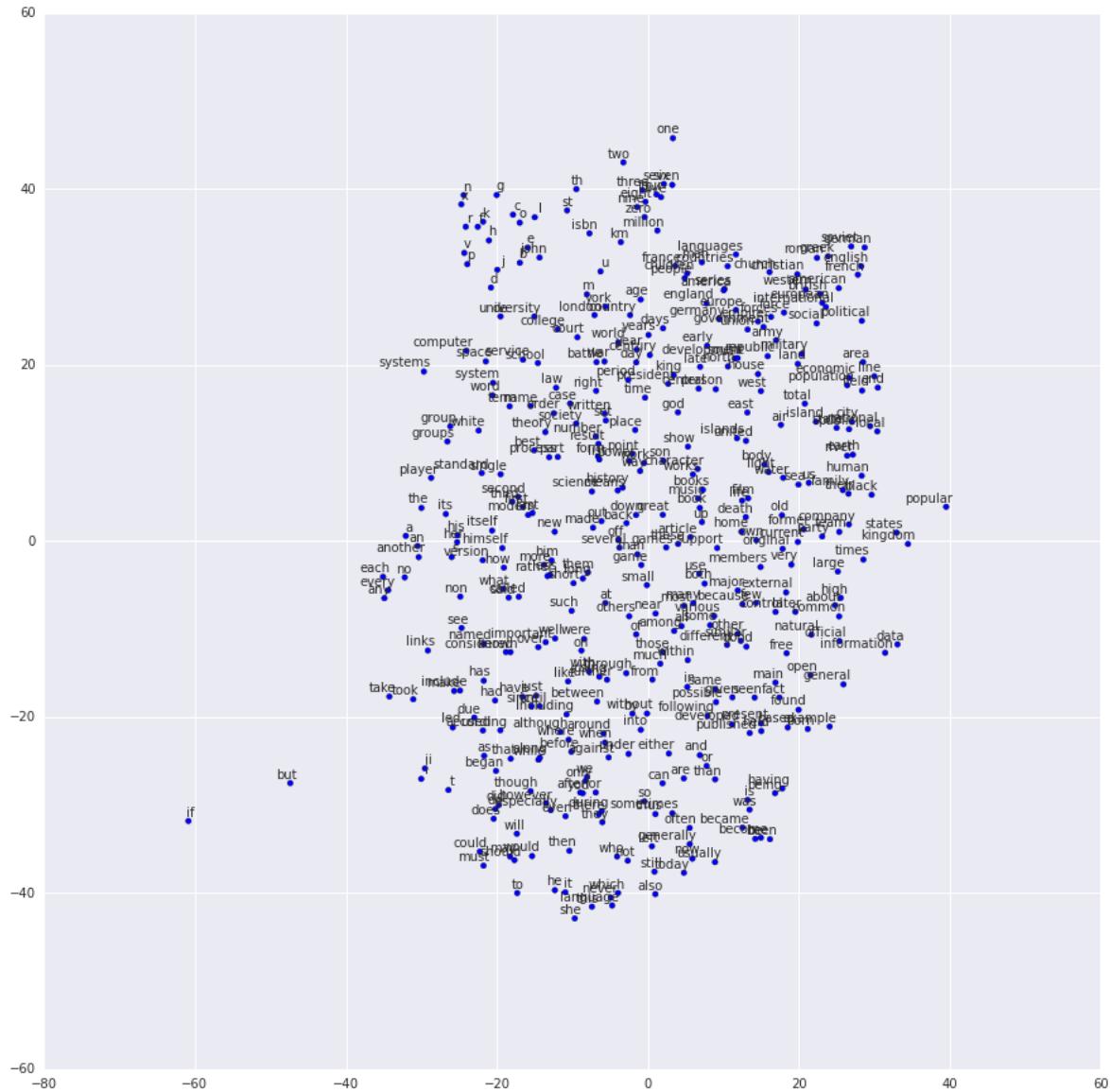
```
1 optimizer = tf.train.GradientDescentOptimizer(learning_rate=1.0).minimize(loss)
```

## 训练过程

训练的过程很简单，只要在循环中使用 `feed_dict` 不断给占位符填充数据，同时调用 `session.run` 即可。

```
1 for inputs, labels in generate_batch(...):
2     feed_dict = {training_inputs: inputs, training_labels: labels}
3     _, cur_loss = session.run([optimizer, loss], feed_dict=feed_dict)
```

嵌套学习结果可视化



Et voilà! 与预期的一样，相似的单词被聚类在一起。对 word2vec 模型更复杂的实现需要用到 TensorFlow 一些更高级的特性，具体是实现可以参考[word2vec.py](#)

嵌套学习的评估：类比推理

词嵌套在 NLP 的预测问题中是非常有用且使用广泛地。如果要检测一个模型是否是可以成熟地区分词性或者区分专有名词的模型，最简单的办法就是直接检验它的预测词性、

语义关系的能力，比如让它解决形如 king is to queen as father is to ? 这样的问题。这种方法叫做类比推理，可参考 Mikolov and colleagues，数据集下载地址为:[questions-words.txt](#)。To see how we do this evaluation 如何执行这样的评估，可以看 build\_eval\_graph() 和 eval() 这两个函数在下面源码中的使用 [word2vec.py](#)

超参数的选择对该问题解决的准确性有巨大的影响。想要模型具有很好的表现，需要有一个巨大的训练数据集，同时仔细调整参数的选择并且使用例如二次抽样的一些技巧。不过这些问题已经超出了本教程的范围。

## 优化实现

以上简单的例子展示了 TensorFlow 的灵活性。比如说，我们可以很轻松得用现成的 tf.nn.sampled\_softmax\_loss() 来代替 tf.nn.nce\_loss() 构成目标函数。如果你对损失函数想做新的尝试，你可以用 TensorFlow 手动编写新的目标函数的表达式，然后用控制器执行计算。这种灵活性的价值体现在，当我们探索一个机器学习模型时，我们可以很快地遍历这些尝试，从中选出最优。

一旦你有了一个满意的模型结构，或许它就可以使实现运行地更高效（在短时间内覆盖更多的数据）。比如说，在本教程中使用的简单代码，实际运行速度都不错，因为我们使用 Python 来读取和填装数据，而这些在 TensorFlow 后台只需执行非常少的工作。如果你发现你的模型在输入数据时存在严重的瓶颈，你可以根据自己的实际问题自行实现一个数据阅读器，参考 新的数据格式。对于 Skip-Gram 模型，我们已经完成了如下这个例子 [word2vec.py](#)。

如果 I/O 问题对你的模型已经不再是个问题，并且想进一步地优化性能，或许你可以自行编写 TensorFlow 操作单元，详见 添加一个新的操作。相应的，我们也提供了 Skip-Gram 模型的例子 [optimized.py](#)。请自行调节以上几个过程的标准，使模型在每个运行阶段有更好的性能。

### 2.7.7 RNN

此教程将展示如何在高难度的语言模型中训练循环神经网络。该问题的目标是获得一个能确定语句概率的概率模型。为了做到这一点，通过之前已经给出的词语来预测后面的词语。我们将使用 PTB(Penn Tree Bank) 数据集，这是一种常用来衡量模型的基准，同时它比较小而且训练起来相对快速。

语言模型是很多有趣难题的关键所在，比如语音识别，机器翻译，图像字幕等。它很有意思—可以参看 [here](#)。

本教程的目的是重现 [Zaremba et al., 2014](#) 的成果，他们在 PTB 数据集上得到了很棒的结果。

## 下载及准备数据

本教程需要的数据在 data/ 路径下，来源于 Tomas Mikolov 网站上的[PTB 数据集](#)

该数据集已经预先处理过并且包含了全部的 10000 个不同的词语，其中包括语句结束标记符，以及标记稀有词语的特殊符号 (<unk>)。我们在 reader.py 中转换所有的词语，让他们各自有唯一的整型标识符，便于神经网络处理。

## LSTM

模型的核心由一个 LSTM 单元组成，其可以在某时刻处理一个词语，以及计算语句可能的延续性的概率。网络的存储状态由一个零矢量初始化并在读取每一个词语后更新。而且，由于计算上的原因，我们将以 batch\_size 为最小批量来处理数据。

基础的伪代码就像下面这样：

```

1 lstm = rnn_cell.BasicLSTMCell(lstm_size)
2 state = tf.zeros([batch_size, lstm.state_size])
3
4 loss = 0.0
5 for current_batch_of_words in words_in_dataset:
6     output, state = lstm(current_batch_of_words, state)
7
8     logits = tf.matmul(output, softmax_w) + softmax_b
9     probabilities = tf.nn.softmax(logits)
10    loss += loss_function(probabilities, target_words)

```

## 截断反向传播

为使学习过程易于处理，通常的做法是将反向传播的梯度在（按时间）展开的步骤上照一个固定长度 (num\_steps) 截断。通过在一次迭代中的每个时刻上提供长度为 num\_steps 的输入和每次迭代完成之后反向传导，这会很容易实现。

一个简化版的用于计算图创建的截断反向传播代码：

```

1 words = tf.placeholder(tf.int32, [batch_size, num_steps])
2
3 lstm = rnn_cell.BasicLSTMCell(lstm_size)
4 initial_state = state = tf.zeros([batch_size, lstm.state_size])
5
6 for i in range(len(num_steps)):
7     output, state = lstm(words[:, i], state)
8
9     # ...

```

```
11 final_state = state
```

下面展现如何实现迭代整个数据集：

```
1 numpy_state = initial_state.eval()
2 total_loss = 0.0
3 for current_batch_of_words in words_in_dataset:
4     numpy_state, current_loss = session.run([final_state, loss],
5         feed_dict={initial_state: numpy_state, words: current_batch_of_words})
6     total_loss += current_loss
```

## 输入

在输入 LSTM 前，词语 ID 被嵌入到了一个密集的表示中（查看 矢量表示教程）。这种方式允许模型高效地表示词语，也便于写代码：

```
1 # embedding_matrix 张量的形状是: [vocabulary_size, embedding_size]
2 word_embeddings = tf.nn.embedding_lookup(embedding_matrix, word_ids)
```

嵌入的矩阵会被随机地初始化，模型会学会通过数据分辨不同词语的意思。

## 损失函数

我们想使目标词语的平均负对数概率最小  $loss = -\frac{1}{N} \sum_{i=1}^N \ln p_{target_i}$  实现起来并非很难，而且函数 `sequence_loss_by_example` 已经有了，可以直接使用。

论文中的典型衡量标准是每个词语的平均困惑度（perplexity），计算式为

$$e^{-\frac{1}{N} \sum_{i=1}^N \ln p_{target_i}} = e^{loss}$$

同时我们会观察训练过程中的困惑度值（perplexity）

## 多个 LSTM 层堆叠

要想给模型更强的表达能力，可以添加多层 LSTM 来处理数据。第一层的输出作为第二层的输入，以此类推。

类 `MultiRNNCell` 可以无缝的将其实现：

```
1 lstm = rnn_cell.BasicLSTMCell(lstm_size)
2 stacked_lstm = rnn_cell.MultiRNNCell([lstm] * number_of_layers)
3
4 initial_state = state = stacked_lstm.zero_state(batch_size, tf.float32)
5 for i in range(len(num_steps)):
6     # 每次处理一批词语后更新状态值。
7     output, state = stacked_lstm(words[:, i], state)
```

```
8  
9      # 其余的代码.  
10     # ...  
11  
12 final_state = state
```

## 编译并运行代码

首先需要构建库，在 CPU 上编译：

```
1 bazel build -c opt tensorflow/models/rnn/ptb:ptb_word_lm
```

如果你有一个强大的 GPU，可以运行

```
1 bazel build -c opt --config=cuda tensorflow/models/rnn/ptb:ptb_word_lm
```

运行模型：

```
1 bazel-bin/tensorflow/models/rnn/ptb/ptb_word_lm \  
2   --data_path=/tmp/simple-examples/data/ --alsologtosterr --model small
```

教程代码中有 3 个支持的模型配置参数：“small”，“medium” 和“large”。它们指的是 LSTM 的大小，以及用于训练的超参数集。

模型越大，得到的结果应该更好。在测试集中 small 模型应该可以达到低于 120 的困惑度 (perplexity)，large 模型则是低于 80，但它可能花费数小时来训练。

# Chapter 3

## Tensorflow 进阶

### 3.1 模型存储和加载

- 生成 checkpoint 文件, 扩展名一般为.ckpt, 通过在 tf.train.Saver 对象上调用 Saver.saver() 生成。它包含权重和其他程序中定义的变量, 不包含 图的结构。如果需要在另一个程序中使用, 需要重建图形结构, 并告诉 Tensorflow 如何处理这些权重。
- 生成 (graph proto file), 这是一个二进制文件, 扩展名一般是.pb, 用 tf.train.write\_graph() 保存每, 只包含图形结构, 不包含权重, 然后使用 tf.import\_graph\_def() 加载 图形。

## 3.2 用 GPU

在 Tensorflow 中 CPU,GPU 用字符串表示

- "cpu:0": 机器上的 CPU
- "gpu:0": 机器上的 GPU
- "gpu:1": 机器上的第二块 GPU

如果 TensorFLow 操作有 GPU 和 CPU 实现，GPU 将被优先指定，例如 matmul 有 CPU 和 GPU 内核，在系统上有 cpu:0 和 gpu:0,gpu:0 将优先运行 matmul。布置采集设备

找到你的操作和 tensor 上的设备，创建一个会话 log\_device\_placement 配置设置为 True

```

1 import tensorflow as tf
2 a = tf.reshape(tf.linspace(-1.,1.,12),(3,4))
3 b = tf.reshape(tf.sin(a),(4,3))
4 c = tf.matmul(a,b)
5 with tf.Session() as sess:
6     print(sess.run(c))
```

输出参数:

```

[[ 0.87280041  0.44710392  0.00666773]
 [ 0.43973413  0.44710392  0.4397341 ]
 [ 0.00666779  0.44710392  0.87280059]]
```

### 3.2.1 手工配置设备

如果你想将你的操作运行在指定的设备中而不由 tensorflow 是自动为你选择，你可以用 tf.device 创建一个设备，左右的操作将在同一个设备上指定。

```

1 import tensorflow as tf
2 with tf.device('/cpu:0'):
3     a = tf.constant([1.,2.,3.,4.,5.,6.],shape = (2,3),name = 'a')
4     b = tf.reshape(a,shape=(3,2))
5     c = tf.matmul(a,b)
6     with tf.Session(config = tf.ConfigProto(log_device_placement=True)) as sess:
7         print(sess.run(c))
```

```

Device mapping:
/job:localhost/replica:0/task:0/gpu:0 -> device: 0, name: TITAN Xp, pci bus id: 0000:06:00.0
/job:localhost/replica:0/task:0/gpu:1 -> device: 1, name: TITAN Xp, pci bus id: 0000:05:00.0
Reshape: (Reshape): /job:localhost/replica:0/task:0/cpu:0
MatMul: (MatMul): /job:localhost/replica:0/task:0/cpu:0
Reshape/shape: (Const): /job:localhost/replica:0/task:0/cpu:0
a: (Const): /job:localhost/replica:0/task:0/cpu:0
[[ 22.  28.]
 [ 49.  64.]]
```

正如你看到的 a,b 被复制到 cpu:0, 因为设备没有明确指定, Tensorflow 将选择操作和可用的设备 (gpu:0)

### 3.2.2 允许 GPU 的内存增长

默认情况下 Tensorflow 将映射所有的 CPUs 的显存到进程上, 用相对精确的 GPU 内存资源减少内存的碎片化会更高效。通常有些程序希望分贝可用内存的一部分, 或者增加内存的需要两。在会话中 tensorflow 提供了两个参数 控制它。第一个参数是 allow\_growth 选项, 根据运行情况分配 GPU 内存: 它开始分配很少的内存, 当 Session 开始运行 需要更多 GPU 内存是, 我们同感 Tensorflow 程序扩展 GPU 的内存区域。注意我们不释放内存, 因此这可能导致更多的内存碎片。为了开启这个选项, 可以通过下面的设置

```

1 config = tf.ConfigProto()
2 config.gpu_option.allow_growth = True
3 sess = tf.Session(config=config, ...)
```

第二种方法是 per\_process\_gpu\_memory\_fraction 选项, 决定 GPU 总体内存中多少应给被分配, 例如你可以告诉 Tensorflow 分配 40% 的 GPU 总体内存。

```

1 config = tf.ConfigProto()
2 config.gpu_option.per_process_gpu_memory_fraction = 0.4
3 sess = tf.Session(config = config)
```

如果你想限制 Tensorflow 程序的 GPU 使用量, 这个参数是很有用的。

在多 GPU 系统是使用 GPU

如果你的系统上有超过一个 GPU, 你的 GPU 的抵消的 ID 将被默认选中, 如果你想运行在不同的 GPU 上, 你需要指定 你想要执行运算的 GPU

```

1 import tensorflow as tf
2 with tf.device('/gpu2:0'):
3     a = tf.constant([1., 2., 3., 4., 5., 6.], shape=(2, 3), name='a')
4     b = tf.reshape(a, shape=(3, 2))
5     c = tf.matmul(a, b)
6     with tf.Session(config=tf.ConfigProto(log_device_placement=True)) as sess:
7         print(sess.run(c))
```

如果你指定的设备不存在, 你将个到一个 InvalidArgumentError:

```
InvalidArgumentError (see above for traceback): Cannot assign a device for operation 'Reshape': Operation was explicitly assigned to /device:GPU:2 but available devices are [ /job:localhost/replica:0/task:0/cpu:0, /job:localhost/replica:0/task:0/gpu:0, /job:localhost/replica:0/task:0/cpu:1 ]. Make sure the device specification refers to a valid device.
[[Node: Reshape = Reshape[T=DT_FLOAT, Tshape=DT_INT32, _device="/device:GPU:2"](a, Reshape/shape)]]
```

如果你想 Tensorflow 在万一指定的设备不存在时自动选择一个存在的设备，你可以在创建会话时配置中设置 allow\_soft\_placement 为 True

```
1 with tf.device('/gpu:2'):
2     a = tf.constant([1., 2., 3., 4., 5., 6.], shape=[3, 2], name='a')
3     b = tf.constant([1., 2., 3., 4., 5., 6.], shape=[2, 3], name='b')
4     c = tf.matmul(a, b)
5 with tf.Session(config=tf.ConfigProto(allow_soft_placement=True,
6                                     log_device_placement=True)) as sess:
7     print(sess.run(c))
```

```
Device mapping:
/job:localhost/replica:0/task:0/gpu:0 -> device: 0, name: TITAN Xp, pci bus id: 0000:06:00.0
/job:localhost/replica:0/task:0/gpu:1 -> device: 1, name: TITAN Xp, pci bus id: 0000:05:00.0
Reshape: (Reshape): /job:localhost/replica:0/task:0/gpu:0
MatMul: (MatMul): /job:localhost/replica:0/task:0/gpu:0
Reshape/shape: (Const): /job:localhost/replica:0/task:0/gpu:0
a: (Const): /job:localhost/replica:0/task:0/gpu:0
[[ 22.  28.
   49.  64.]]
```

## 用多 GPU

如果你想在多张 GPU 上运行 Tensorflow，你可以在 multi-tower fashion 上构造你的模型，每个 tower 被指定到不同的 GPU 上。例如：

```
1 c = []
2 for d in ['/gpu:0', '/gpu:1']:
3     with tf.device(d):
4         a = tf.constant([1.0, 2.0, 3.0, 4.0, 5.0, 6.0], shape=[2, 3])
5         b = tf.constant([1.0, 2.0, 3.0, 4.0, 5.0, 6.0], shape=[3, 2])
6         c.append(tf.matmul(a, b))
7     with tf.device('/cpu:0'):
8         sum = tf.add_n(c)
9     # Creates a session with log_device_placement set to True.
10 sess = tf.Session(config=tf.ConfigProto(allow_soft_placement=True,
11                     log_device_placement=True))
12     # Runs the op.
13     print(sess.run(sum))
14 sess.close()
```



### 3.3 如何利用 Inception 的最后一层重新训练新的分类

现代的认知模型可能有上百万个参数可能需要花几周训练，Transfer 学习是通过完整的像 ImageNet 一样的模型通过已经存在的权重简化数周工作分类的技术。在这个例子中我们将创新训练最终层不修改其它层。详细信息你可以查看[这篇论文](#)。

尽管不完整的训练，但是对于一些应用却惊人的高效，可以在笔记本上训练 30 分钟，不要求 GPU。这个导航将显示给你如何在自己的图像运行示例脚本解释一些控制训练需要的脚本。

#### 3.3.1 训练花

在开始训练前你需要设置图像教网络你想认识的新的类别。接下来的张杰解释如何准备你的图像，但是我们创建一个授权的归档的花的文件使得训练更轻松。为了得到花的图像，运行下面的代码：

```
1 cd ~  
2 curl -O http://download.tensorflow.org/example_images/flower_photos.tgz  
3 tar xzf flower_photos.tgz
```

当你有图像后，你从你的 TensorFlow 源文件目录构建重新训练器

```
1 bazel build --config opt tensorflow/examples/image_retraining:retrain
```

可以通过下面运行：

```
1  bazel build tensorflow/examples/image_retraining:retrain
```

如果你有一个机器支持 AVX 设备集 (最近几年的常用的 x86 CPUs) 你可以通过架构提高 building 运行速度

```
2  bazel build --config opt tensorflow/examples/image_retraining:retrain
```

训练器可以是这样:

```
3  bazel-bin/tensorflow/examples/image_retraining/retrain --image_dir ~/flower_photos
```

这个脚本载入先前 inception v3 模型, 删除顶层, 在新的 flower photos 训练新的模型。在原始 ImageNet 类中没有一种花完整网络被完整的训练过了, transfer 学习是低层已经被训练好 区别不修改任何不同对象。

### 3.3.2 瓶颈

训练花费 30 分钟甚至更长时间取决于你的机器的速度。第一个时期分析所有磁盘上的图像和计算他们的瓶颈, 瓶颈是一个信息对术语 我们经常在最后一层前一层, 倒数第二层已经训练区别输出要求分类的值, 这意味着这必须是有意义的, 因此对于分类器它必须包含足够的信息在一些小的值得集合中做选择, 这意味着我们的最终层训练可以在新的类中工作证明在 ImageNet 中 1000 类对于区别新的对象是有用的。因为每个图像在训练和计算花费时间瓶颈时被多次使用, 他的速度达到缓存起的瓶颈因此不能被重复计算。默认他们存储在/tmp/bottleneck 陌路, 如果你仍然会脚本他们将被重用, 因此你不是必须再次等待这部分。

### 3.3.3 训练

当瓶颈计算完成时, 实际顶层训练开始。你讲看到输出, 显示精度, 可用精度, 交叉熵。训练精度显示在当前训练批中多少被分类正确, 验证训练精度从图像数据集随机选中精度的值, 不同之处在于训练精度基于网络已经学习到的参数, 在训练中可能过拟合到为噪声。验证精确度用不在训练集中的数据性能测量精确度, 如果训练精确度很高, 测试精确度很低说明网络过拟合训练图像存储的部分参数没有用。交叉熵损失函数查看学习进程处理的增么样, 训练对象使得损失尽可能小, 因此你可以分辨出如果学习起作用, 忽略损失噪声损失保持下降的趋势。默认脚本运行 4000 步, 每一步从训练集中随机选择 10 张图像找到缓冲器的瓶颈, 输入数据仅最终层预测。预测然后比较实际 label 和真实值差距反向传递误差。当你继续的时候你应该看到精确度的提高。你应该能看到精确度在 90% 到 95% 之间, 通过提取值将随机的一次次训练, 这个数完全训练好模型后基于在给定测试集中正确标签的百分比。

### 3.3.4 用 TensorBoard 可视化

包含 TensorBoard 总结的脚本吗很容易理解，调试，优化。例如，你可以可视化图和统计，例如在训练中权重和精度变化：

```
4 tensorboard --logdir /tmp/retrain_logs
```

TensorBoard 运行后导航到 localhost:6006 查看 TensorBoard，脚本将默认采集 TensorBoard 总结到 /tmp/retrain\_logs，你可以通过 summaries\_dir 标志指定采集目录。

### 3.3.5 用重新训练的模型

脚本将用训练好的最后一层写出你的一个 Inception v3 版本到 /tmp/output\_graph.pb 文件在 /tmp/output\_labels.txt 包含标签，两个格式见 [C++ and Python image classification examples](#)，因此你可以立即开始新的模型。你去带最顶层，你将需要在脚本中指定新的名字，例如你用 label\_image，用 output\_layer=final\_result。你可以用下面的代码重新训练图：

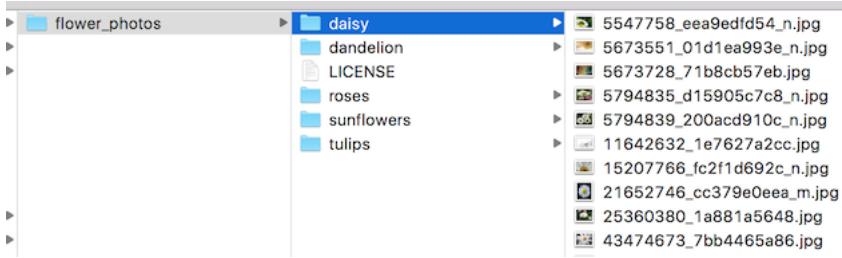
```
5 bazel build tensorflow/examples/image_retraining:label_image && \
6 bazel-bin/tensorflow/examples/image_retraining/label_image \
7 --graph=/tmp/output_graph.pb --labels=/tmp/output_labels.txt \
8 --image=$Home/flower_photos/daisy/21625746_cc379eea_m.jpg
```

你应该看到花的标签的列表在大说书情况下 daisy 在顶层（尽管重新训练的模型被可能会一点不同），你可以用在你的图片上--image 参数用 c++ 代码作为模板整合你自己的应用。如果你想在自己的 Python 程序中用你训练好的模型，上面的 [label\\_image script](#) 如果你发现默认的 Inception v3 模型对你的应用太大或者太慢，看看 [Other Model Architectures section](#)

### 3.3.6 在你自己的分类上训练

如果你已经成功的让脚本在花的例子上工作，你可以教他认识其他你想它认识的东西。理论上你需要设置一个子文件夹，命名分类，每个文件夹包含分类的图像。如果你传递子文件夹的根文件夹作为参数给--image\_dir，脚本像上面训练花一样训练。

实际上它会花一点时间得到你想要的精度，下面是一些常见的问题。



### 3.3.7 创建一个训练图像集合

首先我们需要查看收集到的图像，常见的问题是训练过程中数据的输入。

为了训练能起作用，每个你想要识别的图像你必须至少手机 100 张图片，你收集到的图片越多，训练的精确度可能越好。例如你拍摄一些蓝色的房间，另一些是绿色的房间模型的预测最终基于背景颜色，没有对象特征被考虑。为了避免这种情况，拍摄不同颜色的，没有一些实际能看到的的特征。如果你想了解更多这类问题你需要读[tank recognition](#) 如果你想考虑你用的分类。分隔大的数据集发现一些不同的物理形式为小的可以通过视觉区分的数据集，例如你可以用'vehicle' 可以用来替代'car'，'motobike' 和'truck'，考虑你有一个开放的世界还是封闭的世界将是很有价值的，在封闭的世界你唯一需要考虑的是识别已有的对象，例如一个植物识别的 app 你应该知道用户可能拍摄的花的图片，英雌你必须决定花的种类，相比之下一个巡逻机器人可能通过摄像头看到不同的事物。在这种情况下你想要分类器报告是否确认他看到的，这可能很难，但是你经常收集一些典型的和主体对象不相关的背景图像，，你可能会让它增加一些图片文件夹中未知的分类。检查确保你的图像被正确的标记也是很重要的。经常用生成的标签对于你的目的来说是不可靠的，例如你用 #daisy 命名一个叫 Daisy 的人。如果你想你的图像如果你了解你的图像，扫除任何错误将可能导致最后精确度提高。

### 3.3.8 训练步骤

如果你为你的图片感到高兴，你可以通过修改学习进程中的细节提升你的结果。最简单的方法是用`-how_many_training_steps`。默认是 4000. 但是如果你增加到 8000，他的训练时间将增加到两倍。精确度提高的比率显示你训练的越长一些点将停止，但是你可以试验什么时候达到你的模型的限制。

### 3.3.9 扭曲

随机通过变形，剪裁，变化输入图像的亮度是一个提高结果的常用方法，这样扩展了训练数据的大小，帮助网络学习 真是生活分类器所有的扭曲，在脚本中使用扭曲最大的缺点是缓冲瓶颈不再有用，因此输入图像将不能重用。这意味着训练京城可能花费更多时间，因此我推荐当着作为一个调节方法调节你的模型到合理。你可以传递`-random_crop`,

random\_scale 和 random\_brightness 给脚本扭曲图片。百分比值用来控制图片上扭曲用多少部分。, 合理的值时 5 或者 10.-flip\_left\_right 将在水平方向随机的镜像图像的一半, 有助有你的应用能理解翻转的图像。例如如果你想识别字母这将不是一个好的办法, 因为翻转它们会毁掉原来的含义。

### 3.3.10 超参数

你可以调整一些参数查看是否对你的结果有帮助, -learning\_rate 控制最终层训练更新的幅度。直观理解, 如果这个值变小训练时间将变长, 但是他可能对精度有帮助, 你需要小心试验得到查看什么对于你的 case 生效了。-train\_batch\_size 控制每一训练步多少图像被检查, 因为学习率应用到每批上, 如果你有更大的批得到相同的效果你将需要减小它。

### 3.3.11 训练, 验证, 测试集

当你为你的脚本指定图像文件夹时, 文件夹被分成不同的数据集。最大的数据及是训练集, 训练集包含用于训练网络的数据, 用于更新权重。你也许很想知道为什么我们不用所有的图像训练? 一个道德潜在的问题是当我们做机器学习算法时我们的模型会记住接近正确答案的不相关信息, 你可以想想你的图像可能记住了一些照片的背景, 通过标签匹配对象, 它在训练时所有的图像可能产生一个好的结果, 但是不能再新的图像产生好的结果因为他不能泛化对象的特征, 仅仅在训练图像的时候记住了一些不重要的特征。这个问题被称为过拟合, 为了避免过拟合我们保持我们的一些数据不再训练进程中, 因此模型不能记住他们, 我们用这些图像作为检察确保过拟合没有发生, 当我们在看模型在这些数据上有一个好的精度说明过拟合没有发生, 通常 80% 的数据被用来作为训练集 10% 的数据集用来验证最后 10% 的数据用做测试集预测分类器在真实世界的性能, 通过-testing\_percentage 和-validation\_percentage 标志用来控制比例。通常你应该能留下一些值作为默认, 不应该找到任何好处训练调整他们。注意这个脚本用图象的文件名区分训练集, 验证集, 测试集中的图像 (不是一个随机的函数), 这样保证运行时图片不会再训练集和测试集之间移动, 因为当用于训练模型的图像被验证集中的图像取代时可能会出现一些问题。你也需注意到了在迭代过程中验证正确度的波动。多数波动是验证集的子集的随机性引起的, 选择的验证集用来验证精确度。波动能被最大程度减少, 花费的训练时间增长, 通过选择-validation\_batch\_size=-1 用整个验证集计算精度。当训练结束后你将能检查测试集中错误分类图像, 这可以通过增加-print\_misclassified\_test\_images 标记, 这对于找到那些什么类型的图片让模型困惑 (很难区别的) 是很有帮助的例如你也许发现了一些种类一些常见的图像角度是特别难识别的, 这样是鼓励你增加更多类型的分类训练子类, 检查催眠五分类图片也指出输入数据中的错误, 向错误标签, 其质量魔术的照片。然而, 你应该避免测试集固定点单个误差, 因为他们仅仅反映在训练集中更多的问题。

