

西南财经大学经济信息工程学院

深度学习与文本挖掘

Deep Learning for Text Mining

研究生课程讲义



邱江涛
2019, Tensorflow1.12 版

目录

目录	2
第一章：前言	5
第一节：深度学习的兴起	5
第二节：深度学习框架介绍	11
第三节：关于本课程	12
第二章：神经网络基础	14
第一节：神经网络的发展	14
第二节：神经网络的结构	16
第三节：神经网络的训练	20
第四节：反向传播算法	24
第三章：TensorFlow	27
第一节：基本概念	27
第二节：Variable 的创建、初始化、存储与装载	37
第三节：初识 TensorFlow：手写数字识别	42
第四节：使用 TensorFlow 构建神经网络的步骤	47
第五节：TensorFlow 练习（1）：实现感知机	47
第六节：TensorFlow 练习（2）：曲线拟合	50
第七节：TensorBoard	52
第四章：深度神经网络	53
第一节：为什么需要深度神经网络？	53

第二节：梯度消失、梯度爆炸与解决方法	56
第三节：构建深度前馈神经网络	60
第四节：训练深度神经网络	61
第五节：Tensorflow 常用的激活函数	69
第六节：Tensorflow 常用的损失函数	71
第七节：Tensorflow 常用的优化函数	76
第四章：卷积神经网络	80
第一节：卷积	80
第二节：卷积神经网络的结构	83
第三节：卷积神经网络示例	92
第四节：Dropout	95
第五节：用 TensorFlow 实现 CNN	97
第五章：神经语言模型和词的表示学习	104
第一节：背景知识	105
第二节：word2vec	108
第三节：TensorFlow 的词表示学习	111
第六章：基于 CNN 的文本分类	119
第一节：CNN 文本分类模型	119
第二节：TensorFlow 实现 CNN 文本分类模型	122
第七章：循环神经网络	130
第一节：RNN 结构	130
第二节：使用 TensorFlow 构建 RNN	135
第三节：LSTM	149
第四节：LSTM 的变体	154

第五节：用 Tensorflow 构建词级 LSTM 语言模型.....	155
第六节：训练 RNN 的一些技巧	160
第九章：基于 LSTM 的文本情感分析.....	161
第十章：注意力机制	165
第十一章：RNN Encoder-Decoder 和生成新闻标题	166
第十二章：基于 BiLSTM 的问答系统	168
第十三章：最新发展	169
Appendix A: 常用 TensorFlow 函数	170
1. tf.split	170
2. tf.reshape.....	170
3. tf.concat.....	171
4. tf.nn.xw_plus_b	172
5. tf.nn.softmax_cross_entropy_with_logits_v2	173
6. tf.contrib.seq2seq.sequence_loss.....	174
7. tf.argmax	175
8. tf.nn.embedding_lookup	175
9. tf.reduce_mean.....	176
10. tf.truncated_normal	177
11. tf.random_uniform	177
12. tf.nn.dropout.....	178
13. tf.nn.max_pool.....	178
14. tf.slice	178
Appendix B: TensorFlow 的多线程.....	180

第一章：前言

第一节：深度学习的兴起

2016 年人工智能界最激动人心的一件事就是 Google 旗下的 DeepMind 公司开发的 AlphaGo 以 4 : 1 的比分击败了韩国围棋九段棋手李世石。2017 年则应该是排名第一的柯洁被 AlphaGo 以 3:0 血洗。



图 1-1. AlphaGo 的人机对弈

AlphaGo 的核心技术是 Deep Learning+Reinforcement Learning 技术。AlphaGo 的胜利证明了 Deep Learning 技术的强大，为本来已经在学术界已经很火的 Deep Learning 更是添了一把火。

Reinforcement Learning 不是本课程的授课内容。下面简单介绍一下其原理。

Reinforcement Learning 通常翻译做强化学习。被称作是机器学习领域有监督学习和无监督学习外的第三种学习。

我们人的成长过程中，有一种很重要的方式是和自然交互的过程中学习。例如，小孩碰了火一下，感觉到它是烫的，就知道了火是危险的。无论我们是开车或与人交谈，我们实际上是在感知环境然后做出反应行为。强化学习就是以目标为导向的在与环境交互的过程中学习。它有三要素：（1）感知环境的状态；（2）采取行动影响环境的状态；（3）有一个和环境状态相关的目标。

为什么说强化学习是第三种机器学习，因为前两种都面对这一个数据集。而强化学习面对的是一个环境，而且这个环境是可交互的。

强化学习的产生灵感来自“行为心理学”。它研究的问题是怎样让 Agent（有翻译作主体）在一个环境中采取的行为最大化“累积奖励”（*cumulative reward*）。一个增强学习的 Agent 在离散的时间点（时刻）和它的环境交互。在每个时间 t ，agent 接收一个观察 o_t ，它包括一个奖励（reward） r_t 。Agent 然后从一系列行为（action）中选择一个行为（action） a_t ，其后这项选择被送回到环境。环境于是转移到一个新的状态 s_{t+1} ，与这一转移相关联的 reward r_{t+1} 被确定。增强学习 agent 的目标是尽可能的收集奖励（reward）

Agent 为了采取最优的行动，必须推理论它的行为的长期结果。例如，为了最大化我的未来收入，最好的选择是进学校学习，虽然，这一行为为当前带来的是金钱是损失。增强学习特别适合这一类问题，它们需要在长期和短期利益之间做权衡。增强学习已经在许多领域取得了成功，如机器人控制、电梯调度和游戏领域。

深度学习取得的成就

按照 Nature 文章 Human-level control through deep reinforcement learning，该文章中提出的增强深度学习方法在机器玩游戏的任务中取得了 23/43/49 的性能。即，49 项游戏中，有 43 项可以基本平均的人类选手，在 23 项游戏中击败人类顶尖高手。

在语音识别领域，按照微软语音识别研究组的报告，自从 2010 年采用深度学习技术以来，语音识别领域有了突破性的进展，的错误率有了大幅的下降。

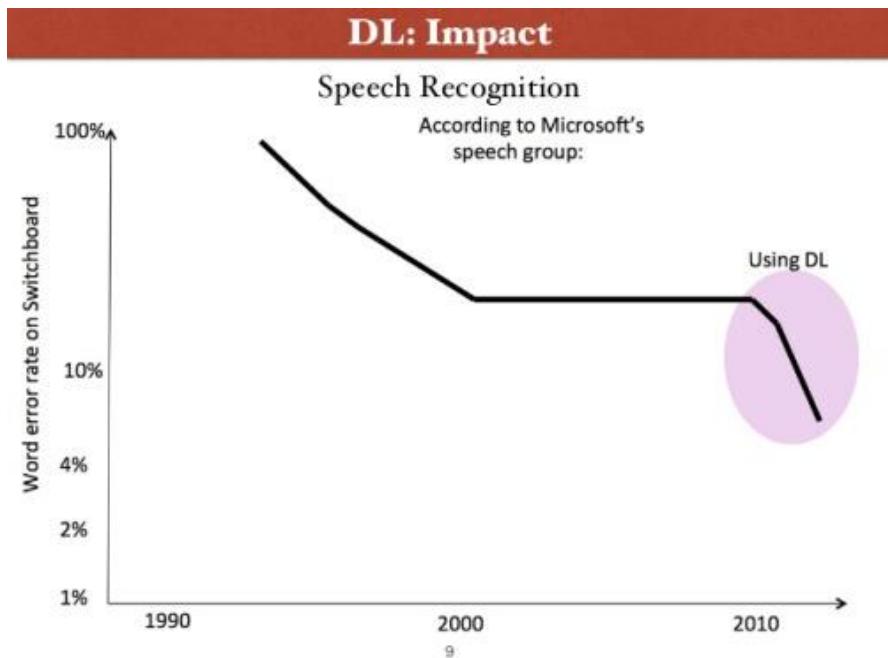


图 1-2：语音识别的错误率变化

按照在 MNIST (手写数字数据集) 测试集上的研究报告。采用纯神经网络的准确率是 96.59%，采用支持向量机 SVM，当采用 LibSVM 的默认参数，准确率是 94.53%；采用优化了参数的 SVM 准确率达到 98.56%；而采用卷积神经网络准确率达到了 99.79%。MNIST 有训练集有 6 万张图片，测试集有 1 万张图片。卷积神经网络仅有 21 张图片未能正确分类。如图 1-3 所示。



图 1-3. 未被识别的手写数字

深度学习技术被 MIT Technology Review 评为了 2013 年 10 大突破技术之一。其评价语是

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

可以访问一个深度学习的演示网站：playground.tensorflow.org 了解深度学习的强大功能

神经网络与深度学习的关系

有人评价神经网络是最优美的编程范式 (programming paradigms) 之一。传统的编程方法中，我们告诉计算机做什么，将大问题分解为小的计算机能处理的任务。对比之下，在神经网络中我们不告诉计算机怎样解决问题。神经网络会自己从观察到的数据学习，并计算出解决方案。

从数据中自动学习听起来很诱人。然而直到 2006 年，研究人员才知道怎样超越传统的方法来训练神经网络。这一年在深度神经网络进行学习的技术被发明，这就是现在称之为的**深度学习**。随后这些技术被进一步开发。今天，深度神经网络和深度学习解决在计算机视觉、语言识别和自然语言处理等领域的问题达到了非常优秀的性能。一些商业公司如 Google, Microsoft, Facebook 已经开发和大规模部署了深度学习框架。

神经网络是由生物学启发的编程机器学习模型（有称是编程范式 programming paradigm），由此计算机可以从观察的数据进行学习。深度学习是一个非常有力的在神经网络上进行学习的技术集合。神经网络和深度学习当前对图像处理、语言识别和自然语言处理中的许多问题提供了最好的解决方案。

深度学习比起传统神经网络的优势在于：

1. 采用了新的激活函数 ReLu，可以使得网络的层次超过三层
2. 采用 Dropout, Maxout 和随机池化技术 (Stochastic Pooling) 解决过拟合问题
3. 可以采用 GPU，解决多层网络训练速度慢的问题。

知识获取与深度学习¹

在专家系统时期，领域专家（主要是语言学家）直接提供显性知识。显性知识人类可以直接构建和理解的知识。详见图 1-4。

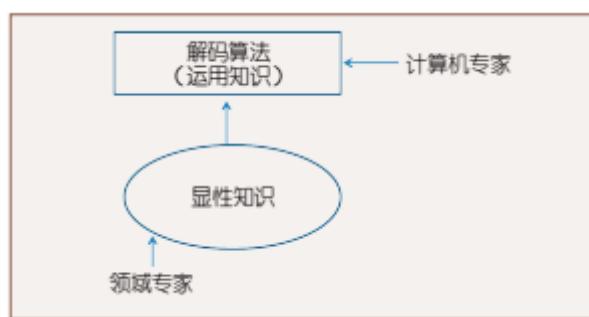


图 1-4. 专家系统

¹刘挺 车万翔，自然语言处理中的知识获取问题，中国计算机学会通讯，第 13 卷 第 5 期 2017 年 5 月

当进入语料库方法主导的时期，领域专家提供的显性知识仍然起到关键作用，例如决定识别对象的特征等。此时标注人员标注的小规模数据中所蕴含的知识更全面、更精确、更加量化。详见图 1-5。

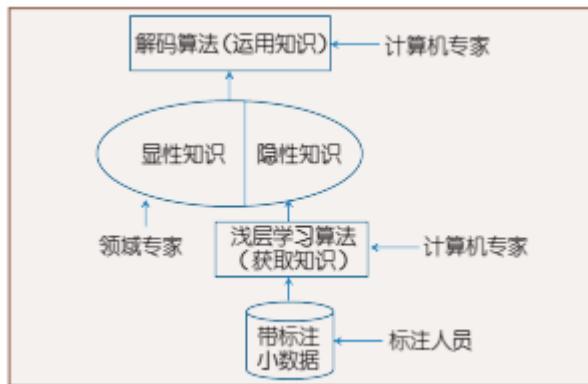


图 1-5. 语料库方法

深度学习“端到端”的方法兴起后，有些简单问题，如果能够找到足够多的大数据，则会按照图 1-6 的模式获取知识，即显性知识完全退场。**深度学习系统几乎不需要领域专家提供的元知识，由机器自己确定特征，并分配各层的功能**，尽管这些功能的划分是隐性的，难以解释的。

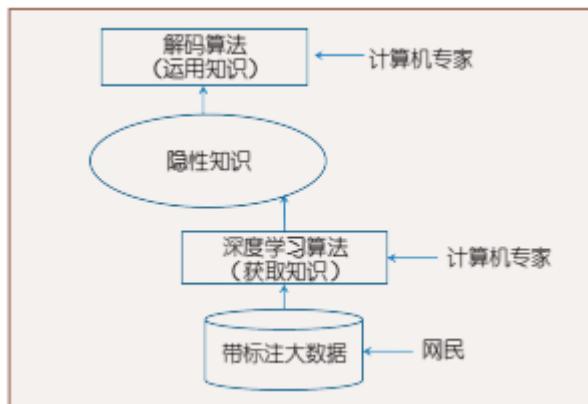


图 1-6. 基于大数据的深度学习。

需要强调的是，这里所谓“深度学习”的方法，关键在于“端到端”，即把从输入到输出的全部工作交给机器去处理，而不再人为地分层。下面试以“信息抽取”为例加以说明。信息抽取有两种做法：一是先做句法分析，再做信息抽取；二是直接做信息抽取，后者就是所谓的“端到端”。在“端到端”的模型中，也是分层的，但是由机器自己去分层处理，各层的含义不是直观可以理解的。当用于端到端的训练数据不足时，就需要人为的帮助，比如把信息抽取的过程分成两步去做：第一步，先做出句法树（这一步增加了显性知识，但也引入了误差）；第二步，实施信息抽取。当用于端

到端的训练数据充足时，就可以一步到位——直接做信息抽取，而且性能更好。与信息抽取类似的还有情感二分类、句间关系分类、问答对匹配等。

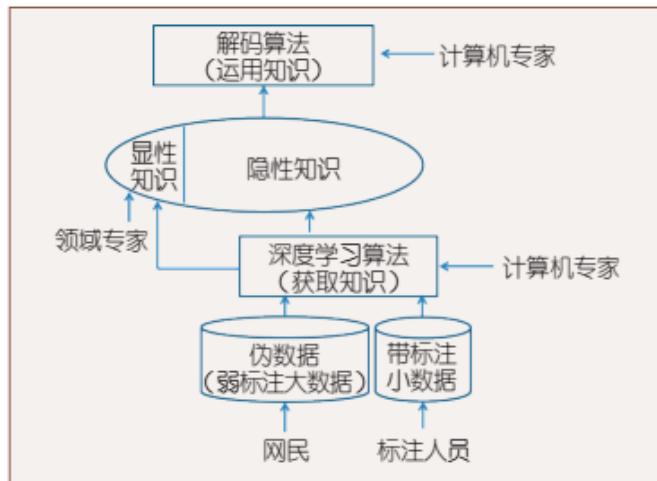


图 1-7. 基于伪数据的深度学习

但是，对很多问题而言，带标注数据是有限的，甚至是有限的，因此图 1-6 所示的理想情况并不多。图 1-7 是比较现实的解决方案。将大规模“伪数据”与人工标注的小数据结合在一起，同时接受领域专家提供的部分元知识，如此，在数据不充分的情况下，借助人工的力量，追求最优的系统性能。

大数据增加深度学习的威力

深度学习不是万能的，当数据不足时，深度学习的效果大打折扣。图 1-8 显示了有用不能获得充足的数据，即使在简单的问题上，深度学习也没有明星的超越传统方法。在复杂问题上甚至有劣势。因为问题复杂，而数据量不够时，学习工具越强大就越容易形成过拟合，所以效果自然不好。

	简单问题	复杂问题
小数据	无明显优势 (例如：词性标注)	有劣势 (例如：深层语义分析)
大数据	优势最明显 (例如：语言模型)	优势较明显 (例如：机器翻译)

图 1-8. 深度学习应用于自然语言处理的效果

从上述分析可以看出，一旦拥有大规模的训练数据，深度学习的威力是巨大的，可以在短时间内以摧枯拉朽的气势替代原有技术。

但是深度学习有它的问题：深度学习目前还停留在实验科学的阶段，其严格的数学解释还未完全建立。Geometric Understanding of Deep Learning 一文从几何的角度理解

深度学习，为深度学习提供严密的数学论证。NIPS2018 有论文从数学角度尝试解释 Dropout 的作用，深入探究 dropout 的本质。

第二节：深度学习框架介绍

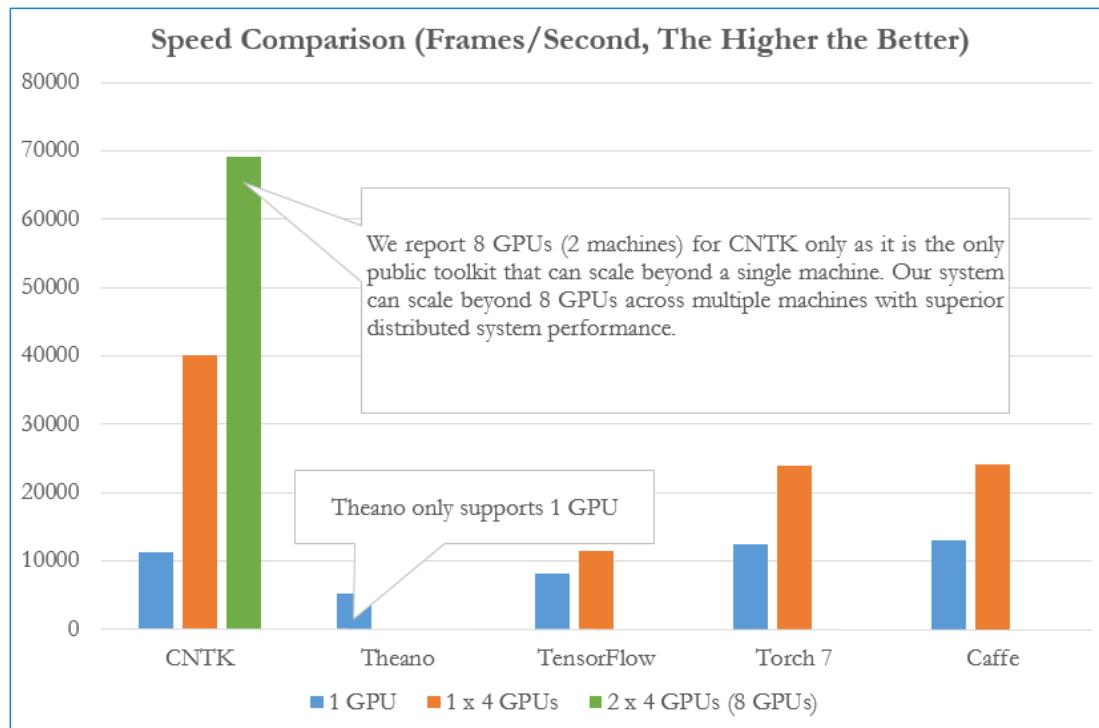


图 1-9：流行的深度学习框架

现在流行的深度学习框架有很多。包括微软开发的 CNTK，Google 的 TensorFlow，蒙特利尔大学开发的 Theano, facebook 开发的 Caffe 和开放社区的 Torch (几个人联合开发，无公司背景) 和由中国人发起的 MXNet。图 1-9 是微软公司发布的各种深度学习框架性能对比图。

framework	base language	multi-GPU?	pros	cons
TensorFlow	Python and C++	yes		
Torch	Lua	yes	1.) easy to set up 2.) helpful error messages 3.) large amount of sample code and tutorials	can be somewhat difficult to set up in CentOS
Caffe	C++	yes		general
Theano	Python	By default, no. Can use more than one but requires a workaround	1.) expressive Python syntax 2.) higher-level spin-off frameworks 3.) large amount of sample code and tutorials	error messages are cryptic

图 1-10：几种深度学习框架简介

微软的 CNTK 是采用 C++ 开发，提供了 C++ 和 Python 的接口。Torch 有个 Python 版本 Pytorch。相比 Tensorflow 它有个优势是调试方便。图 1-10 对几种深度学习框架做一个简单的描述。

第三节：关于本课程

在本课程中，我们先介绍神经网络的基础知识，然后介绍 Deep Learning，进一步介绍实现 Deep Learning 的工具 TensorFlow，并用 TensorFlow 实现几个 Deep Learning 的模型。Deep Learning 在 NLP 领域取得成功，我们将介绍相应的模型，并开发这些模型。最后我们用 Deep Learning 去解决文本挖掘的实际应用问题。

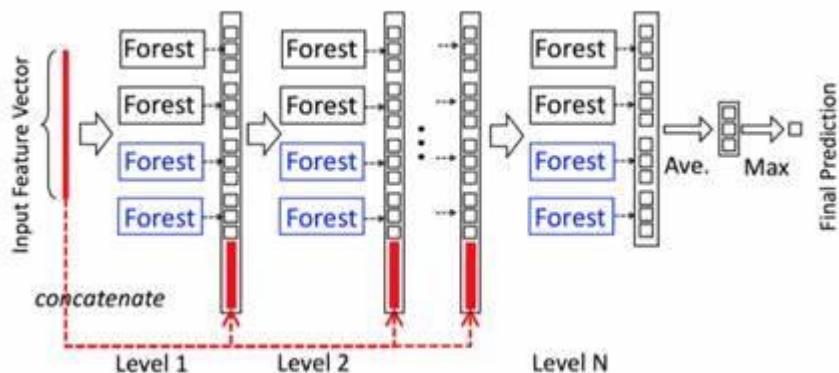
由于深度学习的发展非常迅猛，而本课程只有 9 周，另外由于深度学习对硬件要求的越来越高。因此本课程面向的是深度学习的基础知识和实践操作。更多的新知识理论会在最后一章的最新发展中介绍。

拓展与思考：

深度学习等价于深度神经网络吗？

SIAM，即国际工业与应用数学学会，的报纸 SIAM news 在 2018 年的 6 月份头版上文章说**深度学习是机器学习中使用深度神经网络的子领域**。这也是我们通常对深度学习的认知。

南京大学周志华团队推出了 gcForest，即 multi-Grained Cascade Forest。它是一个基于决策树的集成模型，是一个级联树结构的深度模型。



该模型的每一层中包含多个随机森林。该模型据周志华团队介绍，“它也可以在除了大规模图像数据之外的任务中获得和深度神经网络相似的表现。在大规模图像数据任务中不能取胜，其中原因是目前没有适当的硬件训练足够大的模型”。

第二章：神经网络基础

第一节：神经网络的发展

神经网络NN(Neural Networks)或人工神经网络ANN (Artificial Neural Network) 这一术语的起源于试图发现人脑进行信息处理的数学描述。现在神经网络的概念已经扩展和迁移到由生物神经网络灵感激发的一种计算模型。1958年麻省理工学院的Frank Rosenblatt创建了感知机 (perceptron) , 感知机也叫单层神经网络。它是一个二分类模型。如图2-1所示。

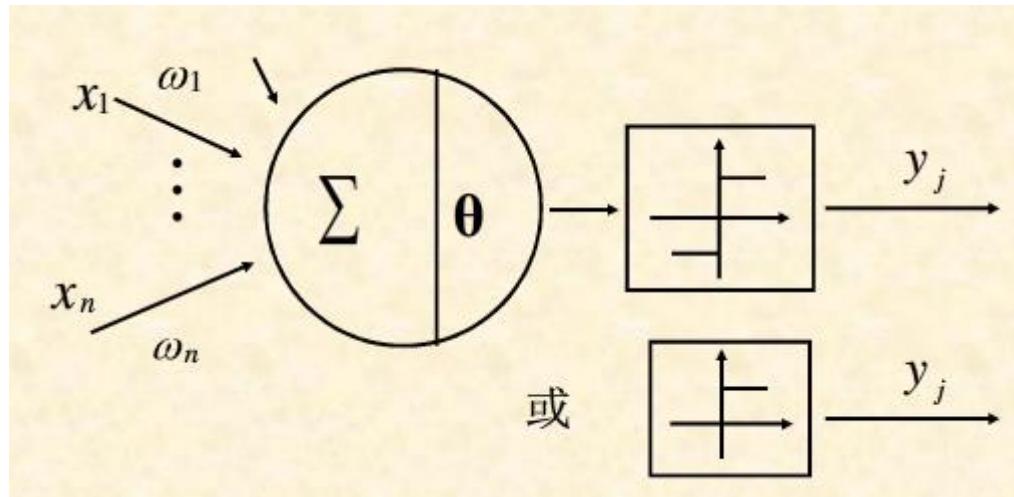


图 2-1 : 感知机

感知机的输入是向量 $x=(x_1, \dots, x_n)$, 输出是一个 0,1 或 -1,1 的值 y_j

$$s_j = \sum_{i=1}^n \omega_{ji} x_i - \theta_j$$

$$s_j = \sum_{i=0}^n \omega_{ji} x_i, \text{ where } (x_0 = \theta_j, \omega_{j0} = -1)$$

$$y_j = f(s_j)$$

其中 θ_j 为偏置 , f 为激活函数 (activation function)

$$f(s) = \begin{cases} 1, & \text{if } s > 0 \\ 0, & \text{otherwise} \end{cases}$$

或

$$f(s) = \begin{cases} 1, & \text{if } s > 0 \\ -1, & \text{otherwise} \end{cases}$$

然而 Minsky and Papert 在 1969 年的一篇论文中指出感知机的重大缺陷，它只能解决线性可分问题，它不能解决 XOR 问题。从此，神经网络的研究陷入了停滞。到 1975 年随着多层神经网络的诞生和反向传播算法的发明，神经网络的研究又迎来了一个高峰。

看图 2-2，有四个点对应输入 $x=\{(0,0), (1,0), (0,1), (1,1)\}$ 。每个点对应的类别标签由他们的颜色确定。感知机不可能找一条线，将这四个点分成两类。

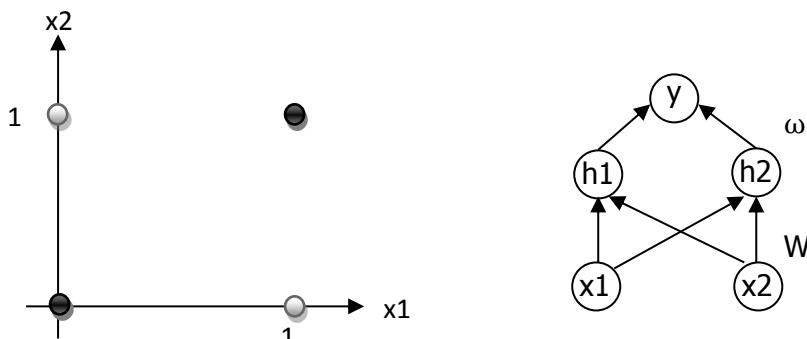


图 2-2：异或问题和加入隐层的多层感知机

而当神经网络加入隐层后

$$\text{设 } W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, c = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \omega = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, b = 0$$

隐层和输出层都使用 Relu 激活函数 $\omega^T \max\{0, W^T x + c\} + b$

考虑对应输入 $x=\{(0,0)^T, (1,0)^T, (0,1)^T, (1,1)^T\}$ ，输出 y 得到什么结果？

可以看到，对应的 y 是 0,1,1,0。该神经网络可以将输入数据正确分类。这是因为该神经网络加入了隐层后，将原始空间中的二维数据进行了非线性变换，映射到了一个新的空间。如图 2-3 所示。

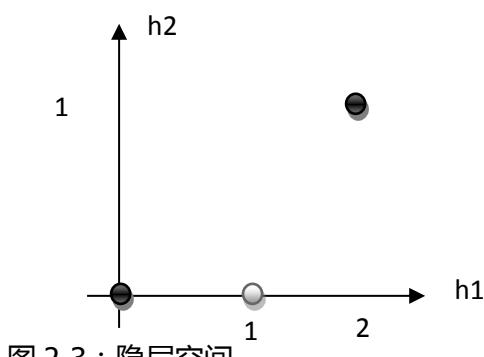


图 2-3：隐层空间

然而，到 90 年代神经网络的研究有开始缓慢，因为神经网络的一些缺陷。直到 2006 随着 Deep Learning 的诞生，神经网络又一次新生。

反向传播 BP 算法是神经网络最受欢迎的训练算法，但是当神经网络的层次增加时，BP 算法不能很好的训练模型。当层次增加时，训练神经网络的目标函数是一个非凸（non-convex）的函数。BP 算法通常陷入局部最优。而且当层数增加的越多，问题越严重。因此 2006 年以前的神经网络通常是浅层神经网络，它们最多包含两层非线性的特征转换（隐层）。在 2006 年，Hinton 提出了受限波兹曼机 RMB (restricted Boltzmann machines)，它有效的解决了更多层神经网络的学习问题。深度学习由此而来。

第二节：神经网络的结构

神经网络的结构包括神经元（neuron），连接神经元的有向有权重的边，

2.1.1 神经元

神经网络最基本的处理单元是神经元。一个单输入神经元如图 2-4 所示。标量输入 p （这里是以标量输入作为讨论，但实际中输入很多情况下是向量）乘上标量权重 w 得到 wp ，再将其送入累加器。另一个输入 1 乘上偏置 b ，再将其送入累加器。累加器输出 n 通常称为净输入，它被送入到一个激活函数 f ，在 f 中产生神经元的标量输出 a 。

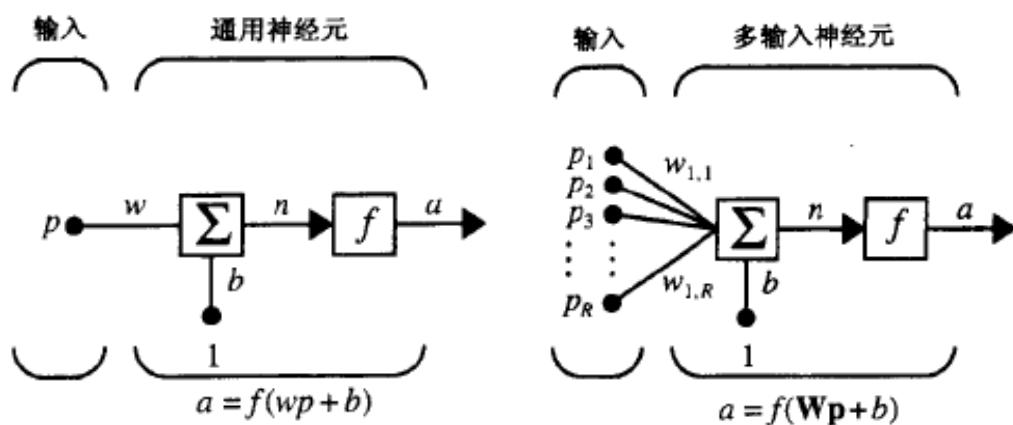


图 2-4：单输入神经元和多输入神经元

神经元按照下式计算

$$a = f(wp + b)$$

实际输出取决于激活函数。 w 和 b 是神经元待学习的参数。

通常神经元有不止一个输入。具有 R 个输入的神经元如图 2-4 所示。其输入 p_1, \dots, p_R 分别对应权重矩阵 W 的元素 $W_{1,1}, W_{1,2}, \dots, W_{1,R}$ 。该神经元只有一个偏置值 b，它与所有输入的加权和累加，从而形成净输入 n。

$$n = W_{1,1}p_1 + \dots + W_{1,R}p_R + b$$

这个表达式写成矩阵为

$$n = Wp + b$$

其中单个神经元的权重矩阵只有一行元素

神经元的输出可以写成

$$a = f(Wp + b)$$

神经网络通常可以用矩阵来描述

2.1.2 网络结构

一般来说，有多个输入的神经元并不能满足实际应用的要求，在实际操作中需要有多个并行操作的神经元，这些并行神经元组成的集合称为层。神经元的层是由 S 个神经元组成的单层网络。注意 R 个输入中的每一个均与每个神经元相连，权重矩阵现有 S 行。

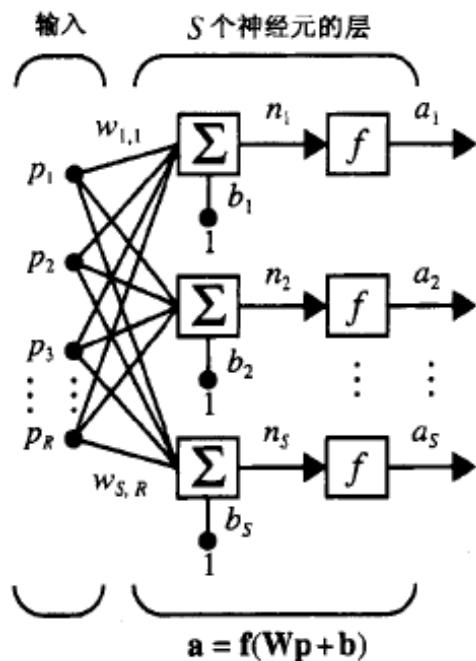


图 2-5：层

通常每层的输入个数并不等于该层中神经元的数目（即 $R \neq S$ ）。同一层中的神经元不必有相同的激活函数。输入向量（这里是指多个标量输入，组成一个输入向量）通过权重矩阵 W 进入网络。

多层神经元

现在考虑有几层神经元的网络，每层都有自己的权重矩阵 W ，偏置向量 b 、净输入向量 n 和一个输出向量 a 。图 2-6 是一个三层网络（这里输入未算作是一个层）

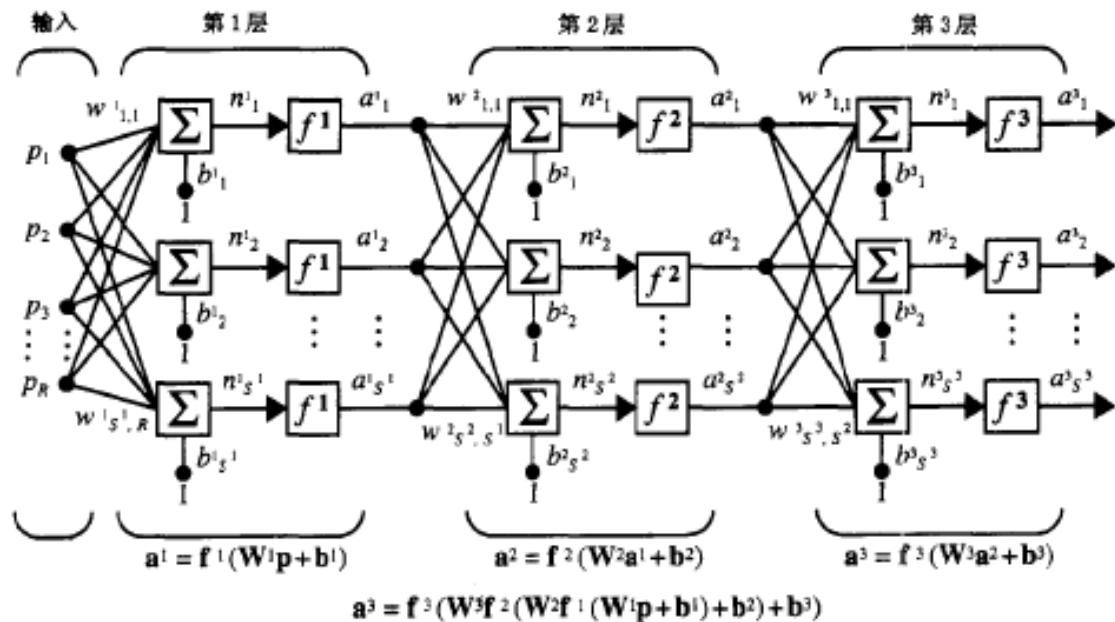


图 2-6：三层神经网络

图 2-6 所示第一层有 R 个输入， S^1 个神经元，第二层有 S^2 个神经元。不同层可以有不同的数目的神经元。第一层和第二层的输出分别是第二层和第三层的输入。如果某层是网络的输出，那么层该层为**输出层**。和其他层称为**隐层**。

多层网络的功能要比单层网络的功能强大许多。例如，如果第一层有 S 形激活函数，第二层具有线性传输函数的网络，经过训练可对大多数函数达到任意精度的逼近。而单层网络做不到这一点。

如此，决定一个网络的层数和神经元个数非常重要。首先，网络的输入和输出是有问题所定义的。所以，如果有 4 个外部变量作为网络输入，那么网络就有 4 个输入。同样如果网络又 7 个输出，则网络的输出层就有 7 个神经元。最后，输出信号所期望的特征有助于选择输出层的激活函数。如果一个输出是 -1 或 1，那么该输出神经元就可以用对称硬极限激活函数。对于确定多层网络中其他层的神经元个数并没有明确的确定方法。对于层数，普遍是小余 3 层。

对于偏置。是否使用偏置是可以选择的。偏置给网络提供了额外的变量，从而使网络有了更强的能力。

2.1.3 激活函数 (Activation Function)

激活函数也有人称为，活化函数或传输函数。激活函数可以是 n 的线性或非线性函数。可以用特定的激活函数满足神经元要解决的特定问题。

最简单的激活函数是 binary step function (有翻译做硬极限函数)。如果输入超出了一个预设的阈值，该函数的输出从一个值转变到另一个值，否则保持原值不变。

线性传输函数。它的输出等于输入 $a=n$ 。

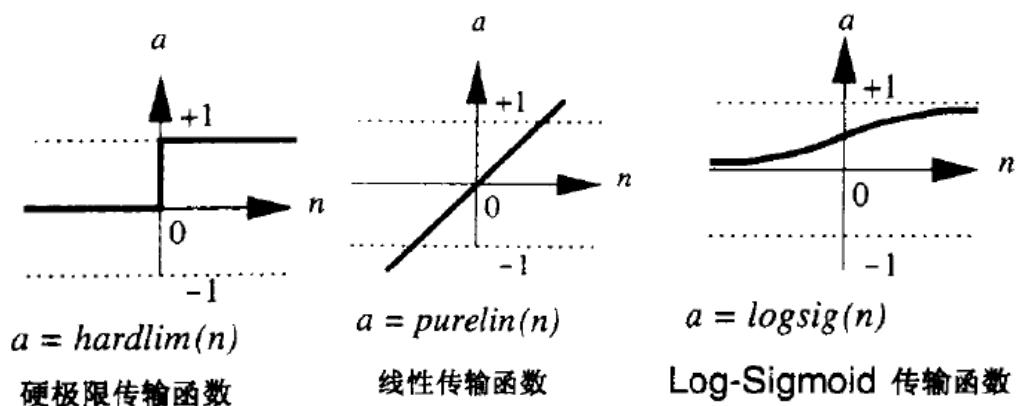


图 2-7：激活函数

对数 S 形激活函数，又称 log-sigmoid 或 sigmoid 函数。该输入在 $(-\infty, +\infty)$ 之间，输出则在 0,1 之间。其数学表达式为 $s(x) = \frac{1}{1+e^{-x}}$

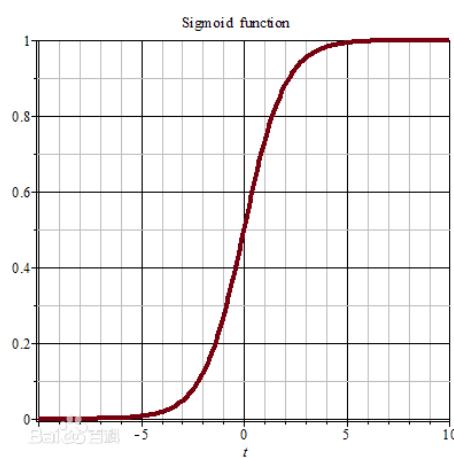
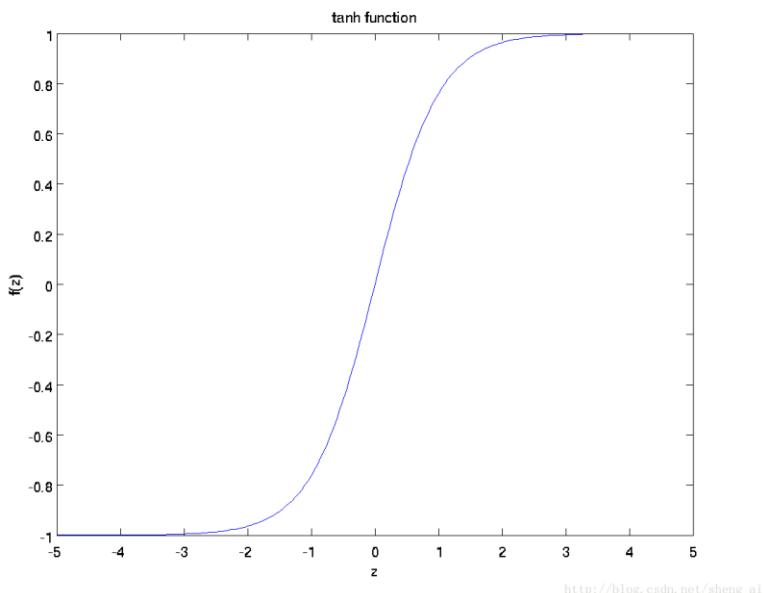


图 2-8：sigmoid 激活函数

从某种程度上说，正是由于 sigmoid 函数是可微的，所以用于反向传播算法训练的多层网络使用了该函数。

tanh 激活函数，将输出值压缩到-1~1 之间。



Softmax 函数，将一个任意实数值的 k 维向量 z 归一化到一个实数值在[0,1]的 k 维向量 $\sigma(z)$ ，满足向量元素的和为 1。

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

2.1.4 前馈神经网络

神经网络包含多种拓扑结构，如前馈神经网络，循环神经网络，自组织映射网络（Self Organizing Map, SOM）等。这里我们只讨论最流行的前馈神经网络。前馈神经网络（FeedForward Network）中各神经元从输入层开始，接收前一级输入，并输出到下一级，直至到输出层。整个网络中无反馈。它包含感知机（最简单的前馈网络），BP（反向传播）神经网络，RBF（Radial Basis Function，径向基函数）网络等。前面我们讨论的神经网络就是前馈神经网络。

第三节：神经网络的训练

机器学习通常模型训练的方法是：

We have a way of evaluating the **loss**, and now we have to **minimize** it. We'll do so with gradient descent. That is, we start with random parameters, and **evaluate** the gradient of the loss function with respect to the parameters, so that we know how we should change the parameters to decrease the loss.