### 3.3.12 更对模型架构

这个脚本默认用 Inception v3 模型架构作为预先训练脚本。这是一个好的开始的地方，因为它提供了高精度的训练结果，但是如果你想部署你的模型到手机设备或者其他资源限制的环境你也许想要这种精确度换区更小的文件尺寸和更快的速度。为了帮助这个retrain在[移动架构](#)上支持 30 个不同的变量。这里有一些比 Inception v3 更小精度的，但是可以得到更小的文件大小(下载小于兆字节)运行快乐几倍。为了训练这个模型，传递—architecture 标志，例如：

```
1 python tensorflow/examples/image_retraining/retrain.py \
2   --image_dir ~/flower_photos -- architecture mobilenet_0.25_128_quantized
```

这将在/temp/创建一个 941KB 模型文件 output\_graph.pb.Mobilenet 的 25% 的参数，占据  $128 \times 128$  大小的输入图像，权重在磁盘中量化为 8 位，你可以选择'1.0','0.75','0.50','0.25'控制权重参数的数量，因此文件尺寸(和一些扩展速度),'224','192','160'或者'128'对于输入图像的尺寸，更小的尺寸更快的速度，选项'\_quantized'预示着是否文件应该包含 8 位或者 32 位浮点权重。速度和大小好处带来的是精确度的损失，但是对于一些用途来说是不重要的，他可以通过训练数据提高、例如用扭曲在花数据集允许你得到得到 80% 的精度，甚至 0.25/128、quantized 图。如果你在你的程序或者 label\_image 中用 Mobilenet 模型，你讲需要一个输入一个指定大小的图像转换一个浮点让位到'input' tensor，典型的 24 位乳香范围 [0,255] 你必须用 (image-128.)/128 转化它到 [-1,1] 范围。

## 3.4 TF layer 向导：建立一个卷积神经网络

TensorFlow[layers module](#)是一个用于轻松建立神经网络的高级 API，它提供了一个方法促进创建 dense(全连接) 层和卷积层，增加激活函数，应用 dropout 规则。在这个导航中，你讲学习如何用 layers 建立一个卷积神经网络模型识别手写体数据集。**手写体数据集**包含 0-9，60000 个训练样本 10000 个测试样本，图像格式为  $28 \times 28$

### 3.4.1 开始

创建文件 cnn\_mnist.py，在手写体程序中添加如下代码：

```
1 from __future__ import absolute_import
2 from __future__ import division
3 from __future__ import print_function
4
5 # Imports
6 import numpy as np
7 import tensorflow as tf
```

```

8   tf.logging.set_verbosity(tf.logging.INFO)
9
10 # Our application logic will be added here
11
12 if __name__ == "__main__":
13     tf.app.run()

```

正如你看到的，你将增加，构造，训练，评估卷积神经网络，最终代码可以点击[这里](#)

### 3.4.2 介绍卷积神经网络

卷积神经网络是当前最先进的用于图像分类任务的模型架构。CNNs 应用一些滤波器从原始的图像像素中提取高级特征，这个模型可能被用在分类。CNN 包含三个组件：

- **卷积层** 应用指定数量的卷积滤波器在图像上。对于每一个子区域，layer 执行一系列数学操作生成一个单个值在输出 feature map，卷积层然后应用 relu 激活函数输出非线性。
- **池化层** 下采样卷积层的图像数据，减小 feature map 的维度从而减小处理时间。常用池化算法是最大池化 (提取 feature map 子区域) 保留最大值，丢掉其它值。
- Dense layers(**全连接层**) 在通过卷积层和下采样层特征提取执行分类。在全连接层，每一个节点连接到前面的节点。

通常 CNN 有一个卷积模块组成，每个层有卷积模块和池化模块组成。最新的卷积模块有一个或者更多的全连接层链接执行分类。最终 CNN 的全连接层包含每个目标类的一个单个节点 (所有模型可能预测的类)，用 softmax 函数生成一个 0-1 的值 (所有值的和维 1)。我们可以解释给定图像和目标的相似情况。

### 3.4.3 建立 CNN MNIST 分类器

用 CNN 架构建立模型分类 MNIST 数据集。

1. 卷积层 1: 应用  $5 \times 5$  卷积核 (提取  $5 \times 5$  像素的区域)，用 relu 激活函数。
2. 池化层 1: 执行最大池化  $2 \times 2$  stride=2(指定的池化区域不重叠)
3. 卷积层 2: 应用 64 个  $5 \times 5$  的卷积核，激活函数为 relu。
4. 池化层 2: 再次执行最大池化操作 (卷积核  $2 \times 2$ ) stride=2。
5. Dense 1: 1024 个神经元，dropout=0.4。

### 6. Dense2:10 个神经元 0-9。

打开 `cnn_mnist.py` 增加下面的符合 TensorFlow's Estimator api 接口的 `cnn_model_fn` 函数。`cnn_mnist.py` 接受 `mnist` 特征数据, 标签, [模型](#)作为参数, 配置 CNN, 返回预测, 损失, 训练操作。

下面的章节函数深入 `tf.layers` 代码创建每一层, 如何计算 loss, 配置训练操作, 生成预测。auguries 你已经体验过 CNN 设 TensorFlow Estimators, 你可以跳到[Training and Evaluating the CNN MNIST Classifier](#)

#### 3.4.4 输入层

这个方法为二维图像数据创建见卷积和池化, 输入 tensor 的形状为 [batch\_size,image\_width,image\_height]

- `batch_size`: 在训练过程执行提图下降的样本数据的子集大小。
- `image_width`: 样本图像的宽。
- `image_height`: 样本图像的高。
- `channels`: 样本图像的颜色通道, 对于彩色图想, 通道为 3, 对于单色图像通道为 1.

在这里, 我们的 MNIST 数据集由  $28 \times 28$  像素的单色照片组成, 因此输入层的形状为 `[batch_size,28,28,1]`, 转变我们的 feature map 到这个形状, 你可以执行操作:

```
1 input_layer = tf.reshape(features[“x”],[-1, 28, 28, 1])
```

这里的-1 表示输入的 `features[“x”]` 的值的 batch size 应该被动态计算, 保持所有的其它维度为常数。这允许我们将 `batch_size` 作为一个可以调节的超参数。例如, 如果我们输入样本到我们的 batchs 是 5 的模型, `features[“x”]` 将包含  $3920(5 \times 28 \times 28)$  值 (每一个值代表一个像素点), `input_layer` 形状将为 `[5,28,28,1]`, 类似的如果我们样本的 batchs 是 1000, `features[“x”]` 将包含 78400 个值, `input_layer` 形状将为 `[100,28,28,1]`。

#### 3.4.5 第一层卷积层

在我们的卷积层我想用 32 个  $5 \times 5$  的卷积核到输入层, 用 ReLU 激活函数, 我们一可用 `conv2d()` 方法创建这个层:

```
1 conv1 = tf.layers.conv2d(
2     inputs=input_layers,
3     filters=32,
4     kernel_size=[5,5],
5     padding="same",
```

```

6     activation=tf.nn.relu
7 )

```

inputs 参数指定我们的输入 tensor(形状为 [batch\_size,image\_width,image\_height,channels]), 这里, 我们链接我们的第一个吉安基层到输入层, 形状为 [batch\_size,28,28,1] 注意: 如果传递参数 data\_format=channels\_first,conv2d() 接受 [channels,batch\_size,image\_width,image\_height] 形状的数据。

filter 参数制定卷积核的个数, 这里卷积核为 32 个。kernel\_size 制定卷积核的维度为 [width,height] (这里 [5,5]) padding 参数制定两个值:valid(默认), 和 same。制定输出 tensor 应该有和输入特征是偶然相同的形状, 我们设置 padding=same, 说明 TensorFlow 增加 0 值到输出 tensor 的边缘波啊池宽度和高度为 28(没有 padding $5 \times 5$  卷积  $28 \times 28$  将生成  $24 \times 24$ tensor, 在  $28 \times 28$  用  $5 \times 5$  提取出  $24 \times 24$  个位置)。activation 参数指定应用到输出的激活函数, 这里我们只顶 tf.nn.relu。conv2d() 的输出形状为 [batch\_size,28,28,32]: 和输入有相同的宽度和高度, 但是有 32 个通道保持每个卷积核的输出。

### 3.4.6 池化层 1

链接我们创建的卷积层和池化层, 我们在 layers 中用 max\_pooling2d() 方法构造执行最大池化, 卷积核 filter 大小为  $2 \times 2$ , stride 为 2。

```

1 pool1 = tf.layers.max_pooling2d(inputs=conv1, pool_size=[2, 2], strides=2)

```

再次, inputs 制定输入 tensor, 形状为 [batch\_size,image\_width,image\_height,channels], 这里我们的输入 tensor 是第一层卷积层的输出 conv1, 形状为 [batch\_size,28,28,32]

pool\_size 指定最大池化 filter 的大小作为 [width,height] (这里是 [2,2]) 如果两个维度相等你可以指定 pool\_size=2。strides 参数制定 stride 的大小, 这里我们设置 strides 为 2, 表示通过 filter 提取子区域的时候宽度和高度都是 2 像素。如果你想设置不同的 width 和 height, 你可以制定一个元祖或者列表。

我们的输出特征是偶然和 max\_pooling2d(pool1, 形状为 [batch\_size,14,14,32]) 相乘:  $2 \times 2$  减少宽度和高度到 50%。

### 3.4.7 二层卷积和池化

我们用 conv2d() 和 max\_pooling2d() 链接卷积和池化。对于卷积层 2, 我们配置 64 个  $5 \times 5$  的卷积核, 激活函数为 ReLU, 池化层 2, 我们用和池化层一眼个间隔:

```

1 conv2 = tf.layers.conv2d(
2     inputs=pool1,
3     filters=64,

```

```

4     kernel_size=[5, 5],
5     padding="same",
6     activation=tf.nn.relu)
7
8 pool2 = tf.layers.max_pooling2d(inputs=conv2, pool_size=[2, 2], strides=2)

```

卷积层用 pool1 作为输入，生成 tensor conv2。conv2 形状为 [batch\_size, 14, 14, 64]，和 pool1 的宽和高相等，64 个通道因为 64 个卷积核。

池化层 2 那 conv2 作为输入，生成 pool2 作为输出，pool2 形状 [batch\_size, 7, 7, 64]（减少 conv2 50% 的宽度和高度）

### 3.4.8 Dense layer

我们添加 dense 层（1024 个神经元和 ReLU 激活函数）到 CNN 生成卷积/池化层提取的特征分类，我们将 flatten 我么呢 feature map(pool2) 到形状 [batch\_size, features]，因此我们的 tensor 有两维，上面的形状变成了 [batch\_size, 7 × 7]：

```

1 pool2_flat = tf.reshape(pool2, [-1, 7 * 7 * 64])

```

现在哦我们用 dense 方法链接我们的 dense:

```

1 dense = tf.layers.dense(inputs=pool2_flat, units=1024, activation=tf.nn.relu)

```

inputs 参数制定输入 tensor：我们的 flattened 的 feature map pool2\_flat。units 参数指定 dense 层的神经元的数量。activation 参数获取激活函数，这里我们依然是用 tf.nn.relu。为了改进我们的模型，我们也应用 dropout 方法正则化 dense 层。

```

1 dropout = tf.layers.dropout(
2     inputs=dense, rate=0.4, training=mode == tf.estimator.ModeKeys.TRAIN)

```

inputs 参数和上面一样，rate 参数制定 dropout 比率，这里用 0.4 表示 40% 的元素将在训练中被随机丢弃。training 参数得到一个 bool 行值制定是否模型在训练模式下运行，dropout 仅仅在 training 为 True 时执行。这里我们检查是否 mode 传递给我们 cnn\_model\_fn 的模型函数是 TRAIN 模式。输出形状为 [batch\_size, 1024]

### 3.4.9 Logits Layers

在我们神经网络的最后一层是 logits 层，然会预测的原始值。我们用 10 个神经元创建一个 dense layers，激活函数哦认为线性激活函数。

```

1 logits = tf.layers.dense(inputs=dropout, units=10)

```

我们最终输出 CNN 的 tensor，logits 形状为 [batch\_size, 10]。

### 3.4.10 常见的预测

logits 层返回我们预测的原始值（形状 [batch\_size,10]）。让我们转化这些原始值到我们的模型函数能返回的两种个不同的格式。

- predicted class: 数字 0-9。
- probabilities: 对于每个可能的目标类的概率。

对于更定的例子，我们的预测类是在相关行 logits 列有最大的值。我们可以用该 tf.argmax 函数找到这个元素的索引。

```
1 tf.argmax(input=logits, axis=1)
```

input 参数制定需要提取最大值的 tensor，axis 参数制定输入 tensor 沿着哪个轴寻找最大值。这里我们写着 1 轴寻找最大值。我们可以用 softmax 生成概率。

```
1 tf.nn.softmax(logits, name="softmax_tensor")
```

我们融合我们的预测到一个字典中，返回一个 EstimatorSpec 对象。

```
1 predictions = {
2     "classes": tf.argmax(input=logits, axis=1),
3     "probabilities": tf.nn.softmax(logits, name="softmax_tensor")
4 }
5 if mode == tf.estimator.ModeKeys.PREDICT:
6     return tf.estimator.EstimatorSpec(mode=mode, predictions=predictions)
```

### 3.4.11 计算 Loss

对于训练和评估阶段，我们需要定义损失函数衡量我们的模型的预测如何接近目标类。对于想 MNIST 的多个分类问题，cross entropy 是典型的被用做损失度量。下面的代码计算交叉熵返回 TRAIN 或者 EVAL 模式：

```
1 onehot_labels = tf.one_hot(indices=tf.cast(labels, tf.int32), depth=10)
2 loss = tf.losses.softmax_cross_entropy(
3     onehot_labels=onehot_labels, logits=logits)
```

我们的 labels tensor 包含一个预测列表，像 [1,9,...]，为了计算交叉熵，你需要转换 labels 为相关的one-hot encoding

```
1 [[0, 1, 0, 0, 0, 0, 0, 0, 0, 0],
2  [0, 0, 0, 0, 0, 0, 0, 0, 0, 1],
3  ...]
```

womenyoingtf.one\_hot 函数执行转换。tf.one\_hot() 有两个参数：

- one-hot tensor 有值的位置，如上面 1, 表示位置索引为 1 的地方有 1
- depth:one-hot tensor 的深度，目标类的数量，这里 depth 为 10,

下面的代码为我们的 labels 创建一个 one-hot tensor, onehot\_labels:

```
1 onehot_labels = tf.one_hot(indices=tf.cast(labels, tf.int32), depth=10)
```

因为 labels 包含值从 0-9,indices 是我们的 labels tensor, 值变为证书。depth 是 10 因为我们有 10 个可能的目标类。下一步我们计算 onehot\_labels 的交叉熵和我们的 logits 层的 softmax 预测。tf.losses.softmax\_cross\_entropy() 得到 onehot\_labels 和 logits 作为参数。在 logits 上执行 softmax 激活函数，返回损失的标量 tensor:

```
1 loss = tf.losses.softmax_cross_entropy(
2     onehot_labels=onehot_labels, logits=logits)
```

### 3.4.12 配置训练操作

在先前的操作中我们为我们的 CNN 定义了损失作为 logits 层和 layers 的 softmax cross-entropy。让我们配置我们的模型在训练落成中优化 loss。我们将用 0.001 学习率和 SGD 作为优化算法

```
1 if mode == tf.estimator.ModeKeys.TRAIN:
2     optimizer = tf.train.GradientDescentOptimizer(learning_rate=0.001)
3     train_op = optimizer.minimize(
4         loss=loss,
5         global_step=tf.train.get_global_step())
6     return tf.estimator.EstimatorSpec(mode=mode, loss=loss, train_op=train_op)
```

### 3.4.13 增加评估度量

为了增加度量到我们的模型，我们在 EVAL 定义了 eval\_metric\_ops 字典:

```
1 eval_metric_ops = {
2     "accuracy": tf.metrics.accuracy(
3         labels=labels, predictions=predictions[ "classes" ])}
4 return tf.estimator.EstimatorSpec(
5     mode=mode, loss=loss, eval_metric_ops=eval_metric_ops)
```

## 3.5 训练评估 CNN MNIST 分类器

我们已经构建了 MNIST CNN 模型函数，现在我们准备训练评估它。

### 3.5.1 载入训练和测试数据

增加 main() 函数到 cnn\_mnist.py 载入训练数据和测试数据。

```

1 def main(unused_argv):
2     # Load training and eval data
3     mnist = tf.contrib.learn.datasets.load_dataset("mnist")
4     train_data = mnist.train.images # Returns np.array
5     train_labels = np.asarray(mnist.train.labels, dtype=np.int32)
6     eval_data = mnist.test.images # Returns np.array
7     eval_labels = np.asarray(mnist.test.labels, dtype=np.int32)
```

我们存储训练数据 train\_data(55000 张原始图像的像素值) 训练 train\_labels(每张图片 0-9) 作为 numpy 数组。类似的我们存储评估数据 (10000 张) eval\_data 和 eval\_labels。

### 3.5.2 创建 Estimator

下一步创建一个 Estimator(一个用于执行高级模型训练, 评估, 推理的 TensorFlow 类), 增加下面代码到 main() 中。

```

1 # Create the Estimator
2 mnist_classifier = tf.estimator.Estimator(
3     model_fn=cnn_model_fn, model_dir="/tmp/mnist_convnet_model")
```

model\_fn 参数指定用于训练, 评估, 预测的模型函数, 我们传递 cnn\_model\_fn, models\_dir 参数制定模型数据的保存目录为 /tmp/mnist\_convnet\_model。

### 3.5.3 建立 Logging Hook

因为 CNN 可能花一会训练, 让我们设置一些采集以至于我们在训练时能跟踪进层。我们用 TensorFlow 的 tf.train.SessionRunHook 创建一个 tf.train.LoggingTensorHook 采集从 softmax 层来的概率值, 增加下面代码到 main():

```

1 # Set up logging for predictions
2 tensors_to_log = {"probabilities": "softmax_tensor"}
3 logging_hook = tf.train.LoggingTensorHook(
4     tensors=tensors_to_log, every_n_iter=50)
```

我们存储一个我们想要采集进 tensors\_to\_log 的 tensor 词典。每个 key 是我们选择的 label, 将在采集输出被打印, 相关的 label 是 TensorFlow 图的 Tensor 的名字, 这里我们的概率可以在 softmax\_tensor 中找到, 我们给我们 softmax 操作的名字在 cnn\_model\_fn 生成概率。

下一步我们创建 LoggingTensorHook, 传递 tensor\_to\_log 到 tensors 参数, 我们设置 every\_n\_iter=50, 制定训练的时候每 50 步采集概率。

### 3.5.4 选练模型

现在我们准备好训练我们的模型，我们通过创建 train\_input\_fn 和在 mnist\_classifier 调用 train()，增加下面到 main()

```

1 # Train the model
2 train_input_fn = tf.estimator.inputs.numpy_input_fn(
3     x={"x": train_data},
4     y=train_labels,
5     batch_size=100,
6     num_epochs=None,
7     shuffle=True)
8 mnist_classifier.train(
9     input_fn=train_input_fn,
10    steps=20000,
11    hooks=[logging_hook])

```

在 Numpy\_input\_fn 调用的时候，我们传递训练特征数据和标签给 x 和 y。我们设置 batch\_size 是 100（模型训练的时候每次最小批次是 100 个样本）。num\_epochs=None 意味着模型将训练直到指定步数到达。我们也设置 shuffle=True 打乱训练数据，在训练调用的时候，我们设置 steps=20000(这意味着模型总共训练 20000 次) 我们传递 looging\_hook 去 hooks 参数，以至于它将在训练期间被触发。

### 3.5.5 评估模型

当训练结束是我们想要在测试及评估我们的模型，我们可以调用 evaluate 方法，在 model\_fn 指定 eval\_metric\_ops 参数度量方法:

```

1 # Evaluate the model and print results
2 eval_input_fn = tf.estimator.inputs.numpy_input_fn(
3     x={"x": eval_data},
4     y=eval_labels,
5     num_epochs=1,
6     shuffle=False)
7 eval_results = mnist_classifier.evaluate(input_fn=eval_input_fn)
8 print(eval_results)

```

为了创建 eval\_input\_fn，我们设置 num\_epochs=1，因此模型评估在一个时期评估数据返回结果。我们也设置 shuffle=False 通过数据序列迭代。

### 3.5.6 运行模型

下面是采集的输出:

```
1 INFO:tensorflow:loss = 2.36026, step = 1
2 INFO:tensorflow:probabilities = [[ 0.07722801  0.08618255  0.09256398, ...]]
3 ...
4 INFO:tensorflow:loss = 2.13119, step = 101
5 INFO:tensorflow:global_step/sec: 5.44132
6 ...
7 INFO:tensorflow:Loss for final step: 0.553216.
8
9 INFO:tensorflow:Restored model from /tmp/mnist_convnet_model
10 INFO:tensorflow:Eval steps [0,inf) for training step 20000.
11 INFO:tensorflow:Input iterator is exhausted.
12 INFO:tensorflow:Saving evaluation summary for step 20000: accuracy = 0.9733, los
13 {'loss': 0.090227105, 'global_step': 20000, 'accuracy': 0.97329998}
```

我们在测试集上获得了 97.3% 的精确度。



# Chapter 4

## 扩展

这个章节解释开发者如何增加功能到 TensorFlow。

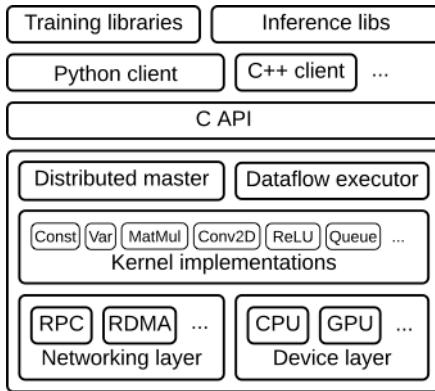
### 4.1 TensorFlow 架构

我们设计 TensorFlow 是为了大规模分布式训练和推理，但是它也能灵活的支持一些新的机器学习模型实验和系统级别的优化。

这个文件描述了这个系统架构使得结合这个规模和灵活度成为可能。假设你熟悉 TensorFlow 基本的一些概念，像计算图，操作绘画。这个文档适合于那些想用当前 API 不支持的一些方法扩展 TensorFlow，想要优化 TensorFlow 的硬件工程师，在法规莫分布式系统上实现机器学习系统或者是任何想要了解 TensorFlow 的 hood 的人。读完它后你应该能读和修改 TensorFlow 核心代码。

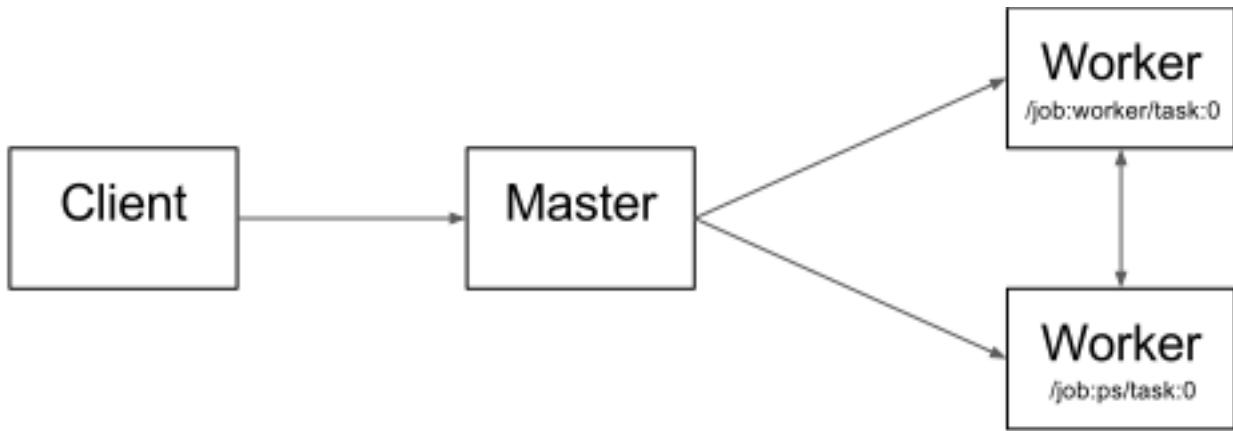
### 4.2 概述

TensorFlow 运行时是一个跨平台的库，下图画出了常用的架构，C API 分隔用户代码和核心代码。



- Client
- 定义计算作为数据流图。
- 用 session 初始化图。
- Distributed Master
  - 从图中修剪一个子图作为定义的参数给 Session.run()
  - 分开不同的子图为多个部分在不同的进程和设备上运行。
  - 分配图块到 worker service。
- Worker services
  - 调度图上的操作在可用的硬件平台 (CPUs, GPUs) 上执行。
  - 发送和接收 worker service 的操作结果。
  - 内核实现。
  - 执行单个图操作的计算。

下图说明逐渐的交互。”job:worker/task:0” 和”/job:ps/task:0” 两个任务在 workers 上。”PS” 代表”parameter server”: 一个负责存储更新模型参数的任务。另一个任务优化参数时发送更新到这些参数，类似的在任务之间的分隔是不被要求的，但是它通常用于分配的训练。



注意 Distributed Master 和 Worker Service 仅仅存在于分布的 TensorFlow。, 单进程版本的 TensorFlow 包含一个特别的 Session 实现能做任何 Distributed master 能做的不仅仅是和本地进程通信。

下面的章节表述了 TensorFlow 的核心。

#### 4.2.1 Client

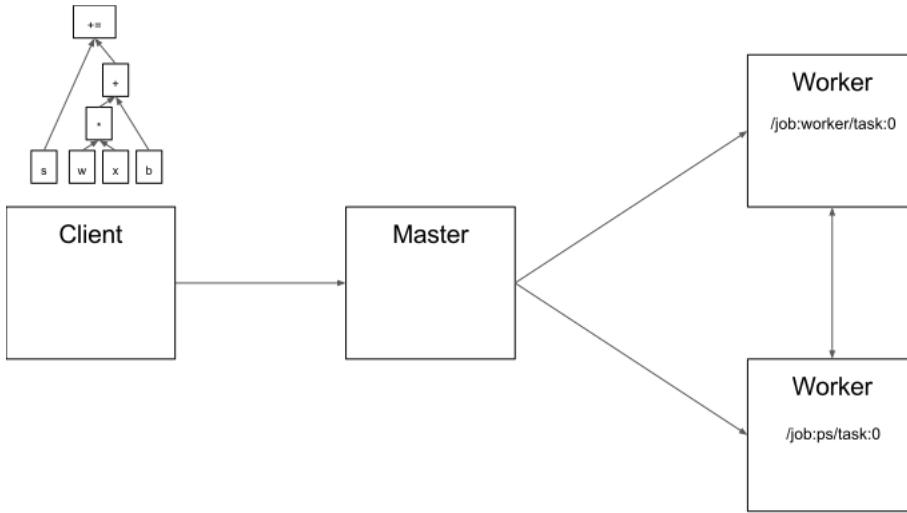
用户写 TensorFlow 程序构造计算图。这个图给需既可以组成单个操作或者用一个像 Estimators API 的方便的库组成神经网络乘和其他高级抽象。TensorFlow 支持多种用户语言, 但是我们优先使用 Python 和 C++, 仅仅是因为我们的内部用户熟悉他们。当特征被建立好后我们将他们接入 C++。因此用户可以得到一个对所有语言优化的实现。大多数的训练库仅仅支持 Python, 但是 C++ 支持更高效的推理。用户创建一个绘画, 发送图的定义到 distributed master 作为 tf.GraphDef 协议缓冲区。然后客户评估图上的一个节点或者多个节点, 评估触发一个 distributed master 的调用初始化计算。

在下图中, 客户建立一个图, 应用权重 ( $w$ ) 到特征向量 ( $x$ ), 增加偏置 ( $b$ ) 保存结果。

#### 4.2.2 Distributed master

- 修剪图得到子图计算用户的节点请求。
- 对于每一个加入的设备, 分隔图获得子图。
- 缓存这些块以至于他们能用在自序列中。

因为 master 查看每一步的计算, 它用像常用的子表达式消除和常数折叠的标准的优化。它然后执行优化的子图。下图显示一个可能的分隔。distributed master 有组合的模型参数为了放置他们在参数服务器上。这里图的边缘被分隔, distributed master 发送接收节点在不同的任务间传送信息。下面的 distributed master 传输子图到分布的任务。



#### 4.2.3 Worker Service

任务中的 worker service。

- 处理 master 的请求。
- 调度内核执行包含本地子图的操作
- 任务间的直接通信。

我们优化 worker service 为了能用更小的花销以支持大的图。我们当前的实现能实现每秒执行上万张子图，使得大量的副本快速的训练。worker service 布置内核到本地设备上然后通过利用多 CPU 多 GPU 尽可能的并行执行。我们为源和目的设备指定发送和接收操作。

- 用 cudaMemcpyAsync() API 在本地 CPU 和 GPU 之间转换，覆盖计算和数据的转化。
- 用对等的 DMA 在不同的本地 GPU 之间转化避免通过主 CPU 的高昂代价。

对于任务间的转化，TensorFlow 用多个协议，报错：

- gRPC over TCP
- RDMA over Converged Ethernet

我们对于 NVIDIA 的多 GPU 通信 NCCL 库有初步的支持，查看 `tf.contrib.nccl`

## 4.3 内核实现

运行包含超过 200 个标准操作白扩数学，数组操作，控制流，状态管理操作。每一个操作对不同的设备有优化，一些操作内核用 Eigen::Tensor 实现，用 C++ 模板生成在多核 CPU 和 GPUs 上生成高效的并行代码，然而我们优先像像 CuDNN 这类更高效实现的库。我们也实现了量化，能在移动设备和高流通数据中心应用上更快地推理，用 gemmlowp 地精读矩阵库加速量化计算。

如果很难或者抵消的表达子计算作为操作的组成，用户可以注册额外的京城通过 C++ 提供更高效的实现，，我们推荐你 duit 一些重要的操作像 ReLU 和 Sigmoid 和相关的梯度注册你的融合内核，XLA 变压器有意额实验性是实现自动内核融合。