BP 神经网络就是采用反向传播算法训练的前馈神经网络。讨论反向传播算法前，我们先看怎样训练一个感知机。

我们用一个例子来解释一个感知机的训练过程。问题描述如下：每天你去餐馆里吃早餐。每天的早餐点三样食物：小菜、包子和饮料。每天你点的这三样食物量不一样，收银员仅仅告诉你总价。几天后，可以用一个感知机来推算出各食物的价格。我们以一个线性激活函数的单神经元的感知机（即线性回归）为例。

$$y = \sum_i w_i x_i = W^T X$$

这里 y 是每天吃的早餐的价格。 W 是三样食物的价格， x 是每样食物的份量。训练模型，即得到模型的参数是一个优化过程。优化过程中通常需要建立一个目标函数。目标函数的建立有很多种，误差平方和是其中一种。

第一步，建立损失函数（目标函数）

$$E = \frac{1}{2} \sum_{n \in N} (t^{(n)} - y^{(n)})^2$$

$t^{(n)}$ 是训练集中一顿早餐的价格， $y^{(n)}$ 是用模型估计的价格。

第二步，用梯度下降方法来训练模型，目标函数对各参数求偏导获得梯度。对参数（当前是权重 w_i ）求偏导

$$g_i = \frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_n \frac{\partial y^{(n)}}{\partial w_i} \frac{dE}{dy^{(n)}} = - \sum_n x_i^{(n)} (t^{(n)} - y^{(n)})$$

第三步，更新权重。用梯度乘上学习率 η 来更新参数。通过多次迭代，最终算法收敛，得到最终的估计参数。

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i} = \eta \sum_n x_i^{(n)} (t^{(n)} - y^{(n)})$$

$$w_i \leftarrow w_i + \Delta w_i$$

感知机训练算法

Steps:

1. 初始化权重 W
2. Loop:
3. 初始 $\Delta W \leftarrow 0$
4. 对于训练集中的每个实例 n 完成下面的步骤

5. 计算输出 $y^{(n)} = f(Wx^{(n)})$
6. 累加 $\Delta w_i = \Delta w_i + x_i^{(n)}(t^{(n)} - y^{(n)})$
7. 更新权重 $W = W + \eta \Delta W$
8. END

这里 η 是学习率。

Python 代码如下：

```

eta=0.005
ws=[50,50,50]
train=((2,5,3),850),((1,4,7),1050),((2,3,5),950),((3,6,9),16
50),((7,4,1),1350))

for _ in range(500):
    y=[]
    d=[]
    delta_ws=[0,0,0]
    for xs,t in train:
        yn=sum([w*x for w,x in zip(ws,xs)])
        dn=t-yn
        delta_ws=[xi*dn+dw for xi,dw in zip(xs,delta_ws)]
    ws=[w+eta*delta_w for w,delta_w in zip(ws,delta_ws)]

print(ws)

```

运行结果：

[149.99854969104385, 50.00249323213941, 99.9987175643975]

下面我们再介绍随机梯度下降算法 (Stochastic Gradient Descent, SGD)。梯度是损失函数对每个参数求偏导得到的一个向量。梯度的每个元素是梯度的方向。SGD 在每次算法迭代中根据偏导值反向调整参数的变化。

SGD 的原理是，我们对一个函数求二阶导数。当 $f''(x) = 0$ 时可以获得该函数的极值。然而计算机只有通过数值计算的方式求解二阶导数。即，先随机初始化各参数；计算函数对各参数的一阶偏导 $u = f'(x_i)$ ，求得梯度；根据梯度反方向调整参数值 $x_{i+1} \leftarrow x_i - \eta u$ ；当 η 足够小， $f(x_{i+1})$ 比 $f(x_i)$ 更小。重复这个过程将可以发现最优的 x_i 。

随机梯度下降算法

输入：学习率 η

1. 初始化参数 θ
2. While (未达到停止迭代的标准) do

- a. 从训练集 $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$ 中抽样 m 个实例
 - b. 计算梯度 $\hat{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(x^{(i)}; \theta), y^{(i)})$
 - c. 更新参数 $\theta \leftarrow \theta - \eta \hat{g}$
3. End While

$f(x^{(i)}; \theta)$ 表示参数 θ 确定时将输入 $x^{(i)}$ 带入模型计算的结果。 $L(f(x^{(i)}; \theta), y^{(i)})$ 表示计算值和真实值的差值。当随机梯度下降算法中 m 取值 $m > 1$ 算法称为 minibatch SGD；当 $m=1$ 称为 Online GD。上面的步骤 2.b 求了样本的均值 ($1/m$)，求与不求均值效果是一样的，只需调整学习率 η 的大小。与前面感知机训练算法相比，随机梯度下降算法关键就是抽样 minibatch。

可以看到，对于 online GD，考察每条训练数据实例后，就计算损失函数的梯度，紧接着就更新参数。而前述的一般梯度下降算法是，考察完所有实例后计算损失函数的梯度，然后再更新参数。一个 Online GD 的例子如下：

当设定初始权重为 $[50, 50, 50]$ 。学习率 $\eta = 1/35$ 。当一天的早餐是：小菜 2 份，包子 5 个，饮料 3 杯，计算价格是 500，但实际价格是 850。差值是 $\text{error}=850-500=350$ 。计算 $\text{delta} = \eta(t^n - y^n)x^n = [1/35*350*2, 1/35*350*5, 1/35*350*3] = [20, 50, 30]$

则更新权重为 $[70, 100, 80]$ 。

Online GD 的 python 代码如下：

```

eta=1/35.0
ws=[50,50,50]

train=((2,5,3),850),((1,4,7),1050),((2,3,5),950),((3,6,9),1650),((7,4,1),1350))

for _ in range(100):
    for xs,t in train:
        y=sum([w*x for w, x in zip(ws,xs)])
        delta_ws=[eta*x*(t-y)for x in xs]
        ws=[w+delta_w for w,delta_w in zip(ws,delta_ws)]

print(ws)

```

运行结果如下：

`[149.9999854696171, 50.00004609118093, 99.99997795027406]`

关于随机梯度下降算法的一些讨论：

(1) 学习过程最终会得到完美答案吗？

很有可能得到的结果不是最优的。

(2) 权重收敛到正确值的速度有多快 ?

跟你的训练集有一定关系。如果输入向量的某些维度高度相关，会收敛的很慢。例如，前面的例子中，每天买的早餐包子、稀饭、小菜的比例都是相同的。几乎不会得到正确结果。

(3) online、minibatch 和标准梯度下降算法性能上有什么区别 ?

标准梯度下降每一轮迭代需要所有样本参与，对于大规模的机器学习应用，经常有 billion 级别的训练集，计算复杂度非常高。因此，有学者就提出，反正训练集只是数据分布的一个采样集合，我们能不能在每次迭代只利用部分训练集样本呢？这就是 minibatch 算法。

我们这里对 online GD 的解释是 minibatch 中的 $m=1$ 的情况。也有人将 online GD 描述为一条训练数据仅使用一次。随着互联网行业的蓬勃发展，数据变得越来越“廉价”。很多应用有实时的，不间断的训练数据产生。在线学习 (Online Learning) 算法就是充分利用实时数据的一个训练算法。Online GD 于 mini-batch GD/SGD 的区别在于，所有训练数据只用一次，然后丢弃。这样做好处是可以最终模型的变化趋势。

可以想象。标准梯度下降使用所有数据，还要迭代多次。如果有 10 万条数据，迭代 10 次。算法要计算 100 万次。Online ($m=1$) 每次只使用一条数据。迭代 10 万次，也才计算 10 万次。可见 online GD 的效率很高。然而随机梯度下降易受到噪声干扰，可能陷入局部最优。Minibatch GD 是两者的一个折中。

在 4.3 节我们对优化算法还有更多的讨论。

第四节：反向传播算法

2.2 节讲述的是训练一个感知机的算法。该算法并不适用于多层网络。Paul Werbos 在他 1974 年的论文中提出一个多层神经网络算法，称为反向传播算法。这里我们不讨论数学推导，只讲述算法的执行过程。多层网络中的某一层的输出是下一层的输入。以一个二层网络为例。

注：按照 Deep Learning 一书。反向传播只是一种计算梯度的方法，而实际的多层感知机的训练算法还是采用诸如前面讲的 SGD 等，优化算法。

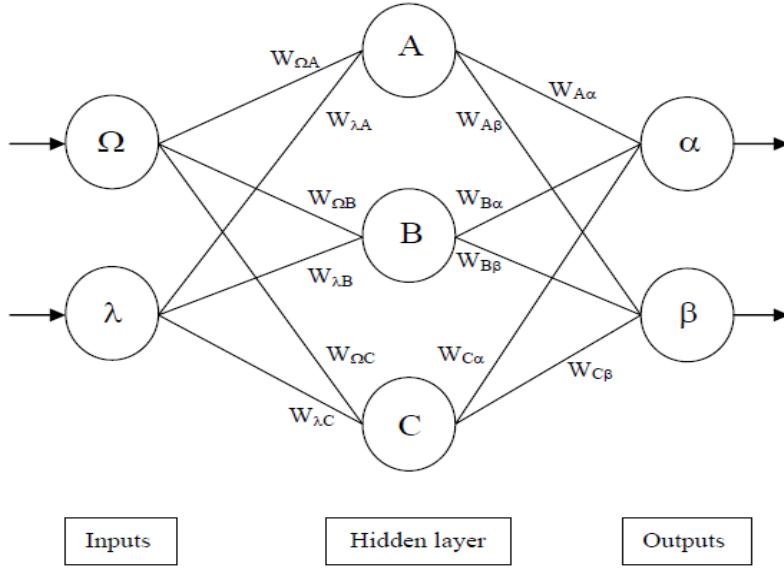


图 2-9：一个多层次神经网络

反向传播算法工作步骤如下：

1. 计算输出神经元的误差的梯度。

注意当激活函数是 sigmoid 函数时，采用下面的方法计算误差。

$$\delta_\alpha = \text{out}_\alpha(1 - \text{out}_\alpha)(\text{Target}_\alpha - \text{out}_\alpha)$$

$$\delta_\beta = \text{out}_\beta(1 - \text{out}_\beta)(\text{Target}_\beta - \text{out}_\beta)$$

如果激活函数是 binary step 函数则直接用 $\delta_\alpha = \text{Target}_\alpha - \text{out}_\alpha$

当采用梯度下降算法来优化参数，梯度等于误差函数对参数求偏导，例如

$$\frac{\partial E}{\partial W_{A\alpha}} = 2\delta_\alpha \text{out}_A$$

2. 改变输出层权重

η 是学习率，权重的更新函数如下

$$\begin{aligned} W_{A\alpha}^+ &= W_{A\alpha} + \eta \delta_\alpha \text{out}_A & W_{A\beta}^+ &= W_{A\beta} + \eta \delta_\beta \text{out}_A \\ W_{B\alpha}^+ &= W_{B\alpha} + \eta \delta_\alpha \text{out}_B & W_{B\beta}^+ &= W_{B\beta} + \eta \delta_\beta \text{out}_B \\ W_{C\alpha}^+ &= W_{C\alpha} + \eta \delta_\alpha \text{out}_C & W_{C\beta}^+ &= W_{C\beta} + \eta \delta_\beta \text{out}_C \end{aligned}$$

3. 计算隐层的误差梯度（反向传播）

先计算隐层的误差

$$\delta_A = \text{out}_A(1 - \text{out}_A)(\delta_\alpha W_{A\alpha} + \delta_\beta W_{A\beta})$$

$$\delta_B = \text{out}_B(1 - \text{out}_B)(\delta_\alpha W_{B\alpha} + \delta_\beta W_{B\beta})$$

$$\delta_C = \text{out}_C(1 - \text{out}_C)(\delta_\alpha W_{C\alpha} + \delta_\beta W_{C\beta})$$

$\delta_\alpha W_{A\alpha} + \delta_\beta W_{A\beta}$ 表示隐层的误差是由输出层的误差反向传播（乘上边的权重）得到的。

隐层误差的梯度

$$\delta_A \text{in}_\lambda$$

4. 改变隐层的权重

$$W_{\lambda A}^+ = W_{\lambda A} + \eta \delta_A \text{in}_\lambda \quad W_{\Omega A}^+ = W_{\Omega A} + \eta \delta_A \text{in}_\Omega$$

$$W_{\lambda B}^+ = W_{\lambda B} + \eta \delta_B \text{in}_\lambda \quad W_{\Omega B}^+ = W_{\Omega B} + \eta \delta_B \text{in}_\Omega$$

$$W_{\lambda C}^+ = W_{\lambda C} + \eta \delta_C \text{in}_\lambda \quad W_{\Omega C}^+ = W_{\Omega C} + \eta \delta_C \text{in}_\Omega$$

W^+ 代表更新后的权重

神经网络里面的计算单元，最重要的激活函数是连续的、可微的。比如常用的 sigmoid 函数，它是连续可微的。这使得可以容易地进行梯度计算，如此就可以使用 BP 算法来训练。通过这样的算法神经网络已经取得了非常多的胜利。

神经网络的训练依赖梯度。**如果一个问题不可微分的、不可计算梯度的，就无法为它训练一个神经网络。这非常重要。**而且在 2006 年之前，也没有人知道如何训练深度超过 5 层的神经网络，不是因为计算设施不够强大，而是我们无法解决梯度消失的问题。Geoffery Hinton 等人做出了巨大的贡献，他们表明通过逐层训练 (layer-by-layer precision) 和预训练 (pre-training) 我们可以克服梯度消失的问题。介绍这些都是为了说明神经网络需要可微的函数、需要能够计算梯度，这是最根本最重要的。虽然如今有一些研究不可微的函数的，但还是需要转换成某种可微的。

第三章：TensorFlow

在 windows 上安装 TensorFlow , 参见
https://tensorflow.google.cn/install/install_windows

TensorFlow 是实现 Deep Learning 的工具之一。按照官网介绍 : TensorFlow 是一个使用数据流图 (data flow graph) 进行数值计算的开源的软件库。图中的节点表示操作 (它可以是数学运算也可以是赋值等非数学运算操作) , 边表示沟通两个运算的多维数据阵列 (tensor, 又翻译做张量) 。数据流图灵活的结构允许用户通过 API 将计算部署到一个或多个 CPU 或 GPU 。 TensorFlow 最初被 Google 机器智能研究组的 Google Brain Team 的科学家开发 , 用于机器学习和深度神经网络的研究。但该系统可以适用于非常宽广的领域。

第一节：基本概念

要使用 TensorFlow 需要理解它的基本工作原理:

- (1) 用数据流图 (data flow graph , 也叫计算图 , computation graph) 来描述计算。
- (2) 在 Session 的环境下执行图
- (3) 用 tensor 来描述数据。
- (4) 用 Variable 对象来维护状态
- (5) 使用 feed 和 fetch 向任意 operation (运算) 放入数据或从运算得到数据。

TensorFlow 是一套编程系统或编程框架 , 用户可以将他的计算描述成图 , 称为数据流图或计算图。

数据流图

数据流图用一个有边和节点的有向图描述数学运算。节点实施运算 , 但也能描述导入 (feed in) 数据 , 导出运算结果 , 或读写持久变量 (persistent variable) 的端点。边描述了节点间的输入输出关系。边携带了 tensor 数据。TensorFlow 名称来自于 tensor 通过边的流动。节点被分配到可计算装置 (CPU 或 GPU) , 可以异步、并行执行。

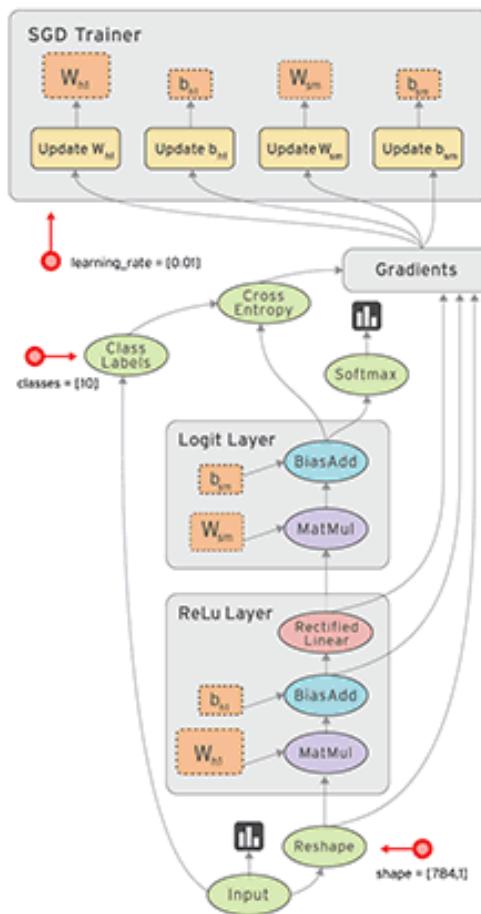


图 3.1 数据流图

在 TensorFlow 库被加载时，它会自动创建一个 Graph 对象，将其作为默认的数据流图。例如：

```
import tensorflow as tf

x=tf.Variable([10],name='foo')
y=x.assign(x*2)

init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init)
r1,r2 = sess.run([x,y])
print(r1,r2)
sess.close()
```

我们也可以明确地创建一个数据流图

```
g = tf.Graph()
```

然后利用 Graph.as_default()方法访问该图的上下文管理器，为该图添加操作。结合 with 语句，可利用上下文管理器通知 TensorFlow 我们需要将一些操作添加到特定的 Graph 对象中。例如：

```
import tensorflow as tf

g = tf.Graph()
with g.as_default():
    x=tf.Variable([10],name='foo')
    y=x.assign(x*2)
    init = tf.global_variables_initializer()

with tf.Session(graph=g) as sess:
    sess.run(init)
    r1,r2 = sess.run([x,y])
    print(r1,r2)
```

在 Graph.as_default()上下文管理器之外定义的任何操作都会被自动放到默认的数据流图中。大多数应用中只使用默认数据流图就够了。当定义多个相互之间不存在关系的模型，则创建多个 Graph 对象十分有用，此时应该不使用默认数据流图。

Operation (操作)

Tensorflow Operation 简称为 op。是计算图上的节点。一个 op 可以处理 0 个或多个 tensor；完成计算产生 0 个或多个 tensor。一个 tensor 是一个多维阵列，例如，可以将一组图片的描述成一个 4-D 浮点数阵列。其维度为[batch, height, width, channels]。从技术上讲，有些 Op 既无任何输入，也无任何输出。Op 的功能并不只限于计算，它还可用于如状态初始化这样的任务。

Session

一个 TensorFlow 的图是一个计算的描述。在计算前，该图需要在一个 session 中执行计算。session 可以理解成 tensorflow 在纯 python 环境中建立的一个计算环境。一个 session 将图的 operation (运算) 放置到 device (装置) 上，如 CPU，GPU 并提供执行运算的方法 (编程概念中的方法)。这些方法将运算结果即 tensor。纯 python 的环境是 session 外的环境。可以在 session 外获得运算结果，它们是以 python 的 numpy 包的 ndarray 对象返回。可以在 session 外的环境通过 feed 操作，向 session 的计算环境传递数据。

用 TensorFlow 范式写程序，实际上是向计算图上添加操作，数据。最后执行 session 的 run () 以后才会进行计算操作。

由 tf.Session()函数创建一个 Session 对象。它的构造方法包含三个参数：

target 指定了所要使用的执行引擎。对于大多数应用，该参数是默认的空串。在分布式设置中使用 Session 对象时，该参数用于连接不同的 tf.train.Server 的实例（即分布式中的服务器）；graph 参数指定了将要在 session 对象中加载的 Graph 对象，其默认值为 None，表示将使用当前默认数据流图。当使用多个数据流图时，最好的方式是显示的传入你希望运行的 Graph 对象；config 参数允许用户指定配置 session 对象所需的选项。如限制 CPU 或 GPU 的使用数目，为数据流图设置优化参数等。典型的 TensorFlow 程序中创建 session 对象无需改变任何默认参数。下面两个创建 session 对象的方式是等价的。

```
tf.Session()
```

```
tf.Session(graph=tf.get_default_graph)
```

一旦创建完 Session 对象，该对象的 run 方法可以计算图上的操作。run 方法参数设置如下：

```
run(  
    fetches,  
    feed_dict=None,  
    options=None,  
    run_metadata=None  
)
```

run 方法运行 fetches 参数中设置的 operation。Fetches 可以单独的一个 operation，可以是任意的嵌套的 list，tuple 或 dict. Feed_dict 是对应的输入元素。

在程序的最后应该调用 session 对象的 close 方法，关闭该对象，释放资源。也可以将 Session 对象作为上下文管理器加以使用，这样代码离开其作用于后，该 session 对象将自动关闭。

```
with tf.Session() as sess:  
    # 运行数据流图  
#session 对象自动关闭
```

Tensor (张量)

TensorFlow 使用一个 tensor 数据结构描述所有数据。在计算图上传递的只能是 tensor。一个 tensor 有一个静态类型，一个 rank，和一个 shape。

(1) Rank (阶)

在 TensorFlow 的计算系统中，rank 是 tensor 的一个维度单位。该 rank 不同于矩阵的 rank (秩)。Tensor rank (有时称为 order, degree or n-dimention) 是 tensor 维度的数量。0 阶张量即标量。例如，下面的 tensor (用 python 的 list 数据结构定义) 的 rank 为 2。

```
t = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
```

一个 rank 为 0 的 tensor 即一个标量；一个 rank 为 1 的 tensor 即一个向量；一个 rank 为 2 的 tensor 即一个矩阵。对于 rank 为 2 的 tensor，可以通过 $t[i, j]$ 访问它的元素；对于一个 rank 为 3 的 tensor t ，可以通过 $t[i, j, k]$ 来访问它的元素。

(2) Shape

TensorFlow 使用三类符号来描述 tensor 的维度。Rank, shape 和维度数。Shape 有的文章中文翻译为“形状”。我们还是使用英文。下表显示了三者的关系。

Rank	Shape	Dimension number	Example
0	[]	0-D	A 0-D tensor. A scalar.
1	[D0]	1-D	A 1-D tensor with shape [5].
2	[D0, D1]	2-D	A 2-D tensor with shape [3, 4].
3	[D0, D1, D2]	3-D	A 3-D tensor with shape [1, 4, 3].
n	[D0, D1, ..., Dn-1]	n-D	A tensor with shape [D0, D1, ..., Dn-1].

(3) 数据类型

除了维度，tensor 还有一个数据类型。用户可以创建下面的数据类型的 tensor。

Data type	Python type	Description
DT_FLOAT	tf.float32	32 bits floating point.
DT_DOUBLE	tf.float64	64 bits floating point.
DT_INT8	tf.int8	8 bits signed integer.
DT_INT16	tf.int16	16 bits signed integer.
DT_INT32	tf.int32	32 bits signed integer.
DT_INT64	tf.int64	64 bits signed integer.
DT_UINT8	tf.uint8	8 bits unsigned integer.

Data type	Python type	Description
DT_STRING	tf.string	Variable length byte arrays. Each element of a Tensor is a byte array.
DT_BOOL	tf.bool	Boolean.
DT_COMPLEX64	tf.complex64	Complex number made of two 32 bits floating points: real and imaginary parts.
DT_COMPLEX128	tf.complex128	Complex number made of two 64 bits floating points: real and imaginary parts.
DT_QINT8	tf.qint8	8 bits signed integer used in quantized Ops.
DT_QINT32	tf.qint32	32 bits signed integer used in quantized Ops.
DT_QUINT8	tf.quint8	8 bits unsigned integer used in quantized Ops.

Variables

Variables maintain state across executions of the graph. 当训练一个模型时，用 Variable 来保存和更新参数（ Variable 也可以设置为是某个全局变量，而不是模型训练的参数，详见 3.2 节的 Tips ）。使用 tf.Variable() 函数来创建一个 Variable 对象。

```
state = tf.Variable(0, name="counter")
```

state 是一个 Variable 对象，它是对那个名为“ counter ”的 tensor 的引用（注： name 参数是可选的，不写时 TensorFlow 会自动为它分配一个唯一的名称）。即内存从创建了一个 tensor 对象，然后分配了一个名字“ counter ”， state 保存这个对象的在内存中的地址。我们还可以根据名字使用 get_variable("counter") 获得该 tensor 。第二节的最后讲到 tf.get_variable() 时有个例子，两个 Variable 对象指向同一个 tensor 。

来看一段代码

```
import tensorflow as tf

# 创建一个Variable对象,该对象指向一个名为counter的tensor。这个
# tensor将被初始化到标量值0。State也是图上的一个Op
state = tf.Variable(0, name="counter")

# 创建一个操作,它完成对state这个Variable对象的加1
```

```

one = tf.constant(1)
new_value = tf.add(state, one)
# 创建一个操作，完成对state这个Variable对象赋值new_value
update = tf.assign(state, new_value)

# 在运行图之前，所有Variable应该被初始化，下面创建一个“初始化”操作
init_op = tf.global_variables_initializer()

# 运行图
with tf.Session() as sess:
    # 首先运行初始操作
    sess.run(init_op)
    # 获得state这个Variable对象的值
    s = sess.run(state)
    print(s)
    # 运行操作update和state，打印 'state'这个Variable对象的值
    for _ in range(3):
        r1, r2=sess.run([update, state])
        print(r1, r2)

```

练习：

(1) 分析一下上面程序的运行结果是什么？

把上面的代码更改成

```

for _ in range(3):
    r1=sess.run(state)
    r2=sess.run(update)
    print(r1, r2)

```

理解一下，运行结果

思考怎么理解这段代码：

```

x=tf.Variable(21)
y=x*2
print(x)
print(y)
init_op= tf.global_variables_initializer()
with tf.Session() as sess:

```

```
sess.run(init_op)
r=sess.run(y)
print(r)
```

这段代码的含义是：在默认的数据流图上创建了一个 Variable 对象 x 和 op 操作：y 和 init_op。x 是一个 variable 对象；y 则是一个操作，将 Variable 对象 x 的值乘上 2，并得到一个 tensor。r=sess.run(y) 表示运行操作 y 的运行结果（一个 tensor）放到了 r 对象中。此时的 r 才是 python 传统概念中的变量。print(x) 和 print(y) 两条语句在控制台上显示

```
<tf.Variable 'Variable:0' shape=() dtype=int32_ref>
Tensor("mul:0", shape=(), dtype=int32)
```

可以看出，x 是一个 Variable 对象（系统自动分配该对象对应到的 tensor 名为 Variable : 0）；y 是一个 tensor

本文中，若涉及‘变量’是指 python 传统意义中的变量，若谈到 Variable，是指 TensorFlow 中的 Variable 类的对象。

在使用 tensorflow 建模时，一个模型的参数需要定义为 Variables。例如，应该把神经网络的权重以 Variable 的形式存储成 tensor。在训练期间，可以通过重复的运行（run）训练图来更新这个 Variable。

上面的代码，session 那一段也可以这么写

```
with tf.Session() as sess:
    # Run the 'init' op
    sess.run(init_op)
    # Run the op that updates 'state' and print 'state'.
    fetches = {}
    fetches['update'] = update
    fetches['state'] = state
    for _ in range(3):
        r = sess.run(fetches)
        print(r['update'], r['state'])
```

这段代码把 operation 放到一个 dict 对象 fetches 中，返回的结果，也是一个 dict 对象 r。返回结果 r 的键的名称对应 fetches 键的名称，不过 r 中存储的值是 operation 运行的结果。

更新一个 Variable 的值，不能直接使用“=”，而应该使用 assign 函数。该函数的第一个参数是待更新的 Variable，或 Variable 的某个部分，第二个参数是值。

```
a = tf.Variable([[1,2,3],[4,5,6]], trainable=False)
update = tf.assign(a[0,0], 10)
```

```
init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init)
r = sess.run(update)
print(r)
```

Fetches

要获得运行结果，在一个 session 对象上用 run() 来执行图中的操作，来获得一个或多个 tensor (操作的运行结果)。前面的例子是获得了一个 tensor 的运行结果，下面的例子是获得了两个 tensor 的运行结果 (输出有两个结果)。

```
input1 = tf.constant([3.0])
input2 = tf.constant([2.0])
input3 = tf.constant([5.0])
# intermed = input2 + input3
intermed = tf.add(input2, input3)
mul = tf.multiply(input1, intermed)

with tf.Session() as sess:
    r1, r2 = sess.run([mul, intermed])
    print(r1)
    print(r2)
```

sess.run() 中其实只给出 mul 操作，intermed 操作也会被计算，但 intermed 的计算结果不会在 sess.run() 的返回结果中。上面把 intermed 也列为了 sess.run() 要运行的操作，因此，intermed 运算结果也输出了。

All the ops needed to produce the values of the requested tensors are run once (not once per requested tensor). (**这句话是说，上面的操作中，运行 mul 的操作会先运行 intermed 操作。但在运行 sess.run([mul, intermed])，intermed 操作不会再运行第二次，而是获得 intermed 运行的结果**)

Placeholder 和 Feeds

上面的例子通过存储数据在 Constant 和 Variables，引入数据到计算图 (Computation Graph)。TensorFlow 也提供了一个 feed 机制可以将数据在运行时喂给模型，即放到图中的操作中。前面的代码中定义一个 variable，并赋值一个常数

```
input1 = tf.constant([3.0])
```

在 tensorflow 中这是一个 operation。该操作在创建计算图时该值就固定了。现在可以用一个临时的对象替换它(这里是 tf.placeholder)，在实际运算时再赋值，即 feed。图

中定义了 placeholder，如果在 session 的 run 函数中运行的操作涉及到 placeholder，就必须在 run 函数中提供数据，否则会产生错误。（如果 run 函数运行的操作不涉及到 placeholder，即使图中定义了 placeholder，不会产生错误）

```
input1 = tf.placeholder(tf.float32,[None])
input2 = tf.placeholder(tf.float32, [None])
output = tf.multiply(input1, input2)

with tf.Session() as sess:
    r = sess.run(output, feed_dict={input1:[7.,3.],
input2:[2.,2.]})
    print(r[0])
    print(r[1])
```

Placeholder 是一个简单的 Variable，将在稍后的时候分配数据。它允许在没有数据的情况下先创建“操作”，建立“计算图”。然后可以通过这些 placeholder 向计算图 feed 数据。此时的

```
sess.run(output, feed_dict={input1:[7.,3.], input2:[2.,2.]})
```

需要两个参数，第一个是 Operation，第二个是 feed 数据。第二个参数的参数名是 feed_dict，它需要的值是 dictionary 对象。该 dictionary 对象中的“键”是 placeholder 名，对应的值就是要 feed 给计算图的值。

创建 placeholder 时，需要给出它的 shape。例如，

```
x = tf.placeholder("float", 3)
y = x * 2

with tf.Session() as session:
    result = session.run(y, feed_dict={x: [1, 2, 3]})
```



```
x=tf.placeholder(tf.int32, [None,3])
y=x*2

with tf.Session() as session:
    result=session.run(y,feed_dict={x:[[1,2,3]]})
    print(result)
```

第一段程序的 tensor 的 shape 是一维的。且向量长度为 3。在第二段程序，None 表示该 tensor 的 shape 是二维的，但第一个维度不定，根据实际输入的数据来定。

在 feed_dict 两段程序的 feed 数据是不一样的。

运算符

TensorFlow 对常见的运算符进行了重载。使用运算符的表达式也是数据流图上的一个操作。

运算符	相关 TensorFlow 运算	描述
$-x$	<code>tf.neg()</code>	返回 x 中每个元素的相反数
$\sim x$	<code>tf.logical_not()</code>	返回 x 中每个元素的逻辑非。只适用于 <code>dtype</code> 为 <code>tf.bool</code> 的 tensor 对象
$\text{abs}(x)$	<code>tf.abs()</code>	返回每个元素的逻辑值
$x+y$	<code>tf.add()</code>	将 x 和 y 逐个元素相加
$x-y$	<code>tf.subtract()</code>	将 x 和 y 逐个元素相减
$x*y$	<code>tf.multiply()</code>	将 x 和 y 逐个元素相乘
x/y	<code>tf.divide()</code>	将 x 和 y 逐个元素相除, 返回浮点小数
x/y	<code>tf.div(x,y)</code>	将 x 和 y 逐个元素相除, 返回整除
$x\%y$	<code>tf.mod()</code>	取模
$x^{**}y$	<code>tf.pow(x, y)</code>	逐一 x 中的元素为底, y 相应元素为指数
$x < y$	<code>tf.less()</code>	逐元素的计算 $x < y$ 的真值表
$x \leq y$	<code>tf.less_equal()</code>	逐元素的计算 $x \leq y$ 的真值表
$x > y$	<code>tf.greater()</code>	逐元素的计算 $x > y$ 的真值表
$x \geq y$	<code>tf.greater_equal()</code>	逐元素的计算 $x \geq y$ 的真值表
$x \& y$	<code>tf.logical_and()</code>	逐个元素的逻辑与运算, <code>dtype</code> 必须是 <code>bool</code>
$X y$	<code>tf.logical_or()</code>	逐个元素的逻辑或运算, <code>dtype</code> 必须是 <code>bool</code>
x^y	<code>tf.logical_xor()</code>	逐个元素的异或运算, <code>dtype</code> 必须是 <code>bool</code>

第二节：Variable 的创建、初始化、存储与装载

更详细的内容请参考 TensorFlow 的 API。

创建和初始化

使用 `tf.Variable` 类的构造方法创建 Variable 对象。当用户创建一个 Variable 时，需要初始化该 variable。可以传递一个 tensor 到构造方法 `Variable()` 进行初始化。例如：

```
W = tf.Variable(tf.zeros([784, 10]), name="foo")
```

```
b = tf.Variable(tf.zeros([10]))
```

`tf.zeros([784, 10])`就是在产生一个 784×10 的 tensor，值均为 0。上面的语句给这个 tensor 分配了一个名称为“foo”，W 是一个 Variable 对象（也可以理解为图上的一个操作），它是对这个名为 foo 的 tensor 的引用。

TensorFlow 提供一个产生 tensor 的函数集合，可以产生并初始化 tensor.

Constant value tensor	描述
<code>tf.zeros(shape, dtype=tf.float32, name=None)</code>	产生一个所有元素为 0 的 tensor
<code>tf.zeros_like(tensor, dtype=None, name=None)</code>	给定一个 tensor 作为参数，返回一个和该 tensor 有相同类型和 shape，但值为 0 的 tensor
<code>tf.ones(shape, dtype=tf.float32, name=None)</code>	创建一个 tensor 所有的元素为 1
<code>tf.ones_like(tensor, dtype=None, name=None)</code>	给定一个 tensor 作为参数，返回一个和该 tensor 有相同 shape，但值为 1 的 tensor
<code>tf.fill(dims, value, name=None)</code>	创建一个 tensor 它的元素用一个标量值填充
<code>tf.constant(value, dtype=None, shape=None, name='Const')</code>	创建一个 tensor，它的值和类型和维度，由参数设定
Sequence	
<code>tf.linspace(start, stop, num, name=None)</code>	从 start 的值开始，到 stop 结束，产生 num 个值
<code>tf.range(start, limit=None, delta=1, name='range')</code>	从 start 开始，以 delta 为步长，到 limit 结束（不包括）创建一个整数序列。如果只是 <code>tf.range(n)</code> 则创建从 0 开始，步长为 1，到 n(不包括)的序列
Random Tensors	创建不同分布的随机数
<code>tf.random.normal(shape, mean=0.0, stddev=1.0, dtype=tf.float32, seed=None, name=None)</code>	产生符合正态分布的随机数
<code>tf.random.uniform(shape, minval=0, maxval=None, dtype=tf.float32, seed=None, name=None)</code>	产生均匀分布的随机数

上表未列举所有的函数，详见

https://tensorflow.google.cn/api_guides/python/constant_op

例 1：创建一个 tensor, 初始值是正太分布的随机数；shape 是 $[1, 3]$, 并显示第一个元素值

```
w= tf.random.normal([1,3], mean=0.0, stddev=1.0,
dtype=tf.float32)
```

```
sess=tf.Session()
rw=sess.run(w)
print(rw[0,0])
```

例 2：创建一个序列，该序列从 1 到 4 产生 5 个数，因为如此产生的是实数，因此给的产生是 1.0 和 4.0，而不是 1 和 4

```
sess = tf.Session()
w=tf.linspace(1.0,4.0,5)
rw=sess.run(w)
print(rw)
```

创建 variables 时，作为参数的 tensor 的 shape 就是 Variable 的 shape。Variable 的 shape 通常是固定的，但 TensorFlow 也提供了函数进行修改，这里不讨论。创建了 Variable 的操作，需要再创建初始化操作来初始化 Variable。看下面代码：

```
import tensorflow as tf

weights = tf.Variable(tf.random.normal([784, 200],
                                      stddev=0.35),
                      name="weights")
init = tf.compat.v1.global_variables_initializer()
with tf.compat.v1.Session() as sess:
    sess.run(init)
    print(sess.run(weights))
```

创建了一个定义 Variable 的操作，同时还需要定义一个初始化 operation。然后用 session 的 run () 方法来运行这些操作。由 tf.Variable() 返回的值实际上是 tf.Variable 类的实例，即 tensor 对象一个 Variable 可以放到一个指定的装置（CPU 或 GPU）上，此处我们不做讨论。

创建 Variable 时，也可以从其他 Variable 来产生初始值，即通过其他 Variable 的 read_value() 方法获得值

```
import tensorflow as tf

weights = tf.Variable(tf.random.normal([784, 200], stddev=0.35),
                      name="weights")
# Create another variable with the same value as 'weights'.
w2 = tf.Variable(weights.read_value(), name="w2")
# Create another variable with twice the value of 'weights'
```

```
w_twice = tf.Variable(weights.read_value() * 2.0, name="w_twice")
init = tf.compat.v1.global_variables_initializer()

sess = tf.compat.v1.Session()
sess.run(init)
r=sess.run(w_twice)
print(r)
```

Variable 的赋值和修改

在 tensorflow 中，一个 Variable 对象的修改是通过该变量的 assign 方法。例如

```
import tensorflow as tf

input1 = tf.Variable(tf.constant([3.0]))
input2 = tf.constant([2.0])
input3 = tf.constant([5.0])
intermed = tf.add(input2, input3)
mul = tf.multiply(input1, intermed)
init = tf.compat.v1.global_variables_initializer()
op=input1.assign([2])
with tf.compat.v1.Session() as sess:
    sess.run(init)
    sess.run(op)
    r1, r2 = sess.run([mul, intermed])
    print(r1)
    print(r2)
```

需要记住的是，`op=input1.assign([1])`也是一个操作。需要用 session 的 run 方法来执行，才能完成赋值操作。

Variable 对象也有自增和自减操作：`assign_add(1)`, `assign_sub(1)`.

```
x=tf.Variable([2])
y=x.assign_add([3])

init_op= tf.global_variables_initializer()
with tf.Session() as sess:
    sess.run(init_op)
    r1,r2=sess.run([y,x])
    print(r1, r2)
```

输出结果是

```
[5] [5]
y=x.assign_add([3])
```

y 是一个操作，做对 Variable 对象 x 加 3 的自增。y 的操作结果是一个 tensor, [5]。x 的值也变成了[5]。

修改一个 tensor 的部分值

Tips:

当建立机器学习模型时，通常把模型的参数建成 Variable，它是模型要训练的部分。有时我们也需要全局的 variable。例如，对学习的步数进行计数的 Variable。它不能作为模型的参数。这时在创建 Variable 时，加上一个参数 trainable=<bool>。如果参数为 True，则新的 variable 被加到 graph collection；如果为 False，则不会把 Varialble 加入。前面讲的都是作为模型参数的 Variable，下面是一个全局变量的 Variable。

```
global_step = tf.Variable(0, name="global_step", trainable=False)
```

存储与装载

最简单的存储一个模型的方法是使用 tf. train. Saver 对象

关于 tf.Variable_scope(), tf.name_scope(), tf.Variable(), tf.get_variable()

命名空间是 tensorflow 按照体系结构组织 Variable 对象和操作名的方式。

tf.name_scope 在默认图上为“操作” 创建命名空间

tf.variable_scope 在默认图上为“Variable”和“操作”创建命名空间

tf.get_variable 创建一个新的 Variable，或者获得一个已经创建的 Variable

tf.Variable 创建一个新的 Variable。tf.name_scope 是在使用 tf.Variable()创建变量时，给变量分配命名空间，例如：

```
with tf.name_scope('conv1'):
    weights1 = tf.Variable([1.0, 2.0], name='weights')
print (weights1.name)
```

运行结果是：

conv1/weights:0

代码 `weights1 = tf.Variable([1.0, 2.0], name='weights')` 中，
weights1 是 tensorflow 数据流图上的一个操作 op，该操作定义了一个变量，名字是
weights，该变量的 shape=[1.0, 2.0]

上面的代码含义是在命名空间 conv1 下创建变量 weights。

`tf.get_variable` 和 `tf.variable_scope` 一起使用，可以创建新 Variable 或获取已经创建的
Variable (Variable 共享)。

```
with tf.variable_scope('conv2') as scope:  
    weights2 = tf.get_variable("weights", shape=[4.0, 2.0])  
print (weights2.name)  
  
with tf.variable_scope('conv2', reuse=True) as scope:  
    w3=tf.get_variable("weights", shape=[4.0, 2.0])  
print(w3.name)
```

运行结果是：

```
conv2/weights:0  
conv2/weights:0
```

操作 weight2 和 w3 获得两个相同的 Variable 的 tensor。

第三节：初识 TensorFlow：手写数字识别

参看 `mnist4beginner.py`。

MNIST 是一个手写数字的数据集。它包含的图片如下图：

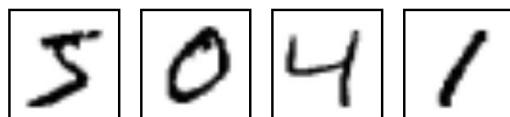


图 3.2 mnist 数据集中的图片示例

每个图片也被分配了一个标签，即图片对应的数字。在本节我们将建立一个预测模型，给定一张图片预测它对应的数字。本节不讨论如何训练一个性能最优的分类器（像是第一节所介绍的），只是探讨如何用 TensorFlow 完成该工作。这里将构建 4。

MNIST 数据集可以在 Yann LeCun's website 的网站下载
<http://yann.lecun.com/exdb/mnist/>。但 TensorFlow 已经预装载了该数据集，使用下

面的代码即可获得该数据集。（注：这段代码是从网络上下载数据集，如果网速不好会影响运行）

```
from tensorflow.examples.tutorials.mnist import input_data
mnist = input_data.read_data_sets("MNIST_data/", one_hot=True)
```

获得的该数据集包含三个部分：55000条训练数据（`mnist.train`），10000条测试数据（`mnist.test`）和5000条校验数据集（`mnist.validation`）。一条mnist数据包含两个部分：手写数字图片和对应的标签。见图3.3。这里称图片`xs`，标签`ys`。

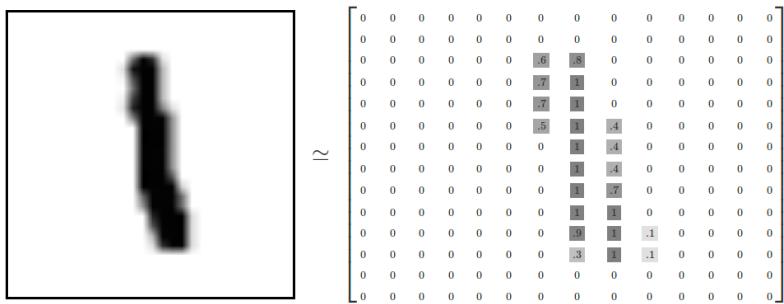


图 3.3 图片的矩阵描述

`mnist.train.images` 就是 `xs`；`mnist.train.labels` 就是 `ys`。每张图片是一个 28*28 像素的矩阵。它们被转换成了 $28 \times 28 = 784$ 长度的向量。

因此，`mnist.train.images` 是一个 `shape=(55000, 784)` 的 tensor。第一个维度是图片的编号，第二个维度是每个图片的像素。像素值在 0-1 之间取值。

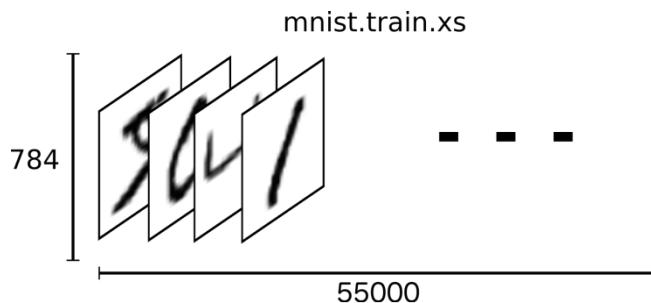


图 3.4

MNIST 的标签 `label` 数据是一个向量，描述 0-9 中的一个值，称为 one-hot 向量。one-hot 向量中只有一个元素的值是 1，其他都是 0。‘1’ 对应该图片对应的数字。例如，标签 3 的 one-hot 向量 [0,0,0,1,0,0,0,0,0]。相应的 `mnist.train.labels` 是一个 [55000, 10] 的矩阵。

Softmax Regression

每张 mnist 的图片对应 0-9 中的一个数字。预测模型应该对输入的图片给出对应每个数字的概率，例如，给出一张图片是 ‘9’ 的概率是 80%，是 ‘8’ 的概率是 5%。在神经网络的多分类任务中，输出层经常选用 softmax 激活函数，因为它可以给出输入对应每个类的概率。

在当前例子中，Softmax Regression 的网络描述如图 3.5：（此处假设，输入向量的维度是 3，输出的类别个数也是 3）

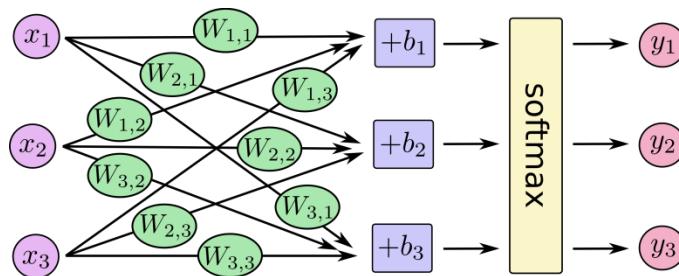


图 3.5 softmax regression 模型

Softmax 其实就是将一组数据求指数后规范化的一个操作，其数学描述是：

给定一组数据 $z = \{z_1, z_2, \dots, z_n\}$

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j \in n} \exp(z_j)}$$

上图的数学描述就是

$$y = \text{softmax}(Wx + b)$$

W 是权重矩阵， b 是偏置向量， x 是输入向量。

实施 softmax 模型

要使用 TensorFlow 首先 import 该包

```
import tensorflow as tf
```

因为在运行计算图时，需要将训练数据 feed 给模型，因此创建 placeholder

```
x = tf.placeholder(tf.float32, [None, 784])
```

因为，图片数据已经被转换成了向量，shape 中的列数是 784。此处 None 表示不限定输入数据的个数。

该模型中需要权重和偏置（bias）。因此，创建 Variable。Variable 可以理解为值可更改的 tensor。图计算的过程中使用，并会更改它。在使用 TensorFlow 进行机器学习时，模型的参数必须设置成 Variable。

```

W = tf.Variable(tf.zeros([784, 10]))
b = tf.Variable(tf.zeros([10]))

```

`tf.zeros([784, 10])`是创建一个 tensor 它的 shape 是`[784, 10]`，它作为 Variable W 的初始值。这里的权重的 Shape 是`[784,10]`，因为有 784 个输入，10 个输出。偏置 Variable b 的 shape 是 10，因为给 10 个输出，每个加上一个偏置。

现在可以实施模型：

```

y = tf.nn.softmax(tf.matmul(x, W) + b)

```

`tf.matmul`是矩阵乘。加上偏置后，进行 softmax 的操作。得到的 y 就是每个输入对应到每个输出的概率矩阵。

训练模型

要训练模型，首先需要定义什么样的模型是好的或者差的，即定义代价或损失函数。
交叉熵是一种损失函数，其定义为

$$H_{y'}(y) = - \sum_i y'_i \log(y_i)$$

y 是预测的概率分布，y' 是真实的概率分布。简单的说，交叉熵定义了模型是多么的无效率。实施交叉熵，需要定义一个 placeholder，label 数据将在运行阶段放在该 placeholder。

```

y_ = tf.placeholder(tf.float32, [None, 10])

```

然后实施交叉熵

```

cross_entropy = tf.reduce_mean(-tf.reduce_sum(y_ * tf.log(y),
reduction_indices=[1]))

```

`y_*tf.log(y)`进行两个矩阵的点乘（注意，两个矩阵相乘用的是 `mul` 函数）。
`reduction_indices=[1]`规定了 `tf.reduce_sum` 在矩阵点乘的结果的列的方向进行求和。
计算的结果是一个 `shape=[n, 1]` 的 tensor。即每个输入被计算了一个交叉熵。
`Reduce_mean` 对这 n 个交叉熵求平均值，得到模型的损失值。

定义了损失函数，就可以继续定义优化操作。TensorFlow 可以根据定义的计算图，进行优化操作训练模型，从而得到估计的模型参数。因此，训练操作就是一个优化操作，我们将优化器的定义和优化操作写在一起，来定义训练操作如下：

```

train_step =
tf.train.GradientDescentOptimizer(0.5).minimize(cross_entropy)

```

此处我们采用的是梯度下降的方法训练模型，学习率是 0.5。TensorFlow 也提供了其他的优化器。

在上面定义操作的背后，实际上是把“操作”添加到一个默认的计算图，它实施反向传播算法（默认实施）和梯度下降优化算法来训练模型。（第二章介绍的BP算法没介绍数学推导过程。实际上神经网络将输入先正向传播；神经网络在每个层计算梯度，但此时需要一个敏感度，敏感度的计算是从输出反向传播过来的；每一层有计算的梯度更新权重）

在开始正式训练模型之前，需要一个初始化操作。

```
init = tf.initialize_all_variables()
```

并在运行训练操作之前，运行初始化操作。

```
sess = tf.Session()  
sess.run(init)
```

训练模型的过程如下

```
for i in range(1000):  
    batch_xs, batch_ys = mnist.train.next_batch(100)  
    sess.run(train_step, feed_dict={x: batch_xs, y_: batch_ys})
```

在每一次训练的迭代中，仅从训练数据集抽去100条训练数据，称为一个batch。将一个batch的训练数据中的数据和label分别feed到训练操作中。它们实际上是feed给了训练操作中用到的placeholder。

使用部分训练集的数据训练模型称为，随机训练（stochastic training），此例中即是第二章中我们提到的随机梯度下降训练。理想情况下，可以一次使用所有的训练数据在每一次迭代中训练模型，但这种方法代价太大。采用随机训练，方法简便，需要的代价小，但训练效果不变。

评估模型

tf.argmax给出一个tensor中，沿着某个维度最高值的索引下标。例如，
tf.argmax(y, 1)是我们的模型认为输入最可能属于某个类的标签。tf.argmax(y_, 1)是正确的标签。使用tf.equal来检测两个值是否相等

```
correct_prediction = tf.equal(tf.argmax(y, 1), tf.argmax(y_, 1))
```

这里又是再定义operation。其中的y是在前面已经定义了的操作。correct_prediction存储了每条输入数据是否被正确预测的判断，其值是[True, False, True]的数据。

进一步我们可以计算精确率

```
accuracy = tf.reduce_mean(tf.cast(correct_prediction, tf.float32))  
tf.cast将True转换成1，False转换成0。
```

最后，输出预测结果。这里又是运行定义的操作。

```
res = sess.run(accuracy, feed_dict={x: mnist.test.images, y_: mnist.test.labels})
print(res)
```

此时运行 accuracy 操作时，accuracy 操作运行了 correct_prediction 操作；而 correct_prediction 操作又运行了 y 操作；y 操作

```
y = tf.nn.softmax(tf.matmul(x, W) + b)
```

y 操作又运行了矩阵乘，此时的 W 保存的是已经训练出的模型参数。

第四节：使用 TensorFlow 构建神经网络的步骤

1. 定义训练数据和标签数据
2. 为训练数据和标签数据定义 placeholder
3. 考虑好有几个隐层，然后定义相应的权重 Variable
4. 定义从输入，到隐层，最后到输出的数据矩阵乘操作。
5. 定义损失计算操作，该操作就是将标签数据与输出层的结果进行损失计算
6. 定义优化器
7. 定义训练操作，就是用优化器优化损失操作
8. 创建 Session 的对象，用该对象的 run 方法运行训练操作。并将训练数据和标签 feed 给训练操作

第五节：TensorFlow 练习（1）：实现感知机

这一节我们通过用 TensorFlow 来实现 2.2 节的感知机例子，来认识 TensorFlow。

用 TensorFlow 开发机器学习算法，关键要定义几个关键 ops: 损失函数，优化器和训练操作。损失函数，实际上是定义了从输入得到最终输出结果的操作，然后定义预测结果和实际结果的损失函数，如可以用误差平方和。设 iPh 是输入，weight 是边权重矩阵，y 是模型计算的输出，lPh 是训练集标签

```
y=tf.matmul(iPh, weight)
loss=tf.reduce_mean(tf.square(y-lPh))
```

优化器可以选择 TensorFlow 提供的优化器

```
optimizer = tf.train.GradientDescentOptimizer(eta)
```

训练操作就是将损失函数带入优化器

```
train_op = optimizer.minimize(loss)
```

训练的过程就是不断的迭代执行 train 操作。迭代过程中，参与运算的 Variable 被当做是模型的参数，如程序中的 weights，会在迭代过程中安装相应优化算法的更新公式更新。此例中，就是 weight<-weight + 梯度

```
import tensorflow as tf
import numpy as np

train=((2.0,5.0,3.0),850.0),((1.0,4.0,7.0),1050.0),((2.0,3.0
,5.0),950.0),
      ((3.0,6.0,9.0),1650.0),((7.0,4.0,1.0),1350.0))
labels=[]
dat=[]
eta=0.005 # learning rate
w=[50.0,50.0,50.0]

for x, l in train:
    labels.append(l)
    dat.append(x)

labels=np.reshape(labels,[-1,1])
size=len(dat)
dim=len(dat[0])

# training
with tf.Graph().as_default():
    lPh=tf.placeholder(tf.float32,[None,1])
    iPh=tf.placeholder(tf.float32,[None,dim])

    weight=tf.Variable(tf.constant(w,shape=[dim,1]),dtype=tf.flo
t32)

    y=tf.nn.relu(tf.matmul(iPh, weight))
    loss=tf.reduce_mean(tf.square(y-lPh))

    optimizer = tf.train.GradientDescentOptimizer(eta)
    train_op = optimizer.minimize(loss)
```

```

init = tf.global_variables_initializer()
sess=tf.Session()
sess.run(init)

for _ in range(1,1000):
    feed_dict={iPh:dat,lPh:labels}
    _,val=sess.run([train_op,loss],feed_dict=feed_dict)
    print(val,sess.run(weight))

d=sess.run(weight)
print(d)

```

使用 TensorFlow 的 Tips :

- (1) 用 TensorFlow 开发机器学习程序时，通常会有多次迭代，迭代过程中参与运算的 Variable 的值被当做模型的参数，它的值会改变。
- (2) 在调用 session 的 run 方法进行训练模型前，定义一个 Variable 初始化操作，然后用 session 的 run 方法执行该操作
- (3) 要想查看一个 Variable 的内容，调用 session 的 run 方法运行它，其返回结果就是内容

对“默认图”的理解：

我们创建 TensorFlow 程序时，默认的会有个 Graph，只要我们创建操作 op，都是在向这个图添加操作（TensorFlow 中把创建 Variable，Constant 也理解成是 operation）。上面的 with tf.Graph().as_default():语句是创建一个图，作为默认图（我们不要这条语句程序也可以允许）。如果需要创建多个图时，需要使用该方法。

1. 创建一个图，并作为默认图

```

g = tf.Graph()
with g.as_default():
    c = tf.constant(5.0)
    assert c.graph is g

```

2. 再构建一个图作为默认图

```

with tf.Graph().as_default() as g:
    c = tf.constant(5.0)
    assert c.graph is g

```

我们定义的操作，都会添加到当前的默认图上的。

第六节：TensorFlow 练习（2）：曲线拟合

该例子中，我们将构建具有一个隐层的神经网络，进行曲线拟合（或函数逼近）。在神经网络中，函数逼近通常隐层采用 sigmoid 激活函数，而输出层采用线性输出函数。曲线是一段正弦波曲线，并加上了随机噪声。

程序如下：

```
import numpy
import tensorflow as tf
import math

# generate data
x=numpy.arange(0,6.3,0.1)
y=[math.sin(val) for val in x]
s = numpy.random.normal(0, 0.1, len(x))
z=[sum(x2) for x2 in zip(y,s)]
x=numpy.array([x])
z=numpy.array([z])
size=len(z)    # the number of instances
idim=1          # the dim of input
wsize=7          # the number of nodes in hidden layer
eta=0.06

with tf.Graph().as_default():
    iph=tf.placeholder(tf.float32, [idim, None])
    lph=tf.placeholder(tf.float32, [idim, None])
    w1=tf.Variable(tf.random_uniform((wsize,idim), 0, 1, dtype=tf.float32))
    w2=tf.Variable(tf.random_uniform((idim, wsize), 0, 1, dtype=tf.float32))
    b1 = tf.Variable(tf.zeros([wsize,1]))
    b2 = tf.Variable(tf.zeros([1]))

    hidden1=tf.nn.sigmoid(tf.matmul(w1, iph)+b1)
    modle =tf.matmul(w2, hidden1)+b2

    loss=tf.reduce_mean(tf.square(modle-lph))
    step = tf.Variable(0, trainable=False)
    rate = tf.train.exponential_decay(0.15, step, 1, 0.9999)
    optimizer = tf.train.AdamOptimizer(rate)

    train_op = optimizer.minimize(loss)

    init_op=tf.initialize_all_variables()
    sess=tf.Session()
    sess.run(init_op)

for step in range(1,80000):
    feed_dict={iph:x,lph:z}
    _,val=sess.run([train_op,loss],feed_dict=feed_dict)
    if step % 1000 == 0:
        print(val)
```

在上面的程序中，我们将训练数据和标签数据转换成了 numpy 的 array 数据结构，如此就可以将数据直接 feed 给 placeholder。用普通 Python 的 list 数据结构则不能。如 3.3 节的程序所示。

TensorFlow 中可以设置动态调整学习率。如，此例子中

```
step = tf.Variable(0, trainable=False)  
rate = tf.train.exponential_decay(0.15, step, 1, 0.9999)
```

动态调整学习率可以使得收敛速度更快

不同的优化算法对学习性能有很大影响，例如，试一试下面两种优化算法

```
optimizer = tf.train.GradientDescentOptimizer(rate)  
optimizer = tf.train.AdamOptimizer(rate)
```

使用 TensorFlow 的 Tips：

(4) 我们在此程序中没有显性的使用反向传播算法来训练神经网络，但实际上 TensorFlow 会自动使用反向传播算法。

数据如下图所示。其中曲线 1 是标准正弦曲线；曲线 2 是加上了噪声的曲线；曲线 3 是神经网络拟合的曲线。我们可以看到神经网络可以克服噪声的干扰，较好的拟合了正弦曲线。对于数据中的噪声，几乎没有过拟合。

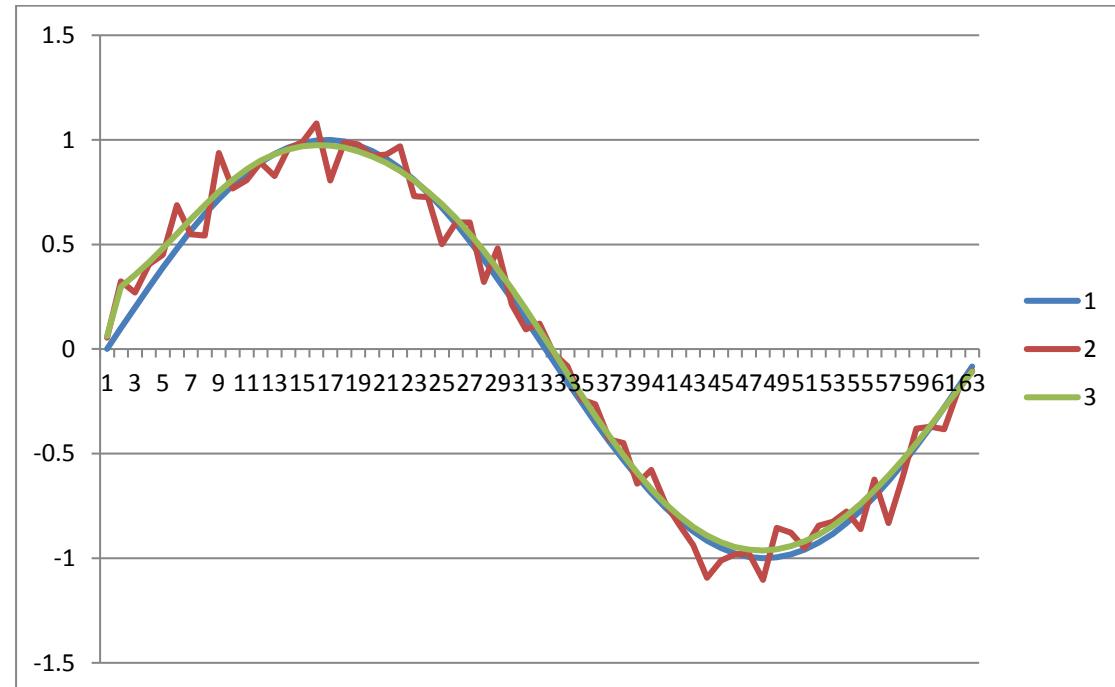


图 3.6 拟合的曲线

第七节：TensorBoard

第四章：深度神经网络

深度前馈神经网络、深度卷积神经网络和深度循环神经网络是当前深度学习中最常用的深度神经网络模型。更有相当多的研究采用这些模型的组合。本章讨论这些深度神经网络的一些共性问题。

第一节：为什么需要深度神经网络？

下面内容来自周志华 IJCAI2019 年的大会报告。

深度神经网络和传统的神经网络的区别是什么？简单来说，就是深度神经网络的层数比传统神经网络会多很多。在 2012 年深度学习、卷积神经网络刚刚开始受到大家重视的时候，那时候 ImageNet 竞赛的冠军是用了 8 层的神经网络。那么到了 2015 年是用了 152 层，到了 2016 年是 1207 层。如今，数千层深的网络非常常见。这是个非常庞大非常巨大的系统，把这么一个系统训练出来难度非常大。

1. 为什么深度模型是有效的？

为什么深的模型要比浅的模型表现好那么多？到今天为止，学术界都还没有统一的看法。周志华从模型的复杂度的角度来讨论。

一个机器学习模型，它的复杂度实际上和它的容量有关，而容量又跟它的学习能力有关。所以就是说学习能力和复杂度是有关的。机器学习界早就知道，如果我们能够增强一个学习模型的复杂度，那么它的学习能力能够提升。

注：

一个模型的容量（capacity）是指模型拟合一个宽范围的函数的能力。模型的容量低，它很可能会欠拟合（underfit），如果模型的容量大它更容易过拟合（overfit）

对神经网络这样的模型来说，提高模型复杂的有两种方法：一是把模型变深，另一种是把模型变宽。如果从提升复杂度的角度，变深是会更有效的。当变宽的时候，只不过是增加了一些计算单元，在变深的时候不仅增加了个数，其实还增加了模型提取数据特征的能力。所以从这个角度来说，应该尝试去把它变深。

如此，大家可能会问，既然早就知道要建立更深的模型？为什么现在才开始做？这就涉及到另外一个问题，把机器学习的学习能力变强了，这其实未必是一件好事。因为机器学习一直在斗争的一个问题就是过拟合。也即，给定一个数据集，机器学习的目的要能够学习数据集里面的一般规律，能够用来预测未来的事情。但是有时候可能把这个数据本身的一些独特特性（或者说是噪声）学出来了，而不是一般规律。错误地把它当成一般规律来用的时候，会犯巨大的错误。这种现象就是所谓的过拟合，就是因为模型的学习能力太强了。所以以往通常不太愿意用太复杂的模型。

那现在我们为什么可以用很复杂的模型？是因为现在设计了许多方法来对付过拟合，比如神经网络有 dropout、early-stop 等。但有一个因素非常简单、非常有效，那就是用很大的数据。比如说手上如果只有 3000 个数据，那学出来的特性一般不太可能是一般规律，但是如果有 3000 万、30 亿的数据，那这个数据里面的特性可能本身就已经是一般规律。所以使用大的数据是缓解过拟合的一个关键的途径。第二，今天有了很多很强大的计算设备，这使得能够使用大规模数据训练模型。第三，通过这个领域很多学者的努力，有了大量的训练这样复杂模型的技巧、算法，这使得我们使用复杂模型成为可能。总结一下就是：第一我们有了更大的数据；第二我们有强力的计算设备；第三我们有很多有效的训练技巧。这导致我们可以用高复杂度的模型，而深度神经网络恰恰就是一种很便于实现的高复杂度模型。

2. 深度神经网络的有效性

如果从复杂度这个角度去解释的话，我们没法说清楚为什么扁平的（flat），或者宽的网络做不到深度神经网络的性能？实际上我们把网络变宽，虽然它的效率不是那么高，但是它同样也能起到增加复杂度的能力。

实际上只要有一个隐层，加无限多的神经元进去，它的复杂度也会变得很大。但是这样的模型在应用里面怎么试都发现它不如深度神经网络好。所以从复杂度的角度可能很难回答这个问题，我们需要一点更深入的思考。所以我们要问这么一个问题：深度神经网络里面最本质的东西到底是什么？

今天我们的回答是，本质是表征学习的能力。这已经成为了学术界的新的共识，甚至有了专门的会议 ICLR。以往我们用机器学习解决一个问题的时候，首先我们拿到一个数据，比如说这个数据对象是个图像，然后我们就用很多特征把它描述出来，比如说颜色、纹理等等。这些特征都是我们人类专家通过手工来设计的，表达出来之后我们再去进行学习。而今天我们有了深度学习之后，现在不再需要手工去设计特征了。你把数据从一端扔进去，结果从另外一端就出来了，中间所有的特征完全可以通过学习自己来解决。所以这就是我们所谓的特征学习，或者说表征学习。我们都认可这和以往的机器学习技术相比可以说是一个很大的进步，这一点非常重要。我们不再需要依赖人类专家去设计特征了。

这个过程中的关键点是什么呢？是逐层计算（layer-by-layer processing）。

这就回答了为什么增加深度比增加宽度有效。

当拿到一个图像的时候，如果把神经网络看作很多层，首先它在最底层，好像我们看到的是一些像素这样的东西。当一层一层处理的时候，慢慢的可能有边缘，再到高的层上可能有轮廓，甚至对象的部件等等。当然这实际上只是个示意，在真正的神经网络模型里面不见得会有这么清楚的分层。但是总体上当逐渐往高的层走，它确实是不断在对对象进行抽象。我们现在认为这好像是深度学习为什么成功的关键因素之一。因为扁平神经网络能做很多深层神经网络能做的事，但是有一点它是做不到的。当它是扁平的时候，它就没有进行这样的一个深度的加工。所以深度的逐层抽象这件事情，可能是很浅层神经网络和深层神经网络之间的关键区别。当然了，这也是一种猜测，我们目前还无法从数学上证明。

逐层计算在机器学习里面也不是新东西。比如说决策树就是一种逐层处理，这是非常典型的。决策树模型已经有五六十年的历史了，但是它为什么做不到深度神经网络这么好呢？答案很简单。首先它的复杂度不够，决策树的深度，如果我们只考虑离散特征的话，它最深的深度不会超过特征的个数，所以它的模型复杂度是有限的；而在神经网络中，当我们想要增加模型复杂度的时候，我们增加任意数目的层，没有任何的限制。第二，也是更重要的，在整个决策树的学习过程中，它内部没有进行特征的变换，从第一层到最后一层始终是在同一个原始特征空间里面进行的，这非常重要。相信这两点对于深度神经网络是非常重要的。

而当我们考虑到这两件事情的时候，我们就会发现，其实深度模型是一个非常自然的选择。有了这样的模型，我们很容易就可以做上面两件事。但是当我们选择用这么一个深度模型的时候，我们就会有很多问题，它容易 overfit，所以我们要用大数据；它很难训练，我们要有很多训练的 trick；这个系统的计算开销非常大，所以我们要有非常强有力的计算的设备，比如 GPU 等等。

实际上所有这些东西是因为我们选用了深度模型之后产生的一个结果，它们不是我们用深度学习的原因。所以这和以往的思考不太一样，以往我们认为有了这些东西，导致我们用深度模型。其实现在我们觉得这个因果关系恰恰是反过来，因为我们要用它，所以我们才会考虑上面这些东西。这曾经是使用浅层网络的原因，如今也可以是使用很深的网络的原因。