# Chapter 5

## Performance

这个导航包含一个优化你的 TensorFlow 代码的集合。对于 Tensorflow 用户来说这是最好的应用，正如在这个文档中最好的时间，高性能模式为在不同的硬件上创建模型文档链接到示例代码。

### 5.1 最好的实践

尽管优化实现不同类型的模型可能不同，下面是通过 tensorflow 实现性能的几个最好的方式，尽管这些暗示在基于图像的模型，我们将增加一些技巧到所有类型的模型。下面列出了最好实践的关键：

- 从原来码编译安装
- 利用队列读取数据
- 在 CPU 上预处理
- 用 NCHW 图像格式
- 在 GPU 上放共享参数
- 用融合的批处理规范

下面章节时处理的详细信息。

### 5.2 从源代码创建安装

为了安装最优化的 TensorFlow 版本，通过源代码编译安装 Tensorflow。从原来码编译优化目标硬件确保最新的 CUDA 平台和 CuDNN 库被用高性能安装。

对于多数稳定的实验，从最新版的[latest release](#)分支编译。为了得到最新性能改变接受一些稳定性风险，从[master](#)编译。

如果你需要在不同的目标硬件平台上编译 TensorFlow，交叉编译最优化目标平台。下面的目录是一个例子高数 bazel 为指定平台编译

```
1 # This command optimizes for Intel's Broadwell processor
2 bazel build -c opt --copt=-march="broadwell" --config=cuda //tensorflow/tools/
               pip_package:build_pip_package
```

### 环境，构建，安装技巧

- 编译最高级别的[GPU 支持](#)，e.g. P100: 6.0, Titan X (pascal): 6.2, Titan X (maxwell): 5.2, and K80: 3.7.
- 安装最新版的 CUDA 平台和 cuDNN 库
- 确保你的 gcc 版本支持对目标 cpu 所有的优化，推荐最小的 gcc 版本为 4.8.3
- TensorFlow 在启动时检查是否已经在 cpu 上编译优化过，如果优化不被包含，TensorFlow 将 chxuan 警告，e.g. AVX, AVX2 和 FMA 设备不被包含。

#### 5.2.1 利用队列读取数据

在利用 GPUs 时性能很差或者没有设置高效的 pipeline 导致缺乏数据，确保设置输入 pipeline 高效利用队列和流数据，一种识别 GPU 处于饥饿状态的方法时生成和查询时间线。一个相信的时间线指南不存在，但是一个快速生成时间线的例子在[XLA JIT](#)部分存在，另一个检查是否 GPU 被充分使用时运行 nvidia-smi 查看，如果 GPU 利用没有达到 100% 这样 GPU 没有足够的快的得到数据。

除非指定一个特殊的情形或者示例代码，没有从 Python 变量给予数据到会话，e.g.

```
1 # Using feed_dict often results in suboptimal performance when using large
      inputs.
2 sess.run(train_step, feed_dict={x: batch_xs, y_: batch_ys})
```

#### 5.2.2 在 CPU 上的预处理

将预处理操作放在 CPU 上可能对性能提升很重要，当预处理发生在 GPU，数据流使从 CPU->GPU(预处理)->CPU->GPU(训练)。这数据被限制在 CPU 和 GPU 之间，当预处理被放在 CPU 上，数据流是 CPU(预处理)->GPU(训练)。另一个好处是在 CPU 上预处理释放 GPU 时间让其集中训练。

将预处理放在 CPU 上可能导致对 sample/sec 处理速度 6 倍以上的处理性能增加，将导致训练时间缩短为原来的  $\frac{1}{6}$ ，确保预处理在 CPU 上，按照如下操作：

```
1 with tf.device('/cpu:0'):
2     # function to get and process images or data.
3     distorted_inputs = load_and_distort_images()
```

### 5.2.3 用大文件

在一些情形下，CPU 和 GPU 可能通过 I/O 操作获取数据时对数据处于饥饿状态。如果你正用一些小文件形成输入数据集，你也许被你的文件系统限制了速度。如果你在 SSD 上而不是 HDD 上存储你的输入数据你的训练循环运行更快。如果是这样你应该通过创建一些大的 TFRecord 文件预处理你的输入数据。

### 5.2.4 用 NCHW 图像数据格式

图像数据格式涉及到图像的批量表示。TensorFlow 支持 NHWC(TensorFlow 默认) 和 NCHW(cuDNN 默认)，N 时图像的批数，H 时图像垂直方向的像素数量，W 是水平方向的像素，C 时图像的通道数，尽管 cuDNN 能处理上面两种格式，但是它处理默认格式更快。最好的实现是用 NCHW 和 NHWC 构建模型正如通常在 GPU 上用 NCHW 训练然后在 CPU 上用 NHWC 推断。

TensorFlow 用这两个格式是的一个简单的历史因为它在 CPUs 上运行快点，然后 TensorFlow 团队发现当 NVIDIA cuDNN 库时 NCHW 运行更好。当即用户推荐在他们的模型中支持两种格式，在很长一段时期，我们计划重写图转化两种格式。

### 5.2.5 用融批规范

当用批规范 `tf.contrib.layers.batch_norm` 设置属性 `fused=True`:

```
1 bn = tf.contrib.layers.batch_norm(
2     input_layer, fused=True, data_format='NCHW'
3     scope=scope, **kwargs)
```

在没有融合批规范计算几个单独的操作。融合批规范结合单个操作进入内核，运行更快。

### 5.3 性能向导

### 5.4 好性能模型

### 5.5 Benchmark

## 5.6 如何用 TensorFlow 量化神经网络

现代神经网络已经被开发出了，最大的挑战是让他们工作！这意味着在训练中的精确度和速度被优先考虑，浮点时是保留精确度的最简单的方法，GPUs 擅长简爱素这些计算，因此没有太多的注意被放在其它数据格式上。

这些天我们做了一些模型部署在商业应用上，训练的计算要求随着研究人员的数量增加，对于推断的需要正在扩张。这意味着推断效率变成的一些团队最麻烦的问题。

这是量化出现了，它覆盖了一些存储数字和计算执行在更多兼容的 32bit 浮点数。我们将关注固定点下面我将说宁更多细节。

#### 5.6.1 为什么做量化工作

训练神经网络通过对权值小的推动，这些小的推动需要浮点精度工作。

预先训练模型和运行推断有很大不同，一个深度网络的神奇的量化时他们像是复制高级别的噪声在他们的输入。如果考虑识别一个你拍摄照片中的的对象，网络必须在它和训练样本中忽视 CCD 上的噪声，光线改变其它不重要的差异在它和训练样本被看到前，之一梨放在重要的类似的事上。这个能力意味着他们需要退待地精读计算作为另一个源噪声，产生精度结果升值数值格式抓住更少信息。

#### 5.6.2 为什么量化

神经网络模型可能占据一些磁盘空间，原始的 AlexNet 腹地使能数据占据超过 200MB。大多数的这些大小被神经网络连接的权重占据，因为经常单个模型有上百万个神经元。因为他们时有一些不同的浮点数，简单的压缩格式像 zip 不能很好的压缩他们，他们被安排仅一个大的层，每一层的权重趋向于一定范围的正态分布，比如-3.0 到 6.0。

最简单的量化动机是通过存储每一层的最大值和最小值缩小文件大小。然后压缩每个浮点值为 8 位代表最接近到 256 内的真实整数。例如范围-3.0-6.0,0 代表-3,255 代表 6.0,128 代表 1.5。我在之后将进行确切的计算，因此有一些细节，但是这意味着你可以得到缩小文件尺寸 75% 的好处，然后你可以通过导入后在不更改任何存在的浮点数代码然后转换为浮点数。

另一个原因是量化前通过在 8 位输入输出减少你运行前的你需要推理计算资源，获取 8 位值仅仅需要浮点数 25% 的内存带宽，因此你可以更充分使用缓存避免 RAM 存取瓶颈，你也可以用 SIMD 操作在每个时钟周期做更多操作。在 yxiieqingkuangxia 你将有一个 DSP 芯片可以激素 8 为计算得到更多好处。

移动嗯计算到 8 为将帮助你更快地运行模型，用更少的电量，同时它也为不能运行浮点代码的嵌入式系统打开了开了一扇门，因此可以应用到 IoT 世界。

### 5.6.3 为什么不直接训练低精度

我们正在一些更低深度上做了一些实验，结果似乎显示你需要高于 8 位处理反向传播和梯度，这使得实现训练变得更复杂推理混乱。我们已经有一致的浮点数模型使用，因此能直接方便的转换他们。

### 5.6.4 你能如何量化你的模型

TensorFlow 支持生成 8 位计算，它也有一些通过用量化计算推理转换训练模型浮点数到相应的图上。例如这里你可以转换最新的 GoogleLeNet 模型用 8 位版本计算：

```

1 curl http://download.tensorflow.org/models/image/imagenet/
2 inception-2015-12-05.tgz -o /tmp/inceptionv3.tgz
3 tar xzf /tmp/inceptionv3.tgz -C /tmp/
4 bazel build tensorflow/tools/quantization:quantize_graph
5 bazel-bin/tensorflow/tools/quantization/quantize_graph \
6   --input=/tmp/classify_image_graph_def.pb \
7   --output_node_names="softmax" --output=/tmp/quantized_graph.pb \
8   --mode=eightbit

```

你将在原始的操作上运行一个新的模型，但是 8 位计算在内部，所有的权重被量化。如果你查看文件尺寸，你将看到大约是原来的 1/4(23MB 对比 91MB)，你可以用相同的输入输出运行这个模型，你将得到相应的结果，这里是代码：

```

1 # Note: You need to add the dependencies of the quantization
2 # operation to the
3 #       cc_binary in the BUILD file of the label_image program:
4 #
5 #       //tensorflow/contrib/quantization:cc_ops
6 #       //tensorflow/contrib/quantization/kernels:quantized_ops
7
8 bazel build tensorflow/examples/label_image:label_image

```

```

9  bazel-bin/tensorflow/examples/label_image/label_image \
10 --image=<input-image> \
11 --graph=/tmp/quantized_graph.pb \
12 --labels=/tmp/imagenet_synset_to_human_label_map.txt \
13 --input_width=299 \
14 --input_height=299 \
15 --input_mean=128 \
16 --input_std=128 \
17 --input_layer="Mul:0" \
18 --output_layer="softmax:0"

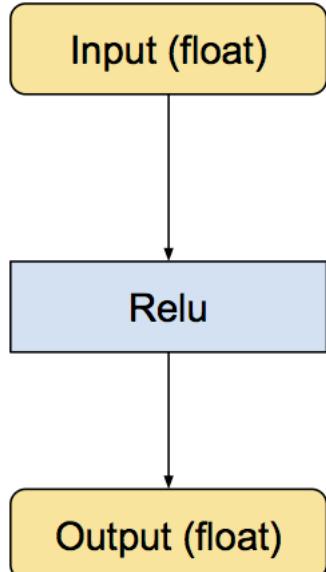
```

你将看到最新的量化图的运行，输出和原始输出十分类似。

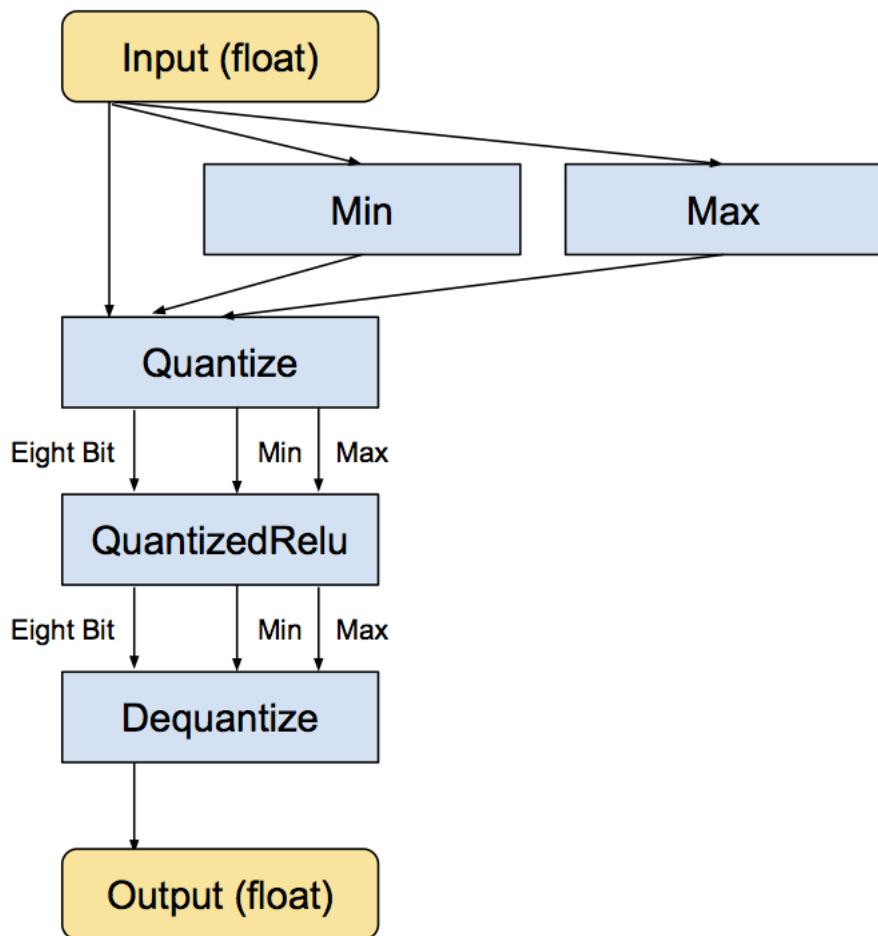
你可以运行相同的处理在你的模型上报春为 GraphDefs，输入输出的名字用在你的网络请求上。我推荐你首先通过 freeze\_graph 脚本，转化检查点为常数存储在文件中。

### 5.6.5 如何量化处理工作

我们在推理过程中通过写 8 位量化本本操作实现两话，这包含卷积，矩阵相乘，激活函数，吃花草做和链接，转化脚本首先对所有的操作量化。有一些小的子图之前有转化函数之后在浮点数和 8 位数之间移动，下面是一个例子，首先原始 Relu 操作输入输出浮点数。

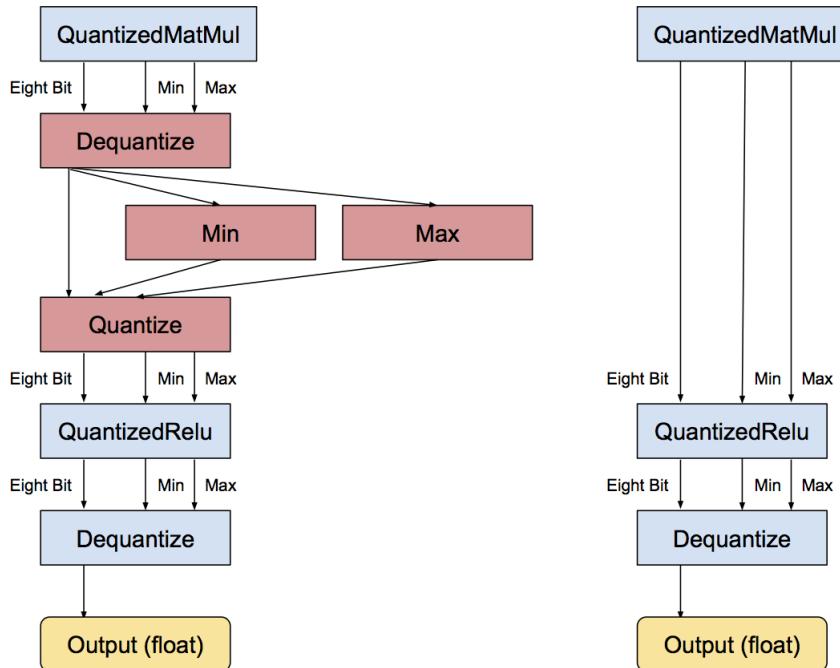


然后相应的转换子图，仍然是浮点输入输出，内部转换完成后以 8 位计算：



最小，最大操作实际上查看输入浮点 tensor 的值，输入他们到量化操作转换 tensor 为 8 位。

当单个操作被转换后，下一步是移除必要的转换到浮点。如果有一个连续的序列操作，将有一些链接 Dequantize/Quantize 操作



应用到大规模模型上时所有的操作已经相应的量化，图上所有的计算用 8 为计算不转换为浮点数。

### 5.6.6 量化 Tensor 将呈现什么

我们通过转化浮点数组为 8 位表达式作为压缩问题。我们知道权重和激活 Tensor 在训练神经网络模型时值的分布在一个小的范围（你也许有一个 -15 到 +15 的权重，-500 到 1000 激活）。在实验中我们了解到神经网络通常在处理噪声时非常健壮，由量化产生噪声类似的误差将不会伤害精度。我们像卷则一个表达式这是容易执行计算，特别是大的矩阵惩罚函数需要运行一个模型的块。

这导致我们选择一个表达式有两个浮点数存储最小值和最大值代表最低和最高量化精度，每个在量化数组中的入口代表一个浮点值范围，现行飞蛾分布在最小值和最大值之间。例如我们有最小值 =-10.0 和最大值 30.0f，和 8 位数据，下面是量化表达式：

量化值	浮点数
0	-10
255	30.0
128	10.

这种表达式的好处是可以代表任一幅度的范围，我们不必退成，它可以代表有符号和无

符号的值，现行扩展使得直接相乘。对转换浮点数前后的一个清晰明确的量化格式定义好处，或者基于调试目的的查看 tensor 在 Tensorflow 上一个是线细节时希望提高将来最小值和最大值需要传递分割开得 Tensor 保持量化值，因此图个变得一点稠密。

最小和最大值范围可以提前计算，权重参数是常数在载入时就知道，因此他们的范围可以作为常数被存储。我们经常知道输入范围例如 (RGB 的值在 0-255)，一些激活函数也知道范围。这可以避免必须分析操作的输出决定范围，我们需要从 8 位输出像卷积或者矩阵乘法这样的做数学操作形成 32Bit 累加结果。

### 5.6.7 下一步

我们发现通过 8 为算法而不是不浮点数可以在移动短和嵌入式设备上得到机器好的性能。你可以看到这个框架我们优化矩阵乘在[gemmlowp](#)，我们仍然需要应用所有的我们需要学习的 TensorFlow 操作去在移动短得到最大型能，但是我们很兴奋正在为此努力。马上，量化实现是一个合理的快和精确度实现我们希望将能在更多的设备上广泛的支持 8 位模型。我们也希望站着时将鼓励社区探索更低精度的神经网络的可能性。



# Chapter 6

## 常用的 python 模块

### 6.1 Argparse

argparse 模块是一个用户友好的命令行接口，当用户每有给定可用的参数时，argparser 能自动生成帮助和使用信息。

```
1 import argparse
2 parser = argparse.ArgumentParser(description='Process some integers.')
3 parser.add_argument('integers', metavar='N', type=int, nargs='+', help='an integer
4 for the accumulator')
5 parser.add_argument('--sum', dest='accumulate', action='store_const', const=sum,
6 default=max, help='sum the integers(
7 default: find the max)')
8 args = parser.parse_args()
9 print(args.accumulate(args.integers))
```

```
hpc@hpc-322:~/TensorFlow_Notebook$ vim code/demo1.py
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo1.py
usage: demo1.py [-h] [--sum] N [N ...]
demo1.py: error: the following arguments are required: N
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo1.py 1 2 3 4
4
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo1.py --sum 1 2 3 4
10
```

代码能根据传入的参数选择相应的函数计算。

- 创建一个 parser
- 增加 arguments
- 解析参数

### 6.1.1 ArgumentParser 对象

```
class argparse.ArgumentParser(prog=None, usage=None, description=None, epilog=None,
parents=[], formatter_class=argparse.HelpFormatter, prefix_chars='-', fromfile_prefix_chars=None,
argument_default=None, conflict_handler='error', add_help=True, allow_abbrev=True)
```

- prog: 程序的名字 (默认为 sys.argv[0])
- usage: 描述程序用法的字符串。 (默认通过 arguments 增加到 parser)
- description: argument 帮助前的文本展示。 (默认为:None)
- epilog: argument 帮助之后的文本展示。 (默认为:None)
- parents: 应该被包含的列表对象。
- formatter\_class: 自定义输出帮助的类。
- prefix\_chars: 参数前面的字符。 (默认为'-')
- fromfile\_prefix\_chars: 应该被读的文件的字符串。
- argument\_default: 参数的全局值。 (default:None)
- conflict\_handler: 解决冲突选项的策略。 (通常不是必需的)
- add\_help: 增加-h/-help 选项到 parser。 (默认为 True)
- allow\_abbrev: 如果缩略不冲突, 可以允许长的选项被缩略。 (默认为 True)

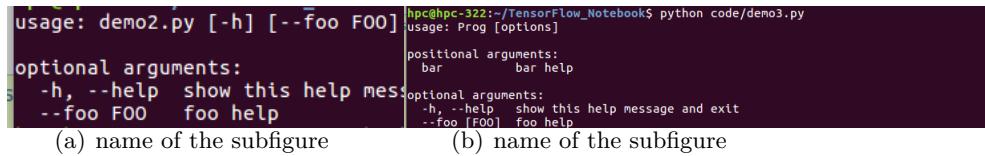
### 6.1.2 prog

默认情况下 ArgumentParser 对象用 sys.argv[0] 决定如何显示程序的名字。

```
1 #filename : arg1.py
2 import argparse
3 parser = argparse.ArgumentParser()
4 parser.add_argument("echo")
5 args = parser.parse_args()
6 print(args.echo)
```

默认情况下 ArgumentParser 从包含用法信息的参数计算 usage message。

```
1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument('--foo', help='foo help')
4 args = parser.parse_args()
```



大多数的 ArgumentParser 构造体用 `description=` 关键字，这个参数给出一个简单的程序说明其如何工作的。在帮助信息中表述在命令行和帮助信息之间。

```
1 import argparse
2 parser = argparse.ArgumentParser(description='A foo that bars')
3 parser.print_help()
```

```
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo4.py
usage: demo4.py [-h]

A foo that bars

optional arguments:
  -h, --help  show this help message and exit
```

一些程序喜欢在参数表述后添加一些额外的信息说明，这些说明可以通过 ArgumentParser 中的 `epilog=` 参数指定。

```
1 import argparse
2 parser = argparse.ArgumentParser(description='A foo that bars',
3 epilog="And that's how you'd foo a bar")
4 parser.print_help()
```

```
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo5.py
usage: demo5.py [-h]

A foo that bars

optional arguments:
  -h, --help  show this help message and exit

And that's how you'd foo a bar
```

有时候一些 parser 共享一些参数，相比于重复定义这些参数，一个单个的 parser 通过传递 `parents` 给 ArgumentParser。`parents=` 参数得到一个 ArgumentParser 对象的列表对象，从中收集所有的位置和选项行为

```
>>> parent_parser = argparse.ArgumentParser(add_help=False)
>>> parent_parser.add_argument('--parent', type=int)

>>> foo_parser = argparse.ArgumentParser(parents=[parent_parser])
>>> foo_parser.add_argument('foo')
>>> foo_parser.parse_args(['--parent', '2', 'XXX'])
Namespace(foo='XXX', parent=2)

>>> bar_parser = argparse.ArgumentParser(parents=[parent_parser])
>>> bar_parser.add_argument('--bar')
>>> bar_parser.parse_args(['--bar', 'YYY'])
Namespace(bar='YYY', parent=None)
```

大多数的 parent parser 指定 `add_help=False`，因此 ArgumentParser 将看到两个帮助选项（一个在 parent 一个在 child）同时报错。你必须在通过 `parsers=` 传递前必须完全初始化

parser, 如果你在 child parser 改变 parent parsers, 改变将不被反映到 child.formatter\_class ArgumentParser.durian 允许指定可用的格式化类自定义格式, 当前有 4 个类:

- argparse.RawDescriptionHelpFormatter
- argparse.RawTextHelpFormatter
- argparse.ArgumentDefaultHelpFormatter
- argparse.MetavarTypeHelpFormatter

RawDescriptionHelpFormatter 和 RawTextHelpFormatter 在如何显示说明上给与更多控制, 默认 ArgumentParser 对 description 和 epilog 在命令终端一行显示。

```

1 import argparse
2 parser = argparse.ArgumentParser(prog='PROG', description='''
3     description was indented wierd
4     but that is okey ''',
5     epilog=''''
6     likewise for this epilog whose whitespace will be
7     cleaned up and whose words will be wrapped
8     across a couple lines ''')
9 parser.print_help()

```

```

hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo/.py
usage: PROG [-h]

this description was indented wierd but that is okey

optional arguments:
  -h, --help  show this help message and exit

likewise for this epilog whose whitespace will be cleaned up and whose words
will be wrapped across a couple lines

```

传递 RawDescriptionHelpFormatter 作为 formatter\_class= 让 description 和 epilog 正确显示。RawTextHelpFormatter 主要维持素有的帮助文本, 值描述的信息。

ArgumentDefaultHelpFormatter: 自动增加关于值的默认信息。

```

1 import argparse
2 parser = argparse.ArgumentParser(prog='Prog',
3     formatter_class = argparse.ArgumentDefaultsHelpFormatter)
4 parser.add_argument('foo', type=int, default=42, help='FOO')
5 parser.add_argument('bar', nargs='*', default=[1, 2, 3], help='BAR!')
6 parser.print_help()

```

```
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo9.py
usage: Prog [-h] [--foo FOO] [bar [bar ...]]

positional arguments:
  bar          BAR! (default: [1, 2, 3])

optional arguments:
  -h, --help    show this help message and exit
  --foo FOO    FOO (default: 42)
```

MatavarTypeHelpFormatter 用 type 显示参数显示值的名字。

```
1 import argparse
2 parser = argparse.ArgumentParser(prog='Prog',
3                                 formatter_class=argparse.ArgumentDefaultsHelpFormatter)
4 parser.add_argument('--foo', type=int, default=42, help='FOO')
5 parser.add_argument('bar', nargs='*', default=[1, 2, 3], help='BAR!')
6 parser.print_help()
```

```
hpc@hpc-322:~/TensorFlow_Notebook$ python code/demo10.py
usage: PROG [-h] [--foo int] float

positional arguments:
  float

optional arguments:
  -h, --help    show this help message and exit
  --foo int
```

prefix\_chars, 大多数命令行参数选项用-, 比如-f/-foo。parsers 需要支持不同的或者说另外的前缀, 像 +f 或者/fo 可以设置 prefix\_chars= 参数指定。prefix\_chars 默认默认为-, 用非-字符能禁用-f/-foo 这种类型的选项。

```
1 import argparse
2 parser = argparse.ArgumentParser(prog='PROG', prefix_chars='+-')
3 parser.add_argument('+f')
4 parser.add_argument('++bar')
5 parser.parse_args('+f X ++bar Y'.split())
```

fromfile\_prefix\_chars, 有时我们处理一个长的参数列表, 将参数保存在文件中比直接在命令行中更容易理解, 如果 fromfile\_prefix\_chars= 参数给 ArgumentParse 结构体, 指定的参数将被作为文件, 被下面的参数取代。例如

```
1 import argparse
2 with open('args.txt', 'w') as fp:
3     fp.write('-f\nbar')
4 parser = argparse.ArgumentParser(fromfile_prefix_chars='@')
5 parser.add_argument('-f')
6 parser.parse_args(['-f', 'foo', '@args.txt'])
```

默认从一个文件读取参数, 上面的表达式 ['-f', 'foo', '@args.txt'] 等于表达式 ['-f', 'foo', '-f', 'bar'], fromfile\_prefix\_chars 参数默认为 None, 意味着参数不被当作文件。argument\_default

通常通过传递 add\_argument 或者通过调用 set\_defaults() 方法指定名字和值对, 然而有时候通过给参数指定一个简单的 parser-wide 是有用的, 这可以通过传递 argument\_default=关键字到 ArgumentParser, 例如调用其全局抑制属性在 parse\_args() 调用, 我们用 argument\_default=SUPPRESS:

```
1 import argparse
2 parser = argparse.ArgumentParser(argument_default=argparse.SUPPRESS)
3 parser.add_argument('--foo')
4 parser.add_argument('bar', nargs='?')
5 parser.parse_args(['--foo', '1', 'BAR'])
6 print(parser.parse_args([]))
```

#### allow\_abbrev

通常我们传递一个参数 liebhiao 给 ArgumentParser 的方法 parse\_args(), 如果选项参数太长的话。特征展示可能通过设置 allow\_abbrev 设置为 False 被禁用。

```
1 import argparse
2 parser = argparse.ArgumentParser(prog='Prog', allow_abbrev=False)
3 parser.add_argument('--foobar', action='store_true')
4 parser.add_argument('--fooley', action='store_true')
5 parser.parse_args(['--foon'])
```

```
hpc@hpc-322:~/TensorFlow_Notebook/code$ python demo14.py
usage: Prog [-h] [--foobar] [--fooley]
Prog: error: unrecognized arguments: --foon
```

#### conflict\_handler

ArgumentParser 对象不允许相同的选项字符串有两个行为, 默认情况下当已经一偶选项字符串使用时尝试穿件一个新的参数 ArgumentParser 对象将报出异常。

```
In [1]: import argparse
In [2]: parser = argparse.ArgumentParser(prog='PROG')
In [3]: parser.add_argument('-f', '--foo', help='old foo help')
Out[3]: _StoreAction(option_strings=['-f', '--foo'], dest='foo', nargs=None, const=None, default=None, type=None, choices=None, help='old foo help', metavar=None)
In [4]: parser.add_argument('--foo', help='new foo help')
      File "<ipython-input-4-b0dbd0131b6e>", line 1
          parser.add_argument('--foo', help='new foo help')
                                         ^
SyntaxError: invalid syntax
```

有时候覆

盖掉就得参数时有用的, 为了得到参数的行为值'resolvce' 可能被应用在 conflict\_handler=参数。

```
1 import argparse
2 parser = argparse.ArgumentParser(prog='PROG', conflict_handler='resolve')
3 parser.add_argument('-f', '--foo', help='old foo help')
4 parser.add_argument('--foo', help='new foo help')
```

```
5 parser.print_help()
```

```
usage: PROG [-h] [-f FOO] [--foo FOO]

optional arguments:
  -h, --help  show this help message and exit
  -f FOO      old foo help
  --foo FOO   new foo help
```

如果所有的选项字符串被覆盖，ArgumentParser 对象仅仅移除一个行为，因此上面的例子中，就得行为-f/-foo 行为保留-f 行为，因为仅仅-foo 选项字符串被覆盖。add\_help

默认情况下 ArgumentParserdurian 增加帮助信息到显示的消息中，例如：

```
1 import argparse
2 parser = argparse.ArgumentParser(description='Process some integers.')
3 parser.add_argument('integers', metavar='N', type=int, nargs='+', help='an integer
                     for the accumulator')
4 parser.add_argument('--sum', dest='accumulate', action='store_const', const=sum,
                     default=max, help='sum the integers(
                     default:find the max)')
5 args = parser.parse_args()
6 print(args.accumulate(args.integers))
```

```
usage: demo1.py [-h] [--sum] N [N ...]

Process some integers.

positional arguments:
  N          an integer for the accumulator

optional arguments:
  -h, --help  show this help message and exit
  --sum       sum the integers(default:find the max)
```

```
1 import argparse
2 parser = argparse.ArgumentParser(description='Process some integers.', add_help=
                                 False)
3 parser.add_argument('integers', metavar='N', type=int, nargs='+', help='an integer
                     for the accumulator')
4 parser.add_argument('--sum', dest='accumulate', action='store_const', const=sum,
                     default=max, help='sum the integers(
                     default:find the max)')
5 args = parser.parse_args()
6 print(args.accumulate(args.integers))
```

```
usage: demo1.py [--sum] N [N ...]
demo1.py: error: the following arguments are required: N
```

### 6.1.3 add\_argument() 方法

ArgumentParser.add\_argument(name or flags..., action][, nargs][, const][, default][, type][, choices][, required][, help][, metavar][, dest]) 定一个一个命令行参数应该被如何解

析，每一个参数自己有自己的详细描述，如下：

- name or flags: 名字或者选项字符串， foo 或者 (-f,--foo)。
- action: 参数出现在命令行后采取的基本的行为。
- nargs: 命令行参数应该被使用的参数的数量。
- const:action 和 nargs 选项要求的常数值。
- default: 缺乏参数的默认值。
- type: 传递参数读取的数据类型。
- choices: 参数的允许值的容器。
- required: 是否命令行选项被忽略。
- help: 简易的参数说明。
- metavar: 在 usage 消息的名字。
- dest: 增加到 parse\_args() 返回对象的属性的名字。

name 或者 flags

当 parse\_args() 被调用的时候。选项参数通过-前缀识别。

```

1 import argparse
2 parser = argparse.ArgumentParser(prog='PROG')
3 parser.add_argument('-f', '--foo')
4 parser.add_argument('bar')
5 print(parser.parse_args(['BAR']))
6 print(parser.parse_args(['BAR', '--foo', 'FOO']))

```

```

hpc@hpc-322:~/TensorFlow_Notebook/code$ python demo16.py
Namespace(bar='BAR', foo=None)
Namespace(bar='BAR', foo='FOO')

```

action

- 'store': 仅仅保存参数的值，例如

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo')
3 parser.parse_args('--foo 1'.split())

```

输出 Namespace(foo='1')

- 'store\_true': 存储 const 参数指定的值，'store\_const' 行为通常用于指定一些 flag。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', action='store_const', const=42)
3 parser.add_argument('--foo')

```

输出:Namespace(foo=42)

- 'store\_true' 和 'store\_false' 指定 'store\_const'。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', action='store_true')
3 parser.add_argument('--bar', action='store_false')
4 parser.add_argument('--baz', action='store_false')
5 parser.parse_args('--foo --bar'.split())

```

输出:Namespace(foo=True, bar=False, baz=True)

- 'append': 一个存储列表，添加每个参数值到列表中，允许选项被多次指定时很有用。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--str', dest='types', action='append_const', const=str)
3 parser.add_argument('--int', dest='types', action='append_const', const=int)
4 parser.parse_args('--str --int'.split())

```

输出:Namespace(type=[<class 'str'>, <class 'int'>])

- 'count': 关键参数出现的次数。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--verbose', '-v', action='count')
3 parser.parse_args(['-vvv'])

```

输出:Namespace(verbose=3)

- help: 打印当前 parser 所有选项的帮助信息， 默认帮助行为被添加到 parser。
- version: add\_argument 调用指定 version= 关键字

```

1 import argparse
2 parser = argparse.ArgumentParser(prog='PROG')
3 parser.add_argument('--version', action='version', version='(%prog) 2.0')
4 parser.parse_args(['--version'])

```

输出 PROG 2.0。

- 你可以通过传递行为子类或者其它对象的接口传递给 action，推荐的方法是扩展 Action，覆盖掉 \_\_call\_\_ 方法和 \_\_init\_\_。

```

1 class FooAction(argparse.Action):
2     def __init__(self, option_strings, dest, nargs=None, **kwargs):
3         if nargs is not None:
4             raise ValueError("nargs not allowed")
5     def __call__(self, parser, namespace, values, option_string=None):
6         print('%r %r %r' % (namespace, values, option_string))
7         setattr(namespace, self.dest, values)
8     parser = argparse.ArgumentParser()
9     parser.add_argument('--foo', action=FooAction)
10    parser.add_argumentParser('bar', action=FooAction)
11    args = parser.parse_args('1 -- foo 2'.split())

```

输出：

Namespace(bar=None,foo=None) '1' None

Namespace(bar=1,foo=None) '2' '--foo'

nargs

- N: 一个整数，命令行下的参数被放到一起成为一个列表：

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', nargs=2)
3 parser.add_argument('bar', nargs=1)
4 parser.parse_args('c --foo a b'.split())

```

输出:Namespace(bar=['c'],foo=['a','b'])

- ?：根据不同情况生成不同的值，如果没有参数指定它的值来自默认生成如果有一个带有-前缀的参数值将被 const 参数生成，如果指定了值将生成指定值。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', nargs='?', const='c', default='d')
3 parser.add_argument('bar', nargs='?', default='d')
4 parser.parse_args(['XX', '--foo', 'YY'])
5 parser.parse_args(['XX', '--foo'])
6 parser.parse_args([])

```

分别输出：

Namespace(bar='XX',foo='YY')

Namespace(bar='XX',foo='x')

Namespace(bar='d',foo='d')

用 nargs='?' 更常用的用法时允许选项输入输出文件:

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('infile',nargs='?',type=argparse.FileType('r'),default
                     =sys.stdin)
3 parser.add_argument('outfile',nargs='?',type=argparse.FileType('w'),
                     default=sys.stdout)
4 parser.parse_args(['input.txt','output.txt'])

```

输出:Namespace(infile=<\_io.TextIOWrapper name='input.txt',encoding='UTF-8'>, outfile=<\_io.TextIOWrapper name='output.txt' encoding='UTF-8'>) parser.parse\_args([])

输出: Namespace(infile=<io.TextIOWrapper name='<stdin>' encoding='UTF-8'>, outfile=<\_io.TextIOWrapper name='<stdout>' encoding='UTF-8'>)

- \*: 所有的命令行参数将被放到一个列表中。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo',nargs='*')
3 parser.add_argument('--bar',nargs='*')
4 parser.add_argument('--barz',nargs='*')
5 parser.parse_args('a b --foo x y --bar 1 2'.split())

```

输出:Namespace(bar=['1','2'],baz=['a','b'],foo=['x','y'])

- +: 所有的命令行参数将被添加到一个列表中, 至少需要一个参数否则将报错。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('foo',nargs='+')
3 parser.parse_args(['a','b'])
4 parser.parse_args([])

```

输出:Namespace(foo=['a',nargs='+'])

usage: PROG [-h] foo [foo ...]

PROG: error: too few arguments

- argparse.REMAINDER: 所有已经存在的参数被添加到一个列表。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('--foo')
3 parser.add_argument('command')
4 parser.add_argument('args',nargs=argparse.REMAINDER)
5 print(parser.parse_args('--foo B cmd --arg1 xx zz'.split()))

```

输出:Namespace(args=['-arg1','XX','ZZ'],command='cmd',foo='B') 如果 nargs 参数没有提供, argument 由 action 决定, 通常这意味着一个的命令行参数被使用一个项目被产生。

#### const

const 参数被用在保存没有被命令行读入的常数来常数值, 两个常见的用法如下:

- 当 add\_argument() 调用的时候设置了 action='store\_const' 或者是 action='append\_const' 通过增加 const 值到一个 parse\_args() 返回的对象的属性。
- 当 add\_argument() 通过选项字符串 (像-f 或者-foo) 和 nargs='?' , 这将穿件一个由 0 行或者一行参数跟着的选项, 当解析命令行时, 如果选项字符串遇到没有命令行参数的时候, 值 const 将被用来替代。'store\_const' 和'append\_const' 行为, const 关键字参数必须给定, 对于其它行为, 默认为 None。

#### default

所有的参数和一些位置的参数在命令行下可能被忽略, add\_argument() 参数 default 的值默认为 None, 指定当没有参数时什么值被使用。没有指定选项字符串, default 的值将取代参数。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', default=42)
3 parser.parse_args(['--foo', '2'])
4 parser.parse_args([])
```

输出: Namespace(foo='2')

Namespace(foo=42)

如果默认值是一个字符串, parser 解析值就好象命令行参数一样, 类似的, parser 应用任何 type 转换参数, 如果在设置属性值前 Namespace 返回值, 否则 parser 用下面的值。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--length', default=42, type=int)
3 parser.add_argument('--width', default=10.5, type=float)
4 parser.parse_args()
```

输出:Namespace(length=10, width=10.5)

对于参数为'?' 或者'\*', 命令行没有值的时候 default 值将被使用

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('foo', nargs='?', default=42)
3 parser.parse_args(['a'])
4 parser.parse_args([])
```

分别输出:

```
Namespace(foo='a')
```

```
Namespace(foo=42)
```

如果 default=argparse.SUPPRESS 如果没有命令行参数将导致没有属性被添加。

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', default=argparse.SUPPRESS)
3 parser.parse_args([])
4 parser.parse_args(['--foo', '1'])
```

分别输出:

```
Namespace()
```

```
Namespace(foo='1')
```

type

默认 ArgumentParser 对象读命令行参数为字符串，然而，经常命令行应该以另一种数据类型解析，像 float, int, add\_argument() 的 type 关键字允许需要的类型检查和转换被执行，常用的内部数据类型和参数可以被作为 type 的值直接使用。

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('foo', type=int)
3 parser.add_argument('bar', type=open)
4 parser.parse_args('2 temp.txt'.split())
```

输出:Namespace(bar=<\_io.TextIOWrapper name='temp.txt' encoding='UTF-8',foo=2) 为了能轻松的使用多种文件类型,argparse 模块提供了工厂 FileType,利用 mode=,bufsize=,encoding= 和 error= 参数，例如 FileType('w') 可以被用来创建一个可写的文件。

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('bar', type=argparse.FileType('w'))
3 parser.parse_args(['output'])
```

输出:Namespace(bar=<\_io.TextIOWrapper name='out.txt' encoding='UTF-8';>) type 能够调用一个字符串参数返回转换过值的参数

```
1 import math
2 import argparse
3 def perfect_square(string):
4     value = int(string)
5     sqrt = math.sqrt(value)
6     if sqrt != int(sqrt):
7         msg = '%r is not a perfect square' % string
8         raise argparse.ArgumentTypeError(msg)
9     return value
```

```

10 parser = argparse.ArgumentParser(prog='PROG')
11 parser.add_argument('foo', type=perfect_square)
12 print(parser.parse_args(['9']))
13 print(parser.parse_args(['7']))

```

输出: Namespace(foo=9)

usage: PROG [-h] foo

PROG: error: argument foo: '7' is not a perfect square

choise

choise 参数在检查值的范围时很方便。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('foo', type=int, choices=range(5, 10))
3 parser.parse_args(['7'])
4 parser.parse_args(['11'])

```

分别输出:Namespace(foo=7)

usage: PROG [-h] 5,6,7,8,9

PROG: error: argument foo: invalid choice: 11 (choose from 5, 6, 7, 8, 9)

choise

一些命令行参数从一些限定值的中选定,可以通过传递 choice 关键字参数给 add\_argument(),当命令行解析的时候,值将被检查如果不在可接受值范围内将显示错误消息。

```

1 parser = argparse.ArgumentParser(prog='game.py')
2 parser.add_argument('move', choices=['rock', 'paper', 'scissors'])
3 parser.parse_args(['rock'])
4 parser.parse_args(['file'])

```

分别输出:

Namespace(move='rock')

usage: game.py [-h] rock,paper,scissors

game.py: error: argument move: invalid choice: 'fire' (choose from 'rock', 'paper', 'scissors')  
choice 选项检查在转化数据类型后进行。

```

1 parser = argparse.ArgumentParser(prog='doors.py')
2 parser.add_argument('door', type=int, choices=range(1, 4))
3 print(parser.parse_args(['3']))
4 print(parser.parse_args(['4']))

```

分别输出:

Namespace(door=3)

usage: doors.py [-h] 1,2,3

doors.py: error: argument door: invalid choice: 4 (choose from 1, 2, 3)

任何支持 in 操作的对象都能被传递给 choise 作为值，因此 dict, set 对象都是常用的的支持的对象。required

通常 argparse 模块假设 flag 像可以被省略的-f 和–bar,, 为了一个选项必需要需要设置 required=True。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', required=True)
3 parser.parse_args(['--foo', 'BAR'])
4 parser.parse_args([])

```

分别输出:

Namespace(foo='BAR')

usage: argparse.py [-h] [-foo FOO]

argparse.py: error: option -foo is required

正如上例，如果 parse\_args() 的 required 被标记，如果不给值将报错。help

help 值包含一些简单的参数说明，当用户要求帮助的时候（通常用-h 或者–help）, help 描述信息将被展示

```

1 parser = argparse.ArgumentParser(prog='frobble')
2 parser.add_argument('--foo', action='store_true', help='foo the bars before
                     frobbing')
3 parser.add_argument('bar', nargs='+', help='foo the bars before frobbed')
4 parser.parse_args(['-h'])

```

输出:

usage: frobble [-h] [-foo] bar [bar ...]

positional arguments:

bar one of the bars to be frobbled

optional arguments:

-h, –help show this help message and exit

–foo foo the bars before frobbing

help 字符串能包含多种格式像程序名字或者默认参数, 可用的指定包含程序的名字,%(prog)s 和多数 add\_argument() 关键字，像%(default)s,%(type)s 等等。

```

1 parser = argparse.ArgumentParser(prog='frobble')
2 parser.add_argument('bar', nargs='?', type=int, default=42, help='the bar to %(prog)
                     s(default:%(default)s)')
3 parser.print_help()

```

输出:

```
usage: frobble [-h] [bar]
```

optional arguments:

```
bar the bar to frobble (default: 42)
```

optional arguments:

```
-h, --help show this help message and exit
```

帮助字符串支持% 格式，如果你想一个% 出现在帮助字符串中，你需要使用%% argparse 对于指定的选项通过设置 argparse.SUPPRESS 设置支持静默帮助。

```
1 parser = argparse.ArgumentParser(prog='frobble')
2 parser.add_argument('--foo', help=argparse.SUPPRESS)
3 parser.print_help()
```

输出:

```
usage: frobble [-h]
```

optional arguments:

```
-h, --help show this help message and exit
```

metavar

当 ArgumentParser 生成帮助消息的时候需要一些方法设计查询每个参数， 默认， ArgumentParser 对象用 dest 值作为每个对象的名字， 默认对于 action 位置的参数， dest 值被直接使用，对于一些选项行为， dest 值时大写的。因此单个位置参数 dest='bar' 将被认做 bar， --foo 应该被跟着一个命令作为 FOO

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo')
3 parser.add_argument('bar')
4 parser.parse_args('X --foo Y'.split())
5 print .print_help()
```

分别输出:

```
Namespace(bar='X', foo='Y')
```

```
usage: [-h] [-foo FOO] bar
```

optional arguments:

bar

optional arguments:

-h, --help show this help message and exit  
 -foo FOO

一个可用的名字被 metavar 指定:

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', metavar='YYY')
3 parser.add_argument('bar', metavar='XXX')
4 parser.parse_args('X -- foo Y'.split())
5 parser.print_help()
```

Namespace(abr='X',foo='Y')  
 usage: [-h] [-foo YYY] XXX

positional arguments:

XXX

optional arguments:

-h, --help show this help message and exit  
 -foo YYY

注意 metavar 仅仅改变显示的名字, parse\_args() 属性的名字仍然由 dest 值决定。不同的 nargs 也许导致 metavar 被多次使用, 提供一个元组给 metavar 指定一个不同的显示。

```

1 parser = argparse.ArgumentParser(prog='prog')
2 parser.add_argument('-x', nargs=2)
3 parser.add_argument('--foo', nargs=2, metavar=('bar', 'baz'))
4 parser.print_help()
```

输出:

usage: PROG [-h] [-x X X] [--foo bar baz]

optional arguments:

-h, --help show this help message and exit  
 -x X X  
 --foo bar baz  
 dest

大多数 ArgumentParser 行为增加一些值作为 parser\_args() 返回值的属性。属性的名字由 dest 决定

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('bar')
3 parser.parse_args(['xxx'])

```

输出:Namespace(bar='xxx')

对于选项参数, dest 的值从选项字符串推断出, ArgumentParser 通过得到长的选项字符串删除初始化-字符串生成 dest 的值, 如果 meiyou 长的选项字符串提供, dest 将通过初始化字符-从第一个短的字符串选项得到。任何内部-字符将被转换为 \_ 字符确保字符串是一个可用的属性名字。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('-f', '--foo-bar', '--foo')
3 parser.add_argument('-x', '-y')
4 parser.parse_args ['-f 1 -x 2'.split()]
5 parser.parse_args('--foo 1 -y 2'.split())

```

分别输出:

Namespace(foo\_bar=1,x='2')

Namespace(foo\_bar='1',x='2')

dest 允许自定义属性的名字:

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', dest='bar')
3 parser.parse_args('--foo XXX'.split())

```

输出:Namespace(bar='XXX')

Action class

Action classes 实现的 Action API, 一个命令行返回的可调的 API。任何这个 API 对象都可以被 zuoweiaction 参数传递给 add\_argument().class argparse.Action(option\_strings, dest, nargs=None, const=None, default=None, type=None, choices=None, required=False, help=None, metavar=None) Action 实力应该是可调用的, 因此子类必须被 \_\_call\_\_ 方法覆盖, 应该接受四个参数:

- parser: 包含这个 action 的 ArgumentParser。
- namespace:parser\_args() 返回的 Namespace 对象, 大多数行为通过 setattr() 增加一个属性到对象。
- varlue: 结合命令行参数和任何转化应用, 类型转换被 type 关键字指定。
- option\_string: 宣告像字符串被用于激活这个 action, option\_string 时一个选项, 将

缺席如果这个 action 和 positional 参数结合。`__call__` 方法也许执行任意行为，但是典型的设置基于 dest 和 value 的 namespace 属性。

`parse_args()` 方法:

`ArgumentParser.parse_args(args=None, namespace=None)` 转换参数字符串为对象指定他们作为 namespace 的属性。之前调用 `add_argument()` 决定决定创建什么对象如何复制，默认 argument 字符串来自 `sys.argv`，一个新的空的 Namespace 对象被创建。Option value syntax

`parse_args` 方法支持多种方法指定选项的值，在最简单的情况下，这个选项和它的值被传递作为两个分开的参数:

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('-x')
3 parser.add_argument('--foo')
4 parser.parse_argument('-x', 'X')
5 parser.parse_args('--foo', 'FOO')
```

分别输出:

`Namespace(foo=None,x='X')`

`Namespace(foo='FOO',x=None)`

对于短的选项，这个选项值可以被链接，多个短选项可以被-前缀连接在一起，只要最新的选项（非空）要求值:

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('-x', action='store_true')
3 parser.add_argument('-y', action='store_true')
4 parser.add_argument('-z')
5 parser.parse_args(['-xyzZ'])
```

输出:`Namespace(x=True,y=True,z='Z')` 不可用的参数

当解析命令行时 `parse_args()` 检查多种错误，包括不明确的选项，不可用的类型，错误的参数为值等等，当出现一个错误，它推出同时打印错误和用法信息。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('--foo', type=int)
3 parser.add_argument('bar', nargs='?')
4
5 # invalid type
6 parser.parse_args(['--foo', 'spam'])
7 usage: PROG [-h] [--foo FOO] [bar]
8 PROG: error: argument --foo: invalid int value: 'spam'
9
10 # invalid option
```

```

11 parser.parse_args(['--bar'])
12 usage: PROG [-h] [--foo FOO] [bar]
13 PROG: error: no such option: --bar
14
15 # wrong number of arguments
16 parser.parse_args(['spam', 'badger'])
17 usage: PROG [-h] [--foo FOO] [bar]
18 PROG: error: extra arguments found: badger

```

## 参数包含

当用户犯错时 `parse_args()` 方法尝试给出错误, 但是一些情况下固有的二义, 例如, 命令行参数-1 可能同时指定一个选项或者尝试提供一个指定位置参数, `parse_args()` 方法导致, 指定位置的参数仅仅用-开始如果他们看起来像负数在 `parser` 没有选像解析看起来像负数:

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('-x')
3 parser.add_argument('foo', nargs='?')
4
5 # no negative number options, so -1 is a positional argument
6 parser.parse_args(['-x', '-1'])
7 Namespace(foo=None, x='-1')
8
9 # no negative number options, so -1 and -5 are positional arguments
10 parser.parse_args(['-x', '-1', '-5'])
11 Namespace(foo='-5', x='-1')
12
13 parser = argparse.ArgumentParser(prog='PROG')
14 parser.add_argument('-1', dest='one')
15 parser.add_argument('foo', nargs='?')
16
17 # negative number options present, so -1 is an option
18 parser.parse_args(['-1', 'X'])
19 Namespace(foo=None, one='X')
20
21 # negative number options present, so -2 is an option
22 parser.parse_args(['-2'])
23 usage: PROG [-h] [-1 ONE] [foo]
24 PROG: error: no such option: -2
25
26 # negative number options present, so both -1s are options
27 parser.parse_args(['-1', '-1'])
28 usage: PROG [-h] [-1 ONE] [foo]

```

```
29 PROG: error: argument -1: expected one argument
```

如果你有一个必须以-开始的参数而且不是负数，你可以插入'-'告诉 parse\_args() 之后的一切：

```
1 parser.parse_args(['--', '-f'])
```

输出:Namespace(foo='f',one=None) 参数缩略 如果缩略没有歧义 parser\_args() 方法默认允许长选项被简写为前缀。

```
1 parser = argparse.ArgumentParser(prog='PROG')
2 parser.add_argument('-bacon')
3 parser.add_argument('-badger')
4 parser.parse_args ['-bac MMM'.split()]
5 Namespace(bacon='MMM', badger=None)
6 parser.parse_args ['-bad WOOD'.split()]
7 Namespace(bacon=None, badger='WOOD')
8 parser.parse_args ['-ba BA'.split()]
9 usage: PROG [-h] [-bacon BACON] [-badger BADGER]
10 PROG: error: ambiguous option: -ba could match -badger, -bacon
```

可能产生多个选项时错误产生，可以通过设置 allow\_abbrev 设置为 False 禁用。Beyond sys.argv

ArgumentParser 通常比 sys.argv 有用，可以穿地一个字符串列表到 parser\_args() 完成，这在测试交互式提示符很有用。

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('integers', metavar='int', type=int, choices=range(100),
3 nargs='+', help='an integer in range 0..9')
4 parser.add_argument('--sum', dest='accumulate', action='store_const',
5 const=sum, default=max, help='sum the integers (default: find the max)')
6 parser.parse_args(['1','2','3','4'])
7 parser.parse_args(['1','2','3','4'], '--sum')
```

输出结果分别为：

Namespace(accumulate=<built-in function max>,integers=[1,2,3,4])

Namespace(accumulate=<built-in function sum>,integers=[1,2,3,4])

Namespace 对象

class argparse.Namespace，简单的 parse\_args() 创建一个对象，保存属性返回它。这个类很简单，仅仅是一个可读表达的对象子类，如果你希望有字典类似的属性，你可以用标准的 python idiom()：

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo')
```

```

3 args = parser.parse_args(['--foo', 'BAR'])
4 var(args)

```

输出:'foo':'BAR'

当 ArgumentParser 指定属性到已经存在的对象时它是很有用的，相比于新的 Namespaced 对象，它可以指定 namespace= 关键参数获得。

```

1 class C:
2     pass
3 C = C()
4 parser = argparse.ArgumentParser()
5 parser.add_argument('--foo')
6 parser.parse_args(args=['--foo', 'BAR'], namespace=C)
7 c.foo

```

输出:'BAR'

子命令: ArgumentParser.add\_subparsers([title][, description][, prog][, parser\_class][, action][, option\_string][, dest][, help][, metavar]) 一些程序分割他们的功能为一个子命令，例如， svn 程序可以有子命令 svn checkout,svn commit,svn update。当程序有一些要求不同类型命令行参数的不同的功能的时候分割功能的方法是一个好的想法，ArgumentParser 支持支持 add\_subparsers 一个子命令，add\_subparsers() 方法调用通常没有参数返回一个特殊的行为对象，这个对象是一个方法，add\_parser() 得到一个命令名字和任何 ArgumentParser 够草体参数，返回一个可以被修改的 ArgumentParser 对象。

- title: 帮助输出 sub-parser 组的标题，如果说明提供了的话默认”subcommands”，否则用参数作为标题。
- description: 在输出帮助中描述 sub-parser 组，默认是 None。
- prog:sub-command 的帮助信息，默认程序的名字和位置上的参数在 subparser 参数前。
- parser\_class: 用于创建一个 sub-parser 实例的类，默认时当 parser。
- action: 在命令行中参数出现厚的基础类型的行为。
- dest:sub-command 下属性的名字将被存储，默认没有值被存储。
- help: 在帮助输出的 sub-parser, 默认为 None。
- metavar: 在 help 中可用的子命令默认是 None 代表子命令 cmd1,cmd2,...

用法:

```

1 # create the top-level parser
2 parser = argparse.ArgumentParser(prog='PROG')
3 parser.add_argument('--foo', action='store_true', help='foo help')
4 subparsers = parser.add_subparsers(help='sub-command help')
5
6 # create the parser for the "a" command
7 parser_a = subparsers.add_parser('a', help='a help')
8 parser_a.add_argument('bar', type=int, help='bar help')
9
10 # create the parser for the "b" command
11 parser_b = subparsers.add_parser('b', help='b help')
12 parser_b.add_argument('--baz', choices='XYZ', help='baz help')
13
14 # parse some argument lists
15 parser.parse_args(['a', '12'])
16 Namespace(bar=12, foo=False)
17 parser.parse_args(['--foo', 'b', '--baz', 'Z'])
18 Namespace(baz='Z', foo=True)

```

注意 parser\_args() 返回的 durian 将包含住 parser 和 subparser 命令行选中的参数，因此在上面的例子中，当一个命令被指定，仅仅 foo 和 bar 被呈现，当 b 被指定，仅仅 foo 和 baz 属性被呈现，类似的，subparser 要求帮助信息，仅仅这个 parser 的帮助信息被打印，帮助信息不包含父或者兄弟 parser 信息。(一个 subparser 命令的帮助消息，然而，可以被 help= 参数增加到上面)

```

1 parser.parse_args(['--help'])
2 usage: PROG [-h] [--foo] {a,b} ...
3
4 positional arguments:
5   {a,b}    sub-command help
6     a      a help
7     b      b help
8
9 optional arguments:
10  -h, --help  show this help message and exit
11  --foo    foo help
12
13 parser.parse_args(['a', '--help'])
14 usage: PROG a [-h] bar
15
16 positional arguments:
17   bar      bar help
18

```

```

19 optional arguments:
20   -h, --help    show this help message and exit
21
22 parser.parse_args(['b', '--help'])
23 usage: PROG b [-h] [--baz {X,Y,Z}]
24
25 optional arguments:
26   -h, --help      show this help message and exit
27   --baz {X,Y,Z}  baz help

```

add\_subparsers() 方法也支持 title 和 description 关键参数, 当两者都呈现的时候在帮助输出 subparser 的命令将出现在自己的组。

```

1 >>> parser = argparse.ArgumentParser()
2 >>> subparsers = parser.add_subparsers(title='subcommands',
3 ...                                         description='valid subcommands',
4 ...                                         help='additional help')
5 >>> subparsers.add_parser('foo')
6 >>> subparsers.add_parser('bar')
7 >>> parser.parse_args(['-h'])
8 usage: [-h] {foo,bar} ...
9
10 optional arguments:
11   -h, --help    show this help message and exit
12
13 subcommands:
14   valid subcommands
15
16   {foo,bar}    additional help

```

更进一步, add\_parser 支持一个 aliases 参数, 允许多字符串访问同一个 subparser, 像 svm, 别名 co 作为 checkout 的简写。

```

1 >>> parser = argparse.ArgumentParser()
2 >>> subparsers = parser.add_subparsers()
3 >>> checkout = subparsers.add_parser('checkout', aliases=['co'])
4 >>> checkout.add_argument('foo')
5 >>> parser.parse_args(['co', 'bar'])
6 Namespace(foo='bar')

```

一个类似的高效处理 sub-commands 结合 add\_subparsers() 方法调用 set\_default() 以至于每个 subparser 知道那个 python 函数应该被执行。

```

1 >>> # sub-command functions
2 >>> def foo(args):

```

```

3 ...     print(args.x * args.y)
4 ...
5 >>> def bar(args):
6 ...     print('((%s))' % args.z)
7 ...
8 >>> # create the top-level parser
9 >>> parser = argparse.ArgumentParser()
10 >>> subparsers = parser.add_subparsers()
11 >>>
12 >>> # create the parser for the "foo" command
13 >>> parser_foo = subparsers.add_parser('foo')
14 >>> parser_foo.add_argument('-x', type=int, default=1)
15 >>> parser_foo.add_argument('y', type=float)
16 >>> parser_foo.set_defaults(func=foo)
17 >>>
18 >>> # create the parser for the "bar" command
19 >>> parser_bar = subparsers.add_parser('bar')
20 >>> parser_bar.add_argument('z')
21 >>> parser_bar.set_defaults(func=bar)
22 >>>
23 >>> # parse the args and call whatever function was selected
24 >>> args = parser.parse_args('foo 1 -x 2'.split())
25 >>> args.func(args)
26 2.0
27 >>>
28 >>> # parse the args and call whatever function was selected
29 >>> args = parser.parse_args('bar XYZYX'.split())
30 >>> args.func(args)
31 ((XYZYX))

```

你可以用 `parse_args()` 在参数解析完成后通过调用合适的函数做这个工作，结合函数和 `action` 像这个像这样典型的轻松的方法处理不同的行为，然而，如果它需要检查 `subparser` 的名字，`dest` 关键值通过 `add_subparsers()` 调用将发挥作用。

```

1 >>> parser = argparse.ArgumentParser()
2 >>> subparsers = parser.add_subparsers(dest='subparser_name')
3 >>> subparser1 = subparsers.add_parser('1')
4 >>> subparser1.add_argument('-x')
5 >>> subparser2 = subparsers.add_parser('2')
6 >>> subparser2.add_argument('y')
7 >>> parser.parse_args(['2', 'frobble'])
8 Namespace(subparser_name='2', y='frobble')