另外还有一点我们要注意的，当我们有很大的训练数据的时候，这就要求我们必须要有很复杂的模型。否则假设我们用一个线性模型的话，给你 2000 万样本还是 2 亿的样本，其实对它没有太大区别。它已经学不进去了。而我们有了充分的复杂度，恰恰它又给我们使用深度模型加了一分。所以正是因为这几个原因，我们才觉得这是深度

模型里面最关键的事情。（这个和另一个流行的观点不一致：当数据量足够大时，什么模型的性能都是差不多的）

这是我们现在的一个认识：**第一，我们要有逐层的处理；第二，我们要有特征的内部变换；第三，我们要有足够的模型复杂度。**这三件事情是我们认为深度神经网络为什么能够成功的比较关键的原因。或者说，这是我们给出的一个猜测。（选自周志华在IJCAI2019的报告）

3. 深度神经网络是万能的吗？

虽然在今天深度神经网络已经这么的流行，这么的成功，但是其实我们可以看到在很多的任务上，性能最好的不见得完全是深度神经网络。比如说 Kaggle 上面的很多竞赛有各种各样的真实问题，有买机票的，有订旅馆的，有做各种的商品推荐等等，还有一些来自企业的真实的工业问题，他们只考虑模型的表现，我们就可以看到在很多任务上的胜利者并不是神经网络，它往往是像随机森林，像 xgboost 等等这样的模型。深度神经网络获胜的任务，往往就是在图像、视频、声音这几类典型任务上，都是连续的数值建模问题。**而在别的凡是涉及到混合建模、离散建模、符号建模这样的任务上，其实深度神经网络的性能可能比其他模型还要差一些。**这也就是我们说的「没有免费的午餐定理」，已经有数学证明，一个模型不可能在所有任务中都得到最好的表现。

第二节：梯度消失、梯度爆炸与解决方法

4.1.1 梯度消失

传统的机器学习需要模型开发者非常清晰的了解应该从数据集抽取什么样的特征。例如，在做文本情感分析时，需要计算词频，词的极性词的情感倾向等工作，**即特征工程在传统的机器学习中很重要。**而对于神经网络来说，把数据喂给网络，网络可以自动的抽取特征。神经网络在层次化非线性特征抽取上的表现出了优秀的能力。为了抽取更多的语义特征，以得到更好的模型性能，很多研究尝试加深模型的深度。然而超过三层的神经网络，我们称之为深度神经网络，在训练时会有一个**梯度消失**的问题（Vanishing gradient problem）。

在机器学习中，梯度消失问题是一种在使用梯度下降法和反向传播训练人工神经网络时出现的难题。在这类训练方法的每个迭代中，神经网络权重的更新值与误差函数梯度成比例，然而在某些情况下，梯度值会几乎消失，使得权重无法得到有效更新，甚至神经网络可能完全无法继续训练。对第 / 层权重更新时计算的梯度

$$\frac{\partial C}{\partial W_l} \propto f'(z^{(l)})f'(z^{(l+1)})f'(z^{(l+2)}) \dots$$

$f'(z^{(l)})$ 是第 l 层激活函数的导数。当所有的 f' 小于 1 时，随着网络层数的增加，梯度 $\frac{\partial C}{\partial W_l}$ 也远远小于 1，越深的层，权重几乎就不会被更新，这就是梯度消失。梯度消失问题在深度前馈神经网络和循环神经网络中表现明显。

举例来说，传统的激活函数，如双曲正切函数 \tanh 的梯度值在 $(0, 1)$ 范围内，而反向传播通过链式法则来计算梯度。这种做法计算前一层的梯度时，相当于将 n 个这样小的数字相乘，这就使梯度（误差信号）随 n 呈指数下降，导致前面的层训练非常缓慢。图 4.1 中，蓝线是 sigmoid 函数，黄线是 sigmoid 函数的导数。从图 4.1 可以观察到，sigmoid 函数的输入值太大或太小，导数值都很小 ($<<1$)。当网络的权重值初始化很差，即在太大的正、负值之间，sigmoid 激活函数会很明显表现出梯度消失问题。

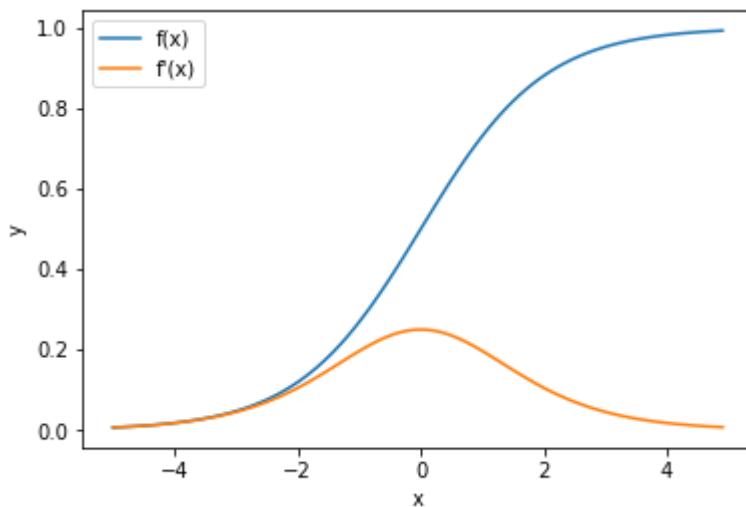
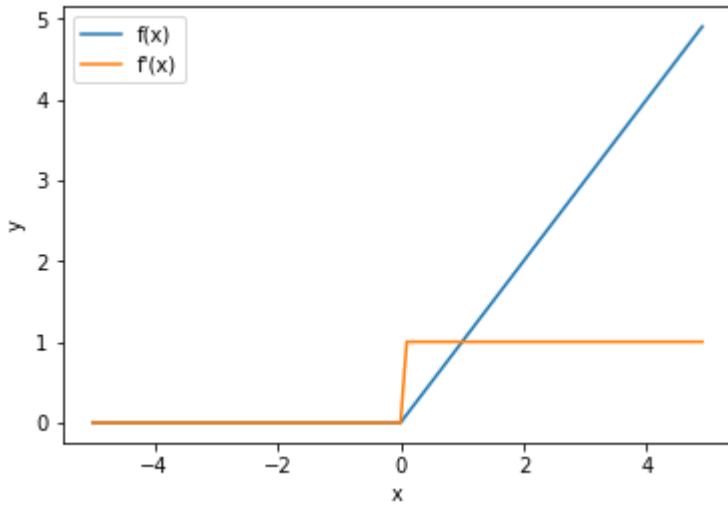


图 4.1 Sigmoid 函数和它的梯度

然而，即使权重的初始值选择的比较好，但随着层数的加深，这一问题仍旧在 sigmoid 激活函数中表现明显。而对于 ReLU 激活函数。当输入大于 0 时，它的导数是 1。否则是 0.



因此，梯度消失问题的解决方案包括：

- (1) 隐层使用 ReLu 激活函数，替代 sigmoid 和 tanh。
- (2) 在 RNN 中，使用 LSTM 或 GRU cell
- (3) 注意权重的初始化。权重采用随机初始化，值不要太大。偏置可以初始化为 0。

4.1.2 观察梯度的消失

4.1.3 梯度爆炸

采用梯度下降算法训练网络时，网络权重更新应该沿着合适的方向和量进行更新。误差梯度就是这个更新的方向和量。**梯度爆炸** (Exploding Gradients) 是指大的误差梯度的累积导致训练网络时权重更新值过大，网络训练时不稳定，网络不能从训练数据学习。极端情况下，导致 NaN 的权重值。

怎样判断是否你的模型出现了梯度爆炸。下面三个特征指示可能存在梯度爆炸：

- (1) 训练时，模型不能收敛，例如，损失函数值很高
- (2) 训练时，模型不稳定，损失函数变化很大
- (3) 得到了 NaN 的损失函数值

下面这些信息显示梯度爆炸的存在：

- (1) 模型的权重值很快的成为很大
- (2) 模型的权重值很变成 NaN
- (3) 训练时，每个层的每个节点的误差梯度值都是大于 1.0

解决梯度爆炸，包括下面的策略

- (1) 重新设计网络结构。减少网络的深度。也可以在训练网络时 batch_size 设置比较小。在 RNN 中可以使用 truncated Backpropagation through time。
- (2) 使用修正的线性激活函数 (Rectified Linear Activation) Relu。在深度前馈神经网络中，sigmoid 和 tanh 激活函数会导致梯度爆炸的产生。使用 ReLu 激活函数可以减小产生梯度爆炸的风险。在隐层采用 ReLu 激活函数是最佳策略。
- (3) 在 RNN 中采用 LSTM cell 或者有 Gate 结构的 Cell 可以减少梯度爆炸。
- (4) 使用梯度剪裁 (Gradient Clipping)。在深度前馈神经网络中如果训练模型的 batch_size 过大，在 LSTM 中如果输入序列很长，都会发生梯度爆炸。如果采用了上面的策略后仍旧发生梯度爆炸，可以在网络训练时限制梯度的大小到一个阈值，这称为 Gradient Clipping。在 TensorFlow 中，在进行优化算法时，不采用 minimize 函数，而是拆开成 compute_gradients 和 apply_gradients 两个方法，在两个方法之间应用 Gradient Clipping 即可（使用 tf.clip_by_value 函数）。例如

```
optimizer = tf.train.AdamOptimizer(learning_rate=learning_rate)
gvs = optimizer.compute_gradients(cost)
capped_gvs = [(tf.clip_by_value(grad, -1., 1.), var) for grad, var in gvs]
train_op = optimizer.apply_gradients(capped_gvs)
```

- (5) 使用权重正则化 (Weight Regularization)。即将网络的权重值作为正则化项加入到损失函数，可以采用 L1 (权重绝对值) 或 L2 (权重平方值) 正则化。一个例子，如果 hidden_weights, hidden_biases, out_weights, 和 out_biases 都是模型的参数，对这些参数应用 L2 正则化，作为正则化项加入损失函数。

```
loss = (tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(
    logits=out_layer, labels=tf_train_labels)) +
    0.01*tf.nn.l2_loss(hidden_weights) +
    0.01*tf.nn.l2_loss(hidden_biases) +
    0.01*tf.nn.l2_loss(out_weights) +
    0.01*tf.nn.l2_loss(out_biases))
```

- (6) 减小优化函数的学习率

可以参考一篇论文详细了解梯度爆炸 <https://arxiv.org/abs/1712.05577>

第三节：构建深度前馈神经网络

深度前馈神经网络的结构包括输出层、隐层和输入层。网络的设计包括输出层选用什么样的激活函数，隐层的深度，每层的神经元的数目等。

1. 结构

按照 Deep Learning 一书 6.4 节，很多任务选择一个隐层就可以很好的完成。但面对一些问题，这个隐层的神经元数需要非常大。这有导致学习参数失败。越深的网络可以每层可以使用更少的神经元（单元 unit）。因此可以通过增加隐层，然后减小隐层的神经元数达到即可以很好的训练参数，又可以达到需要的模型性能。理想的网络结构必须通过试验不断调整参数，结构，监视校验集的误差去发现。

2. 输出层

输出层的设计是面向任务的，即将隐层的输出结果转换成任务需要的输出形式。

(1) 最简单的输出层即线性输出层

$$\hat{y} = W^T h + b$$

线性输出层通常用于产生条件正太分布的均值。

$$p(y|x) = N(y; \hat{y}, I)$$

$N(y; \hat{y}, I)$ 表示学习到的在 y 上的正太分布，均值是 \hat{y} ，方差是单位矩阵 I 。

(2) sigmoid 函数的输出层用于完成产生二元变量 y 的任务，例如二分类任务。

$$\hat{y} = \sigma(W^T h + b)$$

σ 是 Logistics sigmoid 函数。我们也可以理解成输出层是对 $z = W^T h + b$ 的计算结果应用了一个 sigmoid 激活函数，将计算值 z 转换到了一个概率值。

(3) 当有一个具有 n 个可能值的离散变量，我们希望描述它的概率分布时，输出层采用 softmax 函数（或者说给输出层加上一个 softmax 激活函数）。Softmax 经常用于描述一个多类分类器的输出的概率分布。注：softmax 函数基本不用作隐层的激活函数。当一个线性输出层

$$z = W^T h + b$$

z 是一个向量。Softmax 函数将 z 规范化到一个概率分布

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

3. 隐层

这里讨论的隐层设计其实是在讨论隐层的激活函数选择。在学术界隐层的设计非常活跃，但没有一个明确的指导关于怎么设计深度网络的隐层。

(1) ReLU。修正的线性单元 Rectified Linear Units

$$g(z) = \max\{0, z\}$$

如果没有明确的想法，Rectified linear units 是推荐的默认的隐层单元选择。第一节也介绍了在隐层使用 ReLU 激活函数是解决梯度消失和梯度爆炸有效的方法。ReLU 有很多变体，我们不详细讨论。第四节将会介绍一下 tensorflow 提供的 relu 变体。

(2) Sigmoid 和 tanh。在 ReLU 出现之前最常用的激活函数是 sigmoid 和 tanh。这两个激活函数是很相关的，因为 $\tanh(z) = 2\sigma(2z) - 1$ 。在一些必须使用 sigmoid 函数的场合，tanh 表现的更好。Sigmoid 在一些前馈神经网络之外，使用的更频繁。

(3) 其他类型。RBF 函数，softplus, hard tanh。Softplus 可以看做是 ReLU 的平滑版。但 Deep Learning 一书推荐还是使用 ReLU。

dropout 是一个简单的防止深度网络过拟合的方法。它在神经网络的训练期间，随机选择一些神经元的输出。Hinton 的论文中，建议 dropout 层可以放在除了输出层的任何全连接层后，选择概率是 0.5。5.4 节给出了详细的介绍。

Tips:

在 tensorflow 中看到参数 logits，都是指一个全连接层的运算 $W*x+b$ 没有经过激活函数的输出。

第四节：训练深度神经网络

在 2.3 节我们已经初步讨论了使用 SGD 训练一个感知机。这一节我们给出更详细的分析。

tips :

凸函数 convex (凹函数 concave) 和非凸函数 (非凹函数)

一个凸函数是指，它的二阶导数总是非负的。因此给函数有一个全局最小值。凹函数就是说一个函数的二阶导数总是非正的，它有个全局最大值。凸或凹函数都可以容易的使用 SGD 得到它们的全局最小或最大值。而对于非凸和非凹的函数，SGD 可能收敛到一个局部极值，而非全局最值

训练一个深度神经网络，数据集划分成三个部分：训练集（ training set ），校验集（ validation set ），和测试集（ test set ）。三个部分互不重合。当多人进行比赛或我们在考察多个模型时，测试集作为最终评判的数据集合。模型用训练集训练，用校验集检验模型，然后调整模型的超参数。超参数是指那些不能在模型训练中自动学习，需要人工设定的参数，如神经网络的层数，每层的节点数。测试集是在训练过程中看不到的数据。

构建好了深度神经网络，从任务出发我们可以构建损失函数。然后根据损失函数选择合适的优化算法来训练模型。在训练模型时有一些策略可以帮助减少 test error，这有可能以增加 training error 为代价（机器学习模型的泛化能力是我们追求的目标，即在测试集上有小的 test error），这些策略称为 regularization，翻译做正则化。

1. 损失函数 (loss) 或代价函数 (cost)

常见的损失函数有 Zero-one Loss (0-1 损失)，Perceptron Loss (感知损失)，Hinge Loss (Hinge 损失)，Log Loss (Log 损失)，Cross Entropy (交叉熵)，Square Loss (平方误差)，Absolute Loss (绝对误差)，Exponential Loss (指数误差) 等。深度学习中，损失函数的选择是和输出层紧密相关的。我们介绍几种深度学习中最常用的损失函数。

- (1) Mean Squared Error (MSE) **均方误差**是最基本的误差函数，即计算训练数据上每条数据的目标值和预测值的误差的平方，然后对所有数据的计算结果再求和。对于预测任务（线性输出层），通常选用 MSE 计算损失函数。但 MSE 对离群点敏感。
- (2) 交叉熵。对于多分类任务（ softmax 输出层），通常采用 softmax 交叉熵来计算损失函数。对于二分类任务（ sigmoid 输出层）可以使用 binary cross-entropy (二值交叉熵)。交叉熵的计算公式

$$H_{y'}(y) = - \sum_i y'_i \log(y_i)$$

y 是目标分布， y' 是预测分布。 y_i 是目标分布中的一个元素， y'_i 是预测分布的一个元素。二值交叉熵计算公式如下

$$H_{y'}(y) = y' \log(y) + (1 - y') \log(1 - y)$$

y 是一个预测结果（二分类中是标量）。

(3) Hinge loss。可用于最大间隔分类，SVM 采用 Hinge loss 作为目标函数。梯度下降的优化方法训练模型时，不能使用 hinge loss 函数。

Tensorflow 提供的损失函数的详细介绍见第五节。

交叉熵举例：(交叉熵不适合多标签多类分类？)

一个单标签多类分类问题，假设有三个类 A、B、C。训练数据集给出的一条数据记录的标签是 B。即， y 是

Pr(Class A)	Pr(Class B)	Pr(Class C)
0.0	1.0	0.0

softmax 回归预测的每个类别的概率 y_i 是

Pr(Class A)	Pr(Class B)	Pr(Class C)
0.228	0.619	0.153

则计算的交叉熵是

$$H = - (0.0 * \ln(0.228) + 1.0 * \ln(0.619) + 0.0 * \ln(0.153)) = 0.479$$

2. 优化算法

随机梯度下降 (SGD) 算法是最常用的深度神经网络训练算法。2.3 节已经介绍，这里不再重复。传统的 SGD 算法会有些问题。基于 SGD 的改进算法主要包括 momentum (翻译作动量) 和自适应学习率两大类。

(1) 使用动量的 SGD 算法

传统的 SGD 算法学习会很慢。Momentum 算法被设计用于加速学习过程。Momentum 算法对之前计算的梯度累积一个指数衰减的移动平均，并继续沿着梯度的方向移动。

Stochastic Gradient Decent (SGD) with momentum

Require: Learning rate ϵ , momentum parameter α .

Require: Initial parameter θ , initial velocity v .

while stopping criterion not met do

 Sample a minibatch of m examples from the training set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ with corresponding targets $\mathbf{y}^{(i)}$.

 Compute gradient estimate: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$

 Compute velocity update: $v \leftarrow \alpha v - \epsilon g$

 Apply update: $\theta \leftarrow \theta + v$

end while

速度 v 是负梯度的指数衰减平均。超参数 $\alpha \in [0, 1)$ 决定前面的梯度对指数衰减贡献有多大。

Nesterov momentum 是 momentum 算法的变体。它是结合 Nesterov 的加速梯度方法和 momentum 方法的随机梯度下降优化算法。

Stochastic Gradient Descent (SGD) with Nesterov momentum

Require: Learning rate ϵ , momentum parameter α .

Require: Initial parameter θ , initial velocity v .

while stopping criterion not met **do**

 Sample a minibatch of m examples from the training set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ with corresponding labels $\mathbf{y}^{(i)}$.

 Apply interim update: $\tilde{\theta} \leftarrow \theta + \alpha v$

 Compute gradient (at interim point): $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\tilde{\theta}} \sum_i L(f(\mathbf{x}^{(i)}; \tilde{\theta}), \mathbf{y}^{(i)})$

 Compute velocity update: $v \leftarrow \alpha v - \epsilon g$

 Apply update: $\theta \leftarrow \theta + v$

end while

与传统 momentum 算法相比，nesterov momentum 是在应用了当前的 velocity 后计算梯度。这被看做是加了一个校正因子。

(2) 自适应学习率 (adaptive learning rate)

学习率是深度学习最难设置的超参数，它可以显著地影响模型的性能。损失函数经常是对参数的某些方向非常敏感，而对其他方向又不敏感。Momentum 可以一定程度的缓解这一问题，但它引入了新的超参数 velocity v 。自适应学习率的一系列算法为解决这一问题，**为每个参数单独设置学习率**，且在学习的过程中可以自动修改学习率。

AdaGrad 算法为所有的模型参数单独的调整学习率。它在每次迭代中累积参数梯度的平方值。并用该值调整学习率

AdaGrad 算法

Require: Global learning rate ϵ

Require: Initial parameter θ

Require: Small constant δ , perhaps 10^{-7} , for numerical stability

 Initialize gradient accumulation variable $r = \mathbf{0}$

while stopping criterion not met **do**

 Sample a minibatch of m examples from the training set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ with corresponding targets $\mathbf{y}^{(i)}$.

 Compute gradient: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$

 Accumulate squared gradient: $r \leftarrow r + \mathbf{g} \odot \mathbf{g}$

 Compute update: $\Delta \theta \leftarrow -\frac{\epsilon}{\delta + \sqrt{r}} \odot \mathbf{g}$. (Division and square root applied element-wise)

 Apply update: $\theta \leftarrow \theta + \Delta \theta$

end while

运算符 \odot 表示逐个元素相乘。 $g \odot g$ 表示每个参数的梯度平方运算。 $r \leftarrow r + g \odot g$ 表示在迭代的过程中累积每个参数的梯度的平方值。学习率 ϵ 除以 $\delta + \sqrt{r}$ 表示在每次迭代中动态调整学习率。可以看到，AdaGrad 累积梯度，因此学习率在迭代多次后会很小，以至于导致收敛到一个并不理想的 local minimum。因此，AdaGrad 在很多情况下不理想。RMSProp 的提出就改进了该问题。

RMSProp 算法对 AdaGrad 算法进行改进，它改变 AdaGrad 的梯度累积为指数加权移动平均。该算法在非凸环境下表现的更好。

RMSProp 算法

```

Require: Global learning rate  $\epsilon$ , decay rate  $\rho$ .
Require: Initial parameter  $\theta$ 
Require: Small constant  $\delta$ , usually  $10^{-6}$ , used to stabilize division by small numbers.
Initialize accumulation variables  $r = 0$ 
while stopping criterion not met do
    Sample a minibatch of  $m$  examples from the training set  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$  with corresponding targets  $\mathbf{y}^{(i)}$ .
    Compute gradient:  $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$ 
    Accumulate squared gradient:  $r \leftarrow \rho r + (1 - \rho) \mathbf{g} \odot \mathbf{g}$ 
    Compute parameter update:  $\Delta\theta = -\frac{\epsilon}{\sqrt{\delta+r}} \odot \mathbf{g}$ . ( $\frac{1}{\sqrt{\delta+r}}$  applied element-wise)
    Apply update:  $\theta \leftarrow \theta + \Delta\theta$ 
end while

```

指数加权移动平均 (Exponentially Weighted Moving Average , EWMA) 是一种常用的序列数据处理方式。在时间 t , 根据实际的观测值 (或量测值) 我们可以求取 EWMA (t) : $EWMA(t) = \rho Y(t) + (1-\rho) EWMA(t-1)$ for $t = 1, 2, \dots, n$

EWMA (t) : t 时刻的估计值

$Y(t)$: t 时间的量测值 .

从信号处理角度看，EWMA 可以看成是一个低通滤波器，通过控制 ρ ($0 < \rho < 1$) 值，剔除短期波动、保留长期发展趋势提供了信号的平滑形式。

下面是结合 RMSProp 和 momentum 产生了一个新的算法

结合 Nesterov momentum 的 RMSProp 算法

Require: Global learning rate ϵ , decay rate ρ , momentum coefficient α . Require: Initial parameter θ , initial velocity v . Initialize accumulation variable $r = \mathbf{0}$ while stopping criterion not met do Sample a minibatch of m examples from the training set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ with corresponding targets $\mathbf{y}^{(i)}$. Compute interim update: $\tilde{\theta} \leftarrow \theta + \alpha v$ Compute gradient: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\tilde{\theta}} \sum_i L(f(\mathbf{x}^{(i)}; \tilde{\theta}), \mathbf{y}^{(i)})$ Accumulate gradient: $r \leftarrow \rho r + (1 - \rho) \mathbf{g} \odot \mathbf{g}$ Compute velocity update: $v \leftarrow \alpha v - \frac{\epsilon}{\sqrt{r}} \odot \mathbf{g}$. ($\frac{1}{\sqrt{r}}$ applied element-wise) Apply update: $\theta \leftarrow \theta + v$ end while

RMSProp 已经被证明是深度学习中非常有效的优化算法。

Adam 是另外一种自适应学习率优化算法。Adam 是 Adaptive moments 的简写。可以将 Adam 看做是 RMSProp+momentum+新特性的组合。

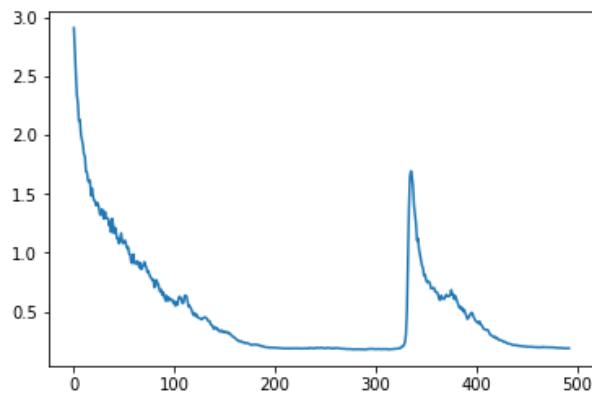
Adam 算法
Require: Step size ϵ (Suggested default: 0.001) Require: Exponential decay rates for moment estimates, ρ_1 and ρ_2 in $[0, 1)$ (Suggested defaults: 0.9 and 0.999 respectively) Require: Small constant δ used for numerical stabilization. (Suggested default: 10^{-8}) Require: Initial parameters θ Initialize 1st and 2nd moment variables $s = \mathbf{0}$, $r = \mathbf{0}$ Initialize time step $t = 0$ while stopping criterion not met do Sample a minibatch of m examples from the training set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ with corresponding targets $\mathbf{y}^{(i)}$. Compute gradient: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$ $t \leftarrow t + 1$ Update biased first moment estimate: $s \leftarrow \rho_1 s + (1 - \rho_1) \mathbf{g}$ Update biased second moment estimate: $r \leftarrow \rho_2 r + (1 - \rho_2) \mathbf{g} \odot \mathbf{g}$ Correct bias in first moment: $\hat{s} \leftarrow \frac{s}{1 - \rho_1^t}$ Correct bias in second moment: $\hat{r} \leftarrow \frac{r}{1 - \rho_2^t}$ Compute update: $\Delta\theta = -\epsilon \frac{\hat{s}}{\sqrt{\hat{r}} + \delta}$ (operations applied element-wise) Apply update: $\theta \leftarrow \theta + \Delta\theta$ end while

首先，Adam 中要为梯度计算一个一阶动量 $s \leftarrow \rho_1 s + (1 - \rho_1) g$ 和一个二阶动量 $r \leftarrow \rho_2 r + (1 - \rho_2) g \odot g$ 。然后用修正的一、二阶动量去调整学习率 $\Delta\theta = -\epsilon \frac{\hat{s}}{\sqrt{\hat{r}} + \delta}$

怎样选择合适的优化算法？没有一个统一的看法。有论文对不同任务时不同算法的性能进行了比较。论文总结，自适应学习率的算法表现更好。本文前面介绍的算法是在深度学习任务中广泛采用的算法，选择哪个算法取决于用户对算法的了解，例如如何设置超参数。《Neural Network Method for NLP》一书的作者指出在他的研究中发现 Adam 算法在大的网络上很有效。

注：使用 Adam 算法的一个问题

一阶动量 s 是梯度均值的指数移动平均，二阶动量 r 是梯度平方的指数移动平均。当训练一个长时间后，参数接近最优值后， r 比 s 减小的更快成为很小，此时梯度会反而增加甚至梯度爆炸。如下图所示损失函数在训练到第 300 多次后突然增加。



对付这个问题，要么换一个优化函数，要么在损失函数突变前结束训练。对于该图中的问题，我尝试使用 GradientDescentOptimizer，确实不会出现损失函数的突变。但，损失函数收敛到 1.4，不然 adam 优化算法可以收敛到 0.1。

应该试一下梯度剪裁可以解决此问题不

3. 参数初始化

深度学习算法被参数的初始值影响。初始值甚至能影响算法是否收敛。对于偏置，通用的做法是设置初始偏置为 0。但对于权重则比较复杂。例如，如果全部初始化为 0， \tanh 激活函数会产生为 0 的梯度；如果权重都一样，隐层单元将产生同样的梯度，它们的行为一致，将浪费模型的 capacity（模型的 capacity 是指模型拟合各种函数的能力，模型越复杂 capacity 越大）。参数初始化的启发规则有很多。下面列举出部分权重初始化的建议：

- (1) 初始权重希望大一些，但不要达到引起梯度爆炸。
- (2) 所有的权重初始化从一个均匀分布 $[-b, b]$ 。就 b 的选择有很多研究。普遍认为，一层的输入越多，权重值应越小。例如，有研究建议有 m 个输入， n 个输出的全连接层初始权重按照均匀分布 $W_{i,j} \sim U\left(-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right)$ 初始化

(3) 有研究使用一个增益因子 (gain factor) g , 仔细选择该增益因子成功训练了有 1000 层的网络。

4. Early Stopping

当训练一个大 capacity 的模型, 它对面对的任务很有可能过拟合。我们可以观察到一个现象, 训练误差稳定的在减小, 但校验集上的误差开始增加。如图 4.2 所示。

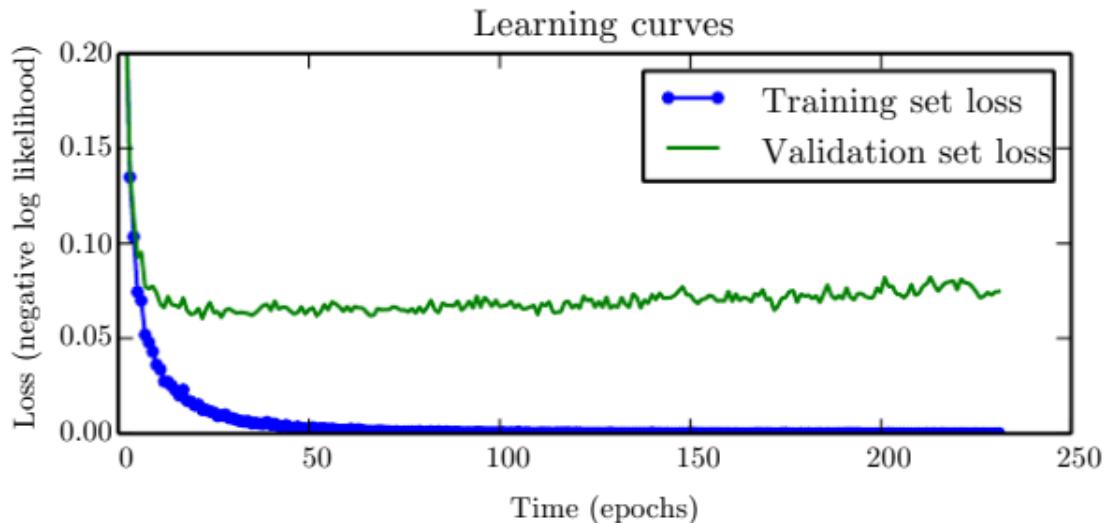


图 4.2 校验集误差的变化

我们希望能够回到具有最好 validation error 时间点的模型 (很可能对应最好的 test error), 采取的做法是, 如果 validation error 在一个时间段内没有变得更小, 则停止模型训练。这称为 early stopping。下面的 early stopping 算法选自 Lutz Prechelt 的 “Early Stopping – But When?”

Early stopping 算法

1. 将训练数据划分成训练集和校验集, 例如 2:1 的比例。(有很多任务中只需划分训练集和测试集, 如果没有校验集则从训练集中划分)
2. 在训练集上训练, 在每几趟训练后在校验集上评估误差, 例如每 5 趟训练 (每几趟后评估是考虑到了训练的效率问题, 很多实践是每趟训练后都在校验集上评估)。.
3. 一旦校验误差在持续了一段时间的高于最低校验误差, 则停止训练。
4. 保存的最低校验误差时的权重作为最终训练得到的网络权重。

Tensorflow 没有提供 early stopping。自己实施的示意代码如下：

```
Saver = tf.train.saver()  
if (loss_value < self.best_loss):  
    self.stopping_step = 0
```

```

    self.best_loss = loss_value
    saver.save(sess,'checkpoint_directory/model_name')
else:
    self.stopping_step += 1
    if self.stopping_step >= FLAGS.early_stopping_step:
        self.should_stop = True
        print("Early stopping is trigger at step: {} loss:{}".format(global_step,
loss_value))
        run_context.request_stop()

```

每次计算校验误差（每几趟训练后计算），检测最低校验误差，并保存对应的参数。

连续的 FLAGS.early_stopping_step 次的计算结果高于最低值时，停止训练。

Tensorflow 使用 tf.train.saver 类的 save 方法可以把训练的模型保存下来。第一个参数是当前的 session 对象，第二个参数是保持到的路径和文件名前缀。想恢复保存的模型时使用 saver.restore(sess, 'checkpoint_directory/model_name') 方法。然后 session 对象里的模型被恢复到保存的模型。此时调用 run 方法执行的就是该模型。

5. 扩大数据集 (data augmentation)

要想模型有更好的泛化能力，就是在更多的数据上训练模型。在现实中，我们能获取的数据总是有限的，因此创建一些假数据把它们添加到数据集中也是常有的一种方法。对于分类任务这种方法可行。在图像处理的任务中，经常将图像进行旋转、变形等处理来 data augmentation。许多模型对数据集中类不平衡 (class imbalance) 问题很敏感。例如，用误差平方和做损失函数，在类不平衡问题很严重的数据集上训练分类模型，多数类将主宰训练过程。对付不平衡数据集的方法有很多研究，最简单的方法是在少数类的数据上重复抽样，扩大到两类数据基本相等。

第五节： Tensorflow 常用的激活函数

Tensorflow 提供了一系列激活函数。详见

https://tensorflow.google.cn/api_guides/python/nn#activation-functions

tf.nn.relu
tf.nn.relu6
tf.nn.crelu
tf.nn.elu
tf.nn.selu
tf.nn.softplus
tf.nn.softsign
tf.nn.dropout
tf.nn.bias_add

tf.sigmoid
tf.tanh

1. tf.nn.relu

relu 激活函数，Computes rectified linear: max(features, 0).

```
tf.nn.relu(  
    features,  
    name=None  
)
```

Feature 是一个 tensor，Name 可选。返回一个和 feature 同结构的 tensor

2. tf.nn.relu6

relu 激活函数。Computes Rectified Linear 6: min(max(features, 0), 6)。其他的同 relu 函数。详见 <http://www.cs.utoronto.ca/%7Ekriz/conv-cifar10-aug2010.pdf>

3. tf.nn.crelu

拼接 relu 激活函数。详见 <https://arxiv.org/abs/1603.05201>

```
tf.nn.crelu(  
    features,  
    name=None,  
    axis=-1  
)
```

将 features 应用一个 relu 激活函数，再和应用在“features”上 relu 得到的结果拼接。

Axis 是拼接的维度。例如

```
x=tf.Variable(tf.random_normal(shape=[2,3]))  
y=tf.nn.crelu(x, axis=-1)  
  
init_op= tf.global_variables_initializer()  
with tf.Session() as sess:  
    sess.run(init_op)  
    r1,r2=sess.run([y,x])  
    print(r1)  
    print(r2)
```

得到的结果是

```

[[0.36861396 0.      1.271101 0.      0.8036915 0.      ]
 [0.      0.      0.      2.0419934 0.43139106 0.4146952 ]]
[[ 0.36861396 -0.8036915  1.271101 ]
 [-2.0419934 -0.43139106 -0.4146952 ]]

```

4. **tf.nn.elu**

Computes exponential linear: $\exp(\text{features}) - 1$ if < 0 , features otherwise. 参见
<http://arxiv.org/abs/1511.07289>

其他同 relu 函数

5. **tf.nn.selu**

Computes scaled exponential linear: $\text{scale} * \alpha * (\exp(\text{features}) - 1)$ if < 0 , $\text{scale} * \text{features}$ otherwise。用于训练深层的前馈神经网络。详见
<https://arxiv.org/abs/1706.02515>

定义如下：

```

def selu(x):
    alpha = 1.6732632423543772848170429916717
    scale = 1.0507009873554804934193349852946
    return scale * K.elu(x, alpha)

```

当要为 selu 的输出应用 dropout 时，使用专门的函数 `tf.contrib.nn.alpha_dropout`。

6. **tf.nn.softplus**

softplus 被看做是 relu 的平滑版。Computes softplus: $\log(\exp(\text{features}) + 1)$.

在 Goodfellow 的 deep learning 一书中指出，relu 的性能比 softplus 更好，不鼓励使用 softplus.

第六节：Tensorflow 常用的损失函数

第三节介绍了深度学习常用的损失函数。本节详细介绍一些 tensorflow 提供的损失函数。https://tensorflow.google.cn/versions/r1.10/api_docs/python/tf/losses

absolute_difference
hinge_loss
mean_pairwise_squared_error
mean_squared_error
sigmoid_cross_entropy
softmax_cross_entropy
sparse_softmax_cross_entropy

```
1. tf.losses.absolute_difference(  
    labels,  
    predictions,  
    weights=1.0,  
    scope=None,  
    loss_collection=tf.GraphKeys.LOSSES,  
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS  
)
```

绝对误差损失函数是最简单的损失函数。它计算目标值和预测值之间误差的绝对值均值。Labels 是目标值，predictions 是预测值，它们是具有相同 shape 的 tensor。Weights 可以给计算结果乘上一个权重。Weights 可以是一个 0 阶的 tensor，或者和 labels 相同 shape 的 tensor。

2. **hinge_loss**

```
tf.losses.hinge_loss(  
    labels,  
    logits,  
    weights=1.0,  
    scope=None,  
    loss_collection=tf.GraphKeys.LOSSES,  
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS  
)
```

Hinge loss 适用于寻找最大间隔的二分类器。Labels 是目标值，其值应该是 0.0 或 0.1，用于指示类别。Logit 是预测值，大于 0 的值认为是分类为正例，小余 0 的值认为分类为反例。

3. **mean_pairwise_squared_error**

```
tf.losses.mean_pairwise_squared_error(  
    labels,  
    predictions,  
    weights=1.0,
```

```

    scope=None,
    loss_collection=tf.GraphKeys.LOSSES
)

```

For example, if `labels=[a, b, c]` and `predictions=[x, y, z]`, there are three pairs of differences are summed to compute the loss: $\text{loss} = [((a-b) - (x-y))^2 + ((a-c) - (x-z))^2 + ((b-c) - (y-z))^2] / 3$

该损失函数的适用场合是什么？

4. mean_squared_error

```

tf.losses.mean_squared_error(
    labels,
    predictions,
    weights=1.0,
    scope=None,
    loss_collection=tf.GraphKeys.LOSSES,
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS
)

```

误差平方和的均值。对 `labels` 和 `predictions` 中的逐对元素计算误差平方和，然后求均值。

5. sigmoid_cross_entropy

```

tf.losses.sigmoid_cross_entropy(
    multi_class_labels,
    logits,
    weights=1.0,
    label_smoothing=0,
    scope=None,
    loss_collection=tf.GraphKeys.LOSSES,
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS
)

```

使用 `tf.nn.sigmoid_cross_entropy_with_logits` 创建的 sigmoid 交叉熵损失函数。度量离散分类任务中概率误差。此时每个类是独立的，不是互斥的。例如**多标签分类**。

multi_class_labels: [batch_size, num_classes] 取 {0, 1}的整数。一行即一条数据在多个类别上的目标值 (0 或 1) 。

Logits : 即神经网络的输出层 , 没经过激活函数的输出 ($W * X + b$) 。类型为 Float , shape=[batch_size, num_classes]。

sigmoid_cross_entropy_with_logits 返回的是一个和 logits 相同 shape 的 tensor。
sigmoid_cross_entropy 返回的是该 tensor 上计算的均值。即长度为 batch_size 的向量。

tf.nn.sigmoid_cross_entropy_with_logits

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

该函数的 labels 和 logits 必须 shape 和 dtype 必须一致。

设 $x = \text{logits}$, $z = \text{labels}$, 该损失函数 (称为 Logistic loss) 的计算公式如下 :

$$\max(x, 0) - x * z + \log(1 + \exp(-\text{abs}(x)))$$

示例代码

```
import tensorflow as tf  
from numpy import mean  
  
x = tf.constant([1,1,0,0])  
x2 = tf.constant([1.0,1.0,0.0,0.0])  
y = tf.constant([0.0,0.0,0.0,1.0])  
loss = tf.losses.sigmoid_cross_entropy(x, y)  
loss2 =  
tf.nn.sigmoid_cross_entropy_with_logits(labels=x2,logits=y)  
with tf.Session() as sess:  
    r1,r2 = sess.run([loss,loss2])  
    print(r1,mean(r2))  
  
#输出 0.8481758 0.8481758
```

代码这么理解：当前只有一条数据（batch_size=1），有四个类别，目标值是类别标签是1, 2。

6. softmax_cross_entropy

```
tf.losses.softmax_cross_entropy(  
    onehot_labels,  
    logits,  
    weights=1.0,  
    label_smoothing=0,  
    scope=None,  
    loss_collection=tf.GraphKeys.LOSSES,  
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS  
)
```

对单标签多类分类计算交叉熵。使用 tf.nn.softmax_cross_entropy_with_logits_v2 来计算。onehot_labels 和 logits 必须有相同的 shape, [batch_size, num_classes]。Onehot_labels 的一个行向量，只能有一个元素的值为 1。即对应的类别。**注意：logits 是神经网络没经过激活函数的输出 ($w^T \cdot X + b$)。该函数会对 logits 做 softmax 规范化。**该函数返回一个长度为 batch_size 的向量。默认的返回值是 logits 和 onehot_labels 逐对计算交叉熵后的和，除非更改 reduction 的参数设置。

```
tf.nn.softmax_cross_entropy_with_logits_v2(  
    tf.nn.softmax_cross_entropy_with_logits_v2(  
        _sentinel=None,  
        labels=None,  
        logits=None,  
        dim=-1,  
        name=None  
)
```

Logits 和 labels 的 shape=[batch_size, num_classes]，它们的 dtype 必须相同。Labels 的每一行是有效的概率分布（累加和为 1），否则梯度的计算将不正确。Logit 是神经网络没经过激活函数的输出 ($w^T \cdot X + b$)。示例代码见下面 sparse_softmax_cross_entropy。

7. sparse_softmax_cross_entropy

```

tf.losses.sparse_softmax_cross_entropy(
    labels,
    logits,
    weights=1.0,
    scope=None,
    loss_collection=tf.GraphKeys.LOSSES,
    reduction=Reduction.SUM_BY_NONZERO_WEIGHTS
)

```

该函数和 softmax_cross_entropy 有相同的效果，只是使用方法不一样。Labels 的 shape= [batch_size]。每个元素给出类别标签[0, num_classes-1]

```

x = tf.constant([1.0,0,0,0])
x2 = tf.constant([1.0,0,0.0,0.0])
x3=[0]
y = tf.constant([0.0,1.0,0.0,2.0])
loss = tf.losses.softmax_cross_entropy(onehot_labels=x,
logits=y)
loss2 =
tf.nn.softmax_cross_entropy_with_logits_v2(labels=x2,logits=y
)
loss3 = tf.losses.sparse_softmax_cross_entropy(x3, y)
with tf.Session() as sess:
    r1,r2,r3 = sess.run([loss,loss2, loss3])
    print(r1,mean(r2),r3)

#输出 2.4938116 2.4938116 2.4938116

```

第七节：Tensorflow 常用的优化函数

Tensorflow 提供了一系列的优化器函数帮助训练模型。这些优化器根据损失函数计算梯度，然后更新参数。参见

https://tensorflow.google.cn/versions/r1.10/api_guides/python/train#Optimizers

tf.train.GradientDescentOptimizer
tf.train.AdadeltaOptimizer
tf.train.AdagradOptimizer
tf.train.AdagradDAOptimizer
tf.train.MomentumOptimizer
tf.train.AdamOptimizer
tf.train.FtrlOptimizer

tf.train.ProximalGradientDescentOptimizer
tf.train.ProximalAdagradOptimizer
tf.train.RMSPropOptimizer

1. tf.train.GradientDescentOptimizer

实施了梯度下降优化算法。它的构造函数主要有两个参数：

学习率 : learning_rate, 一个 tensor 或浮点值。

锁 : use_locking=False, 为 True 时，更新参数时给 Variable 上锁，其他优化器排队。

该优化器有很多方法，最常用的是迷你 minimize。它实际上是把 compute_gradients() 和 apply_gradients() 两个函数组合一起计算。用过想更灵活的定制优化过程时，可以自己使用 compute_gradients() 和 apply_gradients() 定制优化过程。我们这里不介绍这两个函数。

```
minimize(
    loss,
    global_step=None,
    var_list=None,
    gate_gradients=GATE_OP,
    aggregation_method=None,
    colocate_gradients_with_ops=False,
    name=None,
    grad_loss=None
)
```

loss : 是损失函数

global_step : 可选参数，赋给一个 Variable，模型训练时的参数 Variable 更新一次后，global_step 加 1。

var_list: 可选，保存了要被更新的参数 Variable 的 list 或者 tuple。默认的是计算图中的所有参数 tranalbe 为 True 的 Variable。

gate_gradient : 用于控制梯度计算的并行化程度。有三个选项 GATE_NONE, GATE_OP, or GATE_GRAPH。GATE_NONE 提供最大可能的并行。但结果可能导致计算的不可复制性，当多个梯度的计算之间存在关联。GATE_OP 对于每个 Operation，在使用该操作之前，确保所有的梯度已经计算完成。GATE_GRAPH 确保所有 Variable 的所有梯度都计算完成，在使用它们之前。该选项的并行化程度最低。

aggregation_method 设定用于联合 gradient terms 的方法。可以减少内存的占有。参见类 AggregationMethod.

`colocate_gradients_with_ops` , If True, try colocating gradients with the corresponding op.

`grad_loss` 可选项，一个 tensor，保存为 loss 计算的梯度。

2. **tf.train.AdadeltaOptimizer**

Adadelta 优化算法。一个学习率自适应调整的梯度下降算法。它比传统的梯度下降算法有很小的计算开销。它的构造函数的参数包括

`learning_rate=0.001`, 学习率

`rho=0.95`, 衰减率 `decay`

`epsilon=1e-08`, 给梯度更新设定的条件

`use_locking=False`, 见传统的梯度下降算法中描述的该参数

该函数和 **GradientDescentOptimizer** 一样，也有 `minimize` 函数用于对损失函数最小化。也有 `compute_gradients()` 和 `apply_gradients()` 两个函数两个函数可以定制梯度下降的计算过程。`Minimize` 函数详见 `GradientDescentOptimizer` 中的描述。

3. **tf.train.AdagradOptimizer**

Adagrad 优化算法。自适应调整 subgradient 的优化算法。

4. **tf.train.AdagradDAOptimizer**

5. **tf.train.MomentumOptimizer**

6. **tf.train.AdamOptimizer**

Adam 优化算法的详细内容 P60 的介绍。Tensorflow 提供的该函数的参数包括一个默认的学习率 `learning_rate=0.001`,

使用 `AdamOptimizer` 可以直接使用

```
train = tf.train.AdamOptimizer(1e-4).minimize(loss)
```

`loss` , 是定义的模型损失函数。用户也可以将优化过程分成两部分 `compute_gradients` 和 `apply_gradients`。然后在这两步之间，应用用户自定义的操作。例如，梯度爆炸

时，进行梯度剪切（Gradient Clipping），见 P53。也有方法如下显示全局的训练次数。

```
optimizer = tf.train.AdamOptimizer(0.0001)
grads_and_vars = optimizer.compute_gradients(loss)
global_step = tf.Variable(0, name="global_step", trainable=False)
train_op = optimizer.apply_gradients(grads_and_vars, global_step=global_step)
```

- 7. `tf.train.FtrlOptimizer`
- 8. `tf.train.ProximalGradientDescentOptimizer`
- `tf.train.ProximalAdagradOptimizer`
- 9. `tf.train.RMSPropOptimizer`

Tips:

每种优化器的 `learning_rate` 的量级都不太一样。如梯度下降优化器 `GradientDescentOptimizer` 的学习率在 0.5 左右调整。AdamOptimizer 优化器的学习率的默认值是 0.001；`RMSPropOptimizer` 没给默认值，可以设为 0.001

练习：将第三章第二节的手写数字模型扩展到超过三层，争取获得最低的损失率。

第五章：卷积神经网络

卷积神经网络（convolutional neural network，CNN or ConvNet）是一类前馈神经网络，是受生物学启发的多层感知机的变体。生物学家研究猫的视觉皮层发现视觉皮层上的细胞的排列很复杂。这些细胞对视域的一个小的子区域敏感，称作 receptive field 感受野。感受野指听觉系统，视觉系统和本体感觉系统的一些特质。比如在视觉神经系统中，一个神经元的感受野是指视网膜上的特定区域，只有这个区域内的刺激才能激活该神经元。多个子区域排列，覆盖整个视域。这些细胞相当于在输入空间上加入局部滤波器。非常适合展现自然图像的空间上的局部相关性。

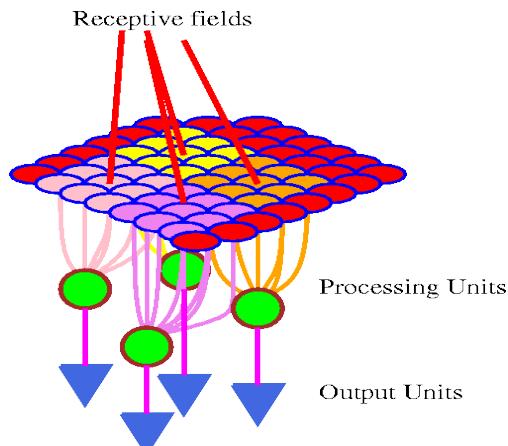


图 5.1：视觉系统上的感受野

卷积神经网络非常适合处理网格拓扑结构的数据，例如，时间序列数据，看做是 1-D 网格；图像数据，看做是 2-D 网格。卷积神经网络有非常的成功例子。

第一节：卷积

卷积是分析数学中一种重要的运算。在泛函分析中，它是通过两个函数 f 和 g 生成第三个函数的一种数学算子。我们这里只考虑离散序列的情况。一维卷积经常用在信号处理中。

例 1：

有两个离散序列：

$$x(n) = \begin{cases} 1, & 0 \leq n \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

$$h(n) = \begin{cases} 1, & 0 \leq n \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

进行卷积计算得到一个新的序列 $y(n)$

$$y(n) = \sum_{i=-\infty}^{\infty} x(i) \cdot h(n-i)$$

我们可以得到 y 的序列

$$y(0)=1, y(1)=2, y(2)=3, y(3)=3, y(4)=3, y(5)=3, y(6)=2, y(7)=1$$

$$\text{其余 } y(n)=0$$

我们也可以按照滤波器的方式来理解离散卷积。给定一个输入信号序列 $x_t, t=1, \dots, n$, 和滤波器 $f_t, t=1, \dots, m$, 一般情况下滤波器的长度 m 远小于信号序列长度 n 。

卷积的输出为：

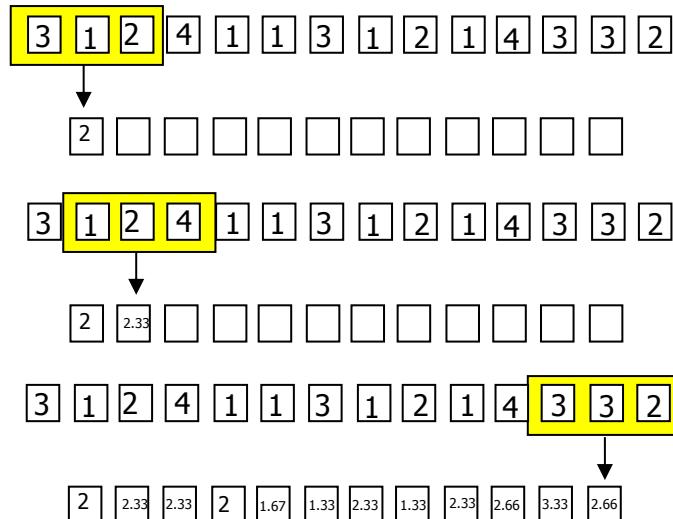
$$y_t = \sum_{k=1}^m f_k \cdot x_{t-k-1}$$

当滤波器 $f_t=1/m$ 时, 卷积相对于信号序列的移动平均。即可以理解为对 x 序列上, 宽为 m 的子序列求平均。

例 2：有一个 x 序列

$$[3] [1] [2] [4] [1] [1] [3] [1] [2] [1] [4] [3] [3] [2]$$

滤波器为 $f_t=1/m, m=3$ 。则产生的 y 序列为



如果对于不在 $[1,n]$ 范围内的 x_t 用零补齐 (zero-padding) , y 序列输出的长度是 $n+m-1$, 称为宽卷积。如果不补零, 输出序列长度是 $n-m+1$, 称为窄卷积。除非特殊声明, 下面所说的卷积默认为窄卷积。

补充 :

设 m 是原始序列长度, n 是滤波器的宽度, s 是 stride。一个序列卷积操作后的长度的计算公式

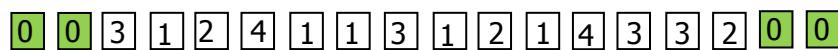
$$l = \left\lfloor \frac{m - n}{s} \right\rfloor + 1$$

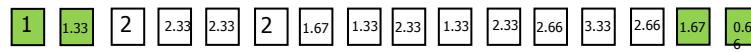
如果有补齐操作, 原始序列长度为 m , 补齐后的长度是 $m + (n - 1) \times 2$

则有补齐的卷积操作的输出长度是

$$l = \left\lfloor \frac{m + n - 2}{s} \right\rfloor + 1$$

上述例子中, 例 1 是宽卷积, 例 2 是窄卷积。例 2 对应的宽卷积如下:





上面例子 2 是一个一维窄卷积。而在图像处理中经常用二维卷积。给定一个图像 x_{ij} , $1 \leq i \leq M$, $1 \leq j \leq N$, 和滤波器 f_{ij} , $1 \leq i \leq m$, $1 \leq j \leq n$, 一般 $m \ll M$, $n \ll N$ 。卷积的输出为:

$$y_{ij} = \sum_{u=1}^m \sum_{v=1}^n f_{uv} \cdot x_{i-u+1, j-v+1}$$

在图像处理中, 常用均值滤波器, 就是当前位置的像素值设为滤波器窗口中所有像素的平均值, 也就是 $f_{uv} = \frac{1}{mn}$

二维卷积就在一个矩阵上面应用一个滤波器的过程。如图 5.2 所示

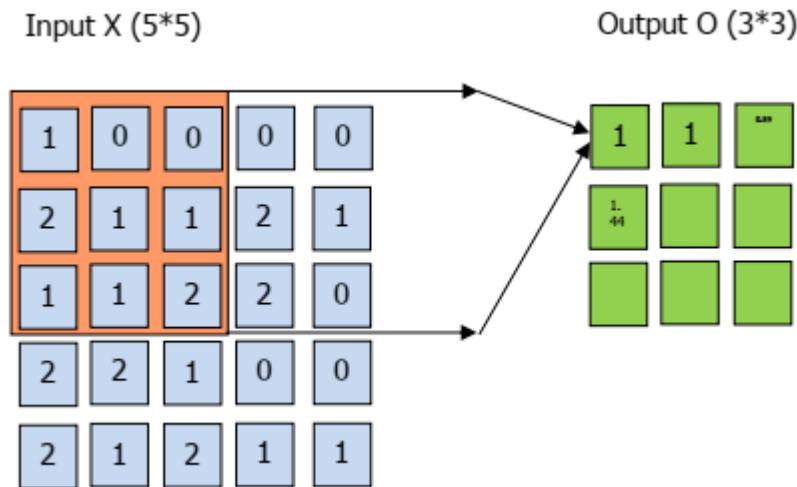


图 5.2 二维卷积

需要说明的是，上面例子中卷积的滤波器都是计算的窗口内元素的均值（移动平均）。下一节可以看到，滤波器也可以有别的运算。

第二节：卷积神经网络的结构

5.2.1 CNN 的特征

根据卷积滤波的思想我们构建卷积神经网络，它具有下面的特性：

1. 局部连接

CNN 利用局部空间上的相关性，它强制在邻近层的神经元之间建立一个局部联通模式。即，在隐层 m 层的输入来自于 $m-1$ 层神经元（单元）的一个子集，一个空间上邻近的单元子集。如图 5.3 所示。

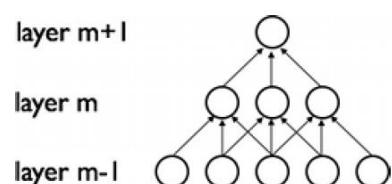


图 5.3：一个 CNN 示例

注：我们前面讲的前馈神经网络是全连接的。

将 $m-1$ 层想象为视网膜，在 m 层的单元在视网膜上有宽度为 3 的感受野，因此仅仅连接到邻近的 3 个神经元（单元）。 $m+1$ 层和其低层（ m 层）也有相似的连接性。我们说， $m+1$ 层的单元相对于 m 层有宽度为 3 的感受野，但相对于输入层（ $m-1$ ）有宽度

为 5 的感受野。每个单元对感受野外的变化不响应。这样的体系结构确保‘滤波器’对空间上的局部输入模式做最强的响应。

然而，我们也可以看到，该结构上如果加的隐层越多导致‘滤波器’成为更加全局化，即对更大的像素空间（输入）进行响应。例如， $m+1$ 隐层的单元可以对宽带为 5 的非线性特征进行编码（encode）。

2. 共享权重

换个角度，可以把 CNN 的滤波器理解为一组边的权重值。滤波器的大小（或组中，值的个数）与局部连接中连接到 m 层中一个神经元的 $m-1$ 层的神经元个数相等。例如，图 5.3 中 $m-1$ 层有邻近的三个神经元连接到 m 层的一个神经元，因此滤波器的大小为 3。CNN 中一个滤波器 h 给连接到隐层 m 每个神经元的边分配权重，因此边共享相同的权重。例如图 5.4 中， m 层有三个神经元， $m-1$ 层的邻近三个神经元连接到 m 层的一个神经元 v_i 时，使用滤波器分配边权重；连接到 m 层神经元 v_{i+1} 的三个边也使用该滤波器分配边权重。进而形成特征映射（feature map）。CNN 中可以设置多个滤波器，进而一个隐层的输出是多个特征映射。

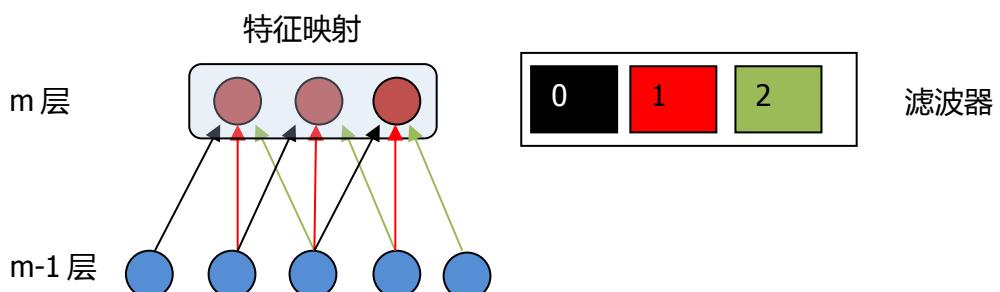


图 5.4：特征映射

图 5.4 中，有一个滤波器，滤波器大小为 3。因此 $m-1$ 隐层每三个相邻的神经元连接到一个 m 层的神经元，边权重由滤波器分配（一组滤波器权重值用红、蓝、绿三色表示）。可以看到隐层 m 的三个单元构成一个 feature map。相同颜色的边共享权重。虽然边共享了权重，但只需做小的算法上的改动，仍可以用梯度下降的方法学习模型参数。一个共享权重的梯度是共享的参数的梯度和。权重共享可以提高模型训练的效率，因为相对的参数数量减少了。卷积操作就来自于共享权重，相当于对一个区域的单元做了算术运算。

我们再用刚才的一维卷积的例子来理解 CNN 的操作。

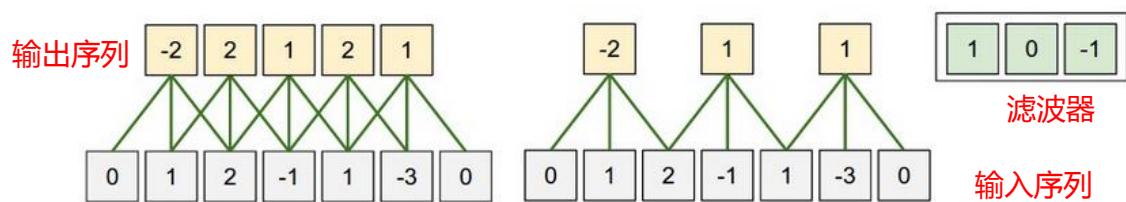


图 5.5 一维卷积

将输入序列理解为神经网络的输入层；滤波器实际上是边的权重。输出序列是神经网络的一层。滤波器实际上表现为了边的权重，即上述的共享边权重。得到的一个输出序列是一个特征映射 feature map。从图中可以看出，将上图看做是神经网络中的两层，卷积操作中需要学习的边权重实际上只有 3 个，即滤波器中的三个值。图 5.5 给了不同步长的两个例子，左边是步长为 1 时的卷积操作；右边是步长为 2 的卷积操作。

再举例：有两个滤波器，滤波器大小为 4。因此， $m-1$ 层每 4 个相邻神经元连接到 m 层的一个神经元。如图 5.6 所示。可以看到应用多个滤波器后可以将原始的数据升维，图中从一维，升维到二维。这样就提供了更丰富的数据处理。我们把输入数据应用滤波器得到的输出称为特征映射（feature map）。

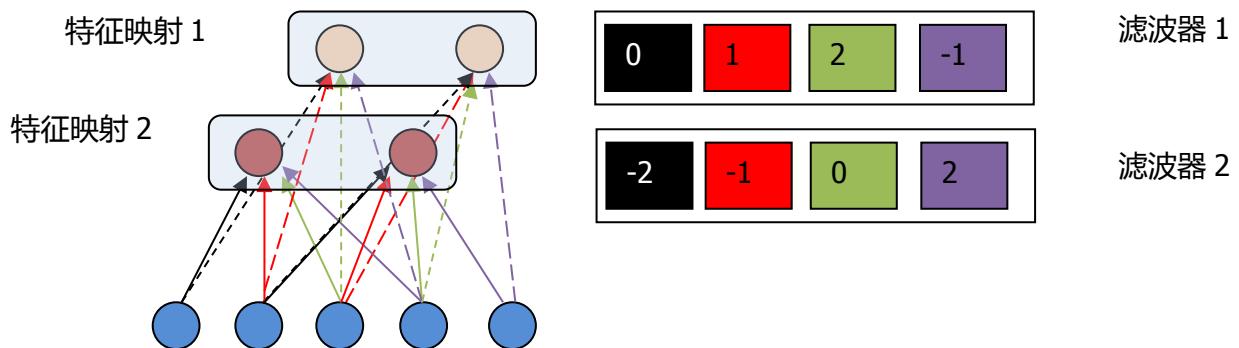


图 5.6 多个滤波器

5.2.2 卷积层

这里的卷积层是指构建卷积神经网络中的一个进行卷积操作的层。

Tips:

为了好理解下面的内容，我们说卷积神经网络的卷积操作运算结果即为神经网络的一层。运算结果（向量或矩阵）的每个元素，就是一个神经元。

1. 一维卷积层

在全连接前馈神经网络中，如果第 l 层有 n^l 个神经元，第 $l-1$ 层有 n^{l-1} 个神经元，连接边有 $n^l \times n^{l-1}$ 个。也就是权重矩阵有 $n^l \times n^{l-1}$ 个参数。当隐层增加，或一层中的神经元增加，权重矩阵的训练参数非常多，训练效率会降低。

如果采用卷积来代替全连接，第 l 层的每一个神经元都只和第 $l-1$ 层的一个局部窗口内的神经元相连，构成一个局部连接网络。第 l 层的第 i 个神经元的输出定义为：

$$a_i^{(l)} = f \left(\sum_{j=1}^m w_j^{(l)} \cdot a_{i-j+m}^{(l-1)} + b^{(l)} \right) = f(W^{(l)} \cdot a_{(i+m-1):i}^{(l-1)} + b^{(l)})$$

其中 $W^{(l)} \in R^{m \times m}$ 为 m 维的滤波器， $a_{(i+m-1):i}^{(l-1)} = [a_{i+m-1}^{(l-1)}, \dots, a_i^{(l-1)}]^T$ ； f 是激活函数。这里 $a^{(l)}$ 的下标从 1 开始。上述的公式也可以写成

$$a^{(l)} = f(W^{(l)} \otimes a^{(l-1)} + b^{(l)})$$

\otimes 表示卷积运算。从该公式可以看出， $W^{(l)}$ 对于所有的神经元都是相同的。这就是卷积层的权重共享特性。这样，在卷积层，只需要 $m+1$ 个参数。**另外，第 $l+1$ 层的神经元个数不是任意选择的，而是满足 P82“补充”中的公式。（此时假设的卷积操作中没有补齐， $stride=1$ ）**

Tips:

- (1) 一个滤波器有一个偏置（标量）
- (2) 滤波器的参数和偏置都是待学习的参数

2. 二维卷积层

上述公式描述的是一维卷积层。在图像处理中，图像是以二维矩阵的形式输入到神经网络中，在文本处理时，很多情况也是将文本转换成了二维矩阵。因此需要二维卷积。假设 $X^{(l)} \in R^{(w_l \times h_l)}$ 和 $x^{(l-1)} \in R^{(w_{l-1} \times h_{l-1})}$ 分别是第 l 层和第 $l-1$ 层的神经元。 $X^{(l)}$ 的每一个元素（一个神经元）为：

$$X_{s,t}^{(l)} = f \left(\sum_{i=1}^u \sum_{j=1}^v W_{i,j}^{(l)} \cdot X_{s-i+u, t-j+v}^{(l-1)} + b^{(l)} \right)$$

注意：这里 W 和 w 含义不同。 $W^{(l)} \in R^{(u \times v)}$ 是第 l 层的二维滤波器。 w 是以矩阵描述一层的神经元的个数时的宽度（因为计算的数据是矩阵形式，**想象一下神经元是按照矩阵排列的**） w_l 是第 l 层的神经元组的宽度， h_l 是排列的列数，第 l 层的神经元个数为 $w_l \times h_l$ ，并且 $w_l = w_{l-1} - u + 1$ ， $h_l = h_{l-1} - v + 1$ 。则上面的公式也可以写为

$$X^{(l)} = f(W^{(l)} \otimes X^{(l-1)} + b^{(l)})$$

为了增强卷积层的表示能力，可以在输入上使用 K 个不同的滤波器来得到 K 组输出。如果把滤波器看做是一个特征提取器，每一组输出都可以看成是输入图像经过一个特征提取后得到的特征。因此，在卷积神经网络中每一组输出也叫作一组**特征映射**（feature map）。不失一般性，我们假设第 $l-1$ 层输出的特征映射组数为 n_{l-1} ，每组特征映射的大小为 $m_{l-1} = w_{l-1} \times h_{l-1}$ 。第 $l-1$ 层的总神经元数 $n_{l-1} \times m_{l-1}$ 。第 l 层的特征映射组数为 n_l （有 n_l 个滤波器）。第 l 层的第 k 组特征映射 $X^{(l,k)}$ 为

$$X^{(l,k)} = f\left(\sum_{p=1}^{n_{l-1}} (W^{(l,k,p)} \otimes X^{(l-1,p)}) + b^{(l,k)}\right)$$

其中， $W^{(l,k,p)}$ 表示第 l 层的第 k 个滤波器。一个滤波器的维度是由输入决定的。输入有 p 个特征映射，每个特征映射是二维的。滤波器则是三维的，前两个维度是对应一个输入特征映射的滤波器的大小，第三个维度大小是 p ，即输入特征映射的数目。

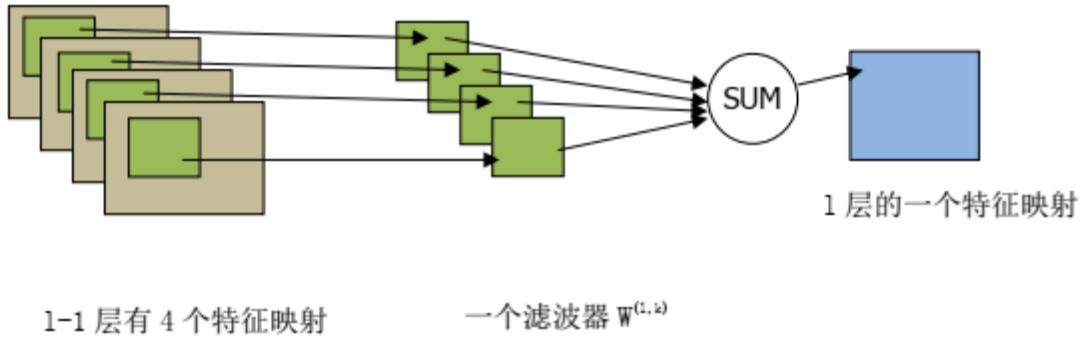


图 5.7 卷积操作示例

下面我们用图来演示一下二维卷积操作：设输入为 5×5 的矩阵（下图假设输入只有一个特征映射），滤波器是 3×3 ，不填充，步长（stride）为 1。输入特征映射经过一个滤波器的运算后，得到一个输出的特征映射。

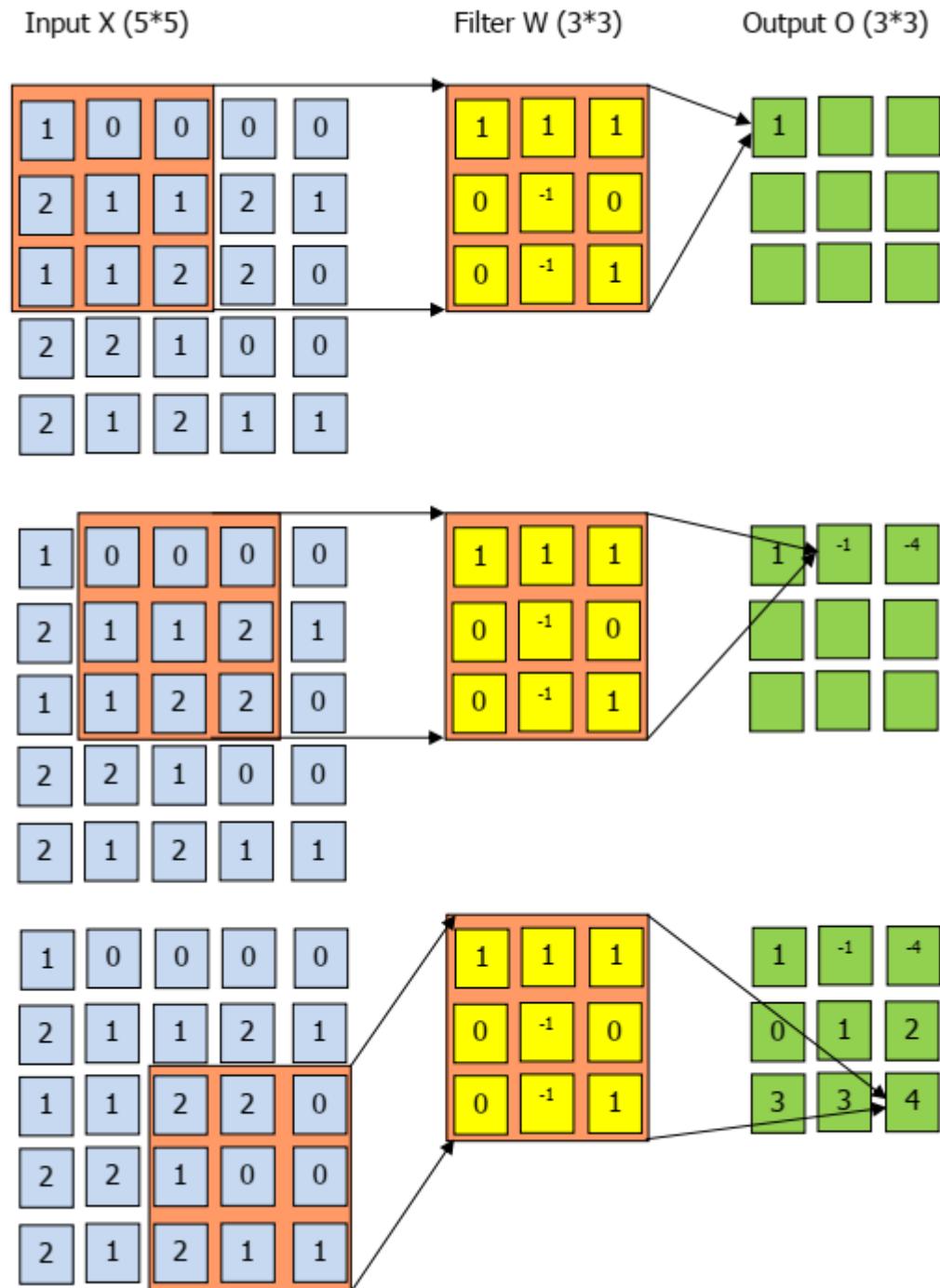


图 5.8 卷积操作示例 2

再举例：输入（ $5 \times 5 \times 3$ ）（假设输入特征映射有 3 个），步长 stride 为 2，滤波器（ $3 \times 3 \times 3$ ）个数为 2，zero-padding 填充量为 1。