```

FileType 对象：

class argparse.FileType(mode='r', bufsize=-1, encoding=None, errors=None) FileType 工厂创建一个能被传递给 ArgumentParser.add\_argument() 的对象。参数有 FileType 对象将用要求的模式打开命令行参数作为文件，换从大小，编码，错误处理。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--raw', type=argparse.FileType('wb', 0))
3 parser.add_argument('out', mtype=argparse.FileType('w', encoding='UTF-8'))
4 parser.parse_args(['--raw', 'raw.dat', 'file.txt'])

```

输出: Namespace(out=<\_io.TextIOWrapper name='file.txt' mode='w' encoding='UTF-8', raw=<\_ioFileIO name='raw.dat' mode='wb'>)

FileType 对象明白伪参数同时自动转换 sys.stdin 为可读的 FileType 对象， sys.stdout 可写的 FileType 对象。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('infile', type=argparse.FileType('r'))
3 parser.parse_args(['-'])

```

输出 Namespace(infile=<\_io.TextIOWrapper name='<stdin>' encoding='UTF-8') Argument group

ArgumentParser.add\_argument\_group(title=None, description=None)

默认情况下, ArgumentParser groups, 当显示帮助信息的时候命令行参数进入对应位置的参数和选项参数。当有一个比默认更好的概念上的参数组, 合适的组能被 add\_argument\_group() 创建:

```

1 parser = argparse.ArgumentParser(prog='PROG', add_help=False)
2 group = parser.add_argument_group('group')
3 group.add_argument('--foo', help='foo help')
4 group.add_argument('bar', help='bar help')
5 parser.print_help()

```

输出: usage: PROG [-foo FOO] bar

group:

bar bar help

-foo FOO foo help

add\_argument\_group() 方法返回一个有 add\_argument() 方法的参数组对象。当一个参数增加到组中, parser 就当它为正常参数,但是在帮助信息中分组显示。add\_argument\_group() 方法接受 title 和 description 参数自定义显示:

```

1 parser = argparse.ArgumentParser(prog='PROG', add_help=False)
2 group1 = parser.add_argument_group('group1', 'group1 description')

```

```

3 group1.add_argument('foo', help='foo help')
4 group2 = parser.add_argument_group('group2', 'group2 description')
5 group2.add_argument('--bar', help='bar help')
6 parser.print_help()

```

usage: PROG [-bar BAR] foo

```

group1:
group1 description

foo foo help

```

```

group2:
group2 description

-bar BAR bar help

```

注意任何不再你的用户定义组中的参数将以对应位置参数和选项参数结束。Mutual exclusion

`ArgumentParser.add_mutually_exclusive_group(required=False)`

创建一个转悠的组，`argparse` 将确保唯一的参数在彼此的组被呈现在命令行。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 group = parser.add_mutually_exclusive_group()
3 group.add_argument('--foo', action='store_true')
4 group.add_argument('--bar', action='store_false')
5 parser.parse_args(['--foo'])
6 parser.parse_args(['--bar'])
7 parser.parse_args(['--foo', '--bar'])

```

分别输出：

`Namespace(bar=True,foo=True)`

`Namespace(bar=False,foo=False)`

`sage: PROG [-h] [-foo | -bar]`

`PROG: error: argument --bar: not allowed with argument --foo`

`add_mutually_exclusive_group()` 方法接受一个 `required` 参数，预示着最新的参数被要求。

```

1 parser = argparse.ArgumentParser(prog='PROG')
2 group = parser.add_mutually_exclusive_group(required=True)
3 group.add_argument('--foo', action='store_true')
4 group.add_argument('--bar', action='store_true')

```

```
5 parser.parse_args([])
```

输出:

```
usage: PROG [-h] (-foo | -bar)
```

```
PROG: error: one of the arguments -foo -bar is required
```

注意当前的 mutually exclusive 参数组不支持 title 和 description 参数。Parser defaults

```
ArgumentParser.set_defaults(**kwargs)
```

大多数时候,parse\_args() 返回的属性对象将被命令行参数和参数行为完全决定。set\_default()  
允许一些额外的属性决定没有命令行增加时的行为:

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('foo', type=int)
3 parser.set_default(bar=42, baz='badger')
4 parser.parse_args(['736'])
```

输出:Namespace(bar=42,baz='badger',fpp=736) 注意 parser 级默认覆盖参数级。

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', default='bar')
3 parser.set_defaults(foo='spam')
4 parser.parse_args([])
```

输出: Namespace(foo='spam') Parser 级别在多个 parser 时特别有用。ArgumentParser.get\_default(dest)  
得到 namespace 属性的默认值, 正如设置 add\_argument() 或者 set\_defaults()

```
1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', default='badger')
3 parser.get_default('foo')
```

输出:'baadger' Printing help

在一些典型的应用中 parse\_args() 将考虑打印用法和错误信息的格式, 然而一些格式  
方法是可用的: ArgumentParser.print\_usage(file=None): 打印 ArgumentParser 应该在命  
令行调用的简单描述, 如果 file 是 None, sys.stdout 被假定。ArgumentParser.print\_help(file=None):  
打印程序的用法信息和 ArgumentParser 参数注册信息, 如果 file 为 None, sys.stdout 被  
假定。ArgumentParser.format\_usage(): 返回在命令行中 ArgumentParser 参数应该被如何  
调用的简要说明字符串。ArgumentParser.format\_help(): 返回一个包含程序用法和 Ar-  
gumentParser 参数注册信息的帮助字符串。Partial parsing

```
ArgumentParser.parse_known_args(args=None, namespace=None)
```

有时候一些脚本也许仅仅解析一些命令行参数, 传递参数到另一个脚本或者程序, 在这种  
情形下, parser.\_known\_args() 方法很有用, 它像 parser\_args() 除了当有额外的参数呈现  
的时候不生成错误, 相反, 它返回一个包含 populated namespace 和保留参数字符串的列  
表的两个元素的元组。

```

1 parser = argparse.ArgumentParser()
2 parser.add_argument('--foo', action='store_true')
3 parser.add_argument('bar')
4 parser.parse_known_args(['--foo', '--badger', 'BAR', 'spam'])

```

输出:(Namespace(bar='BAR', foo=True), ['--badger', 'spam']) Customizing file parsing

ArgumentParser.convert\_arg\_line\_to\_args(arg\_line)

从文件中读入的参数一行读一个, convert\_arg\_line\_to\_args() 能被覆盖。这个方法从参数文件得到一个简单的 arg\_line 字符串, 返回一个参数列表, 每读取一行方法被调用一次。一个有用的覆盖每这个方法是当空格分开的 word 为参数, 下面的例子展示:

```

1 class NyArgumentParser(argparse.ArgumentParser):
2     def convert_tag_line_to_args(self, arg_line):
3         return arg_line.split()

```

Exiting method

ArgumentParser.exit(status=0, message=None): 这个方法终止程序, 以指定的状态推出, 如果参数被给, 打印消息。ArgumentParser.error(message): 这个方法打印包含消息用法信息到标准错误终止程序以状态代码 2。Upgrading optparse code

最初 argparse 模块尝试用 optparse 维持兼容性, 然而 optparse 很难扩展, 特别是改变要求支持新的 nargs= 指定更好的用法消息。当大多数 optparse 已经被复制粘贴过或者 monkey-patched, 它不再尝试维持向后兼容。, argparse 模块在一些方法改进了标准库 optparse:

- 处理位置参数。
- 支持子命令。
- 允许 + 和/前缀。
- 处理 0 或者更多 1 或者更多风格的参数。
- 处理更多的用法消息。
- 提供简单的接口自定义 type 和 action。

optparse 到 argparse 的并行升级

```

1 \item 用 ArgumentParser.add_argument() 调用取代 optparse.OptionParser.add_option()
       调用。
2 \item 用 args=parser.parser_args() 取代 (options, args)=parser.parse_args() 增加
       ArgumentParser.add_argument() 调用给指定
       位置的参数, 记住显现的前向, 现在在
       argparse 上下文称为 args。

```

```

3 \item 用 type 和 action 取代 callback 行为和 callback\_* 关键参数。
4 \item 取代 type 关键字的字符串名字和相关的对象类型（如 int, float, complex 等等）
5 \item 用 Namespace 和 optparse.OptionError, optparse.OptionValueError 取代 optparse.
          Value。
6 \item 用标准那得 Python 语法取代 \%default 或者 \%( default )s 和 \%( prog )s。
7 \item 通过调用 parser.add\_argument( '--version', action='version', version='<the
          version>' ) 取代 OptionParser 结构体 version
          。

```

setting 输入:

```

hpc@hpc322:~/文档/Tensorflow$ python code/arg1.py
usage: arg1.py [-h] echo
arg1.py: error: the following arguments are required: echo
hpc@hpc322:~/文档/Tensorflow$ python code/arg1.py -h
usage: arg1.py [-h] echo

positional arguments:
  echo

optional arguments:
  -h, --help  show this help message and exit
hpc@hpc322:~/文档/Tensorflow$ 

```

add\_argument 方法指定程序需要接受的命令参数，本例中为 echo，此程序运行必须指定一个参数，方法 parse\_args() 通过分析指定的参数返回数据 echo。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("echo", help="show the help information", type=int)
4 args = parser.parse_args()
5 print(args.echo**2)

```

指定参数类型为 int， 默认为 string。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("--verbosity", help="increase output verbosity")
4 args = parser.parse_args()
5 if args.verbosity:
6     print("Verbosity turned on")

```

```

Verbosity turned on
hpc@hpc322:~/文档/Tensorflow$ python code/arg3.py --verbosity a
Verbosity turned on

```

这里指定了--verbosity 程序就显示一些信息，如果不指定程序也不会出错，对应的变量就被设置为 None。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("--verbosity", help="increase output verbosity", action="store_true")

```

```

4 args = parser.parse_args()
5 if args.verbose:
6     print("Verbosity turned on")

```

指定一个新的关键词 action, 赋值为 store\_true。如果指定了可选参数, args.verbose 就赋值为 True, 否则就为 False。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("-v", "--verbose", help="Increase output verbosity", action="store_true")
4 args = parser.parse_args()
5 if args.verbose:
6     print("verbosity turned on")

```

```

hpc@hpc322:~/文档/Tensorflow$ python code/arg4.py --help
usage: arg4.py [-h] [-v]

optional arguments:
  -h, --help    show this help message and exit
  -v, --verbose Increase output verbosity

```

```

1 #args5.py
2 import argparse
3 parser = argparse.ArgumentParser()
4 parser.add_argument("square", type=int, help="display help information")
5 parser.add_argument("-v", "--verbose", action="store_true", help="increase output verbosity")
6 args = parser.parse_args()
7 answer = args.square**2
8 if args.verbose:
9     print("The square of {} equals {}".format(args.square, answer))
10 else:
11     print(answer)

```

输入参数 verbose 和整数 (4) 顺序不影响结果。python args5.py -verbose 4 和 python args5.py 4 -verbose

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("square", type=int, help="display a square of a given number")
4 parser.add_argument("-v", "--verbose", type=int, help="increase output verbosity")
5 args = parser.parse_args()
6 answer = args.square**2
7 if args.verbose == 2:

```

```

8     print("The square of {} equals {}".format(args.square, answer))
9 elif args.verbosity == 1:
10    print("{}^2=={}".format(args.square, answer))
11 else:
12    print(answer)

```

python args6.py 4 -v 0,1,2 通过指定不同的参数 v 为 0,1,2 得到不同的结果。

```

1 #arg7.py
2 import argparse
3 parser = argparse.ArgumentParser()
4 parser.add_argument("square", type=int, help="display the square of a given number")
5 parser.add_argument("-v", "--verbosity", action="count", help="increase output verbosity")
6 args = parser.parse_args()
7 answer = args.square**2
8 if args.verbosity == 2:
9     print("The square of {} equals {}".format(args.square, answer))
10 elif args.verbosity == 1:
11     print("{}^2 == {}".format(args.square, answer))
12 else:
13     print(answer)

```

这里添加参数 action="count", 统计可选参数出现的次数。python arg7.py 4 -v(出现一次), 对应结果为  $x^2 == 16$

python arg7.py 4 -vv(出现两次), 对应出现 The square of 4 equals 16

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("square", type=int, help="display a square of a given number")
4 parser.add_argument("-v", "--verbosity", action="count", default=0, help="increase output verbosity")
5 args = parser.parse_args()
6 answer = args.square**2
7 if args.verbosity>=2:
8     print("The square of {} equals {}".format(args.square, answer))
9 elif args.verbosity>=1:
10    print("{}^2 == {}".format(args.square, answer))
11 else:
12    print(answer)

```

加速让 default 参数。这默认为值 0, 当参数 v 不指定时参数就被置为 None, None 不能

和整型比较。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 parser.add_argument("x", type=int, help="The base")
4 parser.add_argument("y", type=int, help="The exponent")
5 parser.add_argument("-v", "--verbosity", action="count", default=0)
6 args = parser.parse_args()
7 answer = args.x**args.y
8 if args.verbosity >= 2:
9     print("{} to the power {} equals {}".format(args.x, args.y, answer))
10 elif args.verbosity >= 1:
11     print("{}^{} == {}".format(args.x, args.y, answer))
12 else:
13     print(answer)

```

为了让后面的参数不冲突，我们需要使用另一个方法：

```

#args10.py
1 import argparse
2 parser = argparse.ArgumentParser()
3 group = parser.add_mutually_exclusive_group()
4 parser.add_argument("-v", "--verbose", action="store_true")
5 group.add_argument("-q", "--quit", action="store_true")
6 parser.add_argument("x", type=int, help="The base")
7 parser.add_argument("y", type=int, help="The exponent")
8 args = parser.parse_args()
9 answer = args.x**args.y
10 if args.quit:
11     print(answer)
12 elif args.verbose:
13     print("{} to the power {} equals {}".format(args.x, args.y, answer))
14 else:
15     print("{}^{} == {}".format(args.x, args.y, answer))

```

可以输入 `python arg10.py 3 4 -vq` 得到计算结果。

```

1 import argparse
2 parser = argparse.ArgumentParser()
3 group = parser.add_mutually_exclusive_group()
4 group.add_argument("-v", "--verbose", action="store_true")
5 group.add_argument("-q", "--quit", action="store_true")
6 parser.add_argument("x", type=int, help="The base")
7 parser.add_argument("y", type=int, help="The exponent")
8 args = parser.parse_args()
9 answer = args.x**args.y

```

```
10 if args.quit:  
11     print(answer)  
12 elif args.verbose:  
13     print("{} to the power {} equals {}".format(args.x, args.y, answer))  
14 else:  
15     print("{}^{} == {}".format(args.x, args.y, answer))
```

这里参数 v 和 q 不能同时使用。

## 6.2 path

### 6.2.1 函数说明

- `os.path.abspath(path)`: 返回 path 的绝对路径, 在多数平台下, 相当于调用函数 `normpath(join(os.getcwd(),path))`
- `os.path.basename(path)`: 返回 path 的路径 base name, 第二个元素通过传递 path 给 `split()`, 注意这个结果不同于 unix 的 basename 程序, 这里 basename,'foo/bar' 然会 bar, 而 basename() 函数返回空字符串 ("")。
- `os.path.commonpath(paths)`: 返回 paths 队列中最长的 sub-path, 日国路径中包含绝对路径和相对路径的话将报 ValueError 或者如果 paths 是空, 不想 commonprefix(), 这个函数返回一个错的路径。
- `os.path.dirname(path)`: 返回目录的名字, 就是 path 用 split 分割厚的第一个元素。
- `os.path.exists(path)`: 如果春在路径 path 或者一个打开的文件描述返回 True。对于破掉的符号链接返回 False, 在一些平台, 如果权限不允许执行 `os.stat()` 即使存在物理路径这个函数也返回 False。
- `os.path.lexists(path)`: 如果路径存在返回 True, 对 broken 符号链接返回 True, 等效与 exists()。
- `os.path.expanduser(path)`: 在 Unix 和 Windows 上用 ~ 或者 user 取代用户路径的值。在 unix 上一个 被环境变量 HOME 替代 (如果设置了 HOME 环境变量的话), 否则当前用户的 home 目录通过内建模块 pwd 查找, 一个初始化 user 是寻找在 password 目录里面的目录。
- `os.path.expandvars(path)`: 返回环境变量的值, 子字符串形式时 `namename` 被环境变量名取代, 变形的变量名字和参考不存在的变量将不改变。
- `os.path.getatime(path)`: 返回上次访问路径的时间, 返回一个从 epoch 起经历的秒数, 如果文件不存在或不可访问则报 OSError。
- `os.path.getmtime(path)`: 返回最新修改路径的时间, 返回值时一个 epoch 其开始的秒数, 文件不存在或者不可范围跟时报 OSError。
- `os.path.getctime`: 返回系统的 ctime, 在 Unix 上时最新的 metadata 改变的时间, 在 windows 上时 path 创建的时间, 返回一个从 epoch 起经历的秒数, 如果文件不存在或不可访问则报 OSError。

- os.path.getsize(path): 返回字节表示的路径的大小，如果不存在文件或者文件不可范围跟将报出 OSError。
- os.path.isabs(path): 如果路径是绝对路径返回 True。
- os.path.isfile(path): 如果路径是文件将返回 True。
- os.path.isdir(path): 如果存在路径返回 True。
- os.path.islink(path): 如果路径查询一个目录入口时符号链接返回 True，如果 Python 运行时符号链接不支持将返回 False。
- os.path.ismount(path): 如果 path 是一个挂载点，返回 True。
- os.path.join(path,\*paths): 加入一个或者更多的组建，返回值是连接路径和任何成员的路径。
- os.path.normcase(path): 在 Unix, MAX OS 上返回路径不变，在一些敏感的文件系统上将转换路径为小写，在 windows 上将转化斜线为反斜线，如果 path 不是 str 或者 bytes 将报 TypeError。
- os.path.normpath(path): 删去冗余得分和服，因此 A//B,A/B,A./B,A/foo../B 将变为 A/B. 字符串操作也许改变包含符号链接的意义，在 windows 上它转化斜线为反斜线。
- os.path.realpath(path): 返回指定文件名的确定路径，消除路径中出现的任何符号链接。
- os.realpath(path,start=os.curdir): 从当前路径或者 start 路径返回相对的文件路径，这是一个路径计算：文件系统不妨问确定的存在的或者自然的路径或者 start。
- os.path.samefile(path1,path2): 如果 pathname 值访问相同的文件或者目录则返回 True，这有 device 名字和 i-node 数量决定，如果 os.stat() 调用 pathname 失败将报出异常。
- os.path.sameopenfile(fp1,fp2): 如果 fp1 和 fp2 指定的时相同的文件将返回 True。
- os.path.samestat(stat1,stat2): 如果元组 stat1 和 state2 查询的时相同的文件，返回 True，这个结构可需已经被 os.fstate(),os.lstat() 或者 os.stat() 返回，番薯通过 samefile() 和 sameopenfile() 实现基本的比较。
- os.path.split(path): 分割路径为 (head,tail)。tail 不包含斜线，如果以斜线将诶为，tail 将为空，如果没有斜线，头将为空，如果 path 时空，头尾都为空。后面的斜线从 head

删除出位它是 root(一个或者更多的斜线), 在所有的情况下 join(head,tail) 返回一个路径到相同位置作为路径。

- os.path.splitdrive(path): 返回 pathname 到 (drive,tail), 这里 drive 可以使挂载点或者空字符串。在系统上没有用驱动器指定, 驱动器将为空字符串, 在所有的倾向下, drive+tail 将时相同的路径。在 Windows 上, 分割 pathname 成 drive/UNC 共享点和相对路径, 如果路径包含驱动器驱动器将包含冒号 (splitdrive("c:/dir")) 返回 ("c:", "/dir"), 如果路径包含驱动 UNC 路径, 驱动器将包含主机名和 share, 但是不包含四个分隔符 splitdrive("//host/computer/dir")return("//host/computer","/dir")
- os.path.split(path): 分割路径名为 (root,ext) 像 root+ext == path, ext 时空或者以一个周期开头, 导致 basename 被忽略, splitext('.cshrc') 返回 ('.cshrc', '')
- os.path.supports\_unicode\_filenames(): 如果文件名时 unicode 编码的则为 True。

### 6.2.2 例子

#### 1. 获取文件名, 目录, 扩展, 新文路径。

```

1 import os
2 file_path = '~/iris_test.csv'
3 filename = os.path.basename(file_path)
4 new_dir = os.path.join('home', 'hpc', filename)
5 file_dir = os.path.dirname(file_path)
6 dir1 = '~/'
7 fulldir = os.path.expanduser(dir1)
8 sp = os.path.split(new_dir)
9 print('new_dir:', sp[0], 'ext:', sp[1])

```

#### 2. 查看文件同时打开文件

```

1 import os
2 path = '/etc'
3 filename = 'passwd'
4 if os.path.isdir(path):
5     full_path = os.path.join(path, filename)
6     if os.path.isfile(full_path):
7         with open(full_path, 'r') as f:
8             line = f.readlines()
9             for _ in range(len(line)):
10                 print(line)

```

#### 3. 获取文件大小和修改时间

```

1 os.path.getsize('/etc/passwd')
2 os.path.gettime('/etc/passwd')
3 import time
4 time.ctime(os.path.gettime('/etc/passwd'))

```

#### 4. 获取当前目录里面的指定文件的文件名

```

1 dir_name = '/home/hpc/TensorFlow_Notebook/code'
2 pyfile = [name for name in os.listdir(dir_name) if name.endswith('.py')]
3 #or use glob and fnmatch
4 import glob
5 pyfiles = glob.glob(dir_name+'/*.py')
6 from fnmatch import fnmatch
7 pyfiles = [name for name in os.listdir(dir_name) if fnmatch(name, '*.py')]

```

#### 4. 获取指定目录的文件的相关信息

```

1 import os
2 import os.path
3 import glob
4 import time
5 path_name = '/home/hpc/TensorFlow_Notebook/code'
6 pyfiles = glob.glob(path_name+'/*.py')
7 name_sz_data = [(name, os.path.getsize(name), os.path.getmtime(name)) for name in
                  pyfiles]
8 file_metadata = [(name, os.stat(name)) for name in pyfiles]
9 for name, meta in file_metadata:
10     print(name, '\t|', meta.st_size, '\t|', time.ctime(meta.st_mtime))

```

### 6.2.3 常见问题

1. 当你的程序获得目录中的一个文件列表，但是当试着打印文件名的时候文件崩溃，出现 UnicodeEncodeError 异常和一条奇怪的消息—surrogates not allow。

打印位置文件名时使用下面的方法可以避免下面的错误：

```

1 def bad_filename(filename):
2     return repr(filename)[1:-1]
3 try:
4     print(filename)
5 except UnicodeEncodeError:
6     print(bad_filename(filename))

```

默认情况下，Python 假定所有文件名都已经根据 sys.getfilesystemencoding() 的值编码过了。但是，有一些文件系统并没有强制要求这样做，因此允许创建文件名没有正确编码的文

件。这种情况不太常见，但是总会有些用户冒险这样做或者是无意之中这样做了（可能是在一个有缺陷的代码中给 open() 函数传递了一个不合规范的文件名）。当执行类似 os.listdir() 这样的函数时，这些不合规范的文件名就会让 Python 陷入困境。一方面，它不能仅仅只是丢弃这些不合格的名字。而另一方面，它又不能将这些文件名转换为正确的文本字符串。Python 对这个问题的解决方案是从文件名中获取未解码的字节值比如 \xhh 并将它映射成 Unicode 字符 \udchh 表示的所谓的“代理编码”。当你有一个不合格的文件名在目录列表中的是后，python 会将其转化为 unicode 如果你有代码需要操作文件名或者将文件名传递给 open() 这样的函数，一切都能正常工作。只有当你想要输出文件名时才会碰到些麻烦（比如打印输出到屏幕或日志文件等）。特别的，当你想打印上面的文件名列表时，你的程序就会崩溃，崩溃的原因就是字符\udce4 是一个非法的 Unicode 字符。它其实是一个被称为代理字符对的双字符组合的后半部分，因此他是一个非法的 Unicode，所以唯一能称该输出的方法就是遇到不合法文件名时采取相应的补救措施。可以将上述代码修改为：

```

1 for name in files:
2     try:
3         print(name)
4     except UnicodeEncodeError:
5         print(bad_filename(name))

```

或者：

```

1 def bad_filename(filename):
2     temp = filename.encode(sys.getfilesystemencoding(), errors='surrogateescape')
3     return temp.decode('latin-1')

```

## 2. 不关闭一个以打开的文件前提下增加或改变它的 Unicode 编码。

```

1 >>> f = open('sample.txt', 'w')
2 >>> f
3 <_io.TextIOWrapper name='sample.txt' mode='w' encoding='UTF-8'>
4 >>> f.buffer
5 <_io.BufferedReader name='sample.txt'>
6 >>> f.buffer.raw
7 <_io.FileIO name='sample.txt' mode='wb'>
8 >>>

```

在这个例子中，io.TextIOWrapper 是一个编码和解码 Unicode 的文本处理层，io.BufferedReader 是一个处理二进制数据的带缓冲的 I/O 层，io.FileIO 是一个表示操作系统底层文件描述符的原始文件。增加或改变文本编码会涉及增加或改变最上面的 io.TextIOWrapper 层。

detach() 会断开文件最顶层并返回第二层，之后顶层就没什么用了，例如

```

1 >>> f = open('text.txt', 'w')
2 >>> f = io.TextIOWrapper(f.buffer, encoding='latin-1')

```

```

3 >>> b = f.detach()
4 >>> f.write('hello')
5 ValueError
6 <ipython-input-21-0ec9cf64e174> in <module>()
7     f.write('hello')
8
9 ValueError: underlying buffer has been detached

```

一旦断开最顶层后，你就可以给返回结果添加一个新的最顶层，比如：

```

1 >>> f = io.TextIOWrapper(b, encoding='latin-1')
2 <_io.TextIOWrapper name='text.txt' encoding='latin-1'>

```

在文本模式打开的文件中写入原始的字节数据（将数据直接写入缓冲区）

```

1 In [1]: import sys
2
3 In [2]: sys.stdout.write(b'Hello\n')
4 -----
5 TypeError
6 <ipython-input-2-51d3384e9645> in <module>()
7     sys.stdout.write(b'Hello\n')
8
9 TypeError: write() argument must be str, not bytes
10
11 In [3]: sys.stdout.buffer.write(b'Hello\n')
12 Hello

```

类似的，能够读取文本的 buffer 属性来读取二进制数据。I/O 系统以层级结构的形式构建而成。文本文件是通过在一个拥有缓冲的二进制模式文件上增加一个 Unicode 编码/解码层来创建。buffer 属性指向对应的底层文件。如果你直接访问它的话就会绕过文本编码/解码层。

本小节例子展示的 sys.stdout 可能看起来有点特殊。默认情况下，sys.stdout 总是以文本模式打开的。但是如果你在写一个需要打印二进制数据到标准输出的脚本的话，你可以使用上面演示的技术来绕过文本编码层。3. 你有一个对应于操作系统上一个已经打开的 I/O 通道（比如文件，管道，套芥子等）的整形文件描述符，你想将它包装成一个更高层的 Python 文件对象。

一个文件描述符和一个打开的普通文件不一样。文件描述符仅仅是一个操作系统指定的整数，用来指代某系统的 I/O 通道。如果你碰巧有这么一个文件描述符你可以通过 shiyingopen() 函数来将其包装为一个 Python 的文件对象。你仅仅需要使用这个整数值的文件描述符作为第一个参数来替代文件名即可：

```

1 In [4]: import os

```

```

2 In [5]: fd = os.open('text.txt', os.O_WRONLY|os.O_CREAT)
3 In [6]: f = open(fd, 'wt')
4 In [7]: f.write('hello world\n')
5 In [8]: f.close()

```

当高层文件对象被关闭或者破坏的时候，底层文件描述符也会被关闭。如果这个并不是你想要的结果，你可以给 `open()` 函数传递一个可选的 `closefd=False`. 比如:

```
1 f = open(fd, 'wt', closefd=False)
```

在 Unix 系统中，这种包装文件描述符的技术可以很方便的将一个类文件接口作用于一个以不同方式打开的 I/O 通道上，如管道、套接字等。举例来讲，下面是一个操作管道的例子：

```

1 from socket import socket, AF_INET, SOCK_STREAM
2
3 def echo_client(client_sock, addr):
4     print('Got connection from', addr)
5
6     # Make text-mode file wrappers for socket reading/writing
7     client_in = open(client_sock.fileno(), 'rt', encoding='latin-1',
8                      closefd=False)
9
10    client_out = open(client_sock.fileno(), 'wt', encoding='latin-1',
11                      closefd=False)
12
13    # Echo lines back to the client using file I/O
14    for line in client_in:
15        client_out.write(line)
16        client_out.flush()
17
18    client_sock.close()
19
20 def echo_server(address):
21     sock = socket(AF_INET, SOCK_STREAM)
22     sock.bind(address)
23     sock.listen(1)
24     while True:
25         client, addr = sock.accept()
26         echo_client(client, addr)

```

需要重点强调的一点是，上面的例子仅仅是为了演示内置的 `open()` 函数的一个特性，并且也只适用于基于 Unix 的系统。如果你想将一个类文件接口作用在一个套接字并希望你的代码可以跨平台，请使用套接字对象的 `makefile()` 方法。但是如果考虑可移植性的话，那上面的解决方案会比使用 `makefile()` 性能更好一点。

你也可以使用这种技术来构造一个别名，允许以不同于第一次打开文件的方式使用它。例如，下面演示如何创建一个文件对象，它允许你输出二进制数据到标准输出（通常以文本模式打开）：

```

1 import sys
2 # Create a binary-mode file for stdout
3 bstdout = open(sys.stdout.fileno(), 'wb', closefd=False)
4 bstdout.write(b'Hello World\n')
5 bstdout.flush()

```

尽管可以将一个已存在的文件描述符包装成一个正常的文件对象，但是要注意的是并不是所有的文件模式都被支持，并且某些类型的文件描述符可能会有副作用（特别是涉及到错误处理、文件结尾条件等等的时候）。在不同的操作系统上这种行为也是不一样，特别的，上面的例子都不能在非 Unix 系统上运行。

5. 创建临时文件和文件夹，在程序执行完后自动销毁。

```

1 from tempfile import TemporaryFile
2 with TemporaryFile('w+t') as f:
3     # Read/write to the file
4     f.write('Hello world \n')
5     f.write('testing\n')
6     # Seek back to beginning and read the data
7     f.seek(0)
8     data = f.read()
9 # Temporary file is destroyed

```

或者，如果你喜欢，你还可以像这样使用临时文件：

```

1 f = TemporaryFile('w+t')
2 # Use the temporary file
3 ...
4 f.close()
5 # File is destroyed

```

TemporaryFile() 的第一个参数时文件模式，通常来将文本模式使用 w+t，二进制模式使用 w+b。这个模式同时支持读和写操作，在这里很有用，因为当你关闭文件去修改模式的时候，文件实际上已经不存在了。TemporaryFile() 另外还支持内置的 open() 函数一样的参数。比如：

```

1 with TemporaryFile('w+t', encoding='utf-8', errors='ignore') as f:
2     ...