图 5.9 卷积操作示例 3

图 5.8 的例子中个滤波器只是一个二维的矩阵，而图 5.9 的例子中滤波器是一个 cube。即对每个输入的特征映射应用不同的滤波器，且按照图 5.7 对一个滤波器的输出特征映射进行了求和。输出 output 为 $3 \times 3 \times 2$ ，其中的 2 是 output_channels，它由 filter 的个数决定。在 <http://cs231n.github.io/convolutional-networks/> 有个二维卷积操作过程的动画演示。（注，我们的图和该网页的图数据不一样）。我们可以用图 5.10 来描述二维卷积层的从输入到输出的映射关系

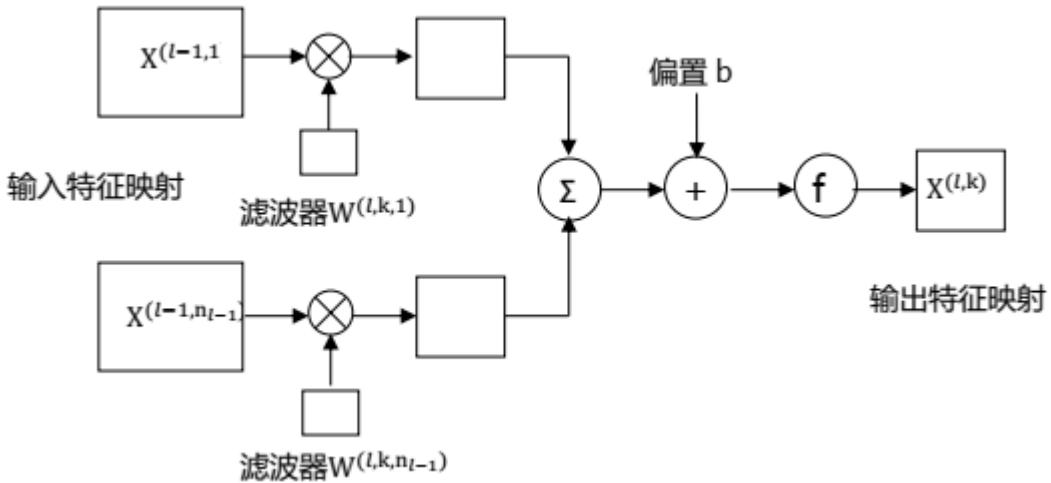


图 5.10 二维卷积层的从输入到输出的映射关系（第 k 个滤波器的例子）

此处的输入的多个特征映射，有 n_{l-1} 个，可以看成是整个 CNN 的输入（如果第一层是卷积层），此时以输入是图像为例，图像的 shape 是 [width, height, channel]，channel 是指图像的通道；如果是 RGB 三色，那么此时输入的是三个特征映射。输入的多个特征映射也可以将当前卷积层看做是 CNN 的第 l 层，则第 l 层的输入特征映射是 $l-1$ 层的输出。此时滤波器的 shape 是 [filter_height * filter_width * in_channels]。图上显示的是第 k 个滤波器。滤波器 $W^{(l,k,n_{l-1})}$ 的含义是第 l 层第 k 个滤波器的第 n_{l-1} 个 channel，它负责对输入的第 n_{l-1} 个特征映射进行卷积操作。这里的 in_channels 等于输入特征映射的个数。之所以此处使用 in_channels 这个表示法，是为了和 TensorFlow 中的参数表示法对应。

我们可以看到一个滤波器中的每个 channel 对一个输入特征映射进行卷积操作，再将各结果求和，再加上偏置，最后得到一个输出映射。有几个滤波器，就有几个输出映射。

5.2.3 子采样层（或池化层）

卷积层虽然可以显著的减少连接的个数，但是每一个特征映射的神经元个数并没有显著减少。这样，如果后面接一个分类器，分类器的输入维数依然很高，很容易出现过拟合。为了解决这个问题，在卷积神经网络一般会在卷积层后再加上一个池化操作

(Pooling) , 也就是子采样 (subsampling) , 构成一个子采样层。子采样层可以大大降低特征的维度 , 避免过拟合。子采样函数 $\text{down}(\cdot)$ 一般是取区域内所有神经元的最大值 (Maximum Pooling) 或平均值 (Average Pooling)

$$\text{down}(\cdot) = \text{pool}_{\max}(R_k) = \max_{i \in R_k} a_i$$

$$\text{down}(\cdot) = \text{pool}_{\text{avg}}(R_k) = \frac{1}{|R_k|} \sum_{i \in R_k} a_i$$

子采样的作用还在于可以使得下一层的神经元对一些小的形态改变保持不变性 , 并拥有更大的感受野。

子采样层 (或池化层) 在输入的 depth 维度上的每个切片进行操作。最常用的形式是一个 pooling 层使用 size 为 $2*2$ 的滤波器 , 在输入立方体的 depth 维度 , 对每个 depth 切片进行步长 stride 为 2 , Pool_{\max} 操作进行子采样。这样有 75% 的神经元被放弃 , depth 维保持不变。

更一般性的描述 pooling 层 :

- (1) Pooling 层的输入是一个 $W_1*H_1*D_1$ 的输入立方体 (Volume)
 - (2) 需要两个超参数: 进行采样的滤波器的 size(滤波器是一个正方形)F ; 步长 S
 - (3) Pooling 层输出立方体的 size: $W_2*H_2*D_2$
- $$W_2=(W_1-F)/S+1; H_2=(H_1-F)/S+1; D_2=D_1$$
- (4) Pooling 层不会使用零填充 zero-padding
 - (5) Pooling 层没有引入参数

图 5.11 展示了一个 pooling 操作的过程。pooling 层空间独立地在输入立方体的每个 depth 切片上下采样立方体。

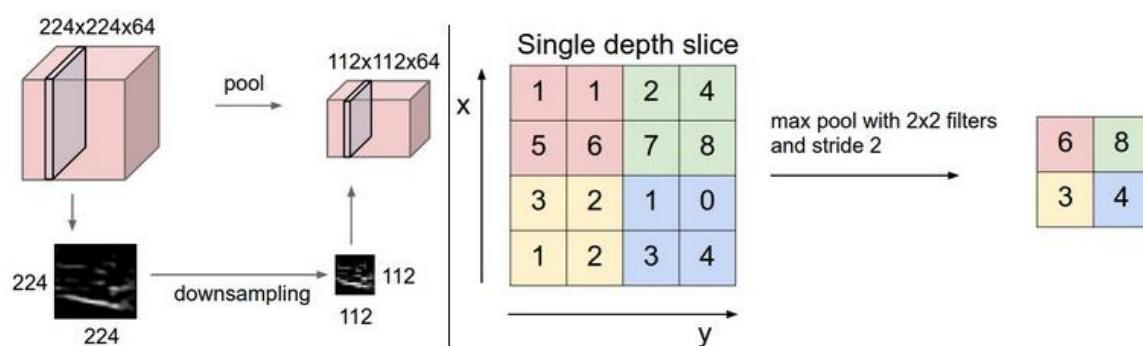


图 5.11 子采样示意图

先看左边，输入是一个 $224 \times 224 \times 64$ 的立方体；它的 depth 维是 64，即有 64 个切片 slice，或者可以理解为输入有 64 个 feature map。Pooling 操作的滤波器 size 是 2×2 （对应前述的 pooling 层超参数 $F=2$ ），操作步长为 $S=2$ 。因此 pooling 层的输出立方体是： $W_2 = (224-2)/2+1 = 112$; $H_2 = 112$; $D_2 = D_1 = 64$ 。

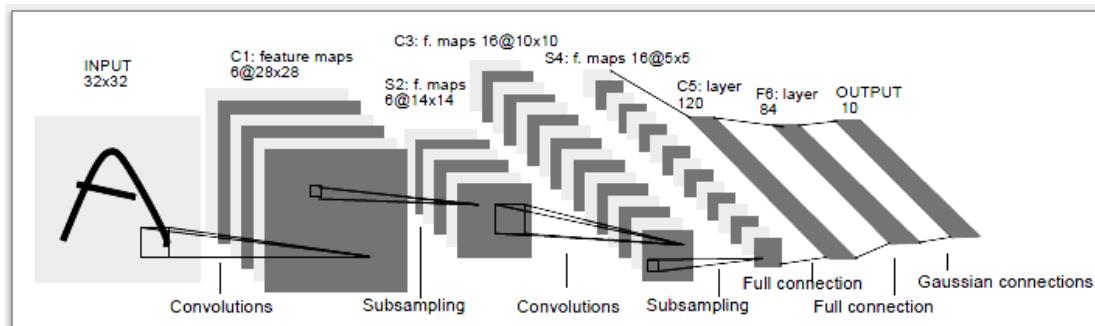
再看右边的 pooling 操作过程。此例中的一个 depth slice(或 feature map)的维度是 4×4 ；滤波器的维度是 2×2 ，步长为 2；pooling 操作采样 MAX 操作，即去滤波器对应区域中的最大值。因此，pooling 操作的输出得到 4 个值。

也有人对 pooling 层有不同意见，他们认为完全可以抛弃 pooling 层。他们建议在卷积层使用更大的步长，就可以有效的减小神经元的数目。

5.2.4 全连接层

卷积神经网络中通常会包含一个全连接层。全连接层的神经元与前一层的所有神经元都有连接，与传统的神经网络一样。

第三节：卷积神经网络示例

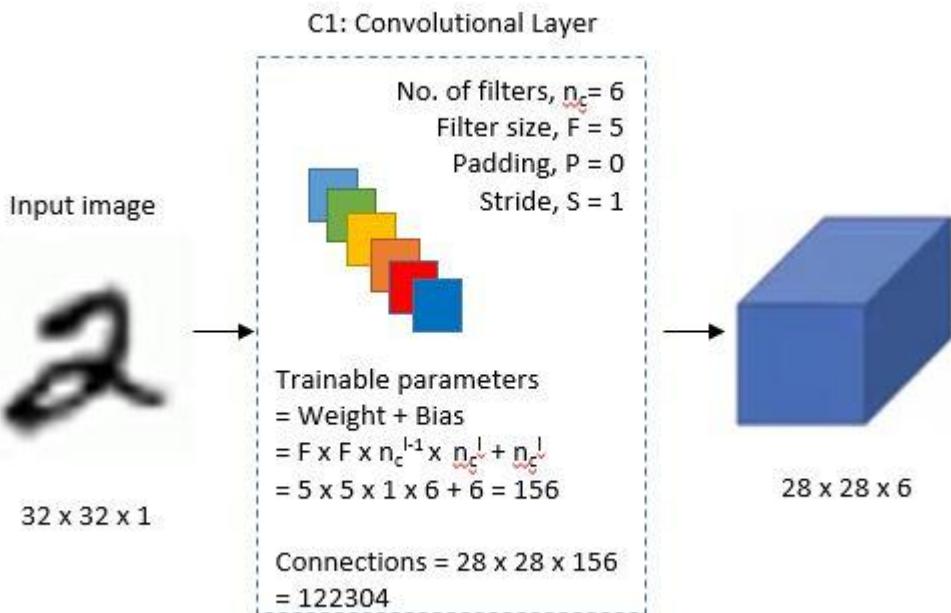


5.12 LeNet-5 结构图

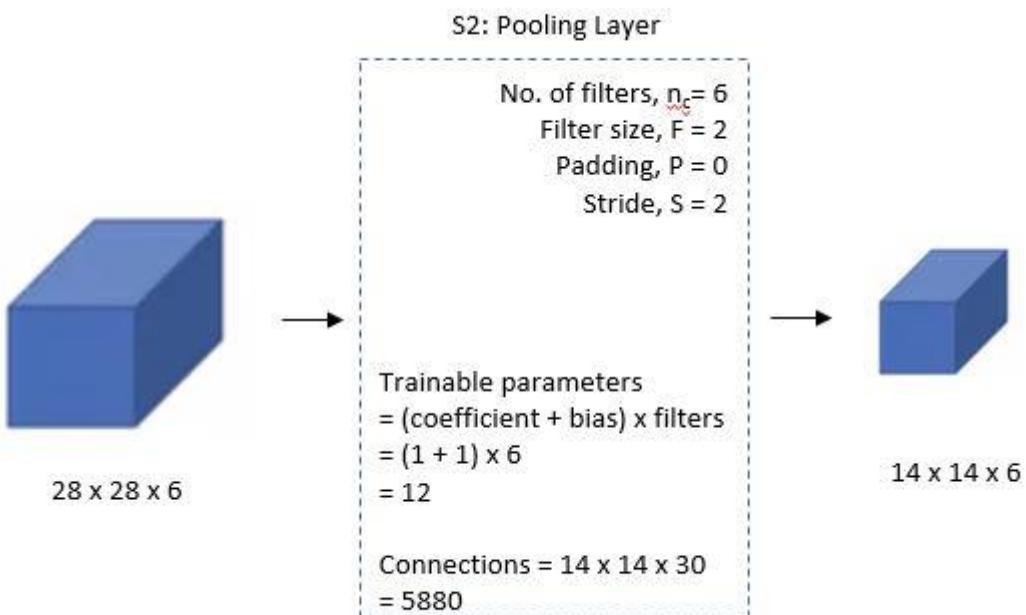
LeNet 是第一个成功应用的卷积神经网络，由 Yann LeCun 在上世纪九十年代开发。LeNet-5 在美国银行系统中已经非常成功的应用在了支票上手写数字的识别。LeNet-5 的网络结构如图 5.12 所示。

不算输入层，LeNet-5 共有 7 层，每一层结构为：

- (1) 输入层：输入图像大小为 $32 \times 32 = 1024$ 。
- (2) C1 层：卷积层。滤波器的大小是 $5 \times 5 = 25$ ，共有 6 个滤波器。得到 6 组大小为 $28 \times 28 = 784$ 的特征映射。因此，C1 层的神经元个数为 $6 \times 768 = 4,704$ 。可训练参数个数为 $6 \times 25 + 6 = 156$ 。连接数为 $156 \times 784 = 122,304$ （包括偏置在内，下同）。

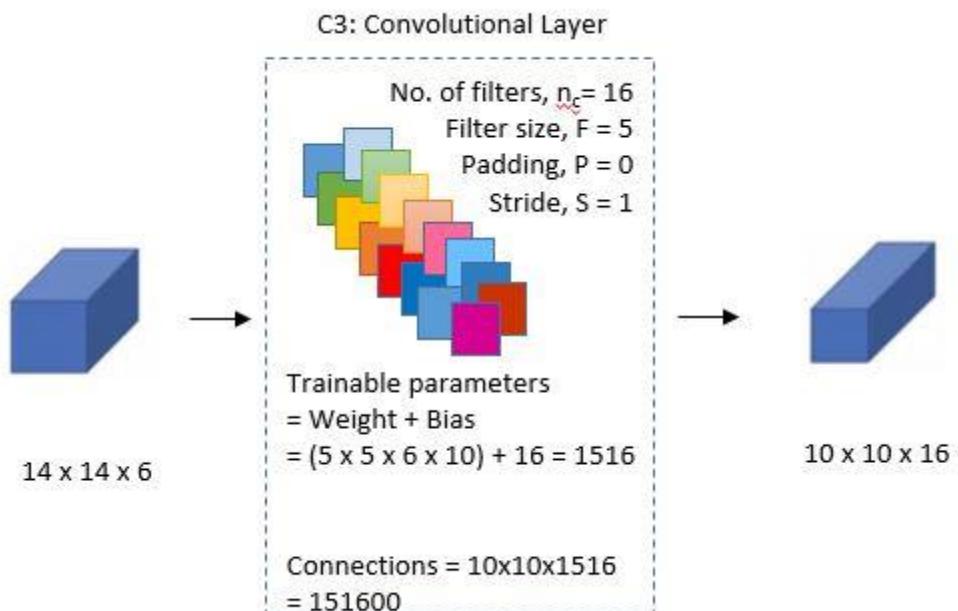


(3) S2 层：子采样层。由 C1 层每组特征映射中的 2×2 领域点按照平均值操作方法次采样 (down sampling) 为 1 个点。这一层的神经元个数为 $14 \times 14 = 196$ 。可训练参数个数为 $6 \times (1+1) = 12$ 。连接数为 $6 \times 196 \times (4+1) = 122,304$ (包括偏置的连接)。

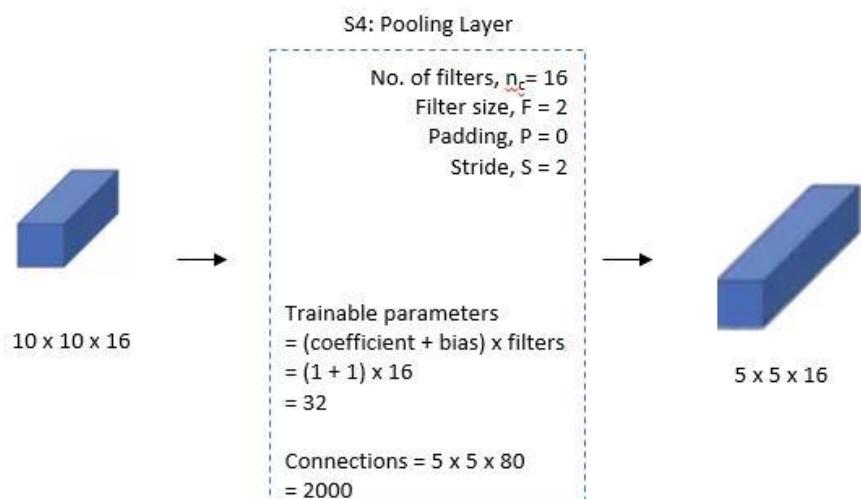


(4) C3 层：卷积层。由于 S2 层也有多组特征映射，需要一个连接表来定义不同层特征映射之间的依赖关系。LeNet-5 的连接表如图 5.13 所示。这样的连接机制的基本假设是：C3 层的最开始 6 个特征映射依赖于 S2 层的特征映射的每 3 个

连续子集。接下来的 6 个特征映射依赖于 S2 层的特征映射的每 4 个连续子集。再接下来的 3 个特征映射依赖于 S2 层特征映射的每 4 个不连续子集。最后一个特征映射依赖于 S2 层的所有特征映射。这样共有 16 个滤波器，大小是 $5 \times 5 = 25$ 。得到 16 组大小为 $10 \times 10 = 100$ 的特征映射。C3 层的神经元个数为 $16 \times 100 = 1600$ 。可训练参数个数是 $6 \times 10 \times 25 + 16 = 1516$ ；连接数为 $1516 \times 100 = 151,600$ 。



(5) S4 层：是一个子采样层，由 2×2 领域点 down sampling 为 1 个点，得到 16 组 5×5 大小的特征映射。可训练参数个数为 $16 \times 2 = 32$ 。连接数为 $16 \times (4+1) \times 5 \times 5 = 2000$ 。



(6) C5 层：卷积层。有 120 个滤波器，大小是 $5 \times 5 = 25$ 。得到 120 组大小为 1×1 的特征映射。每个特征映射与 S4 层的全部特征映射连接。C5 层的神经元个数为 120，可训练参数个数为 $1920 \times 25 + 120 = 48,120$ 。连接数为 $120 \times (16 \times 25 + 1) = 48,120$ 。

(7) F6 层：一个全连接层，有 84 个神经元，可训练参数个数为 $84 \times (120 + 1) = 10,160$ 。连接数和可训练参数个数相同，为 10,164。

(8) 输出层：由 10 个径向基函数 (Radial Basis Function, RBF) 组成。

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X			X	X	X			X	X	X	X	X	X	X	
1	X	X			X	X	X			X	X	X	X	X	X	
2	X	X	X			X	X	X		X		X	X	X	X	
3		X	X	X		X	X	X	X		X		X	X	X	
4			X	X	X		X	X	X	X	X	X	X	X	X	
5			X	X	X		X	X	X	X	X	X	X	X	X	

图 5.23 S2 层到 C3 层的连接表

每列指示 S2 层的哪个特征映射和 C3 层的特征映射的单元相连。

第四节：Dropout

有大量参数的深度神经网络是非常有力的机器学习系统，然而这样的网络过拟合是个很严重的问题。Dropout 是深度学习中防止过拟合的一个简单却非常有效的技巧（参看 Hinton 的文章 Dropout: A Simple Way to Prevent Neural Networks from Overfitting）。术语“dropout”是指在神经网络中 dropping out units（隐层节点）

我们通过 Tensorflow 提供的 dropout 函数来理解它是怎么工作的。

dropout 函数 `tf.nn.dropout(x, keep_prob, noise_shape=None, seed=None, name=None)`

按照概率 `keep_prob` 对输入的元素进行选择输出。选择了的输出，其元素值乘上 $1/\text{keep_prob}$ 进行标定；否则输出为 0。

```
import tensorflow as tf
d=[3.0,1.0,2.0,4.0]
input1 = tf.Variable(tf.constant(d))
drop = tf.nn.dropout(x=input1, keep_prob=0.5)
init_op=tf.initialize_all_variables()
```

```

mycount=len(d)*[0]

sess=tf.Session()
for _ in range(10000):
    sess.run(init_op)
    d=sess.run(drop)
    ind=[i for i, x in enumerate(d) if x==0]
    for x in ind:
        mycount[x] = mycount[x]+1

print(d)
print(mycount)

```

运行结果是：（因为存在随机选择节点，因此每次运行的结果会不同）

```
[ 6. 2. 0. 8.]
[4995, 5009, 5039, 4916]
```

可以看出 `tf.nn.dropout` 函数对输入 (`x=input1`) 按照 `keep_prob` (50%) 的概率来决定每个输入的元素是否作为 `tf.nn.dropout` 的输出。如果决定输出，其输出值是原值乘上 `1/keep_prob`

关于 Dropout，文章中没有给出任何数学解释，Hinton 的直观解释和理由如下：

1. 由于每次用输入网络的样本进行权值更新时，隐含节点都是以一定概率随机出现，因此不能保证每 2 个隐含节点每次都同时出现，这样权值的更新不再依赖于有固定关系隐含节点的共同作用，阻止了某些特征仅仅在其它特定特征下才有效果的情况。
2. 可以将 dropout 看作是模型平均的一种。对于每次输入到网络中的样本（可能是一个样本，也可能是一个 batch 的样本），其对应的网络结构都是不同的，但所有的这些不同的网络结构又同时共享隐含节点的权值。这样不同的样本就对应不同的模型，是 bagging 的一种极端情况。
3. naive bayes 是 dropout 的一个特例。Naive bayes 有个错误的前提，即假设各个特征之间相互独立，这样在训练样本比较少的情况下，单独对每个特征进行学习，测试时将所有的特征都相乘，且在实际应用时效果还不错。而 Dropout 每次不是训练一个特征，而是一部分隐含层特征。
4. 还有一个比较有意思的解释是，Dropout 类似于性别在生物进化中的角色，物种为了使适应不断变化的环境，性别的出现有效的阻止了过拟合，即避免环境改变时物种可能面临的灭亡。

TensorFlow 实现了 Dropout 的操作。4.9 节的第 4 段，给出了 TensorFlow 实施 dropout 的描述

Tips:

实践中使用 dropout 的技巧是。在 CNN 中，最后的卷积层（包括 pooling 层）后加一个全连接层，再加一个 dropout 操作。实验证实可以大幅度的提供模型的性能。至于为什么，没有一个科学的解释。

第五节：用 TensorFlow 实现 CNN

在第 3 章我们已经用 TensorFlow 实现了一个基本的 SoftMax 回归模型来实现手写数字的识别该模型的精确度达到了 91%，实际上这是一个很差的性能。这一节我们将实现一个 CNN 再次进行手写数字的识别，看看精确度的提升。

在 3.3 节我们已知 MNIST 数据集的一张图片是 $28 \times 28 \times 1$ 的数据（灰度图片是单通道）。我们构建的的模型框架如图 5.14 所示：

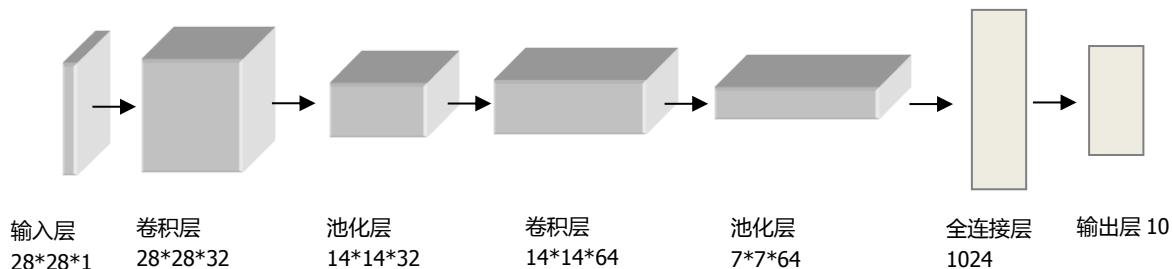


图 5.14 一个用于手写数字识别的卷积神经网络

1. 输入层

首先读入数据，然后创建 placeholder 占位符

```
mnist = input_data.read_data_sets("data/", one_hot=True)

x = tf.placeholder(tf.float32, shape=[None, 784])
y_ = tf.placeholder(tf.float32, shape=[None, 10])
```

此处占位符 x 保存的是长度为 784 的向量，即把输入数据 $mnist$ 保存的数据是把 28×28 的图像转换成了向量的形式。占位符 $y_$ 中的 10 对应输出有 10 种数字，输出是 one-hot 向量。

将读入的以向量形式描述的数据转换成 28×28 的矩阵。

```
x_image = tf.reshape(x, [-1, 28, 28, 1])
```

`tf.reshape` 重新安排 tensor。它的第二个参数就是重新安排的 tensor 的 shape。如果 shape 中的某个维度是 -1，它表示该维度的值不定，由其他维度的值固定后计算来得

到。此处表示转换成维度为 4 的 tensor 时，第一个维度即图像的数量不定，但每张图像的 shape 是固定的 28*28*1。

2. 第一个卷积层和 pooling 层

创建卷积层的第一步是创建滤波器。按照前面讲述的内容，滤波器的参数就是卷积层的权重，是要学习的参数。创建两个 variable，一个是卷积层的权重，即滤波器，一个是偏置变量

```
W_conv1 = weight_variable([5, 5, 1, 32])
b_conv1 = bias_variable([32])
```

创建的权重的 shape=[5,5,1,32]其实是卷积层的滤波器的 size：滤波器的宽和高是 5*5，因为输入数据的通道（特征映射）是 1，这里滤波器的 depth 也是 1，创建 32 个滤波器。每个滤波器有一个偏置，因此偏置变量的设置是 32 个。这里自定义了两个函数

```
def weight_variable(shape):
    initial = tf.truncated_normal(shape, stddev=0.1)
    return tf.Variable(initial)

def bias_variable(shape):
    initial = tf.constant(0.1, shape=shape)
    return tf.Variable(initial)
```

tf.truncated_normal 函数从一个 truncated normal distribution 产生随机数。当产生的随机数超过均值的 2 倍的标准差的范围时，放弃该随机数重新产生。tf.constant 产生常数。

再建立第一个卷积层和池化层

```
h_conv1 = tf.nn.relu(conv2d(x_image, W_conv1) + b_conv1)
h_pool1 = max_pool_2x2(h_conv1)
```

conv2d 是自定义函数，如下

```
def conv2d(x, W):
    return tf.nn.conv2d(x, W, strides=[1, 1, 1, 1], padding='SAME')
```

它调用 tf.nn.conv2d 函数来创建卷积层。该函数是给定一个 4-D 的输入 tensor 和滤波器 tensor，计算一个 2-D 卷积。

```
tf.nn.conv2d(input, filter, strides, padding, use_cudnn_on_gpu=None,
data_format=None, name=None)
```

input 是输入 tensor，它的 shape 是 [batch, in_height, in_width, in_channels]。因为采用的是 minibatch 随机梯度下降 SGD 的优化方法（参见第二章），需要给出一个 batch 参数，即进行训练的一个批次的图像数量。in_height, in_width, in_channels 即

一张图像的高、宽和通道数。一个滤波器（有的文献称作 kernel）tensor 的 shape [filter_height, filter_width, in_channels, out_channels]。此处 in_channels 即对应输入图像的通道个数；out_channels 是该卷积层输出的通道数量，它对应的是滤波器的个数，也即特征映射个数。

Strides 是卷积操作中的步长。Stride 的是一个长度为 4 的向量，我们设 strides = [1, 1, 1, 1]。其中的每个元素对应到输入 tensor 的 shape 的每个维度。如以当前的例子，strides[0] 对应图片数量这个维度；strides[1] 对应一张图片的高度这个维度；strides[2] 对应一张图片的宽度这个维度；strides[3] 对应一张图片的通道维度。Strides 的设置中 strides[0] 和 strides[3] 必须值为 1，strides[1] 和 strides[2] 必须值相等。

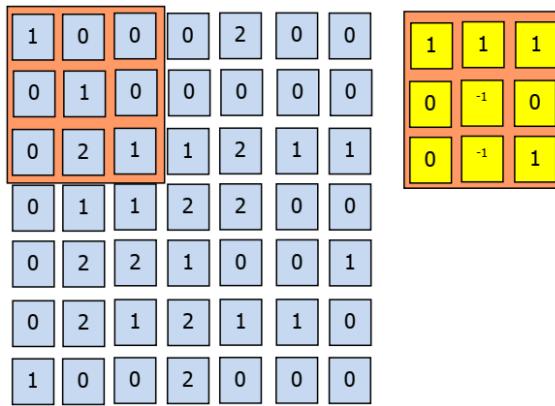
TensorFlow 的补齐方式有些复杂。Padding 参数的设置是两个值的选择：“SAME”和“VALID”。每种选择，通过下面的计算可以算出最后输出 tensor（特征映射组）的 shape。

对于“SAME” padding，我们计算

```
out_height = ceil(float(in_height) / float(strides[1]))
out_width = ceil(float(in_width) / float(strides[2]))
pad_along_height = ((out_height - 1) * strides[1] +
                     filter_height - in_height)
pad_along_width = ((out_width - 1) * strides[2] +
                     filter_width - in_width)
pad_top = pad_along_height / 2
pad_left = pad_along_width / 2
```

分别算出的 pad_along_height 和 pad_along_width 是在高度和宽度的方向补齐几个值。pad_top=pad_along_height/2 是计算出沿高度方向在顶部和底部各补齐多少个值。如果 pad_along_height=5 则在顶部补齐 2 个值，在底部补齐 3 个值。pad_left 类同。

举例：输入 tensor 是 7*7，滤波器 shape 是 3*3，stride 为 2



```

out_height = 4
out_width   = 4
pad_along_height = (4 - 1) * 2 + 3 - 7 = 2
pad_along_width = (4 - 1) * 2 + 3 - 7 = 2
pad_top      = 1 (上下各填充一行)
pad_left     = 1 (左右各填充一列)
补齐以后的输入变成了 9*9 的矩阵，输出是 4*4 的矩阵

```

上例，如果 Stride 为 1，则输出是 7*7 的矩阵。（见 P75 的计算公式）

```

out_height = 7
out_width   = 7
pad_along_height = (7 - 1) * 1 + 3 - 7 = 2
pad_along_width = ((7 - 1) * 1 + 3 - 7 = 2
pad_top      = 1
pad_left     = 1

```

对于 “VALID” padding，我们计算

```

out_height = ceil(float(in_height - filter_height + 1) /
float(strides[1]))
out_width   = ceil(float(in_width - filter_width + 1) /

```

注：所有上面涉及的补齐都是零补齐

则，上例：输入 tensor 是 7*7，滤波器 shape 是 3*3，stride 为 2，输出矩阵是 3*3；如果，输入 tensor 是 7*7，滤波器 shape 是 3*3，stride 为 1，输出矩阵是 5*5。可以看出，实际上 VALID 就是没有进行补齐运算。

卷积层应用了 ReLU 激活函数 `tf.nn.relu(features, name=None)`。它计算 `rectified linear : max(features, 0)`。

Pooling 层 `h_pool1` 的定义中使用的自定义函数定义如下：

```
h_pool1 = max_pool_2x2(h_conv1)
def max_pool_2x2(x):
    return tf.nn.max_pool(x, ksize=[1, 2, 2, 1], strides=[1, 2, 2, 1], padding='SAME')
```

函数 `tf.nn.max_pool` 在输入上进行 max pooling 操作。其参数如下：

`value`: 是一个 4-DTensor 即输入，其 `shape`=[batch, height, width, channels]；数据类型是 `tf.float32`。

`ksize`: 是一个 int 型 list，长度 ≥ 4 。它是池化操作的滤波器的窗口大小，对应输入 tensor 的每个维度。

`strides`: 是一个 int 型 list，长度 ≥ 4 。滑动窗口的步长。

`padding`: 'VALID' 或者 'SAME'。

`data_format`: 'NHWC' 或者 'NCHW'。

`name`: 可选项，分配操作的名称。

3. 第二个卷积层和 pooling 层

再定义第二个卷积层和 Pooling 层的权重

```
W_conv2 = weight_variable([5, 5, 32, 64])
b_conv2 = bias_variable([64])
```

第二个卷积层的权重（即滤波器）的长和宽是 5*5；因为第一个卷积层的滤波器有 32 个，因此输出的特征映射（通道）是 32 个；第二个卷积层使用 64 个滤波器。

定义第二个卷积层和 pooling 层

```
h_conv2 = tf.nn.relu(conv2d(h_pool1, W_conv2) + b_conv2)
h_pool2 = max_pool_2x2(h_conv2)
```

4. 全连接层

定义全连接层的权重

```
W_fc1 = weight_variable([7 * 7 * 64, 1024])
b_fc1 = bias_variable([1024])
```

第二个 pooling 层的输出 tensor 的 shape 是 7*7*64，而全连接层有 1024 个神经元。因此，偏置的数量也是 1024。我们会把第二个 pooling 层输出的 tensor 转换成一个向

量，即 flatten 操作（可以参考传统神经网络中的一层的神经元都是一个向量的排列）。然后进行全连接层的计算，再加上一个 ReLU 激活函数。

```
h_pool2_flat = tf.reshape(h_pool2, [-1, 7*7*64])
h_fc1 = tf.nn.relu(tf.matmul(h_pool2_flat, W_fc1) + b_fc1)
```

对全连接层引入 dropout 操作

```
keep_prob = tf.placeholder(tf.float32)
h_fc1_drop = tf.nn.dropout(h_fc1, keep_prob)
```

dropout 函数 `tf.nn.dropout(x, keep_prob, noise_shape=None, seed=None, name=None)`

按照概率 `keep_prob` 对输入的元素输出，输出的元素值乘上 $1/\text{keep_prob}$ 进行标定；否则输出为 0。

每个元素被保留或放弃是独立的。如果给定了 `noise_shape`，它是一个向量（或 1-D tensor），仅仅满足 `noise_shape[i] == shape(x)[i]` 的维度进行独立的保留或放弃的决策。例如，如果 `shape(x) = [k, 1, m, n]`，并且 `noise_shape = [k, 1, 1, n]`，each batch and channel component will be kept independently and each row and column will be kept or not kept together.

5. 输出层

输出层有 10 个神经元，因为输出是对 10 个数字的判断。全连接层有 1024 个神经元，因此定义的权重和偏置如下：

```
W_fc2 = weight_variable([1024, 10])
b_fc2 = bias_variable([10])
```

输出层对 dropout 的输出和权重进行矩阵相乘，然后使用 softmax 激活函数。

```
y_conv = tf.nn.softmax(tf.matmul(h_fc1_drop, W_fc2) + b_fc2)
```

6. 训练模型

下面是用 CNN 在 MNIST 数据集上进行手写数字识别的实验。

```
cross_entropy = tf.losses.softmax_cross_entropy(logits=y_conv, onehot_labels=y_)
train_step = tf.train.AdamOptimizer(1e-4).minimize(cross_entropy)
correct_prediction = tf.equal(tf.argmax(y_conv, 1), tf.argmax(y_, 1))
accuracy = tf.reduce_mean(tf.cast(correct_prediction, tf.float32))
sess.run(tf.global_variables_initializer())

for i in range(80000):
    batch_xs, batch_ys = mnist.train.next_batch(100)
    _, loss = sess.run([train_step, cross_entropy], feed_dict={x: batch_xs, y_: batch_ys, keep_prob: 0.5})

    if(i%1000==0):
        print(loss)
```

```
print("test accuracy %g"%accuracy.eval(feed_dict={  
    x: mnist.test.images, y_: mnist.test.labels, keep_prob: 1.0}))
```

练习：

(1) 运行上述的程序，看一看 CNN 在 MNIST 数据集上的性能如何。

(2) 实现 LeNet5

第六章：神经语言模型和词的表示学习

在传统的文本挖掘、NLP 领域中，每个词被当做一个原子单元（atomic unit）。例如，一个词就是一个字符串，词之间的相似性只有通过词典的定义来评估；在文本处理时一个词被编码分配一个编号。基于这样思想的 N 元（N-gram）语言模型在 NLP，文本挖掘中仍是非常受欢迎。机器学习领域的研究有这样的共识：简单的模型+非常大量的训练数据>复杂模型+少量数据。N-gram 这些简单的技巧在许多任务中会遇到问题。例如，在语音识别中需要非常大量的领域内相关数据，然而在这样的任务中获取高质量的这样数据的量被限制了。因此在这些领域简单技巧+海量数据的模式被限制了。必须寻找更高级的技巧。

在 NLP 领域采用词的分布式表示（distributed representation of words）和神经网络构建的神经网络语言模型很多研究已经证明其性能超过了 N-gram 语言模型。“word embeddings 词嵌入”或者“word representation 词表示”或者“distributed representations of words 分布式词表示”三个术语都是一个意思，是用一个实数向量表示词，本文后面将其称为词向量。词的表示学习在深度学习出现之前就已经有了。2001 年 Bengio 就在两篇论文中提出了 word embeddings。而和 word embedding 思想相似的关于符号的分布式描述则在更早 1986 年就由 Hinton 提出。但是随着深度学习在许多领域取得成功，将深度学习应用在自然语言处理时需要词的表示学习，因此这几年词的表示学习的研究也火热起来。将词用“词向量”的方式表示可谓将 Deep Learning 算法引入 NLP 领域的一个核心技术。大多数宣称用了 Deep Learning 的论文，其中往往也用了词向量。

一个 word embeddings **W:words→Rⁿ**

是一个函数将词映射到高维向量（可以达到 200 到 500 个维度）。例如

$$W("cat")=(0.2, -0.4, 0.7, \dots)$$

$$W("mat")=(0.0, 0.6, -0.1, \dots)$$

通常建立 word embedding 即词的表示学习是一个学习任务。

注：

这么理解，词的表示学习把词映射到了高维的实数空间。经常在一起出现的词被映射到空间相邻的位置，不经常出现的就映射到比较远的位置。因此，在这个向量空间就可以计算词的语义相似性。

有很多词的表示学习的方法，Word2vec 和 Glove 是其中著名的两个。Word2vec 是基于神经网络的方法，Glove 是基于张量分解的方法。它们基于训练集训练出 word 向量，其维度可以在[50-100]。而且令人惊奇的，在这些新方法获得的词向量上的一些算术操作和它们的语义关系可以对应。例如，

$$\text{vec}(\text{"King"}) - \text{vec}(\text{"Man"}) + \text{vec}(\text{"Woman"}) \approx \text{vec}(\text{"Queen"})$$

$$\text{vec}(\text{"Madrid"}) - \text{vec}(\text{"Spain"}) + \text{vec}(\text{"France"}) \approx \text{vec}(\text{"Paris"})$$

第一节：背景知识

6.1.1 词向量

One-hot Representation

自然语言理解的问题要转化为机器学习的问题，第一步是要找一种方法把这些符号数学化。NLP 中最直观，也是到目前为止最常用的词表示方法是 One-hot Representation。这种方法建立一个词表，给每个词顺序编号，每个词就是一个很长的向量，向量的维度等于词表大小，只有对应位置上的数字为 1，其他都为 0。当然在实际应用中，一般采用稀疏编码存储，主要采用词的编号。例如

“话筒”表示为 [0 0 0 **1** 0 0 0 0 0 0 0 0 0 ...]

“麦克”表示为 [0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 ...]

采用稀疏编码后，话筒记为 3，麦克记为 8（假设从 0 开始记）。这种表示方法也存在一个重要的问题就是“词汇鸿沟”现象：任意两个词之间都是孤立的。光从这两个向量中看不出两个词是否有关系，哪怕是话筒和麦克这样的同义词也不能被识别出语义相似性。

Distributed Representation of Words

前面已经谈论了，“word embeddings 词嵌入”或者“word representation 词表示”或者“distributed representations of words 分布式词表示”三个术语都是一个意思。它们都是关于用实数向量描述一个词，称之为词向量。

词向量的用途

词向量除了在深度学习中使用，还有其他的用途：

(1) 词之间的相似性。当在一个语料库上训练完了词向量。就可以用余弦相似度等来计算两个词之间的相似性。进而，可以做词聚类，发现相似的词的集合。

(2) 短文本的相似性计算。两个文本中的每个词向量的逐对相似度之和的平均值

6.1.2 语言模型、统计语言模型和神经语言模型

语言模型 (language model) 可以完成这样一个任务，它为一种语言中的句子计算概率分布 (产生一个句子的可能性)。它也可以完成这样的任务，计算在一个词序列之后跟随一个给定的词 (会一个词的序列) 的概率。例如，barked这个词跟在the lazy dog之后的概率。目前的语言模型的性能 (智能) 还达不到和人一样的程度，但是机器翻译、语音识别的关键技术。

(1) 统计语言模型

传统的统计语言模型是表示语言基本单位 (一般为句子) 的概率分布函数，这个概率分布也就是该语言的生成模型。一般语言模型可以使用各个词语条件概率的形式表示，我们也称之为n-gram语言模型：

$$p(s) = p(w_1, w_2, \dots, w_N) = \prod_{n=1}^N p(w_n | \text{context})$$

这里 context 是一个词的上下文，即一个词的概率是一个条件概率和它的上下文有关。如果不考虑上下文，则这个语言模型是一元语言模型，即 n-gram 中的 n=1。

$$p(w_1, w_2, \dots, w_N) = \prod_{n=1}^N p(w_n)$$

当 n=2，这是二元语言模型，一个词的条件概率仅考虑它前面的一个词。

$$p(w_1, w_2, \dots, w_N) = \prod_{n=1}^N p(w_n | w_{n-1})$$

因为 n 的值越大，语言模型越复杂。在信息检索和文本挖掘中，一元模型往往已经足够。如果 n>2 更高阶的模型往往过于复杂，得不偿失。但在深度学习中很多研究拓展到更高阶的语言模型。

传统的语言模型的参数估计利用语料库进行最大似然估计。如，

$$p(w_i | w_{i-1}) = \text{count}(w_i, w_{i-1}) / \text{count}(w_{i-1})$$

count(w_i, w_{i-1})是词 w_i 和 w_{i-1} 以这样的次序在语料库中出现的次数。count(w_{i-1})是词 w_{i-1} 在语料库中出现的次数。

(2) 神经语言模型

Bengio 在 2003 年提出神经网络语言模型 NNLM，是用前馈神经网络训练语言模型。他的论文 “A Neural Probabilistic Language Model” 做了详细描述。Bengio 用三层神经网络训练语言模型，如图 6.1 所示。NNLM 的目标是学习一个语言模型 $p(w_1, w_2, \dots, w_T) = \prod_{t=1}^T p(w_t | w_{t-n+1}, \dots, w_{t-1})$

图中最下方的 $w_{t-n+1}, \dots, w_{t-2}, w_{t-1}$ 就是前 $n-1$ 个词。模型根据已知的 $w_{t-n+1}, \dots, w_{t-2}, w_{t-1}$ 这已知的前 $n-1$ 个词预测 w_t 。 $c(w)$ 表示词对应的词向量。通常会有个词查找表，可以根据一个词获取该词的词向量。

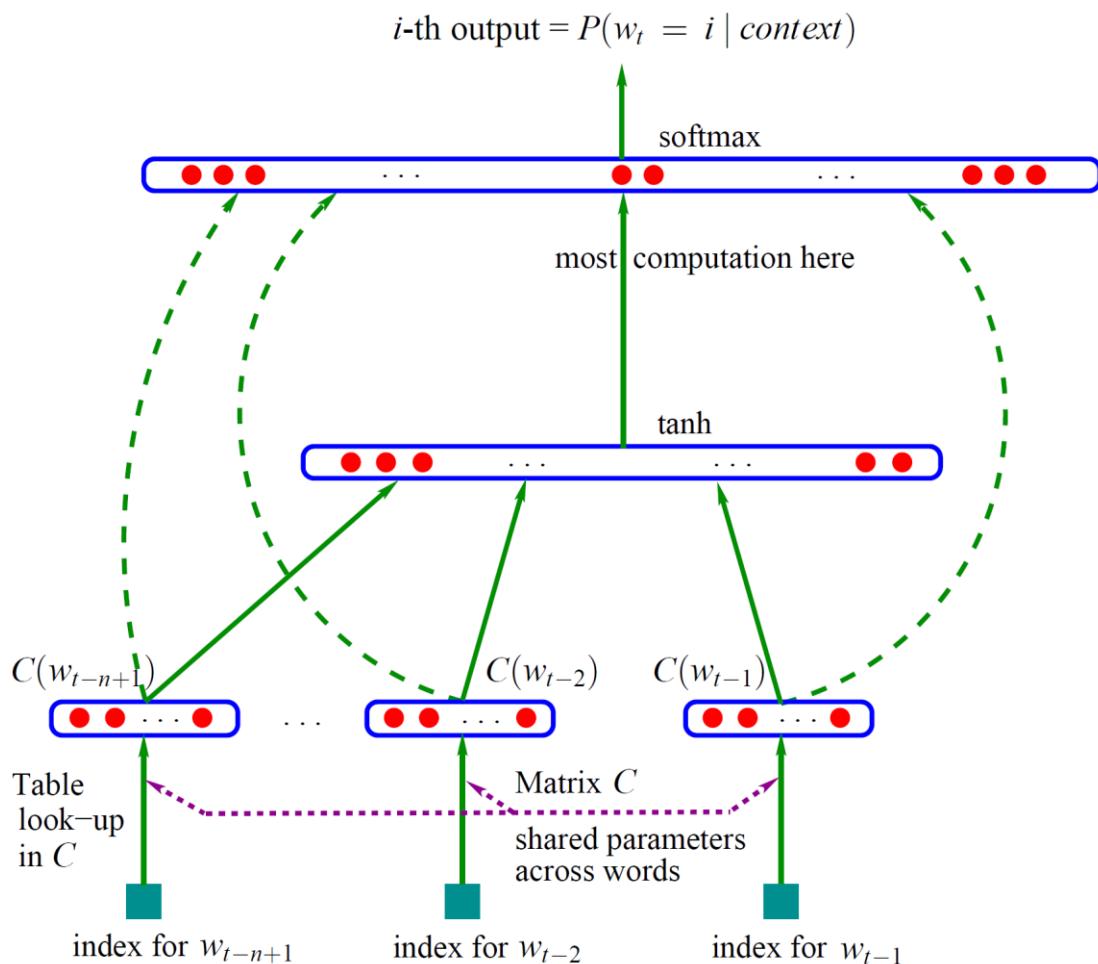


图 6.1. 神经网络语言模型

网络的隐层是将 $C(w_{t-n+1}), \dots, C(w_{t-2}), C(w_{t-1})$ 这 $n-1$ 个向量首尾相接拼起来，形成一个 $(n-1)*m$ 的向量，下面记为 x 。

网络的隐层进行计算 $D + Hx$ 计算得到。 D 是偏置向量， H 是权重向量， x 是输入向量。隐层使用 \tanh 作为激活函数。 \tanh 激活函数，将输出压缩到了 -1 到 1 之间。

网络的输出层一共有 $|V|$ 个节点（ V 是词汇表），节点 y_i 表示下一个词为 w_i 的未规范化log概率。最后使用 softmax 激活函数将输出值 y 规范化成概率。最终， y 的计算公式为：

$$y = b + Wx + U * \tanh(d + Hx)$$

公式中 U （一个 $|V| \times h$ 的矩阵）是隐层到输出层的参数，式子中还有一个矩阵 W ，这个矩阵包含了从输入层到输出层的直连边。直连边就是从输入层直接到输出层的一个线性变换，也是神经网络中的一种常用技巧。如果不需要直连边的话，将 W 置为0就可以了。

需要注意的是，一般神经网络的输入层只是一个输入值，而 NNLM 有个矩阵 C 也是模型学习的参数。 C 是一个词向量的集合构成的矩阵。优化结束之后，词向量有了，语言模型也有了。

6.1.3 词的表示学习

词向量的表示学习有两种形式。一种是专门用工具，如神经语言模型，词向量是模型的参数，训练完成一个语言模型时就获得了词向量。通常我们会在很大的语料库上专门学习词向量，这称为 **pre-trained embeddings**（通用的词向量）。例如 Google 就提供了通用的词向量，用户可以去下载。语言模型可以用于词的表示学习。其实 Bengio 的 NNLM 也可以专门用作训练词向量（语言模型都可以训练词向量），下面介绍的 word2vec 中的 CBOW 就是对 NNLM 的改进。

另一种，词向量是其他任务的副产品。例如，词向量是一个 CNN 文本分类模型待学习的参数，当训练完了语言模型，也就得到了每个词的向量。很多深度学习的文本挖掘任务中就是如此操作，即不需要事先准备词向量。词向量可以从任务的训练过程中，从训练集中学习。

在特殊任务中建议还是使用自己训练的词向量。也有研究两种词向量结合使用。例如，在用 CNN 处理文本时，每种词向量构成输入数据的一个 Channel。第二节讨论训练词向量的两种算法。

第二节：word2vec

word2vec 其实是一个词表示学习的工具箱

（<https://code.google.com/archive/p/word2vec/>），该工具箱的实施基于 Google 公司的 Tomas Mikolov 的两篇论文：“Efficient Estimation of Word Representations in Vector Space” 和 “Distributed Representations of Words and Phrases and their Compositionality”。这个工具箱实施了词表示学习的两个模型 continuous bag-of-

words (CBOW) 和 skip-gram architectures (见上述论文) 。学习的词向量用在进一步的文本挖掘 , NLP 中。例如 , 再用在深度学习中作为深度神经网络的输入。

Word2vec 工具使用语料库作为输入产生 word 向量作为输出。它首先从训练文本数据构造词典 , 然后学习词的向量表示。其结果可以用在许多机器学习的应用中作为特征。

word2vec 非常受欢迎的另一个原因是其高效性 , Mikolov 在论文[2]中指出一个优化的单机版本一天可训练上千万词。我们下面介绍 Word2Vec 工具箱中实现的两个模型。

6.2.1 CBOW

CBOW 是一种与前馈 NNLM 类似的模型 (简称 FNNLM) , 不同点在于 CBOW 去掉了最耗时的非线性隐层 tanh , 且所有词共享隐层。“词共享隐层” 的含义如下 :

FNNLM 中是拼接 ($n-1$) 个 m 维向量 , 因此隐层的神经元数是 $(n-1)*m$; 而在 CBOW 模型中是把 $(n-1)$ 个 m 维向量取平均值 , 隐层的神经元数是 m 。与 FNNLM 不同的是 , CBOW 中不仅会使用要预测的词前面的 k 个词 , 也会使用之后的 k 个词。

如图 6.2 所示。可以看出 , CBOW 模型是预测 $P(w_t|w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})$ 。在 Mikolov 的论文中指出 , $k=4$ 可以获得更好的性能。这里 w_t 的前 k 个和后 k 个词可以无关顺序 , 但有个连续的窗口 $k*2$, 因此该模型称作 continuous bag of words 模型 CBOW。(就是说 , 在训练文本上使用一个 $k*2$ 的窗口来获得连续的词的序列 , 但实际上在这个词的序列内部实际上是无关次序的 , 因为都是要求平均)

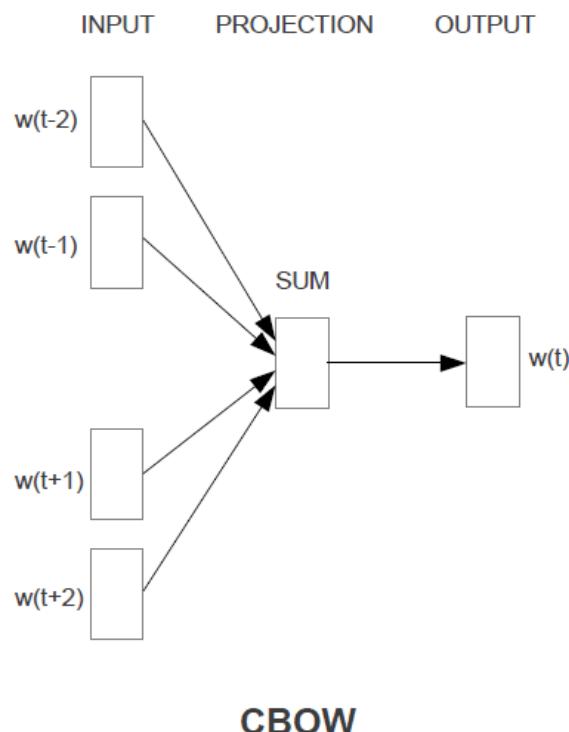


图 6.2. CBOW 模型

从输入层到隐层所进行的操作实际就是上下文向量的求和。图中没画出 NNLM 中的矩阵 C，即词向量集合。C 也是 CBOW 的参数。比 FNNLM，CBOW 还少了直连操作。另外，图 6.2 是 TensorFlow 介绍的 CBOW 模型，各输入向量做了求和操作，不是求平均，这两种方法没有本质的不同。

6.2.2 Skip-gram

Skip-gram 也是 Mikolov 在和 CBOW 同一篇论文中提出的。它与 CBOW 相似，但不是基于上下文预测当前词 w_t 的概率 $p(w_t | \text{context})$ ，skip-gram 使用当前的词向量来预测该词之前和之后各 k 个词的概率。即预测概率 $p(w_i | w_t)$ ，其中 $t-c \leq i \leq t+c$ 且 $i \neq t$ ，参数 c 决定窗口大小。假设存在一个 $w_1, w_2, w_3, \dots, w_T$ 的词组序列，Skip-gram 的目标是最大化：

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq t} \log(p(w_{t+j} | w_t))$$

概率的计算使用 softmax 函数，定义如下：

$$p(w_0 | w_1) = \frac{\exp(v'_{w_0}^T v_{w_1})}{\sum_{t=1}^W \exp(v'_{w_t}^T v_{w_1})}$$

v_w 和 v'_w 是词 w 的输入和输出的词向量表示； W 是词汇表中的大小。在具体的实施中，该模型很不实用，因为计算梯度 $\nabla \log(p(w_{t+j} | w_t))$ 的代价太大（要和词汇表中的每个词进行计算，word2vec 采用了 Negative Sampling 技巧解决此问题）。

在 Milolov 的论文“Distributed Representations of Words and Phrases and their Compositionality”给出了改进的模型。

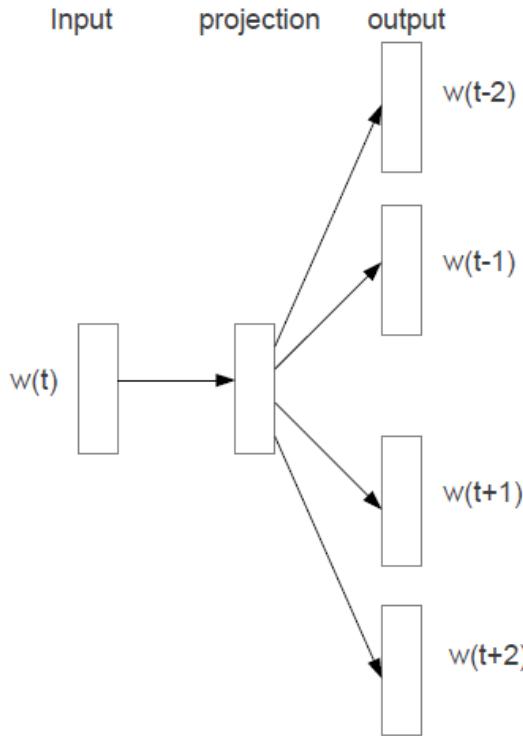


图 6.3. Skip-gram

该模型之所以叫 skip-gram，是在一个窗口内， w_t 不止和它的前一个词 w_{t-1} 和后一个词 w_{t+1} 计算概率 $p(w_{t-1}|w_t)$ 、 $p(w_{t+1}|w_t)$ ，而是和窗口内所有的词，这就是 skip。这样“白色汽车”和“白色的汽车”都会被识别为相同的短语。而且 skip-gram 是一个对称模型，即

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_t|w_{t+j})$$

Tips:

CBOW 和 skip-gram 本身都是语言模型，但他们的目的不是建立语言模型，而是使用语言模型来产生词向量。CBOW 是计算给的一个词的序列，中心词为 w 的概率。Skip-gram 是给定一个词 w_i ，在一个窗口内可以看到另一个词 w_j 的概率。

第三节：TensorFlow 实现词表示学习

TensorFlow 的 tutorial 提供 word2vec 的实现，包括 CBOW 和 skip-gram。算法上这两个模型是相似的，除了 CBOW 从 context(上下文)（如，the cat sits on the）预测目标词（如，mat），而 skip-gram 做相反的预测过程，从目标词预测上下文。从统计学上看，CBOW 更适合小数据集，skip-gram 在大数据集下表现的更好。

为了方便了解 TensorFlow 的 word2vec 的实施，我们再用 tensorflow tutorial 中的描述，介绍一下用神经网络实现的语言模型（和前面的 NNLM 不同），下面简称神经语言模型。神经网络语言模型如下

$$p(w_t|h) = \text{softmax(score}(w_t, h)) = \frac{\exp\{\text{score}(w_t, h)\}}{\sum_{\text{word } w' \text{ in } V} \exp\{\text{score}(w', h)\}}$$

V 是词汇表； $\text{score}(w_t, h)$ 计算词 w_t 和 context（或称作 history, 简写 h ）的相匹配性，通常采用两个向量里的点乘计算。最大似然法训练该模型，最大化目标函数（ \log 似然）（注：负 \log 似然等价于交叉熵损失函数）

$$J_{ML} = \log P(w_t|h) = \text{score}(w_t, h) - \log \left(\sum_{\text{word } w' \text{ in } V} \exp\{\text{score}(w', h)\} \right)$$

然而，该模型的计算代价很高。在训练的每一步需要为 h 计算词汇表 V 中所有其他词 w' 的 score。

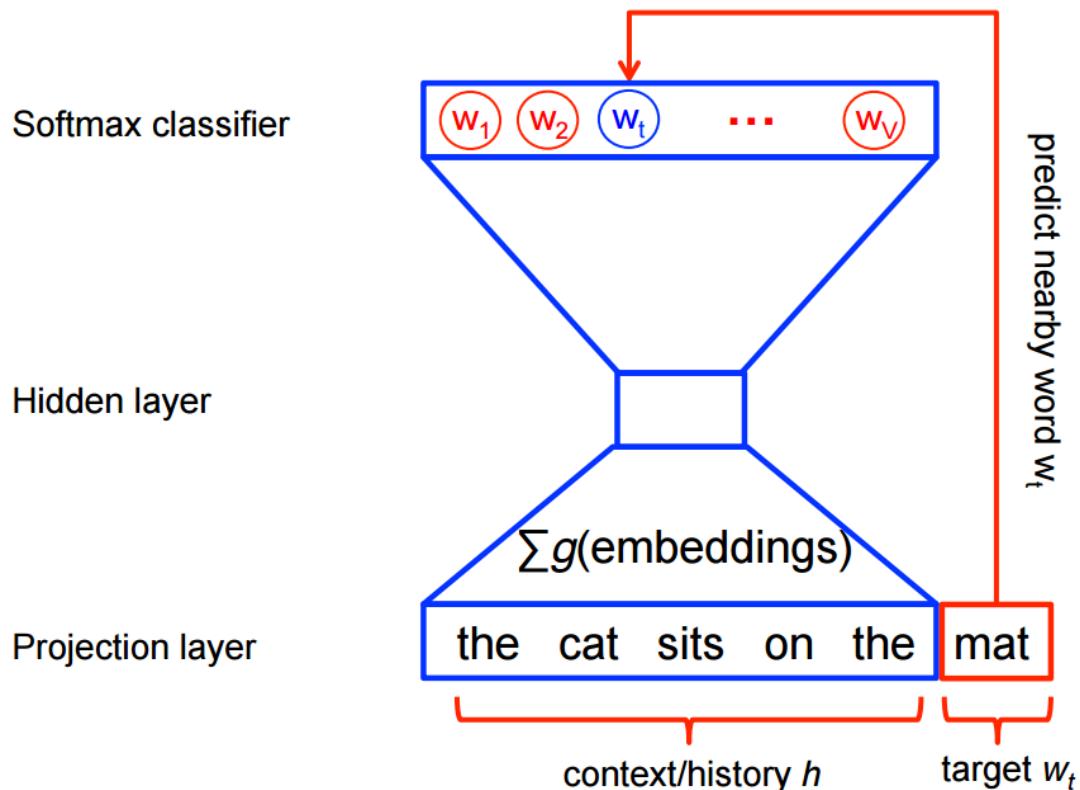


图 6.4. 神经网络语言模型

在 word2vec 的模型学习中，不需要上图的全概率模型。CBOW 和 skip-gram 使用了一个二分类器（logistics 回归）在相同 context 下，区分真实的目标词 w_t 的向量和 k 个噪声向量 $\tilde{w} \in P$ 。图 6.5 是一个 CBOW 模型。Skip-gram 模型可以理解就是把该图倒置。CBOW 模型的目标函数是

$$J_{\text{NEG}} = \log Q_{\theta}(D = 1 | w_t, h) + \sum_{\tilde{w} \in P} \log Q_{\theta}(D = 0 | \tilde{w}, h)$$

$Q_{\theta}(D = 1 | w_t, h)$ 是 Logistics 回归概率，即在数据集 D 中给定上下文 h 看见词 w_t 的概率。该值按照学习到的词向量 θ 来计算。 $Q_{\theta}(D = 0 | \tilde{w}, h) = 1 - Q_{\theta}(D = 1 | \tilde{w}, h)$ 。优化目标就是最大化 $Q_{\theta}(D = 1 | w_t, h)$ ，而最小化 $Q_{\theta}(D = 0 | \tilde{w}, h)$ 。即，给定上下文 h 能在 D 看见目标词 w_t ，而不会看见 k 个噪声词 \tilde{w} 。

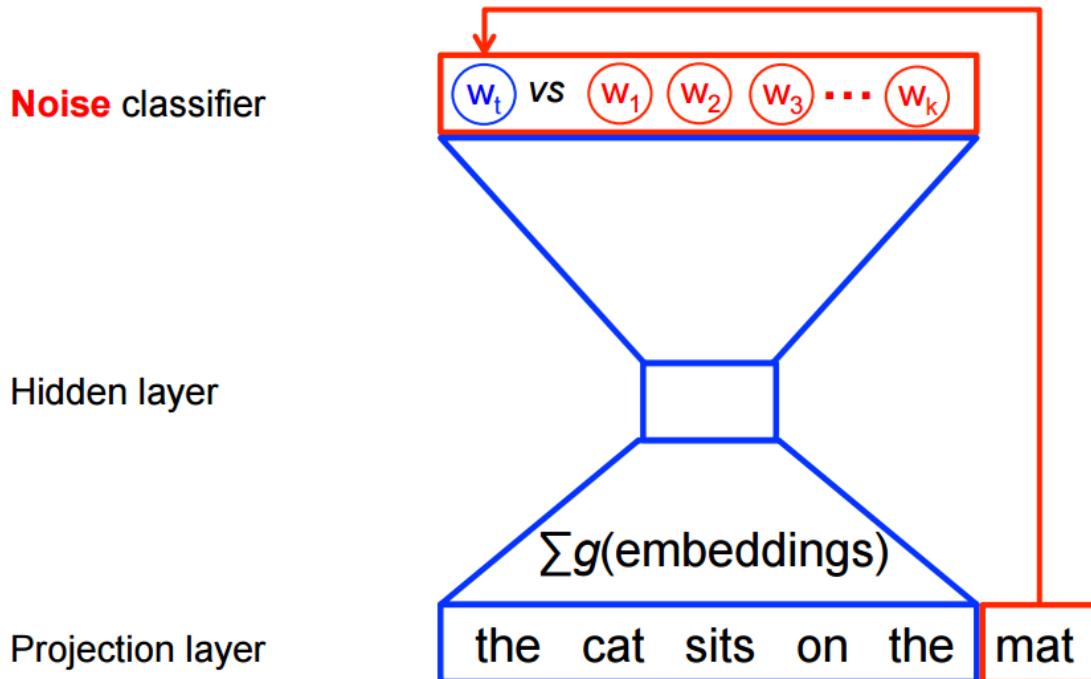


图 6.5. CBOW

在模型的训练过程中，从词汇表随机选择 k 个噪声词，即抽样 k 个负例。这种方法称作 **negative sampling**。因为不是计算所有的词汇表的词，因此 word2vec 训练速度提高很多。上述计算损失函数的方法和 noise-contrastive estimation (NCE) 理论相似。TensorFlow 提供了 NCE 计算损失函数的方法 `tf.nn.nce_loss()`。（ negative sampling, NCE 都属于 Candidates Sampling，
https://tensorflow.google.cn/extras/candidate_sampling.pdf ）

下面我们将介绍 skip-gram 的 tensorflow 实施。看一个例子

The quick brown fox jumped over the lazy dog.

从这个句子我们建立一个目标词和它的 context 的数据集，即 ‘(context, target)’ 对的集合。这里的上下文 context 就是一个目标词左右两边的词。即，一个窗口。设窗口大小为 1。

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

因为 skip-gram 倒置了 context 和 target , 试图预测从目标词预测每个 context 词。以上面句子为例 , 就是从 quick 预测 the 和 brown。因此 , skip-gram 的数据集就是‘(输入 , 输出) 对’集合。

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

目标函数是在整个数据集上定义的 , 但 word2vec 采用随机梯度下降 SGD 来训练模型。batch 的取值是 $16 \leq \text{batch_size} \leq 512$ 。

我们设想一下训练的第 t 步 , 训练数据是 (quick, the) , 目标是从 quick 预测 the。选择 num_noise 数量个负例。为描述的简单 , 假设 num_noise=1 , 选择了 sheep 作为噪声负例。下面为观察的词对 (quick , the) 和负例词对 (quick, sheep) 计算损失。则在 t 步的目标函数是

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

优化的目标是对词向量 θ 做更新 , 来最大化该目标函数。首先需要获得目标函数的梯度 $\frac{\partial J_{\text{NEG}}}{\partial \theta}$ (我们不需要自己计算梯度 , TensorFlow 提供了很方便的方式来计算梯度) 。然后沿着梯度的方向更新词向量。当在整个数据集上进行重复进行训练模型 , 其效果将是每个词移动词向量 , 直到模型可以成功的区分真实的词和噪声词。

用 TensorFlow 实施 skip-gram 时 , 首先定义词向量的矩阵 (或称作查找表 lookup) , 这是一个很大的随机矩阵。

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

因为词向量就是模型要学习的参数 , 因此定义成 TensorFlow 的 Variable。初始化每个元素的值在 [-1, 1]。

noise-contrastive estimation 损失函数按照 logistics 回归模型定义。为此 , 需要为词汇表中的每个词定义权重和偏置。也称作与输入词向量对应的输出权重。

```
nce_weights = tf.Variable(tf.truncated_normal([vocabulary_size, embedding_size], stddev=1.0 / math.sqrt(embedding_size)))
```

```
nce_biases = tf.Variable(tf.zeros([vocabulary_size]))
```

对于词汇表中的每个词 , 分配了一个整数编码。Skip-gram 模型的输入是一个 batch 的整数集合 , 即 context 词的集合 ; 另一个是目标词。输入被定义成 placeholder

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

紧接着需要从查找表，将输入的词的整数编号映射成词向量（embeddings）

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

TensorFlow 的 NCE 函数 `tf.nn.nce_loss` 封装了图 6.6 中从词向量到输出的计算过程，包括 negative sampling 的过程，并计算了损失函数。这使得 TensorFlow 实施 skip-gram 过程简化了。

```
tf.nn.nce_loss(weights, biases, inputs, labels, num_sampled, num_classes,  
num_true=1, sampled_values=None, remove_accidental_hits=False,  
partition_strategy='mod', name='nce_loss')
```

重要的参数如下：

- `weights`: A Tensor of shape [num_classes, dim], or a list of Tensor objects whose concatenation along dimension 0 has shape [num_classes, dim]. The (possibly-partitioned) class embeddings.
- `biases`: A Tensor of shape [num_classes]. The class biases.
- `inputs`: A Tensor of shape [batch_size, dim]. The forward activations of the input network.
- `labels`: A Tensor of type int64 and shape [batch_size, num_true]. The target classes.
- `num_sampled`: An int. The number of classes to randomly sample per batch.
- `num_classes`: An int. The number of possible classes.