```

在大多数系统上，同感 TemporaryFile() 创建的文件都是匿名的，甚至连目录都没有。如果你想打破这个限制，可以使用 NamedTemporaryFile() 来代替。比如：

```

1 In [16]: with NamedTemporaryFile('w+t') as f:
2     ...:     print('filename is:', f.name)
3 from tempfile import NamedTemporaryFile
4 with NamedTemporaryFile('w+t') as f:
5     print('filename is:', f.name)
6 filename is: /tmp/tmp4dwoxytf

```

这里被打开的文件的 `f.name` 属性包含了临时文件的文件名。当你需要将文件传递给其它代码来打开这个文件的 `scipio`, 这个就很有用了, 和 `TemporaryFile()` 一样, 结果文件关闭时会被自动删除调。如果你不想这么做呢, 可以传递一个关键字参数 `delte=False` 即可。比如:

```

1 with NamedTemporaryFile('w+t', delete=False) as f:
2     print('filename is:', f.name)
3 ...

```

为了创建一个临时目录, 可以使以哦嗯 `tempfile.TemporaryDirectory()`。比如:

```

1 from tempfile import TemporaryDirectory
2
3 with TemporaryDirectory() as dirname:
4     print('dirname is:', dirname)
5     # Use the directory
6 ...
7 # Directory and all contents destroyed

```

`TemporaryFile()`、`NamedTemporaryFile()` 和 `TemporaryDirectory()` 函数应该是处理临时文件目录的最简单的方式了, 因为它们会自动处理所有的创建和清理步骤。在一个更低的级别, 你可以使用 `mkstemp()` 和 `mkdtemp()` 来创建临时文件和目录。比如:

```

1 In [19]: tempfile.mkstemp()
2 Out[19]: (13, '/tmp/tmp6heplg63')
3
4 In [20]: tempfile.mkdtemp()
5 Out[20]: '/tmp/tmpcd70_9po'

```

但是, 这些函数并不会做进一步的管理了。例如, 函数 `mkstemp()` 仅仅就返回一个原始的 OS 文件描述符, 你需要自己将它转换为一个真正的文件对象。同样你还需要自己清理这些文件。

通常来讲, 临时文件在系统默认的位置被创建, 比如 `/var/tmp` 或类似的地方。为了获取真实的位置, 可以使用 `tempfile.gettempdir()` 函数。比如:

```

1 In [20]: tempfile.mkdtemp()
2 Out[20]: '/tmp/tmpcd70_9po'
3

```

```
4 In [21]: tempfile.gettempdir()
5 Out[21]: '/tmp'
```

所有和临时文件相关的函数都允许你通过使用关键值参数 prefix,suffix 和 dir 来自定义目录以及命名规则，比如：

```
1 In [24]: from tempfile import NamedTemporaryFile
2 In [25]: f = NamedTemporaryFile(prefix='mytemp', suffix='.txt', dir='/tmp')
3 In [26]: f.name
4 '/tmp/mytempw2pxl2v5.txt'
```

最后还有一点，尽可能以最安全的方式使用 tempfile 模块来创建临时文件。包括仅给当前用户授权访问以及在文件创建过程中采取措施避免竞态条件。

你需要将一个 Python 对象序列化为一个字节流，一边将它保存到一个文件，存储到数据库或者通过网络传输它。

用 pickle 模块将一个对象保存在一个文件中

```
1 import pickle
2 data = [1,2,3,4]
3 f = open('sample', 'wb')
4 pickle.dump(data, f)
```

将一个对象保存在一个文件中

```
1 import pickle
2 data = range(10)
3 f = open('temp', 'wb')
4 pickle.dump(data)
```

如果想将一个对象转化为字符串，可以使用 pickle.dumps():

```
1 s = pickle.dumps(data)
```

为了从字节流中恢复一个对象，使用 pickle.load() 或者 pickle.loads() 函数。比如：

```
1 # Restore from a file
2 f = open('somefile', 'rb')
3 data = pickle.load(f)
4
5 # Restore from a string
6 data = pickle.loads(s)
```

对于大多数应用程序来讲，dump() 和 load() 函数的使用就是你有效使用 pickle 模块所需的全部了。它可适用于绝大部分 Python 数据类型和用户自定义类的对象实例。如果你碰到某个库可以让你在数据库中保存/恢复 Python 对象或者是通过网络传输对象的话，那么很有可能这个库的底层就使用了 pickle 模块。

pickle 是一种 Python 特有的自描述的数据编码。通过自描述，被序列化后的数据包含每个对象开始和结束以及它的类型信息。因此，你无需担心对象记录的定义，它总是能工作。举个例子，如果要处理多个对象，你可以这样做：

```

1 >>> import pickle
2 >>> f = open('somedata', 'wb')
3 >>> pickle.dump([1, 2, 3, 4], f)
4 >>> pickle.dump('hello', f)
5 >>> pickle.dump({'Apple', 'Pear', 'Banana'}, f)
6 >>> f.close()
7 >>> f = open('somedata', 'rb')
8 >>> pickle.load(f)
9 [1, 2, 3, 4]
10 >>> pickle.load(f)
11 'hello'
12 >>> pickle.load(f)
13 {'Apple', 'Pear', 'Banana'}
14 >>>

```

还能序列化成函数，类，接口，但是结果数据仅仅将他们的名称编码成对应的代码对象。例如

```

1 >>> import math
2 >>> import pickle.
3 >>> pickle.dumps(math.cos)
4 b'\x80\x03cmath\ncos\nq\x00.'
5 >>>

```

当数据反序列化回来的时候，会先假定所有的源数据时可用的。模块、类和函数会自动按需导入进来。对于 Python 数据被不同机器上的解析器所共享的应用程序而言，数据的保存可能会有问题，因为所有的机器都必须访问同一个源代码。有些类型的对象是不能被序列化的。这些通常是那些依赖外部系统状态的对象，比如打开的文件，网络连接，线程，进程，栈帧等等。用户自定义类可以通过提供 `__getstate__()` 和 `__setstate__()` 方法来绕过这些限制。如果定义了这两个方法，`pickle.dump()` 就会调用 `__getstate__()` 获取序列化的对象。类似的，`__setstate__()` 在反序列化时被调用。为了演示这个工作原理，下面是一个在内部定义了一个线程但仍然可以序列化和反序列化的类：

```

1 # countdown.py
2 import time
3 import threading
4
5 class Countdown:
6     def __init__(self, n):

```

```

7     self.n = n
8     self.thr = threading.Thread(target=self.run)
9     self.thr.daemon = True
10    self.thr.start()
11
12    def run(self):
13        while self.n > 0:
14            print('T-minus', self.n)
15            self.n -= 1
16            time.sleep(5)
17
18    def __getstate__(self):
19        return self.n
20
21    def __setstate__(self, n):
22        self.__init__(n)

```

运行下面结构化代码

```

1 >>> import countdown
2 >>> c = countdown.Countdown(30)
3 >>> T-minus 30
4 T-minus 29
5 T-minus 28
6 ...
7
8 >>> # After a few moments
9 >>> f = open('cstate.p', 'wb')
10 >>> import pickle
11 >>> pickle.dump(c, f)
12 >>> f.close()

```

然后退出 Python 解析器并重启后再试验下:

```

1 >>> f = open('cstate.p', 'rb')
2 >>> pickle.load(f)
3 countdown.Countdown object at 0x10069e2d0>
4 T-minus 19
5 T-minus 18
6 ...

```

你可以看到线程又奇迹般的重生了，从你第一次序列化它的地方又恢复过来。

pickle 对于大型的数据结构比如使用 array 或 numpy 模块创建的二进制数组效率并不是一个高效的编码方式。如果你需要移动大量的数组数据，你最好是先在一个文件中将其保存为数组数据块或使用更高级的标准编码方式如 HDF5 (需要第三方库的支持)。

由于 pickle 是 Python 特有的并且附着在源码上，所有如果需要长期存储数据的时候不应该选用它。例如，如果源码变动了，你所有的存储数据可能会被破坏并且变得不可读取。坦白来讲，对于在数据库和存档文件中存储数据时，你最好使用更加标准的数据编码格式如 XML, CSV 或 JSON。这些编码格式更标准，可以被不同的语言支持，并且也能很好的适应源码变更。

### 6.3 正则表达式介绍

操作符	说明	实例
[]	字符集合, 对单个字符给出取值范围	[abc]表示 a,b,c,[a-z]表示 a 到 z 的单个字符
.	任何单个字符	
[^ ]	非字符集, 对单个字符给出排除范围	[^ abc]表示非 a 或者 b 或者 c 的单个字符
*	前一个字符 0 次或者无限次扩展	abc* 表示 ab,abc,abcc 等
+	前一个字符 1 次或无限次扩展	abc+ 表示 abc,abcc,abccc 等
?	前一个字符 0 次或者一次扩展	abc? 表示 ac,abc
	左右表达式任一个	abc def 表示 abc 或者 def
{m}	扩展前一个字符 m 次	ab{2}c 表示 abc,abbc
{m,n}	扩展前一个字符 m 到 n 次, 包含 n	ab{1,2}c 表示 abc,abbc
^	匹配字符串开头	^abc 表示 abc 且在一个字符串开头
\$	匹配字符串结尾	abc\$ 表示 abc 且在一个字符串的结尾
()	分组标记, 内部只能使用   操作符	(abc) 表示 abc, (abc def) 表示 abc 或者 def
(...)	这是一个扩展的符号, 第一个字符在'?'后面决定了深层的语法。扩展通常没有创建一个新的 group,(?P<name>...) 时该规则惟一的特例)	
(?aiLmsux)	来自集合'a','i','L','m','s','u','x' 的一个或者多个字母, group 匹配空字符串字符给整个正则表达式设置相关的 flags: re.A,re.I,re.L,re.M,re.S,re.X。如果你洗完桑包含 flags 作为正则表达式的一部分而不是传递一个 flag 参数到 re.compile() 函数这就是很有用的, Flasg 应该首先用在表达式字符串。	



操作符 (?:...)	说明	实例
\d	数字等价与 [0-9]	
\D	非数字等价与 [0-9]	
\number	匹配相同 number 的组。组以 1 开始, 例如 (.+) \1 匹配'the the'or'55 55', 但是'the the'(中间需要有空格), 这种特殊的序列仅仅被用来匹配 1 到 99 组。如果第一个数字为 0 或者是 3 为八进制的, 他将被解释为一个 group match, 在字符类 '[' and ']' 中, 所有的数被当作字符。	
A 匹配	匹配字符串的开始	
\b	匹配空字符串, 但是仅仅是单词前面或者后面的空字符串, 单词被定义为一个 unicode 字母数字序列或下划线特征, 因此单词为被空格或者为字母数字预示, 非强跳得字符串, 注意, \b 被定义为 a\w 和 a\W 之间, 或者在\w 和单词开始之间, 这意味着 r'\bfoo\b' 匹配'foo','foo.),(foo)',bar foo baz' 而不是'foobar' 或者'foo3'	
\B	匹配空字符串, 但是仅仅当它不在单词的开头或者结尾时, 这意味着 r'py\B' 匹配'python','py3','py2', 而不是'py','py.' 或者是'py!'\B 和 \b 相反, 因此单词时 unicode 字母数字或者下划线, 尽管这能被 ASCII flag 改变	
\S	匹配不是任何不是空格的 unicode 字符, 和\s 相反, 如果 ASCII flag 被用这因为等于 [^\t\n\r\f\v](但是 flag 影响整个正则表达式, 因此在这种情况下 [^\t\n\r\f\v])	
z	匹配字符串的尾部	
(?imsx-imsx:...)	在字符字母集合'i','m','s','x' 中, '-' 跟着的来自同样字母集合的一个或者更多字母), 对于部分表达式字母集合或者移去相关的 flags:re.I,re.M,re.S,re.X。	
<?P=name>	:对于 group 的一个反向引用, 它匹配之前 name 命名的 group 无论什么文本。	

(?+...)	一个注释，括号里面的内容被简单的忽视	
(?=...)	如果... 匹配下一步，不小于任何字符串。例如 Isaac (?=Asimov) 将匹配'Isacc' 如果它被'Asimov' 跟着的话。	
(?!...)	如果... 不匹配下一个，例如 Isaac (?!Asimov) 将匹配'Isaac'，仅仅是它没有'Asimov' 跟着。	
\w	单词字符，等价与 [A-Za-z0-9_]	

正则表达式的语法实例

P(Y YT YTH YTHO)?N	'PN', 'PYN', 'PYTN', 'PYTHN', 'PYTHON'
PYTHON+	'PYTHON', 'PYTHONN', 'PYTHONNN', ...
PY[TH]ON	'PYTON', 'PYHON'
PY[TH]?ON	'PYON', 'PYaON', 'PYbON', 'PYcON', ...
PY{:3}N	'PN', 'PYN', 'PYYN', 'PYYYN', ...

常用的正则表达式:

^[A-Za-Z]+\$	26 个字母组成的字符串
^[A-Za-z0-9]+\$	由 26 个字母和数字组成的字符串
^-?\d+\$	整数形式的字符串
^[0-9]*[1-9][0-9]* \$	正整数形式的字符串
[1-9]\d{5}	中国境内邮政编码，6 位
[\u4e00-\u9fa5]	匹配中文字符
\d{3}-\d{8} \d{4}-\d{7}	国内电话号码，010-68913536

匹配 IP 地址的正则表达式: \d+\.\d+\.\d+ 或者 \{1,3\}. 精确写法:

0-99:[1-9]?\\d

100-199:1\\d{2}

200-249:2[0-4]?\\d

250-255:25[0-5]

IP 地址的正则表达式:((([1-9]?\\d|1\\d{2}|2[0-4]\\d|25[0-5]).){3}(([1-9]?\\d|1\\d{2}|2[0-4]\\d|25[0-5]))

## 6.4 RE 库的主要功能函数

re.search()	在一个字符串搜索匹配正则表达式的一个位置。
re.match()	从一个字符的开始为值起匹配正则表达式，返回 match 对象。
re.fullmatch()	如果整个字符串匹配正则表达式然会相应的 match 对象，不匹配返回 None，注意这不同于 0 长度匹配
re.findall()	搜索字符串，以列表类型返回全部匹配的字串
re.split()	将一个字符串按照正则表达式匹配结果进行分割，返回列表类型
re.finditer()	搜索字符串，返回一个匹配结果的迭代类型，每个迭代元素时 match 对象
re.sub()	在字符串中替换所有匹配正则表达式的子串，返回替换后的字符串。
re.subn()	执行替换操作凡是返回一个 (new_string,number_of_subs_made) 元组
re.escape(pattern)	转义素有的字符除了 ASCII 字母，数字和下划线，如果你想匹配一个也许有正则表达式在里面的任一字符串这是很有用的。
re.purge()	清除正则表达式缓存

re.search(pattern,string,flags=0): 在一个字符串中搜索匹配正则表达式的一个位置返回 match 对象。

- pattern: 正则表达式的字符串或原声字符串表示。
- string: 待匹配字符串。
- flags: 正则表达式使用时的控制标记。

6.4.1 re 表达式中的 flags	使\w \W\b\B\d\D \s\S 值执行 ASCII 匹配而不是 Unicode 匹配，仅仅对于 Unicode 样式有意义对 Byte 样式忽略。
re.A	
re.DEBUG	显示编译表达式的调试信息
re.I	忽略正则表达式的大小写，[A-Z] 能够匹配小写。
re.L	使得\w \W\b\B\d\D \s\S 依赖于当前现场，当现场机制不可信时不鼓励使用，在不管什么时候它处理一个 cultrue，你应该用 Unicode 匹配，这个 flag 仅仅可以被用在 bytes 样式中。
re.M	正则表达式中的操^ 作能够将给定字符串的每一行当作匹配开始
re.S	正则表达式中的. 操作能够匹配所有的字符，默认匹配除换行外的所有字符
re.VERBOSE(re.X)	这个 flag 通过允许你分割逻辑部分和增加注释允许你写的正则表达式更好，空 pattern 中的空格被忽略特别是当一个字符类或者当有为转义的反斜线时，当一行包含不饿时字符类得 # 和非转义斜线时，所有的左边以 # 开头的字符将被忽略
re.error(msg, pattern=None, pos=None)	<ul style="list-style-type: none"> <li>- msg: 非正式格式的错误消息</li> <li>- pattern: 正则表达式</li> <li>- pos: 在 pattern 编译失败的索引（也许是 None）</li> <li>- lineno: 对应位置的行（也许是 None）</li> <li>- colno: 对应位置的列（也许是 None）</li> </ul>

```

1 import re
2 match = re.match(r'1\d{5}', 'BIT 100081')
3 if match:
4     match.group(0)

```

re.match(pattern,string,flags=0): 从一个字符串的开始位置起匹配正则表达式，返回 match 对象。

```

1 import re
2 match = re.match(r'1\d{5}', '100081 BIT')
3 if match:
4     print(match.group(0))

```

re.findall(pattern,string,flags=0): 搜索字符串，以列表类型返回能匹配的子串。

```

1 import re
2 ls = re.findall(r'1\d{5}', 'BIT 100081 TSU100084')

```

re.split(pattern,string,maxsplit = 0,flags=0): 将字符串按照正则表达式匹配结果进行分割，返回列表类型。

maxsplit: 最大分割数，剩余部分作为最后一个元素输出。

```

1 import re
2 re.split(r'1\d{5}', 'BIT100081 TSU100084')
3 re.split(r'1\d{5}', 'BIT100081 TSU100084', maxsplit=1)

```

re.finditer(pattern,string,flags=0): 搜索字符串，返回一个匹配结果的迭代类型，每个迭代元素时 matchdurian。

```

1 import re
2 for m in re.finditer(r'1\d{5}', 'BIT100081 TSU100084'):
3     if m:
4         print(m.group(0))

```

re.sub(pattern,repl,string,count=0,flags=0) 在一个字符串中替换所有匹配正则表达式的子串返回替代后的字符串。

- repl: 替换匹配字符串的字符串
- string: 待匹配字符串
- count: 匹配的最大替换次数

```

1 import re
2 re.sub(r'1\d{5}', '110', 'BIT100081 TSU100084')

```

Re 库的另一种等价用法:

```

1 rst = re.search(r'1\d{5}', 'BIT 100081')

```

等价于

```

1 pat = re.compile(r'1\d{5}')
2 pat.search('BIT 100081')

```

regex.search	在字符串中搜索匹配正则表达式的第一位置，返回 match 对象
regex.match()	在字符串的开始为值起配置正则表达式，返回 match 对象
regex.findall()	所有字符串，以列表类型返回全部能匹配的子串
regex.split()	将字符串按照正则表达式匹配结果进行分割，返回列表类型。
regex.finditer()	搜索字符串，返回一个匹配结果的迭代类型，每个迭代元素是 match 对象
reg.sub()	在一个字符串中替换所有匹配正则表达式的子串，返回替换后的字符串

Match 对象：一次匹配的结果，包含匹配的很多信息。

```

1 match = re.search(r'1\d{5}', 'BIT 100081')
2 if match:
3     print(match.group(0))
4 type(match)

```

match 对象的属性和方法

.string	待匹配的文本
.re	匹配时使用的 patter 对象 (正则表达式)
.pos	正则表达式搜索文本的开始位置
.endpos	正则表达式搜索文本的结束位置
.group(0)	获得匹配后的字符串
.start()	匹配字符串在原始字符串的开始位置
.end()	匹配字符串的结尾位置
.span()	返回 (.start(),.end())
.expand()	用 sub() 方法返回一个通过在 temple 字符串替代\的像\n 被转换成合适的字符串，数值反向索引 (\1,\2) 和 (\g<1>,\g<name>) 被相应组里面的内容取代\字符串
.__getitem__(g)	允许你轻松的访问一个 match 组
.groupdict(default=None)	返回一个包含所有子组的匹配对象，key 是子组的名字，被用在 groups 的默认参数 默认参数不参加匹配，默认值时 None。
.lastindex	最新匹配的组的整数索引，或者如果没有组被匹配就为 None。例如表达式 (a)b,((a)(b)) 和 ((ab)) 将有 lastindex == 1 如果应用的字符串'ab'，然而表达式 (a)(b) 将有 lastindex == 2，如果与应用在同一个字符串。
.lastgroup	最新匹配名字，如果 group 没有一个名字或者没有 group 就匹配为 None。
.re	正则表达式的 match() 或者 search() 方法生成的 match 实例

Re 库默认采用贪婪匹配，即输出匹配最长的字子串

```

1 match = re.search(r 'PY.*N', 'PYANBNCNDN')
2 match.group(0)

```

通常搜索的时候 PYAN 就能匹配出结果但是根据贪婪匹配，匹配待匹配字符串中最长的字符串。输出最短子串 PYAN。

```

1 match = re.search(r 'PY.*?N', 'PYANBNCNDN')

```

最小匹配操作符

操作符	说明
*?	前一个字符 0 次或者无限次扩展，最小匹配
+?	前一个字符 1 次或者浮现次扩展，最小匹配
??	前一个字符 0 次或者 1 次扩展，最小匹配
{m,n}?	扩展前一个字符串 m 到 n 次(含 n)，最小匹配

```

1 import re
2 m = re.match(r'(\w+ \w+)', 'Isaac Newton, physicist')
3 m.group(0)
4 m.group(1)
5 m.group(2)
6 m.group(1,2)

```

输出：

```

'Isaac Newton'
'Isaac'
'Newton'
('Isaac','Newton')

```

```

1 m = re.match(r'(\d+).(\d+)', '3.1415')
2 m.groups()

```

输出：

```
('3','1415')
```

```

1 m = re.match(r'(?P<first_name>\w+) (?P<last_name>\w+)', 'Malcolm Reynolds')
2 m.groupdict()

```

输出：'first\_name': 'Malcolm', 'last\_name': 'Reynolds'

## 6.5 常用的 sys 函数

- sys.abiflags: 在 POSIX 体同上 Python 用标准的 configure 脚本编译, 包含 PEP3149 指定的 ABI flags。
- sys.argv: 传递给 Python 的命令行参数, argv[0] 是脚本的名字, 在解释器中如果命令行用-c 选项, argv[0] 被设置为'-c'。如果没有脚本名字被传递给 python 解释器, argv[0] 是空字符串。
- sys.base\_exec\_prefix: Python 启动时设置, 在 site.py 之前运行前设置为 exec\_prefix。如果不运行一个虚拟环境, 值保持不变, 如果 site.py 找到的虚拟环境被用了, prefix 和 exec\_prefix 的值将被改变到指向虚拟环境, 由于 base\_prefix 和 base\_exec\_prefix 将任何指向 python 安装的 base 环境 (虚拟换将被创建)。
- sys.base\_prefix: 在 site.py 运行前 python 启动中值和 prefix 相同。如果不运行在虚拟环境中, 值将保持不变 rugosasite.py 找到一个虚拟环境被用, prefix 和 exec\_prefix() 值将被改变到指向虚拟环境, 由于 base\_prefix 和 base\_exec\_prefix 将保留指向 python 安装的 base 环境 (虚拟换将被创建)。
- byteorder: 本地变量的指示器, 这将在 big-endian 平台有一个值'big','title' 在 little-endian 平台。
- sys.vuiltin\_module\_name: 被编译进 Python 解释器的模块的字符串元组。(信息在其他方法下不可用-modules.keys() 仅仅显示导入的模块)。
- sys.call\_tracing(func,args): 调用 func(\*args), 当 trace 使能时。trace 状态被后来保存和恢复。从 checkpoint 文件 debug 去玄幻调试其它代码。
- sys.copyright: 包含 python 解释器版权信息的字符串。
- sys.\_clear\_type\_cache(): 清除内部变量的缓存, 类型缓存用来加速属性和方法的查找这个函数用来降低泄漏 debug 的非比要得查找。
- sys.\_current\_fnames(): 返回映射每个线程的标识符到函数调用时的线程栈的栈顶。注意 traceback 模块中的函数能编译调用被给定一个帧的栈。在调试线程锁时很有用: 这个函数线程锁死操作, 这样线程的调用被冻结和特们的死锁一样长。帧返回一个非死锁的线程也许忍受没有关系到当前这次调用的代码激活的线程检查帧。, 这个函数仅仅被用在内部或者特殊的目的。
- sys.\_debugmallocstats(): 打印 cpython 内存分贝其的低级的信息到标准的错误输出。如果 python 配置了-with-pydebug, 它也只行一些开销巨大的内部组成检查。

- sys.dllhandle: 指定处理 python dll 的整数，在 Windows 上可用。
- sys.displayhook(value): 如果值为 None, 函数打印 rep(value) 到 sys.stdout, 报春之在 builtins.\_\_。如果 repr(value) 时不可编码的 sys.stdout.encoding 和 sys.stdout.error 句柄, 解码 sys.stdout.encoding 和 backslashreplace 错误句柄。sys.displayhook 被调用在计算输入交互式 python 会话表达式的结果, 显示值能通过指定参数被自定义。

```

1 def displayhook( value ):
2     if value is None:
3         return
4     builtins.___ = None
5     text = repr( value )
6     try:
7         sys.stdout.writer( text )
8     except UnicodeEncodeError:
9         bytes = text.encode( sys.stdout.encoding, 'backslashreplace' )
10        if hasattr( sys.stdout, 'buffer' ):
11            sys.stdout.buffer.writer( bytes )
12        else:
13            text = bytes.decode( sys.stdout.encoding, 'strict' )
14            sys.stdout.buffer.writer( text )
15        sys.stdout.write( "\n" )
16        builtins.___ = value

```

- sys.dont\_write\_bytecode: 如果为真, python 不尝试写.pyc 文件到源模块, 值依赖-B 命令行选项和 PYTHONDONTWRITEBYTECODE 环境变量通过设置 True 或者 False 确定, 但是你可以在你自己控制二进制文件生成。
- sys.excepthook(type,value,traceback): 这个函数打印出一个给定的 traceback 和 sys.stderr 异常。当出现异常时, 解释器设置三个参数异常类, 异常实例和 traceback 对象调用 sys.execpthook。在交互式会话中这发生在控制被返回到终端前, 在 Python 程序中仅当程序退出时被调用, 在处理类似顶级异常可以通过指定另一个三个参数函数它哦 sys.exeothook。
- sys.\_\_displayhook\_\_
- sys.\_\_excepthook\_\_: 这个对象在程序的开头包含 displayhook 和 excepthook 的初始值, 他们被保存以至于他们被异常取代时 displayhook 和 excepthook 可以被恢复。
- sys.exc\_info: 这个函数给出关于当前被处理的异常的信息的元组。信息返回被指定到当前线程和当前栈帧, 如果当前栈帧没有处理异常, 信息被调用的栈帧得到, 或者它的调用器得到, 因此直到在处理异常时栈帧被发现, 这里处理一个异常被定义为处理

一个异常发生。对于任何栈帧，仅仅当前异常信息被处理。如果在栈帧中没有异常被处理，返回包含三个 None 的元组。否则返回值为 (type,value,traceback)，他们分别为 d 得到的被处理异常的类型，异常实例，和 traceback 对象（压缩调用栈）。

- sys.exec\_prefix: 一个字符串给 site-specific 目录前缀到 python 文件安装平台之前，默认是’/usr/local’，这可以通过设置 configure 脚本—exec-prefix 参数被设置编译时间，特别是所有的配置文件(像 pyconfig.h 头文件)被安装子啊 exec\_prefix/lib/pythonX.Y/config 和共享库模块被按转子啊 exec\_prefix/lib/pythonX.Y/lib-dynload，这里 X,Y 代表跑一趟好哦那得版本。
- sys.executable: 给 python 解释器一个绝对路径字符串，如果 python 不能获得真是的执行路径，sys.executable 将为 None。
- sys.exit([arg]): 从 python 推出，SystemExit 异常时生成，选项参数可以被给定为整数(默认为 0)或者其它对象类型。如果时一个整数，0 被认为成功终止，任何非零数值被认为异常终止。多数系统要求值在 0-127 之间，否则将产生不确定结果，一些系统约定指定推出代码，但是通常不完善，Unix 程序生成用 2 代表命令行语法错误 1 代表其它错误，如果另一类型的对象被传递，None 相当于传递 0，其他队像被打印到 stderr 和推出代码为 1，类似的 sys.exit("some error message") 是当程序出错一个快速退出程序的方法。因此 exit() 当从主进程退出进程时产生异常。异常不被拦截。
- sys.flags 结构序列 flags 暴露命令行状态

attribute	flag
debug	-d
inspect	-i
interactive	-i
optimize	-O or -OO
dont_write_bytecode	-B
no_user_site	-s
no_site	-S
ignore_environment	-E
verbose	-v
bytes_warning	-b
quit	-q
hash_randomization	-R

- sys.float\_info: 一个结构序列保持 float 类型的信息，它包含精确度和内部表达式低级信息，值符合在头文件 float.h 中定义的浮点常数。

attribute	float.h macro	explanation
epsilon	DBL_EPSILON	1 和大于 1 的最新值之间的差作为浮点数
dig	DBL_DIG	浮点数能带秒的最大精度
mant_dig	DBL_MANT_DIG	浮点精度。base-radix 浮点数的精度
max	DBL_MAX	有限浮点数的最大值
max_exp	DBL_MAX_EXP	radix <sup>(e-1)</sup> 代表的最大整数 e 代表无穷浮点数
max_10_exp	DBL_MAX_10_EXP	最大 $e10^{**2}$ 代表的最大浮点
radix	FLT_RADIX	指数表达式的基数
rounds	FLT_ROUNDS	整数常数代表 round 模式，这反映了系统在解释器启动时 FL

属性 sys.float\_info.dig 需要更进一步扩展，如果 s 时任何字符串表达一个十进制数，然后转换 s 为浮点数将恢复一个字符串表达式。

```

1 import sys
2 sys.float_info.dig
3 15
4 s = '3.14159265358979'      # decimal string with 15 significant digits
5 format(float(s), '.15g')    # convert to float and back -> same value
6 '3.14159265358979'

```

但是对于字符串 sys/float\_info.dig 指定精度，这不总是 true。

```

1 s = '9876543211234567'      # 16 significant digits is too many!
2 format(float(s), '.16g')    # conversion changes value
3 '9876543211234568'

```

- sys.float\_repr\_style: 指示 repr() 函数如何处理入店时的字符串。日国字符串有一个值'short' 然后对于一个有限的浮点数 x, repr(x) 产生一个短字符串 float(repr(x)) == x。否则 float\_repr\_style 有值'legacy' 和 repr(x) 行为正如子啊 python3.1 中的一样。
- sys.getalloctedblock(): 返回解释器当前分配的内存块数量，这个函数在更重和调式内存泄漏时很有用，因为解释器内部换传，结果可能因为调用而不同，你也许可以调用 \_clear\_type\_cache() 和 gc.collect() 得到雨鞋结果。如果 python 编译实现不能合理的计算这些信息，getalloctedbloacks() 允许返回 0。
- sys.getdefaultencoding(): 返回 Unicode 实现的字符串的默认编码的名字。
- sys.getdlopenflags(): 同 dlopen() 返回当前 flag 的值。flag 值的符号名字能被在 os 模块中找到

- `sys.getfilesystemencoding()`: 返回用于转换 Unicode 文件名和 bytes 文件名的编码名字, 为了最好的兼容性, str 应该在所有情况下被用在 filename, 尽管文件名作为 bytes 被支持, 函数接受 fanti 文件名应该支持 str 或者 bytes 内部转换系统偏好的表达。编码总是兼容 ASCII os.fsencode() 和 os.fsdecode() 应该被用于保证正确的编码和错误的模型使用。
  - 在 MAC OS 上编码为 utf-8
  - Unix 编码时 locale 编码
  - 在 windows 上编码也许是'utf-8' 或者是'mbcs', 依赖于用户配置。
- `sys.getfilesystemencodererrors()`: 返回转换 unicode 文件名和 bytes 文件名错误模式的名字, 编码名字有 `getfilesystemencoding()` 指定的便阿妈名字。os.fsencode() 和 os.fsdecode() 用来确保争取的编码和错误模式使用。
- `sys.getrefcount(object)`: 返回 object 对象的引用返回的储量通常高于你认为的呀, 因为它包含临时引用作为 `getrefcount()` 的参数。
- `sys.getsizeof(object,[,default])`: 返回对象的比特大小, 对象可以使任何类型的对象, 所有内建的多项将返回争取的结果, 但是这没有保持新的第三方扩展作为实现。仅仅内存消耗直接属性到对象, 对象访问时没有内存消耗。如果内定默认将返回不提供均值到这个值, 否则, `TypeError` 将产生。`getsizeof()` 调用对象的 `__sizeof__` 方法, 如果对象通过垃圾回收器管理增加一个额外的垃圾回收器。
- `sys.getrecursionlimit()`: 返回循环限制的当前值, 最大的 python 解释器栈深度。这限制阻止由无限循环从 c 栈移除和 python 崩溃, 它可以被 `setrecursionlimit()`。
- `sys.getsizeof(object,[,default])`: 返回对象的比特大小, 对象可以使任何类型, 所有内建的兑奖将被正确返回, 但是不是必须保持 true 给第三方扩展当它的实现被指定, 仅仅对象的直接内存消耗属性, 不是独享引用的内存消耗。如果对象没有给定获取大小, 默认将被返回, 否则 `TypeError` 将被报出。
- `sys.getwitchinterval()` 返回解释器的线程交换区间。
- `sys._getframe([depth])`: 从调用的栈返回一个帧对象, 如果宣讲整数 depth 被给定, 返回栈顶下的帧对象调用。如果 depper 比调用的栈深, `ValueError` 被报出。默认深度为 0, 返回调用栈顶的帧。
- `sys.getprofile()`: 获取 `setprofile()` 设置的 profile 函数。
- `sys.gettrace()`: 得到 `settrace()` 的 trace 函数。

- sys.getwindowsversion(): 返回一个描述当前 windows 版本的描述的名字元组。命名元素时 major,minor,build,platform,service\_pack\_minor,service\_pack\_major,suit\_mask,product\_type 和 platform\_version.service\_pack 包含一个字符串, platform\_version 一个三元组和所有其它值。这个组建可以同感 name 访问, 因此 sys.getwindowsversion()[0] 被等同  
 (VER\_NT\_WORKSTATION) 系统是工作站  
 platform 将被 2(VER\_PLATFORM\_WIN32\_NI) (VER\_NT\_DOMAIN\_CONTROLLER) 是系统是域控制器  
 (VER\_NT\_SERVER) 系统是服务器  
 函数包装 WIN32 GetVersionEx() 函数, windows 可用
- sys.get\_asyncgen\_hooks(): 返回一个类似名称元组的 asyncgen\_hooks 对象, 这里 firsttier 和 expected 均可设为 None 或者获取异步生成器作为参数的函数, 通过时间循环调度异步生成器终止。
- sys.get\_coroutine\_wrapper(): 返回 None 或者一个由 set\_coroutine\_wrapper 包装器.

- |  |  |
|--|--|
| <ul style="list-style-type: none"> <li>• sys.has_info: 数值 hash 实现参数给一个结构序列。</li> </ul> | width hash 值的位宽<br>modulus 用于数值 hash 方案的主要模块 P<br>inf 返回正无穷大的 hash 值<br>nan 返回非数的 hash 值<br>imag 返回复数的虚部<br>algorithm str,bytes 和 memoryview 的 hash 算法的实现<br>hash_bits hash 算法的内部输出大小<br>seed_bits hash 算法的种子值 |
|--|--|
- sys.hexversion: 单个证书的编码版本。这被保证增加, 包括合适的 support non-product release 版本, 例如。测试 Python 解释器时最新的版本 1.5.2 用

```

1 if sys.hexversion >= 0x010502F0:
2     # use some advanced feature
3     ...
4 else:
5     # use an alternative implementation or warn the user
6     ...

```

- sys.
- sys.
- sys.

- sys.
- sys.

## 6.6 collections

## 6.7 base64

## 6.8 struct

## 6.9 hashlib

## 6.10 `itertools`

## 6.11 contextlib

## 6.12 XML

### 6.13 HTMLParser

## 6.14 ZipFile

## 6.15 url

### 6.15.1 urllib.request

## 6.16 requests

### 6.16.1 发送请求

```

1 import requests
2 r = requests.get('https://github.com/timeline.json')
3 r = requests

```

### 6.16.2 requests 库的 7 个主要方法

requests.request()	够找一个请求，支持一下各方法的基础方法
requests.get()	获取 HTML 网页的主要方法，对应于 HTTP 的 GET
requests.head()	获取 HTML 网页头信息的方法，对应于 HTTP 的 HEAD
requests.post()	向 HTML 网页提交 POST 请求的方法，对应于 HTTP 的 POST
requests.put()	像 HTML 网页提交 PUT 请求的方法，对应于 HTTP 的 PUT
requests.patch()	像 HTML 网页提交局部修改请求，对应于 HTMP 的 PATCH
requests.delete()	像 HTML 网页提交删除请求，对应于 HTTP 的 DELETE

requests.get(url,params=None,\*\*kwargs)

- url: 想要获取的网页的 url 链接。
- params:url 中额外的参数，字典或字节流格式，可选
- \*\*kwargs:12 个控制访问的参数。

### 6.16.3 request 对象的属性

属性	说明
r.status_code	HTTP 请求的返回状态，200 表示连接成功，404 表示失败
r.text	HTTP 响应内容的字符串形式，即，url 对应的页面内容
r.encoding	从 HTTP header 中猜测响应的内容编码方式
r.apparent_encoding	从内容分析出的响应内容编码方式（备选编码方式）
r.content	HTTP 响应内容的二进制形式

#### 6.16.4 理解 encoding 和 apparent\_encoding

r.encoding: 从 HTTp header 中猜测的响应内容编码方式, 如果 header 中不存在 charset, 则认为编码为 ISO-8859-1 r.text 根据 r.encoding 显示网页内容

r.apparent\_encoding: 根据网页内容分析出的编码方式, 可以看做是 r.encoding 的备选。

#### 6.16.5 理解 Requests 库的异常

异常	说明
requests.ConnectionError	网络链接错误异常, 如 DNS 查询时白, 拒绝链接
requests.HEEPError	HTTP 错误异常
requests.URLRequired	URL 缺失异常
requests.TooManyRedirects	超过最大重定向次数, 产生重定向异常
requests.ConnectTimeout	连接远程服务器超时异常
requests.Timeout	请求 URL 超时, 产生超时异常
requests.raise_for_status()	如果不是 200, 产生异常, requests.HTTPError

r.raise\_for\_status() 方法在内部判断 r.status\_code 是否等于 200, 不需要额外加 if 语句, 该语句便于利用 try-except 进行异常处理。

#### 6.16.6 HTTP 协议

HTTP:Hypertext Transfer Protocol 超文本传输协议。

HTTP 是一句“请求与响应”模式的, 武装到的应用层协议, HTTP 协议采用 URL 定位网络资源的标志, URL 格式如下:

```
http://host[:port][path]
host: 合法的 Internet 主机域名和 IP 地址。
port: 端口号, 缺省端口为 80
path: 请求资源的路径
```

#### HTTP 协议对资源的操作

方法	说明
GET	请求获取 url 位置的资源
HEAD	请求获取 URL 位置资源的响应消息报告，即获得该资源的头部信息
POST	请求像 URL 位置的资源后附加新的数据
PUT	请求 URL 位置存储一个资源，覆盖原 URL 位置的资源
PATCH	请求局部更新 URL 位置的资源，即改变该处资源的部分内容
DELETE	请求删除 URL 位置存储的资源

head 方法的使用

```

1 In [3]: r = requests.head('http://www.baidu.com')
2 In [4]: r
3 Out[4]: <Response [200]>
4 In [5]: r.headers
5 Out[5]: {'Server': 'bfe/1.0.8.18', 'Date': 'Tue, 08 Aug 2017 11:46:32 GMT', 'Content-Type': 'text/html', 'Last-Modified': 'Mon, 13 Jun 2016 02:50:04 GMT', 'Connection': 'Keep-Alive', 'Cache-Control': 'private, no-cache, no-store, proxy-revalidate, no-transform', 'Pragma': 'no-cache', 'Content-Encoding': 'gzip'}
```

post 方法的使用 (像 URLPOST 一个表单，自动编码为 form)

```

1 In [10]: payload = {'key1': 'value1', 'key2': 'value2'}
2 In [11]: r = requests.post('http://httpbin.org/post', data=payload)
3 In [12]: print(r.text)
4 {
5     "form": {"key2": "value2",
6             "key1": "value1"
7     }
8 }
```

put 方法

```

1 In [18]: r = requests.put('http://httpbin.org/put', data = payload)
2
3 In [19]: print(r.text)
4 {
5     "args": [],
6     "data": "",
7     "files": {},
```

```

8   "form": {
9     "key1": "value1",
10    "key2": "value2"
11  },
12  "headers": {
13    "Accept": "*/*",
14    "Accept-Encoding": "gzip, deflate",
15    "Connection": "close",
16    "Content-Length": "23",
17    "Content-Type": "application/x-www-form-urlencoded",
18    "Host": "httpbin.org",
19    "User-Agent": "python-requests/2.18.3"
20  },
21  "json": null,
22  "origin": "210.47.0.232",
23  "url": "http://httpbin.org/put"
24 }

```

requests.request(method,url,\*\*kwargs)

**\*\*kwargs 控制访问参数**

params	字典或字节序列，作为参数增加到 url 中
data	字典，字节序列或文件对象，作为 Request 的内容
json	JSON 格式的数据，作为 Request 的内容
header	字典,HTTP 定制头
cookies	字典或 CookieJar，Request 中的 cokkie
auth	元组，支持 HTTP 认证功能
file	字典类型，传输文件
timeout	设定草食时间，s 为单位。
proxies	字典类型，设定访问代理服务器，可以增加登录认证
allow_redirects	:True/False, 默认为 True, 重定向开关
stream	True/False, 默认为 True, 获取内容立即下载开关
verify	True/False, 默认为 True, 认证 SSL 整数开关
cert	本地 SSL 整数路径

# Chapter 7

## Bazel

### 7.1 Bazel start

#### 7.1.1 用工作空间

所有的共建操作在你的文件系统上包含有所有构建的软件的源代码目录，正如 build 输出的符号连接的目录（例如 `bazel-bin` 和 `bazel-out`），构建发生在 `workspace`。`workspace` 的位置目录不重要，但是必须包含一个叫最哦 `WORKSPACE` 的顶层目录。构建是可用的 `workspace`。`WORKSPACE` 文件能被用来查询构建输出需要的外部依赖。一个 `workspace` 可以在多个项目中共享。

```
1 touch WORKSPACE
```



Chapter 8

# Tensorflow 技巧

## 8.1 文件读取



# Chapter 9

## Tensorflow API

### 9.1 tf.app.flags

#### 9.1.1 DEFINE\_boolean

DEFINE\_boolean(flag\_name,default\_value,docstring): 定义一个'boolean' 类型的 flag。

- flag\_name:flag 的名字，是一个字符串。
- default\_value:flag 应被看作一个 boolean 的默认值。
- docstring: 用 flag 的一个帮助信息。

#### 9.1.2 DEFINE\_boolean

: 定义一个'boolean' 类型的 flag。

- flag\_name:flag 的名字，是一个字符串。
- flag\_default\_value: 默认的 boolean 类型的值。
- docstring: 用 flag 的一个有用的帮助信息。

#### 9.1.3 DEFINE\_float

: 定义一个浮点数类型的 flag。

- flag\_name: 作为 flag 的名字，应该是字符串。
- default\_value:flag 的默认值，应该是浮点数。
- docstring: 用 flag 的一个有用的帮助信息。

### 9.1.4 DEFINE\_integer

: 定义一个整数的 flag。

- flag\_name:flag 的名字，应该是字符串。
- default\_value:flag 的默认值，应该是一个整数。
- docstring: 用 flag 的一个有用的帮助信息。

### 9.1.5 DEFINE\_string

: 定义一个字符串的 flag。

- flag 的名字，应该是字符串。
- default\_value:flag 的默认值，应该是字符串。
- docstring: 用 flag 的一个有用的帮助信息。

### 9.1.6 tf.squeeze

`tf.squeeze(input, axis=None, name=None, squeeze_dims=None)` 说明: 从指定的 Tensor 中移除 1 维度。

- input:tensor, 输入 Tensor。
- axis: 列表, 指定需要移除的位置的列表, 默认为空列表 [], 索引从 0 开始 squeeze 不为 1 的索引会报错。
- name:caozuoide 名字
- squeeze\_dims: 否决当前轴的参数。
- 返回一个 Tensor, 形状和 input 相同, 包含和 input 相同的数据, 但是不包含有 1 的元素。
- 异常: squeeze\_dims 和 axis 同时指定时会有 ValueError。

```
# 't' is a tensor of shape [1, 2, 1, 3, 1, 1]
shape(squeeze(t)) ==> [2, 3]
# 't' is a tensor of shape [1, 2, 1, 3, 1, 1]
shape(squeeze(t, [2, 4])) ==> [1, 2, 3, 1]
```

## 9.1.7 tf.metrics

```
accuracy(labels,predictions,weights=none,metrics_collections,updates_collections=none,name=None)
```

- labels:tensor, 和 predictions 的形状相同, 代表真实值。
- predictions:tensor, 代表预测值。
- weights:tensor, rank 可以为 0 或者 labels 的 rank, 必须能和 label 广播 (所有的维度必须是 1, 或者和 labels 维度相同)
- metrics\_collection:accuracy 应该被增加的一个 collectiobn 列表选项。
- update\_collections:update\_op 应该添加的选项列表。
- name:variable\_scope 名字选项。
- accuracy: 返回值 tensor, 代表精度, 总共预测对的和总数的商。
- update\_op: 返回值适当增加 total 和 count 变量和 accuracy 匹配。
- valueerror: 异常如果 predictions 和 labels 有不同的形状, 或者 weight 不是 none 它的形状不合 prediction 匹配, 或者 metrics\_collections 会哦这 updates\_collections 不是一个 list 或者 tuple。

## 9.1.8 tf.stack

stack(values, axis=0, name='stack'): stack 一个 n 维 tensor 为 n+1 维 tensor。给定一个长度为 N 的形状为 (A,B,C) 的 tensor, 如果 axis==0 输出 tensor 的形状为 (N,A,B,C), 如果 axis==1, 输出 tensor 的形状为 (A,N,B,C) # 'x' is [1,4]

```
# 'y' is [3,6]
# 'z' is [3,6]
stack([x,y,z])==>[[1,4],[2,5],[3,6]]
stack(x,y,z,axis=1)==>[[1,2,3],[4,5,6]]
tf.stack([x,y,z]) = np.asarray([x,y,z])
```

输出参数:

- 一个 Tensor 列表。
- 整数, 默认为 0, 支持负坐标。
- 操作的名字。

§ 一个 stack 的 Tensor。

§ ValueError: 如果 axis 超过  $[-(R+1), R+1]$

Example

```

1 import tensorflow as tf
2 x = tf.constant([1,4])
3 y = tf.constant([2,5])
4 z = tf.constant([3,6])
5 r1 = tf.stack([x,y,z])
6 r2 = tf.stack([x,y,z],axis=1)
7 with tf.Session() as sess:
8     print(sess.run(r1).shape)
9     print(sess.run(r2).shape)
```

### 9.1.9 tf.reshape

`tf.reshape(tensor, shape, name=None)`

- Tensor: 一个 Tensor。
- shape: 一个列表, 数值类型为 int32 或者时 int64
- name: 操作的名字。

S 指定形状的 Tensor。

```

1 import tensorflow as tf
2 a = tf.linspace(0., 9., 10)
3 b = tf.reshape(a, [2,5])
4 with tf.Session() as sess:
5     a = sess.run(a)
6     b = sess.run(b)
7     print(a.shape)
8     print(b.shape)
```

### 9.1.10 tf.random\_crop

```

1 random_crop(
2     value,
3     size,
4     seed=None,
5     name=None
6 )
```

从 value tensor 随机剪裁一块指定 size 的 tensor, 要求 value.shape 大于参数 size 的值。如果一维不剪切需要传递完整的 size, 例如 RGB 图像可以用 size = [crop\_height,crop\_width,3] 剪切。

输入参数:

- value: 需要被剪切的 tensor
- size: 和 value 有相同 rank 的一维 tensor。
- seed: Python 整数, 用于创建一个随机种子。
- name: 操作的名字。

Retuen 一个剪切的和 value 有相同的 rank, 形状为 size 的 tensor。

#### 9.1.11 tf.random\_gamma

```
1 random_gamma(  
2     shape ,  
3     alpha ,  
4     beta=None ,  
5     dtype=tf.float32 ,  
6     seed=None ,  
7     name=None  
8 )
```

从指定 [Gamma 分布](#) 中获得 shape 样本, alpha 是形状参数, beta 是反向放大参数。

输入参数:

- shape: 一个一维整数 Tensor 或者 Python 数组, 指定 alpha, beta 参数的输出采样数据。
- alpha: 一个 tensor 或者 python 值或者 N 维 dtype 数据类型。alpha 提供形状参数, 必须和 beta 广播。
- beta: 一个 tensor 或 python 值或者 dtype 类型的 n 维数组。默认为 1. beta 提供 gamma 分布的反向放大参数。必须和 alpha 广播。
- dtype: alpha,beta,output 的输出:float16,float32 或者 float64。
- seed: 一个 Python 整数, 用于创建随机种子。
- name: 操作的选项。

Returns : samples: 数据类型为 dtype 的一个形状为 tf.concat(shape,tf.shape(alpha+beta)) 的 tensor。

例子:

```

1 samples = tf.random_gamma([10],[0.5,1.5])#形状为10, alpha为0.5, 1.5, 采样输出形状
   为[10,2]
2 samples = tf.random_gamma([7, 5], [0.5, 1.5]) # 采样输出形状为[7, 5, 2], where
   each slice [:, :, 0] and [:, :, 1] # 代表从两个分布采样的$7\items5$。
3 samples = tf.random_gamma([30], [[1.],[3.],[5.]], beta=[[3., 4.]]) #采样输出形状
   [30, 3, 2], 每$2\times2$ (广播运算) 30个
   采样点

```

### 9.1.12 tf.random\_normal

```

1 random_normal(
2     shape,
3     mean=0.0,
4     stddev=1.0,
5     dtype=tf.float32,
6     seed=None,
7     name=None
8 )

```

输出正态分布随机值。

输入参数:

- shape: 一个一维整数 tensor 或者 Python 数组。表示输出 tensor 的形状。
- mean: 一个 0 维 tensor 和 dtype 的 Python 值，正态分布的均值。
- stddev: 一个 0 维 tensor 和 dtype 的 Python 值，正态分布的标准差。
- dtype: 输出数据类型。
- seed: 一个整数，用于创建一个随机数种子。
- name: 操作的名字。

Returns 制订形状的正态分布值。

## 9.1.13 tf.random\_normal\_initializer

继承于[Initializer](#) 别名:

- Class `tf.contrib.keras.initializers.RandomNormal`
- Class `tf.random_normal_initializer`

输入参数:

- mean: 一个 python 标量或者标量 tensor。生成随机值的均值。
- stddev: 一个 python 标量或者标量 tensor。标生成随机值的标准差。
- seed:python 整数, 用于创建随机种子。
- dtypes: 数据类型, 仅仅支持浮点数类型。

方法:

- `__init__`

```
1 __init__(  
2     mean=0.0,  
3     stddev=1.0,  
4     seed=None,  
5     dtype=tf.float32  
6 )
```

- `__call__`

```
1 __call__(  
2     shape,  
3     dtype=None,  
4     partition_info=None  
5 )
```

- `from_config`: 从配置字典实例化 initializer。

```
1 from_config(  
2     cls,  
3     config  
4 )
```

例如:

```

1  initializer = RandomUniform(-1, 1)
2  config = initializer.get_config()
3  initializer = RandomUniform.from_config(config)

```

- `get_config()`

#### 9.1.14 `tf.random_poisson`

```

1 random_poisson(
2     lam,
3     shape,
4     dtype=tf.float32,
5     seed=None,
6     name=None
7 )

```

从 Poisson 分布中湖区制定形状的样本，`lam` 是描述分布的参数。

输入参数:

- `lam`: 一个 Tensor 或者 Python 值或者 `dtype` 指定的 N 维数组。`lam` 是 Poisson 分布的参数。
- `shape`: 一维整数 tensor 或者 python 数组，输出每个 `rate` 指定的参数的形状。
- `dtype`: `lam` 和输出的数据类型:`float16, float32, float64`.
- `seed`: Python 整数，用于创建一个分布的随机种子。
- `name`: 操作的名字。

Returns : 形状为 `tf.concat(shape, tf.shape(lam))` 的 tensor (值的数据类型如 `dtype` 指定)。

例子:

```

1 samples = tf.random_poisson([0.5, 1.5], [10]) # 采样输出形状为[10, 2]，切片输出
2                                         [:, 0] and [:, 1] 代表不同分布的数据
3 samples = tf.random_poisson([12.2, 3.3], [7, 5]) # 采样输出形状[7, 5, 2]，切片
4                                         [:, :, 0] 和 [:, :, 1] 代表 $7\times 5$ 从两个分布的采样输出。

```

## 9.1.15 random\_shuffle

```

1 random_shuffle(
2     value ,
3     seed=None ,
4     name=None
5 )

```

沿着 tensor 的一维随机打乱数据。像 value[j] 映射一个 tensor 到一个数出 output[i]。例如

```

1 [ [ 1 ,  2 ] ,           [ [ 5 ,  6 ] ,
2   [ 3 ,  4 ] ,  ==>    [ 1 ,  2 ] ,
3   [ 5 ,  6 ] ]           [ 3 ,  4 ] ]

```

输入参数:

- value: 一个应该被打乱的值。
- seed: 一个 python 正数用于创建随机数种子。
- name: 操作的名字。

Returns 一个指定类型的值和形状，沿着某一维读被打乱的 tensor。

## 9.1.16 tf.random\_uniform

```

1 random_uniform(
2     shape ,
3     minval=0 ,
4     maxval=None ,
5     dtype=tf.float32 ,
6     seed=None ,
7     name=None
8 )

```

输出均匀分布的随机值，生成的值的范围在 [minval,maxval)，下界 minval 包含在范围内，上界不再。对于浮点数默认范围是 [0,1)。至少 maxval 必须被明确指定。在整数情况下，会有一些轻微的偏移，除非 maxval-minval 是 2 的次幂。偏移是 maxval-minval 比较小的值相对于输出范围 ( $2^{32}, 2^{64}$ ) 很小。

输入参数:

- shape: 一维整数 tensor 或者 Python 数组，和数出 tensor 的形状相同。
- minval: 0 维 tensor 或者 dtype 指定的 python 值，生成随机数的下界，默認為 0,。

- maxval:0 维 tensor 或者 dtype 指定的 python 值。生成随机数的上界，dtype 是浮点数时，默认为 1。
- dtype: 输出的类型:'float16,float32,float64,int32,int64'。
- seed: 一个 Python 整数。用于创建分布的随机数种子。
- 操作的名字。

Return 返回指定形状的均匀分布值。

Raise ValueError: 如果 dtype 是整数并且 maxval 没有被指定。

### 9.1.17 tf.random\_uniform\_initializer

一个继承自 Initializer 的类。

别名:

- tf.contrib.keras.initializers.RandomUniform
- tf.random\_uniform\_initializer

生成均匀分布 tensor 的 initializer。

输入参数:

- minval:0 维 tensor 或者 dtype 指定的 python 值，生成随机数的下界，默认为 0,。
- maxval:0 维 tensor 或者 dtype 指定的 python 值。生成随机数的上界，dtype 是浮点数时，默认为 1。
- seed: 一个 Python 整数。用于创建分布的随机数种子。
- dtype: 输出的类型:'float16,float32,float64,int32,int64'。

方法:

```

1   minval=0,
2   maxval=None,
3   seed=None,
4   dtype=tf.float32
5 )
```

- \_\_call\_\_

```

1 __call__(
2     shape ,
3     dtype=None ,
4     partition_info=None
5 )

```

- from\_config

```

1 from_config
2
3 from_config(
4     cls ,
5     config
6 )

```

- get\_config

#### 9.1.18 tf.one\_hot

```

1 ne_hot(
2     indices ,
3     depth ,
4     on_value=None ,
5     off_value=None ,
6     axis=None ,
7     dtype=None ,
8     name=None
9 )

```

返回一个 One-hot tensor。indices 表示的下表的值为 on\_value, 所有其它值为 off\_value。

如果 on\_value 不提供, 默认为 dtype 下的 1。off\_value 不提供, 默认值为 0, 如果输入 Incices rank 是 N, 输出 rank 将是 N+1。新的 axis 将在 axis 创建。, 如果 indices 是一个标量, 输出形状将是一个 depth 长度的向量, 如果 indices 是亿个长度为 features 向量输出形状将为:

```

1 feature x depth ifn axis == -1
2 depth x features if axis == 0

```

如果 indices 是一个矩阵, 形状为 [batch,features], 输出形状将为:

```

1 batch x features x depth if axis == -1
2 batch x depth x features if axis == 1
3 depth x batch x features if axis == 0

```

如果 dtype 没有被提供，他将假设 on\_value 或 off\_value 的数据类型，如果一个或者倍多的值呗传递，，如果没有 on\_value,off\_value, 或者 dtype 被提供，dtype 将默认 tf.float32。例子：

假设：

```

1 indices = [0, 2, -1, 1]
2   depth = 3
3   on_value = 5.0
4   off_value = 0.0
5   axis = -1

```

输出形状为  $[4 \times 3]$  输出：

```

1 output =
2   [5.0 0.0 0.0] // one_hot(0)
3   [0.0 0.0 5.0] // one_hot(2)
4   [0.0 0.0 0.0] // one_hot(-1)
5   [0.0 5.0 0.0] // one_hot(1)

```

假设：

```

1 indices = [[0, 2], [1, -1]]
2   depth = 3
3   on_value = 1.0
4   off_value = 0.0
5   axis = -1

```

输出

```

1 output =
2   [
3     [1.0, 0.0, 0.0] // one_hot(0)
4     [0.0, 0.0, 1.0] // one_hot(2)
5   [
6     [0.0, 1.0, 0.0] // one_hot(1)
7     [0.0, 0.0, 0.0] // one_hot(-1)
8   ]

```

用默认 on\_value 和 off\_value：

```

1 indices = [0, 1, 2]
2   depth = 3

```

输出

```

1 output =
2   [[1., 0., 0.],

```

```

3     [0., 1., 0.],
4     [0., 0., 1.])

```

参数:

- indices: 一个索引的 Tensor。
- depth: 一个定义 one hot 维度的深度的标量。
- on\_value: 制定 Indices[j]=i(填入 i), 默认为 1。
- off\_value: 一个标量当 Indices[j]!=i(默认为 0)。
- axis: 填入的轴, 默认为-1。
- dtype: 输出 tensor 的数据类型。

Returns :one-hot tensor。

Raise

- TypeError: 如果 dtype 和 on\_value,off\_value 数据类型不一样。
- TypeError: 如果 on\_value 和 off\_value 不合另一个匹配。

### 9.1.19 tf.unstack

```

1 unstack(
2     value,
3     num=None,
4     axis=0,
5     name='unstack',
6 )

```

unstackrank-R 为 rank-(R-1) 的 tensor。通过 chipping 沿着 axis 从 value unstack num tensor。如果 num 不被指定 (默认) 它从 value 的形状推算。如果 value.shape[axis] 不知道, ValueError 异常。例如一个 tensor 的形状为 (A,B<C<D); 如果 axis ==0 输出第 i 维的切片 value[i,:,:,:], output 的输出每个 tensor 形状为 (B,C,D). (注意沿着维度 unstack) 如果 axis == 1 然后输出第 i 个 tensor 在切片的值 [:,i,:,:] 和每个 output 的每个输出将有形状 (A,C,D)。tf.unstack(x,n)=list(x)

参数:

- value:rank R>0 的 Tensor 能被 unstack。
- num 一个整数。axis 的长度, 如果为 None 将自动推断。
- axis: 一个整数。unstack 沿着的轴。默认是第一维, 支持负的索引。

- name: 操作的名字。

Returns 从 value unstack 的 Tensor 列表。

Raises :

- ValueError: 如果 num 没有被指定且不能推断出。
- ValueError: 如果 axis 超过 [-R,R)。

## 9.2 tf.image

### 9.2.1 tf.image.decode\_gif

```
tf.image.decode_gif(contents,name=None)
```

- contents: 一个字符串 Tensor, GIF 编码的图像。
- name: 操作的名字。
- 返回一个 8 位无符号的 Tensor, 四维形状为 [num\_frames,height,width,3], 通道顺序是 RGB。

### 9.2.2 tf.image.decode\_jpeg

```
tf.image.decode_jpeg(contents,channels=None,ratio=None,fancy_upscaling=None,  
try_recover_truncated=None,acceptable_fraction=None,dct_methed=None,name=None)
```

解码 JPEG 编码的图像为无符号的 8 位整型 tensor。

- contents: 一个字符串 tensor, JPEG 编码的图像。
- channels: 一个整数默认为, 0 代表编码图像的通道数 (JPEG 编码的图像), 1 代表灰度图, 3 带秒 RGB 图。
- ratio: 一个整数, 默认为 1, 取值可以是 1,2,4,8, 表示缩减图像的比例。
- fancy\_upscaling:bool 型, 默认为 True, 表示用慢但是更好的提高色彩浓度。
- try\_recover\_truncated:bool 型, 默认是 False, 如果时 True 尝试从截断的输入恢复图像。
- acceptable\_fraction:float 型, 默认是 1, 可接受的最小的截断输入的因子。
- dct\_methed:string 类型, 默认为”。指定一个解压算法, 默认是”由系统自行指定。可用的值有 [”INTEGER\_FAST”,”INTEGER\_ACCURATE”]
- name: 操作的名字。
- 返回值为一个 8 位无符号整型 Tensor, 3 维形状 [height,width,channels]

### 9.2.3 tf.image.encode\_jpeg

`tf.image.encode_jpeg(image,format=None,quality=None,progressive=None,optimize_size=None,chroma_downsampling=None,density_uint=None,x_density=None,y_density=None,xmp_metadata=None,`

- `image`: 一个 3 维 [height,width,channels]，8 位无符号整型 Tensor。
- `format:string` 类型，可以为 "", "grayscale", "rgb"，默认为 ""。如果 `format` 没有指定或者不为空字符串，默认格式从 `image` 的通道中选，1: 输出灰度图，3: 输出 RGB 图。
- `quality`: 整型，默认值为 95，代表压缩质量值 [0,100]，值越大越好，单速度越慢。
- `optimize_size:bool` 型，默认为 False，如果为 True 用 CPU/RAM 减少尺寸同时保证质量。
- `chroma_downsampling:bool` 型，默认为 True。
- `density_unit`: 一个字符串，可以为 "in", "cm"，指定 `x_density` 和 `y_density.in` 每 inch 的像素，cm 表示每厘米的像素。
- `x_density`: 一个整数，默认为 300，每个 density 单位的水平像素。
- `y_density`: 一个整数，默认为 6300，数值方向上每 density 单位的像素。
- `xmp_metadata:string` 类型，默认为 ""，如果为空，嵌入 XMP metadata 到图像头部。
- `name`: 操作的名字。
- `name`: 操作的名字。
- 返回 0 维字符串型 JPEG 编码的 Tensor。

### 9.2.4 tf.image.decode\_png

`tf.image.decode_png(contents,channels=None,dtype=None,name=None)` 解码 PNG 编码的图像为 8 位或者 16 位无符号整型 Tensor。

- `contents`: 一个 0 维 PNG 编码的图像的字符串的 Tensor。
- `channels`: 整型默认为 0，代表解码图像的通道，0 用 PNG 编码图像数，1: 代表输出灰度图像。3: 代表输出 RGB 图像。4: 代表输出 RGBA 图像。
- `dtype:tf.DType`, 值可以为 `tf.uint8,tf.uint16`, 默认为 `tf.uint8`。
- `name`: 操作的名字。
- 返回 3 维 [height,width,channels] 的 Tensor。

## 9.2.5 tf.image.encode\_png

```
tf.image.encode_png(image,compression=None,name=None)
```

- 一个 8 位或者 16 位的 3 维 Tensor, 形状为 [height,width,channels]
- compression: 一个整数, 默认为-1, 表示压缩等级。
- name: 操作的名字。
- 返回一个 0 维 string 型的 PNG-encoded 的 Tensor。

## 9.2.6 tf.image.decode\_image

```
tf.image.decode_image(contents,channels=None,name=None)
```

- contents: 0 维编码图像的字符串。
- channels: 整数, 默认为 0, 解码图像的通道数。
- name: 操作的名字。
- 返回 JPEG,PNG 的 8 位无符号的形状为 [height,width,num\_channels], GIF 文件的形状为 [num\_frames,height,width,3]
- ValueError: 通道数不正确。

## 9.2.7 tf.image.resize\_images

```
tf.image.resize_images(images,size,method=ResizeMethod.BILINEAR,align_corners=False)
```

- images: 形状为 [batch,height,width,channels]4 维 Tensor,3 为 Tensor, 形状为 [height,width,channels]
- size: 一位 32 整型 Tensor 元素为 new\_height,new\_width, 新的图像尺寸。
- method: ResizeMethod, 默认为 ResizeMethod.BILINEAR
  - ResizeMethod.BILINEAR: 二进制插值。
  - ResizeMethod.NEAREST\_NEIGHBOR:
  - ResizeMethod.BICUBIC:
  - ResizeMethod.AREA:
- align\_corners:bool 型, 如果为真提取对齐四个角, 默认为 False。

- 异常
  - ValueError: 图像形状和函数要求的不一样。
  - ValueError:size 是不可用的形状或者类型。
  - ValueError: 指定的方法不支持。
- 如果图像时 4 维 [batch,new\_height,new\_height,channels], 如果图像是 3 维, 形状为 [new\_height,new\_width,channels]

## 9.3 layer

### 9.3.1 tf.layers.average\_pooling1d

**函数功能:** 一维数据的平均池化。

```

1 average_pooling1d(
2     inputs ,
3     pool_size ,
4     strides ,
5     padding='valid' ,
6     data_format='channels_last' ,
7     name=None
8 )

```

参数:

- 池化的 Tensor, rank 必须是 3。
- pool\_size: 一个正数或者是一个整数列表或元组, 代表池化窗口大小。
- strides: 一个整数, 指定池化操作的步数。
- padding: 一个字符串, padding 的方法, 可以是'valid' 或者是'same'。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,length,channels), channels\_first 形状为 (batch,channels,length)
- name: 字符串, layer 名字。
- 返回值: 输出 tensor, rank 为 3。

### 9.3.2 tf.layers.average\_pooling2d

**函数功能:** 二维数据的平均池化

```

1 average_pooling2d(
2     inputs ,
3     pool_size ,
4     strides ,
5     padding='valid' ,
6     data_format='channels_last' ,
7     name=None
8 )

```

参数:

- inputs: 池化的 Tensor, rank 必须为 4
- pool\_size: 两个元素的正数或者元组。指定池化窗口的大小。可以是单个整数表示，指定所有的空间维度为相同的值。
- padding: 一个字符串，padding 的方法，可以是'valid' 或者是'same'
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,height,width,channels), channels\_first 形状为 (batch,channels,height,width)
- name: 字符串, layer 名字。
- 返回值: 输出 tensor。

### 9.3.3 tf.layers.average\_pooling3d

**函数功能:** 三维输入的平均池化

```

1 average_pooling3d(
2     inputs ,
3     pool_size ,
4     strides ,
5     padding='valid' ,
6     data_format='channels_last' ,
7     name=None
8 )

```

参数:

- inputs: 池化的 Tensor, rank 必须为 5
- pool\_size: 正数或者元组。列表 (3 个元素) 指定池化窗口的大小。可以三个元素的整数，列表或者元组，指定所有的空间维度为相同的值。
- padding: 一个字符串，padding 的方法，可以是'valid' 或者是'same'
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,depth,height,width,channels), channels\_first 形状为 (batch,depth,channels,height,width)
- name: 字符串, layer 名字。
- 返回值: 输出 tensor。

## 9.3.4 tf.layers.batch\_normalization

**函数功能:**batch normalization layer 的函数接口。

```

1 batch_normalization(
2     inputs,
3     axis=-1,
4     momentum=0.99,
5     epsilon=0.001,
6     center=True,
7     scale=True,
8     beta_initializer=tf.zeros_initializer(),
9     gamma_initializer=tf.ones_initializer(),
10    moving_mean_initializer=tf.zeros_initializer(),
11    moving_variance_initializer=tf.ones_initializer(),
12    beta_regularizer=None,
13    gamma_regularizer=None,
14    training=False,
15    trainable=True,
16    name=None,
17    reuse=None,
18    renorm=False,
19    renorm_clipping=None,
20    renorm_momentum=0.99,
21    fused=False
22 )

```

batch normalization layer 的函数[参考](#),当训练的时候 moving\_mean 和 moving\_variance 需要被更新。默认的更新操作在 tf.GraphKeys.UPDATE\_OPS, 因此她们需要被添加到 train\_op, 例如

```

1 update_ops = tf.get_collection(tf.GraphKeys.UPDATE_OPS)
2 with tf.control_dependencies(update_ops):
3     train_op = optimizer.minimize(loss)

```

参数:

- inputs:2 维以上 Tensor, 进行 normalization, 第一个维度是 batch\_size, 如果 data\_format 是 NHWC 的最后一维 data\_format 是 NCHW 的第二维
- decay: 滑动平均的衰退状态, 接近于 1, 典型值是 0.999,0.99,0.9。低的 decay 值是 0.9, 如果模型训练性能好但是在验证和测试性能差, 尝试着设置 zeros\_debias\_moving\_mean=True 提高稳定性。
- center, bool 型为 True beta 偏移到正则化的 tensor, 如果为 False, beta 被忽略。

- scale:bool 型, 如果为 True, 乘上 gamma, 如果为佳 gamma 不用。当下一层为线性(nn.relu)时 scaling 可能被下一层使用而禁用。
- epsilon: 小的浮点数证驾到方差项防止除零错误。
- activation\_fn: 激活函数, 默认设置为 None 跳过激活函数使用线性激活函数。
- param\_initializers:beta,gamma,moving mean 和 moving variance 优化选项
- param\_regularizers:beta 和 gamma 的正则化选项。
- updates\_collections: 收集计算的更新操作。updates\_ops 需要执行 train\_op。如果为 None, 控制依赖将被增加保证更新在适当的位置计算。
- is\_training: 这层是否在训练模式如果在训练状态它将用指数移动平均求和加统计短到 moving\_mean 和 moving\_variance。当他不在在训练模式的时候她将用 moving\_mean 和 moving\_variance 的值。
- reuse: 是否层和它的变量被重用, 重用层的范围必须给定。
- variables\_collections: 变量的选项收集。
- outputs\_collections: 收集到增加输出。
- trainable: 如果为 True 增加变量到图 GraphKeys.TRAINABLE\_VARIABLES 上。
- batch\_weights: 有 batch\_size 形状的 tensor, 包含每个批的频率, 如果设置了值, 批用权重均值和方差归一化(可以用来收集训练选中样本的偏置)。
- fused: 如果为 True 用一个更快的基于 nn.fused\_batch\_norm 的融合实现。如果为 None, 如果可能用 fused 实现。
- data\_format: 一个字符串, NHWC(默认)NVCHW 也支持。
- zero\_debias\_moving\_mean:moving\_mean 用一个 zeros\_debias, 它创建一个新的变量 moving\_mean/biased 和'moving\_mean/local step'
- scope: 变量范围的选项。
- renorm: 是否批重新归一化, 在训练过程中增加额外的变量, 对于这个值得推倒时相同的

- renorm\_clipping: 用剪切重新归一化映射 key' rmax', 'rmin', 'dmax' 到标量 Tensor 的词典，相关的 (r,d) 被用作'corrected\_value = normalized\_value \* r + d', r 被剪切到 [rmin,rmax],d 到 [-dmax,dmax], 不指定 rmax,rmin,dmax,dmin 相应的被设置为 inf 和 0.
- renorm\_decay:momentum 在重新归一化中用于更新移动平均和标准差，太小（将增加噪声）太大（走样的评估）都将影响训练，decay 用于在推理时得到均值和方差。
- 返回一个代表输出操作的 Tensor。
- 异常
  - ValueError: 如果 batch\_weights 不是 None 但是 fused 是 True 时。
  - ValueError: 如果 data\_format 不是 NHWC 或者 NCHW。
  - ValueError: 如果 inputs 的 rank 没有指定。
  - ValueError: 如果输入的 channels 或者 rank 没有指定。

### 9.3.5 conv1d

**函数功能:** 一维卷积层的函数接口, 这个层创建一个卷积核和输出卷积输出一个 tensor。如果 use\_bias 是 True(bias\_initializer 被提供的话), bias 向量被创建添加到输出。最后如果激活函数不是 None, 激活函数也被用到输出。

```

1 conv1d(
2     inputs ,
3     filters ,
4     kernel_size ,
5     strides=1,
6     padding='valid' ,
7     data_format='channels_last' ,
8     dilation_rate=1 ,
9     activation=None ,
10    use_bias=True ,
11    kernel_initializer=None ,
12    bias_initializer=tf.zeros_initializer() ,
13    kernel_regularizer=None ,
14    bias_regularizer=None ,
15    activity_regularizer=None ,
16    trainable=True ,
17    name=None ,
18    reuse=None
19 )

```

- input: 输入 tensor。
- filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 一个整数或者元组或者列表, 指定一维卷积窗的大小。
- strides: 一个单个整数, 列表或者元组, 指定 stride 的长度, 指定任何不等于 1 的常数和指定任何 dilation\_rate 值部位 1 不兼容。
- padding:valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,length,channels), channels\_first 形状为 (batch,channels,length)
- dilation\_rate: 一个整数或者元组或者列表, 指定腐蚀卷积的腐蚀率, 指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- activation: 激活函数, 设置 None 维持线性激活函数。
- use\_bias:bool 型, 是否层用 bias。
- kernel\_initializer: 卷积核初始化器。
- bias\_initializer: 初始化偏置向量, 如果为 None 将没有 bias 被用。
- activation\_regularizer: 输出的正则化函数
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

### 9.3.6 conv2d

**函数功能:** 这层创建一个卷积核和输入相卷积输出。如果 use\_bias 是 True(bias\_initializer 提供了), bias 向量被创建添加到输出, 最后 activation 不是 None, 他被应用到输出。

```

1 conv2d(
2     inputs ,
3     filters ,
4     kernel_size ,
5     strides=(1, 1),

```

```
6     padding='valid',
7     data_format='channels_last',
8     dilation_rate=(1, 1),
9     activation=None,
10    use_bias=True,
11    kernel_initializer=None,
12    bias_initializer=tf.zeros_initializer(),
13    kernel_regularizer=None,
14    bias_regularizer=None,
15    activity_regularizer=None,
16    trainable=True,
17    name=None,
18    reuse=None
19 )
```

- input: 输入 tensor。
- filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 2 个整数元素或者元组或者列表, 指定二维卷积窗的宽和高, 可以用一个整数指定所有空间维度相同。
- strides: 一个整数, 列表或者元组, 指定 stride 的长度, 指定任何不等于 1 的常数和指定任何 dilation\_rate 值不为 1 不兼容。
- padding: valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,height,width,channels), channels\_first 形状为 (batch,channels,height,width)
- dilation\_rate: 两个整数或者元组或者列表, 指定腐蚀卷积的腐蚀率, 可以指定单个整数指定所有的空间维数相等, 指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- activation: 激活函数, 设置 None 维持线性激活函数。
- use\_bias: bool 型, 是否层用 bias。
- kernel\_initializer: 卷积核初始化器。
- bias\_initializer: 初始化偏置向量, 如果为 None 将没有 bias 被用。

- activation\_regularizer: 输出的正则化函数
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

### 9.3.7 conv2d\_transpose

**函数功能:** 二维卷积层的接口函数, 你希望用变形到正常卷积相反的方向你需要转置卷积, 有时候一些卷积的输出形状和输入形状相同但是维持链接样式是兼容的。

```

1  conv2d_transpose(
2      inputs ,
3      filters ,
4      kernel_size ,
5      strides=(1, 1),
6      padding='valid',
7      data_format='channels_last',
8      activation=None,
9      use_bias=True,
10     kernel_initializer=None,
11     bias_initializer=tf.zeros_initializer(),
12     kernel_regularizer=None,
13     bias_regularizer=None,
14     activity_regularizer=None,
15     trainable=True,
16     name=None,
17     reuse=None
18 )

```

- input: 输入 tensor。
- filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 2 个正整数组成的元组或者列表, 指定卷积核的宽和高, 可以一用一个整数指定所有空间维度相同。
- strides: 一个两个正整数组成的列表或者元组, 指定 stride 的长度, 指定任何不等于 1 的常数和指定任何 dilation\_rate 值不为 1 不兼容。

- padding:valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,height,width,channels), channels\_first 形状为 (batch,channels,height,width)
- dilation\_rate: 两个整数或者元组或者列表, 指定腐蚀卷积的腐蚀率, 可以指定单个整数指定所有的空间维数相等, 指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- activation: 激活函数, 设置 None 维持线性激活函数。
- use\_bias:bool 型, 是否层用 bias。
- kernel\_initializer: 卷积核初始化器。
- bias\_initializer: 初始化偏置向量, 如果为 None 将没有 bias 被用。
- activation\_regularizer: 输出的正则化函数
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

#### 9.3.8 conv3d

**函数功能:** 三维卷积的函数接口。这层创建一个卷积核和输入卷积生成输出 tensor。如果 use\_bias 是 True, bias\_initializer 被提供, 偏置向量创建添加到输出, 最终如果激活函数不是 None, 激活函数被用在输出上。

```

1 conv3d(
2     inputs ,
3     filters ,
4     kernel_size ,
5     strides=(1, 1, 1),
6     padding='valid',
7     data_format='channels_last',
8     dilation_rate=(1, 1, 1),
9     activation=None,
10    use_bias=True,
```

```

11   kernel_initializer=None,
12   bias_initializer=tf.zeros_initializer(),
13   kernel_regularizer=None,
14   bias_regularizer=None,
15   activity_regularizer=None,
16   trainable=True,
17   name=None,
18   reuse=None
19 )

```

- input: 输入 tensor。
- filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 3 个正整数组成的元组或者列表, 指定卷积核的深度, 宽和高, 可以用一个整数指定所有空间维度相同。
- strides: 三个正整数组成的列表或者元组, 指定 stride 的深度, 宽, 高, 指定单个整数代表所有空间维度相同, 指定任何 stride 不等于 1 的常数和指定任何 dilation\_rate 值不为 1 不兼容。
- padding:valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,depth,height,width,channels), channels\_first 形状为 (batch,depth,channels,height,width)
- dilation\_rate: 三个整数组成的元组或者列表, 指定腐蚀卷积的腐蚀率, 可以指定单个整数指定所有的空间维数相等, 指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- activation: 激活函数, 设置 None 维持线性激活函数。
- use\_bias:bool 型, 是否层用 bias。
- kernel\_initializer: 卷积核初始化器。
- bias\_initializer: 初始化偏置向量, 如果为 None 将没有 bias 被用。
- activation\_regularizer: 输出的正则化函数
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES

- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

### 9.3.9 conv3d\_transpose

**函数功能:** 三维卷积函数接口

```
1 conv3d_transpose(  
2     inputs,  
3     filters,  
4     kernel_size,  
5     strides=(1, 1, 1),  
6     padding='valid',  
7     data_format='channels_last',  
8     activation=None,  
9     use_bias=True,  
10    kernel_initializer=None,  
11    bias_initializer=tf.zeros_initializer(),  
12    kernel_regularizer=None,  
13    bias_regularizer=None,  
14    activity_regularizer=None,  
15    trainable=True,  
16    name=None,  
17    reuse=None  
18 )
```

- input: 输入 tensor。
- filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 3 个正整数组成的元组或者列表, 可以一用一个整数指定所有空间维度相同。
- strides: 三个正整数组成的列表或者元组, 指定单个整数代表所有空间维度相同。
- padding: valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,depth,height,width,channels), channels\_first 形状为 (batch,depth,channels,height,width)

- dilation\_rate: 三个整数组成的元组或者列表，指定腐蚀卷积的腐蚀率，可以指定单个整数指定所有的空间维数相等，指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- activation: 激活函数，设置 None 维持线性激活函数。
- use\_bias:bool 型，是否层用 bias。
- kernel\_initializer: 卷积核初始化器。
- bias\_initializer: 初始化偏置向量，如果为 None 将没有 bias 被用。
- activation\_regularizer: 输出的正则化函数
- trainable:bool 型，如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

### 9.3.10 dense

**函数功能:** 这个层实现操作:output = activation(input.kernel+bias)，这里 activation 传入 activation 的激活函数 (如果不为 None)，kernel 是一个层创建的权重矩阵，bias 是一个层创建的偏置向量。

```

1  dense(
2      inputs ,
3      units ,
4      activation=None ,
5      use_bias=True ,
6      kernel_initializer=None ,
7      bias_initializer=tf.zeros_initializer() ,
8      kernel_regularizer=None ,
9      bias_regularizer=None ,
10     activity_regularizer=None ,
11     trainable=True ,
12     name=None ,
13     reuse=None
14 )

```

- input: 输入 tensor。

- unit: 整数或者长整数输出空间的维数。
- activation: 激活函数, 设置为 None 表示非线性函数。
- use\_bias:bool 型, 当前层是否使用 bias
- kernel\_initializer: 权重矩阵的初始化函数。
- bias\_initializer: 偏置的初始化函数。
- kernel\_regularizer: 权重矩阵的正则化函数。
- bias\_regularizer: 偏置的正则化函数。
- activation\_regularizer: 输出的正则化函数。
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

### 9.3.11 dropout

**函数功能:** 设置随机丢弃值得比率, 帮助阻止过拟合。这个单位被  $\frac{1}{1-rate}$ , 因此他们的和在训练和推理时不改变。

```

1 dropout(
2     inputs ,
3     rate=0.5 ,
4     noise_shape=None ,
5     seed=None ,
6     training=False ,
7     name=None
8 )

```

- input: 输入 tensor。
- dropout 比率, 值在 0-1 之间, 例如 rate=0.1 表示丢掉输入的 10%。
- noise\_shape:int32 类型的一维 tensor 代表二进制 dropout mask 乘上输入, 例如, 如果你的输入形状为 (batch\_size, timesteps, features), 你想 dropout mask 和所有的 timesteps, 你可以用 noise\_shape=[batch\_size, 1, features]

- seed: 一个 Python 整数用于创建随机数种子。
- training: python bool 或者 TensorFlow bool 标量 tensor(例如 placeholder), 是否在训练模式 (dropout) 或者在推理模式 (返回没修改的输入) 返回输出
- name: layer 的名字。
- 输出 tensor。

### 9.3.12 max\_pool1d

**函数功能:** 一维输入的最大池化层。

```

1 max_pooling1d(
2     inputs ,
3     pool_size ,
4     strides ,
5     padding='valid' ,
6     data_format='channels_last' ,
7     name=None
8 )

```

- inputs: 需要池化的输入 tensor, rank 必须为 3
- pool\_size: 一个整数或者列表或者元组, 代表池化窗口的大小
- strides: 一个整数或者元组或者列表, 指定池化操作的 stride。
- padding: 一个字符串可以为'valid' 或者'same'
- data\_format: 一个字符串, 默认为 channels\_last 或者 channels\_first. 输入维度顺序, channels\_last 相关输入的形状为 (batch, length, channels), channels\_first 输出形状为 (batch, channels, length)
- name: 一个字符串, layer 的名字
- 输出一个三维 tensor。

### 9.3.13 max\_pool2d

**函数功能:** 二维输入的最大池化

```
1 max_pooling2d(  
2     inputs ,  
3     pool_size ,  
4     strides ,  
5     padding='valid' ,  
6     data_format='channels_last' ,  
7     name=None  
)
```

- inputs: 需要池化的输入 tensor, rank 必须为 4
- pool\_size: 两个整数组成的元组或者列表 (pool\_height,pool\_width), 代表池化窗口的大小, 可以为单个整数表示所有空间维度相等。
- strides: 两个整数组成的元组或者列表 (pool\_height,pool\_width), 指定池化操作的 stride, 可以为单个整数表示所有空间维度相等。
- padding: 一个字符串可以为'valid' 或者'same'
- data\_format: 一个字符串, 默认为 channels\_last 或者 channels\_first. 输入维度顺序, channels\_last 相关输入的形状为 (batch, height,width, channels), channels\_first 输出形状为 (batch, channels, height,width)
- name: 一个字符串, layer 的名字
- 输出一个三维 tensor。

#### 9.3.14 max\_pool3d

**函数功能:** 三维输入的最大池化层

```
1 max_pooling3d(  
2     inputs ,  
3     pool_size ,  
4     strides ,  
5     padding='valid' ,  
6     data_format='channels_last' ,  
7     name=None  
)
```

- inputs: 需要池化的输入 tensor, rank 必须为 5

- pool\_size: 三个整数组成的元组或者列表 (pool\_depth,pool\_height,pool\_width), 代表池化窗口的大小, 可以为单个整数表示所有空间维度相等。
- strides: 三个整数组成的元组或者列表 (pool\_height,pool\_width), 指定池化操作的 stride, 可以为单个整数表示所有空间维度相等。
- padding: 一个字符串可以为'valid' 或者'same'
- data\_format: 一个字符串, 默认为 channels\_last 或者 channels\_first. 输入维度顺序, channels\_last 相关输入的形状为 (batch, depth,height,width, channels), channels\_first 输出形状为 (batch, channels, depth,height,width)
- name: 一个字符串, layer 的名字
- 输出一个三维 tensor。

### 9.3.15 separable\_conv2d

**函数功能:** 深度方向分隔 2 维卷积层, 这层执行深度方向上通过 chennel 分开的卷积接着在深度方向上混合通道。如果 use\_bias 是 True 并且 bias 初始化被提供了, 它增加一个偏置向量到输出, 选项应用激活函数生成最终输出。

- inputs: 需要池化的输入 tensor item filters: 整数, 输出空间的维数 (卷积核的数量)。
- kernel\_size: 2 个整数元素或者元组或者列表, 指定二维卷积窗的宽和高, 可以一用一个整数指定所有空间维度相同。
- strides: 两个正整数组成的列表或者元组, 可以一用一个整数指定所有空间维度相同,, 指定任何不等于 1 的常数和指定任何 dilation\_rate 值不为 1 不兼容。
- padding:valid 或者 same。
- data\_format: 一个字符串 channels\_last(默认) 或者 channels\_first, 输入维度的顺序, channels\_last 输入形状为 (batch,height,width,channels), channels\_first 形状为 (batch,channels,height,width)
- dilation\_rate: 两个整数或者元组或者列表, 指定腐蚀卷积的腐蚀率, 可以指定单个整数指定所有的空间维数相等, 指定任何值 dilation\_rate!=1 和任何 strides 值!=1 不兼容。
- depth\_multiplier: 每个输入通道的深度方向卷积输出, 总共的深度方向卷积数等于 num\_filters\_in\*depth\_multiplier

- activation: 激活函数, 设置 None 维持线性激活函数。
- use\_bias:bool 型, 是否层用 bias。
- depthwise\_initializer: 深度方向卷积核的初始化器
- pointwise\_initializer:pointwise 卷积核的初始化器。
- depthwise\_regularizer:depthwise 卷积核的正则化器。
- pointwise\_regularizer:pointwise 卷积核的正则化器。
- bias\_regularizer: 偏置向量的正则化器。
- bias\_initializer: 初始化偏置向量, 如果为 None 将没有 bias 被用。
- activation\_regularizer: 输出的正则化函数
- trainable:bool 型, 如果为 True 增加变量到 GraphKeys.TRAINABLE\_VARIABLES
- name:layer 的名字。
- reuse: 是否重用之前层的相同名字的权重。
- 输出 tensor。

## 9.4 tf.train

提供了训练模型的类和函数。

### 9.4.1 优化器

优化器类提供方法计算损失函数对于变量的梯度的计算方法，子类集合实现了像 Adagrad 和 GradientDescent 等经典算法。

Optimizer

基础的优化类，定义了增加一个操作到训练模型的 API，你不直接需要这个类而是需要它的一些像 GradientDescentOptimizer, AdagradOptimizer, 或者 MomentumOptimizer 的子类。用法

```

1 # Create an optimizer with the desired parameters.
2 opt = GradientDescentOptimizer(learning_rate=0.1)
3 # Add Ops to the graph to minimize a cost by updating a list of variables.
4 # "cost" is a Tensor, and the list of variables contains tf.Variable
5 # objects.
6 opt_op = opt.minimize(cost, var_list=<list of variables>)

```

在训练程序的过程中你需要返回操作。

```

1 # Execute opt_op to do one step of training:
2 opt_op.run()

```

#### 在应用他们之前处理梯度

条用 minimize() 计算梯度，应用它们在变量上。如果你想在应用他们之前处理你可以按照下面的步骤使用优化器。

1. 用 comput\_gradients() 计算梯度。
2. 按照你的希望处理梯度。
3. 用 apply\_gradients() 处理梯度。

```

1 # Create an optimizer.
2 opt = GradientDescentOptimizer(learning_rate=0.1)
3
4 # Compute the gradients for a list of variables.
5 grads_and_vars = opt.compute_gradients(loss, <list of variables>)
6

```

```

7 # grads_and_vars is a list of tuples (gradient, variable). Do whatever you
8 # need to the 'gradient' part, for example cap them, etc.
9 capped_grads_and_vars = [(MyCapper(gv[0]), gv[1]) for gv in grads_and_vars]
10
11 # Ask the optimizer to apply the capped gradients.
12 opt.apply_gradients(capped_grads_and_vars)

```

minimize () he compute\_gradients() 接受一个 gate\_gradients 参数控制 fradient 应用中的并行度。

GATE\_NONE: 并行的计算，应用梯度，在执行过正中最大化并行程度，在结果中一些非重复性的代价。例如两个梯度的矩阵乘法依赖于输入值:GATE\_NONE 可能被应用到输入前其他梯度被计算导致非重复性的结果。

GATE\_OP: 对于每个 Op，在他们使用之前确保所有的梯度被计算了。为了避免 Op 的 race 为多个输入生成梯度 condition，这里梯度依赖于输入。

GATE\_GRAPH: 确保在它们任何一个被使用前所有变量的梯度被计算，提供了最小的并行化但是如果你想在应用他们之前处理所有的梯度这是很有用的。Slots

像 MomentumOptimizer 和 AdagradOptimizer 之类的优化器子类，结合变量训练分配管理额外的变量。这称为 Slots，Slots 有名字，你可以要求优化器它使用的名字。当你有一个 slot 名字你可以对变量要求优化器创建保留 slot 值。当你调试训练算法报告 slots 统计信息时很管用。**方法**

```

__init__
1 __init__(
2     use_locking,
3     name
4 )

```

创建一个新的优化器，他必须通过子类的构造体调用。

参数:

- use\_locking:bool, 如果为 True 用 lock 阻止当前变量更新。
- name: 非空字符串为 optimizer 创建的累加器的名字。
- ValueError: 名字格式不对。

apply\_gradients

```

1 apply_gradients(
2     grads_and_vars,
3     global_step=None,
4     name=None
5 )

```

应用梯度到变量尚，这是 minimize() 的第二部分，他返回应用梯度的 Op。

参数:

- grads\_add\_vars: 返回 compute\_gradients() 的 (梯度, 变量) 对列表。
- global\_step: 变量被更新后此选项变量增加 1.
- name: 返回操作的名字。默认传递名字给优化器构造函数。

S 返回: 应用指定梯度的操作。如果 global\_step 不是 None，操作增加 global\_step

- 异常

- TypeError: 如果 grads\_and\_vars 不对。
- ValueError: 如果变量的梯度为 none。

computer\_gradients

```

1 compute_gradients(
2     loss ,
3     var_list=None ,
4     gate_gradients=GATE_OP ,
5     aggregation_method=None ,
6     colocate_gradients_with_ops=False ,
7     grad_loss=None
8 )

```

计算损失关于 bar\_list 的梯度，第一部分是 minimize(). 返回一个 (gradient,variable) 对这里 gradient 是变量的梯度。注意梯度可以是 tensor, IndexedSlices 或者 None (如果没有变量的梯度)

- loss: 包含需要最小化的值的 tensor。
- var\_list:tf.Variable 更新最小化 loss 的列表或者元组。默认图中的变量列表在 GraphKey.TRAINABLE\_VARIABLES 下。
- gate\_gradients: 如何 gate 梯度计算, 可以是 GATE\_NONE,GATE\_OP 或者是 GATE\_GRAPH。
- aggregation\_method: 指定结合梯度的方法，可用值定义在 AggregationMethod。
- colocate\_gradients\_with\_ops: 如果为 True, 尝试随着相关操作 colocating 梯度。
- grad\_loss: 一个保持住 loss 梯度的 Tensor。

S 返回 (gradient,variable) 对，变量总是被呈现但是梯度可能是 None。