我们建立损失函数

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

上面在计算图上定义了损失函数节点，就需要定义一个优化器来计算梯度和更新参数。这里使用随机梯度下降的优化方法。

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

训练模型的过程，就是用 `feed_dict` 将数据放入 `placeholder`，在循环中调用 `session.run()` 用新数据训练模型。

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

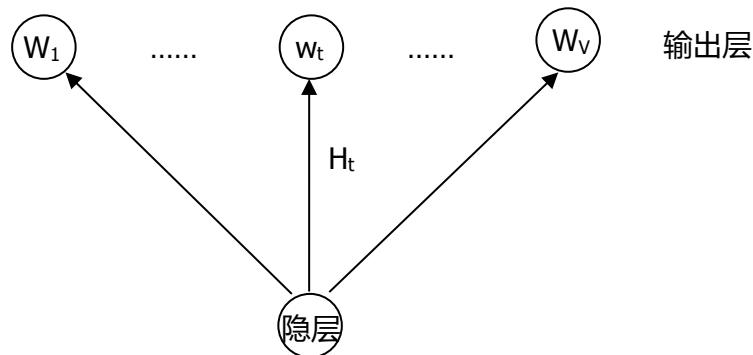
对于如何实施 word2vec 我还是有很多疑惑，原论文很多细节没讲。Tensorflow 的 `tf.nn.nce_loss` 函数封装了实施过程。我理解的 skip-gram 模型如下图 6.6。

(我理解它是这样一个模型：)

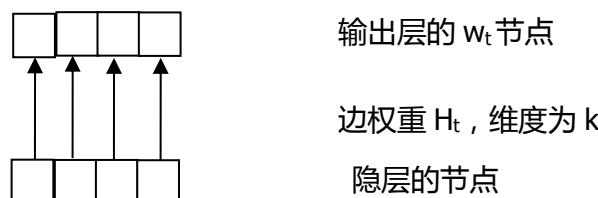
设词汇表 V ，它的大小是 $|V|$ ，词向量的维度是 k 。

训练集是 $(\text{quick}, \text{the}), (\text{quick}, \text{brown}), (\text{brown}, \text{quick}), (\text{brown}, \text{fox}) \dots$ 这样的（输入，输出）对的集合。

我们以 online SGD 为例，即此时模型的输入就是一个（输入，输出）对，然后更新参数。模型的参数有两个：一个是词向量，一个是从隐层到输出层的边的权重 H ，它们的维度都是 $[|V|, k]$ 。 H 对应上面 TensorFlow 代码中的 `nce_weights`。隐层是一个节点，它的输出的维度是 k 维和输出层节点的全连接，但连接是向量对应。



单独的一条边权重是 H_t 是一个 k 维向量。



模型的输出层有 $|V|$ 个节点，激活函数是 logistics 回归，每个节点的输出是一个概率值。（问：这里的 Logistics 回归的系数是不是也是待学习的参数？）

TensorFlow 中 Skip-gram 实施过程如下：为了好讲解，我们假设采用的是 online SGD，即每次用一个实例训练，而实际上采用的是 minibatch。

(1) 建立 `embeddings`，`shape` 为 $[|V|, k]$ ，初始化元素值 $(-1, 1)$ 。`embeddings` 被定义为模型的 `Variable`。因此在训练中它的值会被更新。

- (2) 建立权重矩阵 H , 对应前面代码的 nce_weights。它的 shape 是 $[|V|, k]$ 。它也被定义为模型的 Variable。再建立一个偏置 Variable b , 它的 shape 是 $[|V|, 1]$
- (3) 从训练集取一条训练数据 (input, output) 。 Input, output 是一个词在词汇表中的编号。从查找表 embeddings , 将 input, output 映射成词向量。
- (4) 采用 negative sampling , 抽样 n 个 Output 的对应负样例。
- (5) Input 的词向量 $*H_t+b$, 经过 logistics 回归激活函数 , 算出一个 true logit。
- (6) Input 的词向量 $*H_{neg}+b$, 经过 logistics 回归激活函数 , 算出一个 sampled logit。
- (7) 使用 true logit 和 sampled logit 计算交叉熵损失函数。
- (8) 按照误差反向传播理解 , 由上一步计算的误差修正权重 H_t
- (9) 反向传播修正输入 input 的词向量。在一趟训练中 , 只有目标词和负例抽样的词 (均是输出层的节点) 参与计算 , 所以只有这些词的边的权重被修正 , 也只有这些词的词向量被修正。

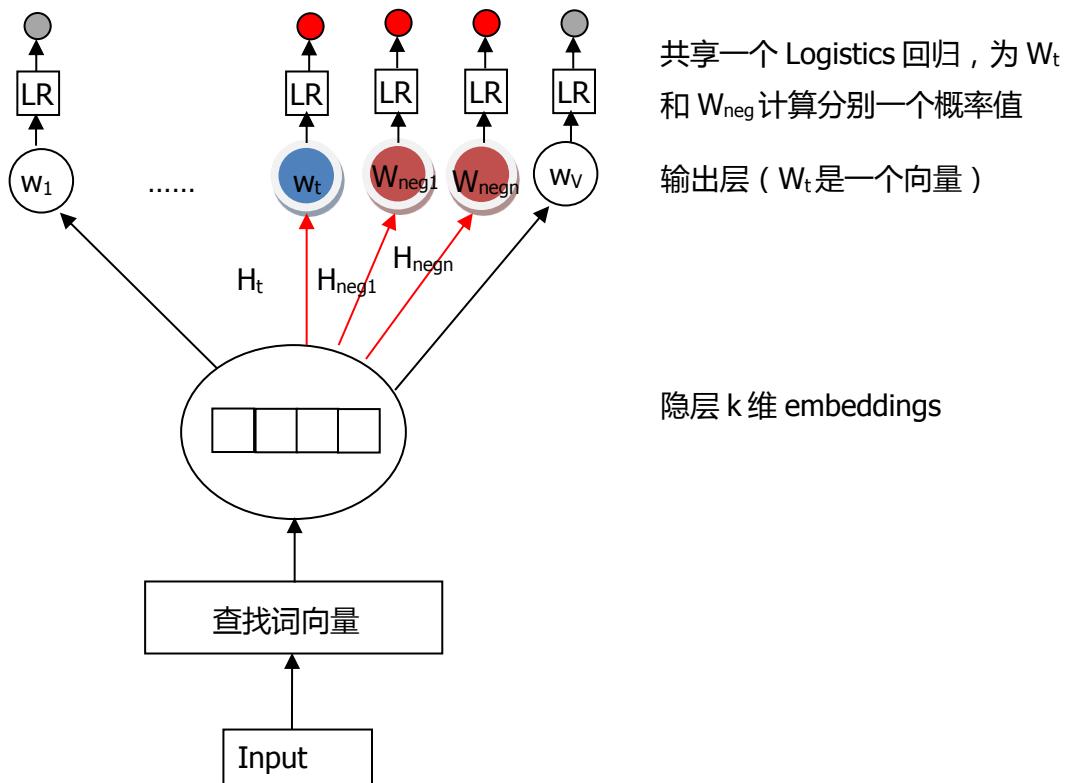


图 6.6 TensorFlow 中实施的 skip-gram 模型

练习：用 TensorFlow 实现 CBOW 模型

注：CBOW 模型和 skipgram 模型相似。在 tensorflow 实施的 skipgram 模型中，一条训练数据是一对词 (w_t, w_{h1})。 w_{h1} 是出现在 w_t 的 context 的一个词。而在 CBOW 模型中，一条训练数据是 $(w_t, (w_{h1}, \dots, w_{hn}))$ ，即一个目标词 w_t 和它的 context。然后把 (w_{h1}, \dots, w_{hn}) 求和得到一个词向量，剩下的就和 skipgram 没有区别了。

第七章：基于 CNN 的文本分类

第一节：CNN 文本分类模型

Yoon Kim 在 EMNLP2014 上发表的论文 Convolutional Neural Networks for Sentence Classification，这篇文章可以说是 cnn 模型用于文本分类的开山之作（其实第一个用的不是他，但是 Kim 提出了几个变体 variants，并有详细的调参）。在这篇论文中构建的 CNN 是针对句子进行分类。这里其实是用句子指代短文本，如评论数据。该论文构建的是一个评论的情感分类器。该深度模型如下：

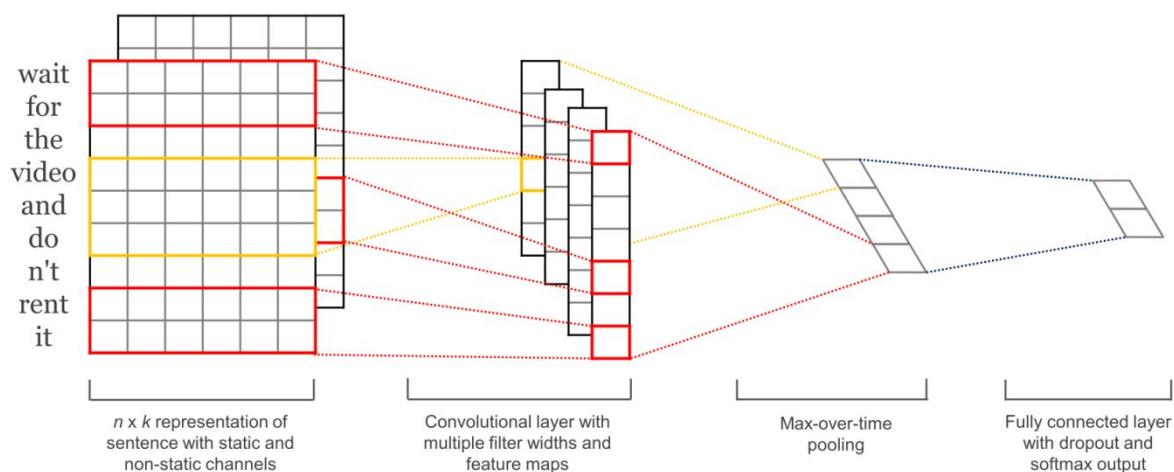


图 7.1 有两个通道的文本分类 CNN 结构(这个图在卷积操作上画的有问题)

(1) 输入层：该模型的输入是一个句子；句子中的每个词用词向量表示。输入的词向量维度为 k 。设 $x_i \in R^k$ 是句子中第 i 个词的 k 维词向量。一个长度为 n 的句子被描述成

$$x_{1:n} = x_1 \oplus x_2 \oplus \dots \oplus x_n$$

这里 \oplus 是拼接操作符，不是连接成长向量，而是拼接成一个矩阵。此文中 $x_{i:i+j}$ 是指词向量 $x_i, x_{i+1}, \dots, x_{i+j}$ 的拼接操作。该文提到 padded where necessary，是说如果句子长度不足 n 用零填充， n 为数据集中最长的文本长度。在前一章讲 CNN 就行图像处理时，一张图片的高宽是固定的，这里也将一个句子转换成了一个高宽固定的矩阵（高是 n ，宽是词向量的长度）。

在前一章中讲到图片有三个通道 (RGB)，因此我们可以描述成输入层有三个特征映射。这里为了处理文本，建立了两个通道或特征映射。输入的句子的每个词的词向量获得有两种方式。一是 Mikolov 使用它的 word2vec 在 Google News dataset 上训练词向量。产生的词向量的维度是 300，包含三百万个词和词组。这个是公开的词向量集合 (<https://code.google.com/archive/p/word2vec/>)。第二个是如前一章所示，将词向量也作为 CNN 模型中的参数，在 CNN 的分类模型中经过训练后得到一组词向量。Yoon Kim 在输入层，或 embeddings 层，建立两种词向量的输入，于是得到输入层有两个特征映射，第一个称为 static channel; 第二个称为 non-static channel。

(2) 卷积层：在输入层上加一个卷积层。一个滤波器 $w \in R^{h*k*d}$ 的一个窗口大小是 $h*k$ ， h 是词的个数， k 是词向量的长度， d 是滤波器的 depth，见图 7.1。从一个滤波器窗口产生一个特征 c_i 。

$$c_i = f(W \cdot x_{i:i+h-1} + b)$$

b 是偏置。如此产生的一个特征映射是一个向量 $c = [c_1, c_2, \dots, c_{n-h+1}]$ 。（注：图 7.1 中卷积操作后第一个向量的长度是 7，应该是 8，当 stride=1 时）

卷积层还有个参数 region size。在每个 region size 上可以设定多个滤波器。如果 region size 是 m ，每个 region size 上滤波器的个数是 n ，则实际卷积层上有 $m*n$ 个滤波器（参考图 7.3）。这里的滤波器又是一个 volume 结构，volume 的 depth 和输入的特征映射的个数相同，在多个输入特征映射上的卷积操作后求和，与第五章描述的卷积层操作相同。（注：图 7.1 没有描述出一个滤波器在两个输入特征映射上的操作）。

图 7.1 中有 $2*2=4$ 个滤波器 ($m=2$, $n=2$, region size 分别是 2 和 3)

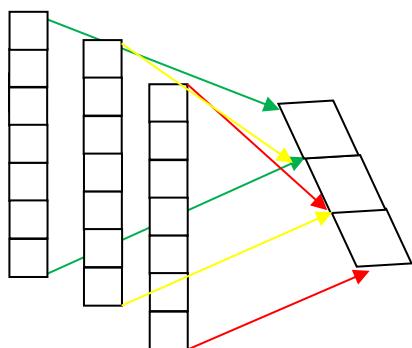


图 7.2 pooling 层

(3) pooling 层：子采样操作应用的是 maxpooling 操作。简单地说就是子采样时的滤波器采样最大值的方法 $\hat{c} = \max(C)$ 进行子采样；滤波器的 shape 是和一个特征映射的

shape 一致。其思想是捕获特征映射中最重要的特征。因此，其结果是一个 feature map 子采样操作后得到的是一个值。每个子采样操作的结果拼接成一个向量，构成 pooling 层的输出。如图 7.2 所示。

(4) softmax 输出层：输出层有两个节点，即二分类的结果。Pooling 层和输出层采用全连接。

(5) dropout：在 pooling 层的输出加 dropout 操作。

在实践中，该 CNN 模型在卷积层给出三种 region size，：[3, 4, 5]。在每种 region size 上建立 100 个滤波器。

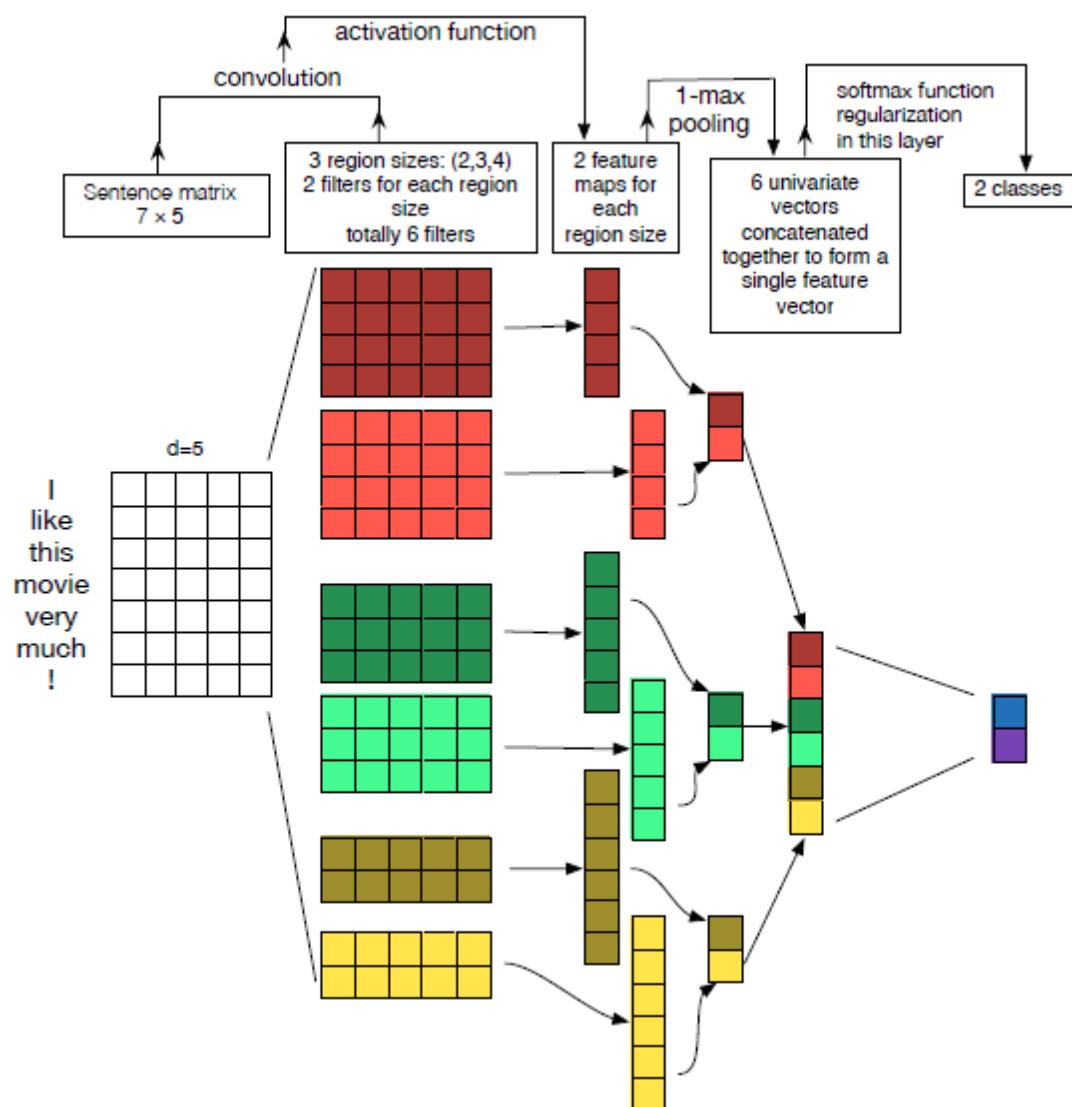


图 7.3 有多个 window 的滤波器的文本分类 CNN, 滤波器的 stride=1

一个详细描述的文本分类 CNN 结构图见图 6.3 (详见论文 A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence

Classification)。图 7.3 的 CNN 的结构是文本分类的结构示例图。它的一些参数不太一样，如滤波器的 region size，滤波器的个数等。

第二节：TensorFlow 实现 CNN 文本分类模型

wildml 对这篇 Yoon kim 的论文 Convolutional Neural Networks for Sentence Classification 有一个 tensorflow 的实现

(<http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/>)。源程序有 4 个文件。为了好理解，便于讲授和理解工作过程。我将源文件中的 train.py 和 text_cnn.py 合并改写了一个 mycnn4text.py 文件，它放弃了一些辅助功能，代码更简单直观。下面以 mycnn4text.py 讲授实现 CNN 文本分类的过程。

(注：它这个版本没有使用 pretrained embeddings)

Tips:

我的构建一个 Tensorflow 深度学习模型的思路是：

1. 准备数据集
2. 超参数
3. 构建数据流图
 - (1) placeholder
 - (2) 查找表
 - (3) 搭建深度网络
4. 损失函数
5. 优化器
6. 其他操作
7. 创建 session，训练模型

1 参数设置

在该版本的模型中，未考虑使用公开词向量建立输入，即只有一个随机初始化的词向量。也即输入的特征映射（通道）是 1。词向量的长度是 128；卷积层滤波器有三种 window 或 region size : [3, 4, 5]。每种 window size 的滤波器是 128 个。Dropout 的 keep_prob 是 0.5；minibatch 的随机梯度下降训练中，batch size 是 64。总结其重要参数设置是：

参数名	值	解释
embedding_dim	128	Dimensionality of character embedding
filter_sizes	[3, 4, 5]	Comma-separated filter sizes
num_filters	128	Number of filters per filter size

keep_prob	0.5	Dropout keep probability
batch_size	64	Batch Size
l2_reg_lambda	0	L2 regularizaion lambda
num_epochs	200	训练模型时迭代次数
num_classes	2	输出层节点数

2 数据准备

使用电影评论数据 sentence polarity dataset v1.0

```
( ( http://www.cs.cornell.edu/people/pabo/movie-review-data/ )
```

该数据集包括两个文件一个正向评论数据和一个负向评论数据。各包含 5331 条评论，每条评论占一行。每条评论视为一个句子。该数据由 Pang/Lee 创建，在 ACL 2005 的论文中使用。因此文本分类（句子分类）的任务即判断一条评论的情感倾向正向或负向。

将两个文件合并产生数据和标签（在 data_helper.py 中）

```
def load_data_and_labels():
    """
    Loads MR polarity data from files, splits the data into words and generates
    labels.
    Returns split sentences and labels.
    """
    # Load data from files
    positive_examples = list(open("./data/rt-polaritydata/rt-polarity.pos",
    "r").readlines())
    positive_examples = [s.strip() for s in positive_examples]
    negative_examples = list(open("./data/rt-polaritydata/rt-polarity.neg",
    "r").readlines())
    negative_examples = [s.strip() for s in negative_examples]
    # Split by words
    x_text = positive_examples + negative_examples
    x_text = [clean_str(sent) for sent in x_text]
    # Generate labels
    positive_labels = [[0, 1] for _ in positive_examples]
    negative_labels = [[1, 0] for _ in negative_examples]
    y = np.concatenate([positive_labels, negative_labels], 0)
    return [x_text, y]
```

对应到输出层有两个节点，标签是一个 list 结构[0,1]是正例标签；[1, 0]是负例标签。

输入层“喂”给模型的数据的 shape 是[batch_size, document_max_size]。

document_max_size 是根据输入数据中最长的句子来定。数据的一行是一条评论的词的序列。每个词已经被编号。（ mycnn4text.py ）

```
x_text, y = data_helpers.load_data_and_labels()
```

```

# Build vocabulary
max_document_length = max([len(x.split(" ")) for x in x_text])
vocab_processor =
learn.preprocessing.VocabularyProcessor(max_document_length)
x = np.array(list(vocab_processor.fit_transform(x_text)))

```

learn.preprocessing.VocabularyProcessor 初始化一个 VocabularyProcessor 实例。它的参数包括

max_document_length: Maximum length of documents. If documents are longer, they will be trimmed, if shorter - padded.

min_frequency: Minimum frequency of words in the vocabulary.

vocabulary: CategoricalVocabulary object.

经 vocab_processor 转换后得到的 numpy array 数据对象 x 是一个[文档集合文本个数 , max_document_length]的二维数据结构。x 的一行是被编码后的一篇文档。

3. 输入层

详见 mycnn4text.py

输入层在源码中称为 embedding layer。创建 placeholder 作为“喂”给模型的输入数据。

```
sequence_length=x_train.shape[1]
```

```

input_x = tf.placeholder(tf.int32, [None, sequence_length], name="input_x")
input_y = tf.placeholder(tf.float32, [None, num_classes], name="input_y")
dropout_keep_prob = tf.placeholder(tf.float32, name="dropout_keep_prob")

```

这里 sequence_length 等于 max_document_length。num_classes 等于输出层的节点数，这里是二分类，所以值为 2。

在输入层，建立所有词向量的 Variable，又称为查找表，命名为 C。

```
C = tf.Variable(tf.random_uniform([vocab_size, embedding_dim], -1.0, 1.0), name="lookuptable")
```

因为，在当前这个版本的文本分类 CNN 中词向量是作为待学习的参数，所以建立 Variable 作为词向量。该词向量 Variable 的 shape=[词汇表的大小，词向量的长度]。

前面建立的 VocabularyProcessor 实例 vocab_processor 可以获得词汇表的大小

```
vocab_size=len(vocab_processor.vocabulary_)
```

tensorflow 提供一个查找表函数，它可以根据 word 编号查找对应的 word embeddings (词向量)。

```
tf.nn.embedding_lookup(params, ids, partition_strategy='mod', name=None, validate_indices=True)
```

重要的两个参数，`params` 是所有的词向量，例如上面的 `W`；`ids` 是词的编号集合。该函数从词向量集合上查找 `ids` 这个集合里的词的词向量。返回的结果是一个 `tensor`，它的 `shape=[? , max_document_length , embeddings_size]`。

输入层创建转换成词项量的输入 `embedded_chars`

```
embedded_chars = tf.nn.embedding_lookup(C, input_x)
```

因为 tensorflow 的卷积操作（详见第 4 章）

```
tf.nn.conv2d(input, filter, strides, padding, use_cudnn_on_gpu=None, data_format=None, name=None)
```

`input` 是输入 `tensor`，它的 `shape` 是 `[batch, in_height, in_width, in_channels]`。因此需要给 `embedded_chars` 增加一个维度

```
embedded_chars_expanded = tf.expand_dims(embedded_chars, -1)
```

`tf.expand_dims(input, dim, name=None)` 函数插入一个维度到一个 `tensor` 的 `shape`。新增加的维度的 `size` 是 1。维度索引 `dim` 从 0 开始。如果给一个负值则是从后面开始插入维度。

举例：

```
# 't'是一个tensor，它的shape=[2]
shape(expand_dims(t, 0)) ==> [1, 2]
shape(expand_dims(t, 1)) ==> [2, 1]
shape(expand_dims(t, -1)) ==> [2, 1]

# 't2'是一个tensor它的shape=[2, 3, 5]
shape(expand_dims(t2, 0)) ==> [1, 2, 3, 5]
shape(expand_dims(t2, 2)) ==> [2, 3, 1, 5]
shape(expand_dims(t2, 3)) ==> [2, 3, 5, 1]
```

4. 卷积层+pooling 层

TensorFlow 创建卷积层使用 `conv2d` 函数

```
tf.nn.conv2d(input, filter, strides, padding, use_cudnn_on_gpu=None, data_format=None, name=None)
```

其参数 `filter` 是一个 `variable`，它的 `shape` 是 `[filter_height, filter_width, in_channels, out_channels]`。此处 `in_channels` 即对应输入图像的通道个数；`out_channels` 是该卷积层输出的通道数量，它对应的是滤波器的个数，也即特征映射个数。

所以，定义滤波器的 `shape`

```
filter_shape = [filter_size, embedding_dim, 1, num_filters]
```

创建卷积层的权重矩阵，即创建滤波器和偏置

```
W = tf.Variable(tf.truncated_normal(filter_shape, stddev=0.1), name="W")
b = tf.Variable(tf.constant(0.1, shape=[num_filters]), name="b")
```

创建卷积层，采用 relu 激活函数

```
conv = tf.nn.conv2d(
    embedded_chars_expanded,
    W,
    strides=[1, 1, 1, 1],
    padding="VALID",
    name="conv")
h = tf.nn.relu(tf.nn.bias_add(conv, b), name="relu")
```

conv 和 h 是 tensor，它们的 shape=[batch_size, sequence_length-filter_size+1, 1, 滤波器个数]。（注：此处第三个维度的 1 来自于 embedding_dim - 滤波器宽度+1，见上面滤波器的宽度设为了 embedding_dim。另外，输入虽然有两个通道，但经过滤波器在两个通道上的卷积结果求了和，因此卷积输出的没有考虑到了通道，见图 5.7）。

再在卷积层上加上一个 pooling 层。pooling 层采用最大值采样，其函数如下：

```
tf.nn.max_pool(value, ksize, strides, padding, data_format='NHWC', name=None)
```

Performs the max pooling on the input.

参数：

value: A 4-D Tensor with shape [batch, height, width, channels] and type tf.float32.

ksize: A list of ints that has length >= 4. The size of the window for each dimension of the input tensor.

strides: A list of ints that has length >= 4. The stride of the sliding window for each dimension of the input tensor.

padding: A string, either 'VALID' or 'SAME'. The padding algorithm. See the comment here

data_format: A string. 'NHWC' and 'NCHW' are supported.

name: Optional name for the operation.

Returns: A Tensor with type tf.float32. The max pooled output tensor.

子采样层如下：

```
pooled = tf.nn.max_pool(
    h,
    ksize=[1, sequence_length - filter_size + 1, 1, 1],
    strides=[1, 1, 1, 1],
    padding='VALID',
    name="pool")
```

sequence_length 是一个句子建立二维矩阵的高度 (即 max_document_length) ; sequence_length - filter_size + 1 就是卷积操作后的特征映射的高度。因此 pooling 操作后得到的 pooled 是一个 tensor , 它的 shape=[batch_size,1,1, 濾波器个数]。

因为一个输入的在卷积+pooling 操作后的所有结果要拼接到一个向量 , 因此在程序的设计上用一个循环语句对所有 size 的濾波器迭代的分别进行卷积+pooling 操作 , 并将操作的结果 pooled 添加到一个 list 结构 pooled_outputs 中。

```
pooled_outputs = []
for i, filter_size in enumerate(filter_sizes):
    # Convolution Layer
    filter_shape = [filter_size, embedding_dim, 1, num_filters]
    W = tf.Variable(tf.truncated_normal(filter_shape, stddev=0.1), name="W")
    b = tf.Variable(tf.constant(0.1, shape=[num_filters]), name="b")
    conv = tf.nn.conv2d(
        embedded_chars_expanded,
        W,
        strides=[1, 1, 1, 1],
        padding="VALID",
        name="conv")

    # Apply nonlinearity
    h = tf.nn.relu(tf.nn.bias_add(conv, b), name="relu")

    # Maxpooling over the outputs
    pooled = tf.nn.max_pool(
        h,
        ksize=[1, sequence_length - filter_size + 1, 1, 1],
        strides=[1, 1, 1, 1],
        padding='VALID',
        name="pool")
    pooled_outputs.append(pooled)
```

进一步然后将 pooled_outputs 转换成一个扁平的结构

```
num_filters_total = num_filters * len(filter_sizes)
h_pool = tf.concat(axis=3, values=pooled_outputs)
h_pool_flat = tf.reshape(h_pool, [-1, num_filters_total])
```

5 . Dropout

```
h_drop = tf.nn.dropout(h_pool_flat, dropout_keep_prob)
```

6. 输出层+softmax

```
W_output_shape=[num_filters_total, num_classes]
W_output = tf.Variable(tf.truncated_normal(W_output_shape, stddev=0.1))
b_output = tf.Variable(tf.constant(0.1, shape=[num_classes]))
```

pooling层和输出层是全连接 , 因此权重参数的shape是[num_filters_total, num_classes]
计算输出 , 因为是二分类所以最终的输出层需要先进行 softmax 操作 , 然后进行 logit 变换 , 将输出转换成 0 或 1 的值。建立的交叉熵损失函数如下 :

```
scores = tf.nn.xw_plus_b(h_drop, W_output, b_output, name="scores")
losses = tf.nn.softmax_cross_entropy_with_logits_v2(logits=scores, labels=input_y)
```

上面的函数 `xw_plus_b` 完成矩阵相乘加上偏置的操作；

`softmax_cross_entropy_with_logits_v2` 将 `softmax`, `logit` 变换，计算交叉熵融合到一个函数里进行操作了。

```
l2_loss = tf.nn.l2_loss(W_output)
l2_loss += tf.nn.l2_loss(b_output)
loss = tf.reduce_mean(losses) + l2_reg_lambda * l2_loss
```

在机器学习中，一个防止过拟合的技巧是加上 regularization，翻译做“正则化”。例如，上面对参数 `W_output` 求 L2 范数，并将结果加入到损失函数中。

7.计算模型的准确率

这里定义一个 `predictions` 操作，是为了检验模型的准确率。通过比较模型计算的输出和训练集的真实标签计算模型准确率。

```
predictions = tf.argmax(scores, 1, name="predictions")
correct_predictions = tf.equal(predictions, tf.argmax(input_y, 1))
accuracy = tf.reduce_mean(tf.cast(correct_predictions, "float"), name="accuracy")
```

8.训练操作

定义优化器，即定义训练操作

```
optimizer = tf.train.AdamOptimizer(0.01)
grads_and_vars = optimizer.compute_gradients(loss)
global_step = tf.Variable(0, name="global_step", trainable=False)
train_op = optimizer.apply_gradients(grads_and_vars, global_step=global_step)
```

这里的优化模型的学习率设为了 0.01。用户可以自己根据需要调节。

`Global_step` 是一个全局变量统计模型被训练的次数。因为该变量的不需要作为模型的参数去学习，而只是完成一个计数功能作为全局变量，因此创建它的 `variable` 时，设置 `trainable=False`。当把该全局变量作为参数传递给优化器进行模型训练时，该全局变量的值会被自动更新。

9.模型训练

在 `data_helper.py` 产生一个迭代器对象，用户可以用该迭代器对象，一次获得一个 batch 的训练数据。

```
batches = data_helpers.batch_iter(list(zip(x_train, y_train)), batch_size, num_epochs)
```

在方法 batch_iter 中， num_epochs 参数可以理解为训练数据要使用多少次来训练模型（即）；batch_size 参数即 minibatch GSD 方法中，一个批次数据的大小。因此，模型实际的迭代次数是 num_epochs*batch_size 次。

最终模型训练的代码如下：

```
for batch in batches:  
    x_batch, y_batch = zip(*batch)  
    feed_dict = {  
        input_x: x_batch,  
        input_y: y_batch,  
        dropout_keep_prob: keep_prob  
    }  
    _, step, summaries, loss_val, acc_val = sess.run(  
        [train_op, global_step, loss_summary, loss, accuracy],  
        feed_dict)  
    print("step {}, loss {:.3f}, acc {:.3f}".format(step, loss_val, acc_val))  
    current_step = tf.train.global_step(sess, global_step)
```

练习：上面的程序未使用 pretrain word embeddings。更改上面的程序，使用 pretrain word embedding，建立第二个通道。可以使用 GloVe 训练的词向量
<https://nlp.stanford.edu/projects/glove/>

详细内容在我的 github 上看相关程序。

第八章：循环神经网络

Recurrent Neural Networks，翻译为循环神经网络，简称 RNN。Recursive Neural Networks，翻译为递归神经网络，也简称 RNN。注意两者的区别。

RNN 是一类神经网络，它的 cell 之间的连接形成一个有向环。RNN 创建了一个网络内部状态，允许展示动态时态行为。不像前馈神经网络，RNN 可以使用内部记忆来处理任意输入序列。如此，RNN 可以应用在建立序列模型的任务。

第一节：RNN 结构

人类思考时并不仅仅是从某个时刻的信息开始思考。当你读文章时，你可以基于前面的词来理解当前的词。当前的思考不是在思考后就放弃，而是作为下个阶段思考的基础。人类的思考具有持久性和连续性。

传统的神经网络不能模拟上述的人类思考过程，这是它的一个主要缺点。RNN 强调这一问题。RNN 是用循环串起来的一组网络，可以持久保存信息。

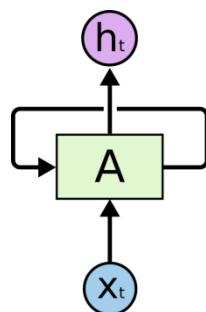


图 8.1 有循环的 RNN

图 8.1 是一个简单结构的 RNN，它的一个 chunk (有翻译做“块”，在 tensorflow 的术语里称为 cell) A 有个输入 x_t (向量)，输出一个值 h_t (向量)。一个循环 (loop) 使得信息被传递从网络的一步到下一步。这里的块 A，内部可以有简单或复杂的结构。简单，如它可以是一个前馈神经网络。复杂的结构可以参看第三节的 LSTM。

这些循环 (loop) 使得 recurrent neural networks 看起来有些神秘。然而，RNN 结构上并不是和传统神经网络有完全的差异。RNN 可以看做是同一个网络的多次复制，每

一次传递一个信息到循环中的下一个网络。我们展开这个 RNN。红框指示的是一个 time step。

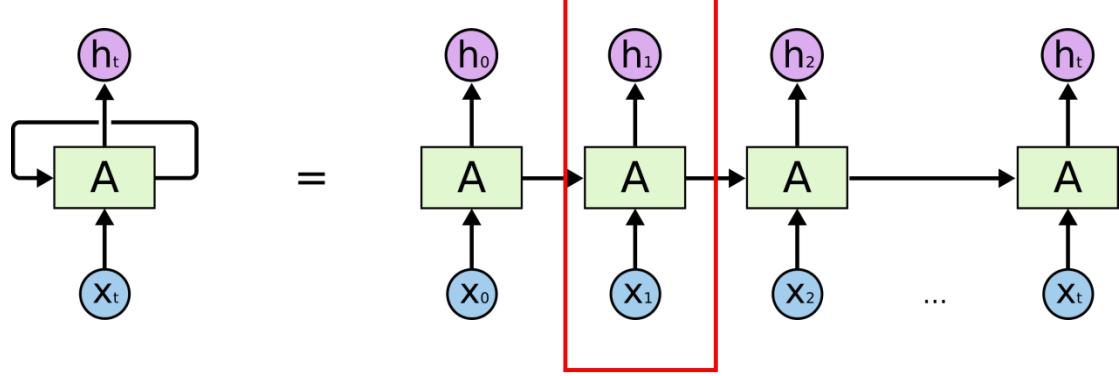


图 8.2 一个展开的 RNN

这种看起来像是链式的状态反映出 RNN 是和“序列”高度相关的。RNN 是神经网络处理序列数据理想的结构。在过去几年里 RNN 在很多问题上取得成功：语音识别、语音模型、机器翻译和图像处理等。可参看一篇文章 “The Unreasonable Effectiveness of Recurrent Neural Networks” <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

RNN 有很大的灵活性。有很多形式的 RNN。如图 8.3 所示。

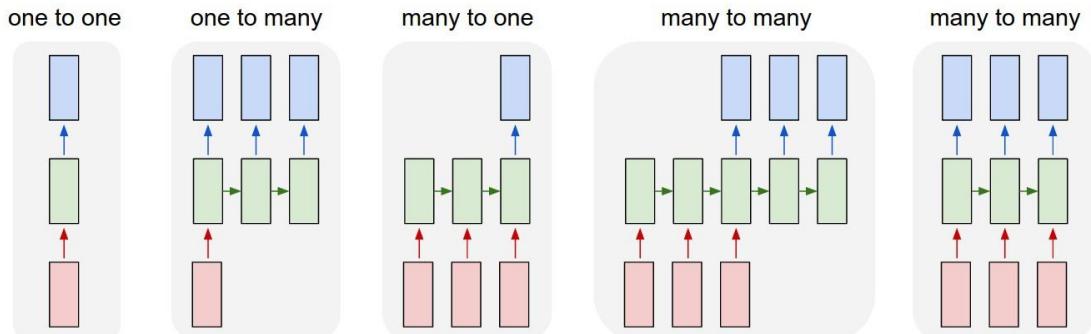


图 8.3 RNN 的各种结构

图中方块表示一个向量，箭头表示函数（例如，矩阵相乘）。输入向量用红色表示，绿色保存 RNN 的状态。One to one 模式称为 vanilla neural network（不算作是 RNN）。它接受固定大小的输入（例如，图像文件），给出固定大小的输出（例如，类别）。而 RNN 可以给定一个向量的序列，而输出可以是一个序列向量或就一个向量。One to many 模式，固定大小的输入，计算一个序列的输出，例如在 image caption 任务中给出一个图片，输出对图像的内容注释的句子。Many to one 是序列输入，计算一个固定大小的输出。例如，输入是一个句子，输出是句子的情感极性。Many to many 可以是序列输入，序列输出。例如在机器翻译中，输入序列是英文句子，输出是中文句子。第二种 many to many 是同步的输入序列到输出序列。例如，对 video 做分类，希望在视频的每一帧上贴标签。注意在上面的每种 RNN 中，都没有

预先规定序列的长度，因为 recurrent transformation (一个绿色方块) 是固定的，能按照我们的要求应用多次。

8.1.1 字符级语言模型：一个 RNN 模型实例

下面我们看一个基本 RNN 的例子，如图 8.4 所示。它是一个字符级的语言模型，采用纯 Python 实施（没有采用 Tensorflow）。模型实现的代码参看

<https://gist.github.com/karpathy/d4dee566867f8291f086>。或我的学习资料中的 min-char-rnn.py

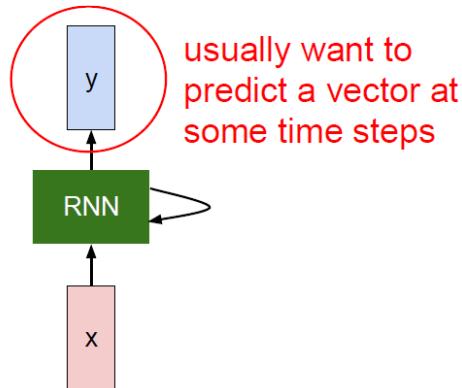


图 8.4 一个循环未展开的基本 RNN 结构

展开后 RNN 的每个时刻的结构，术语称为 time step，翻译做“时间步”。为了描述 RNN 的结构，我们举个最简化的例子：

假设当前字符表只有四个字符 "h,e,l,o"。图 8.5 是一个图 8.4 展开后的 RNN 示例。输入和输出层的维度为 4；隐层有三个神经元。该图显示当 RNN 被“喂”字符 “hell” 作为输入，前向传递被激活。输出层包含 RNN 分配下一个字符（从字符表中选）的确信程度。绿色数值是高值，红色数值是低值。每个 time step 的隐层之间有个状态的传递。

使用该 RNN 构建一个字符集的语言模型。训练集是文本，要求给定一个字符序列 RNN 可以建模下一个字符的概率分布。这个模型可以产生文本，它一次产生一个字符。我们假设有个字母表，仅有 4 个字母 “h”、 “e”、 “l”、 “o”。训练集是一个单词 “hello”。这个序列可以产生四条训练数据：(1) 给定 context “h”，看见 “e”的概率；(2) 在 context “he”， “l” 被看见的概率；(3) 在 context “hel” 看见 “l”的概率；(4) 在 context “hell” 看见 “o”的概率。

具体实施中，每个字母采用 one-hot 编码。给定一个输入字符序列，RNN 的每个时间步“喂”一个字母（one-hot 编码的一个向量）。每个时间步输出一个向量，将该向量转换成对应的字符。由此得到一个输出序列（维度为 4 的向量，每个维度一个字

符）。可以将输出解释为 RNN 当前分配下一个到来的字符的确信程度。如输出 $<1.0, 2.2, -3.0, 4.1>$ 对应输出字符是 helo 的确信程度。因为 2.2 最高，因此输出的是字符是 “e”。

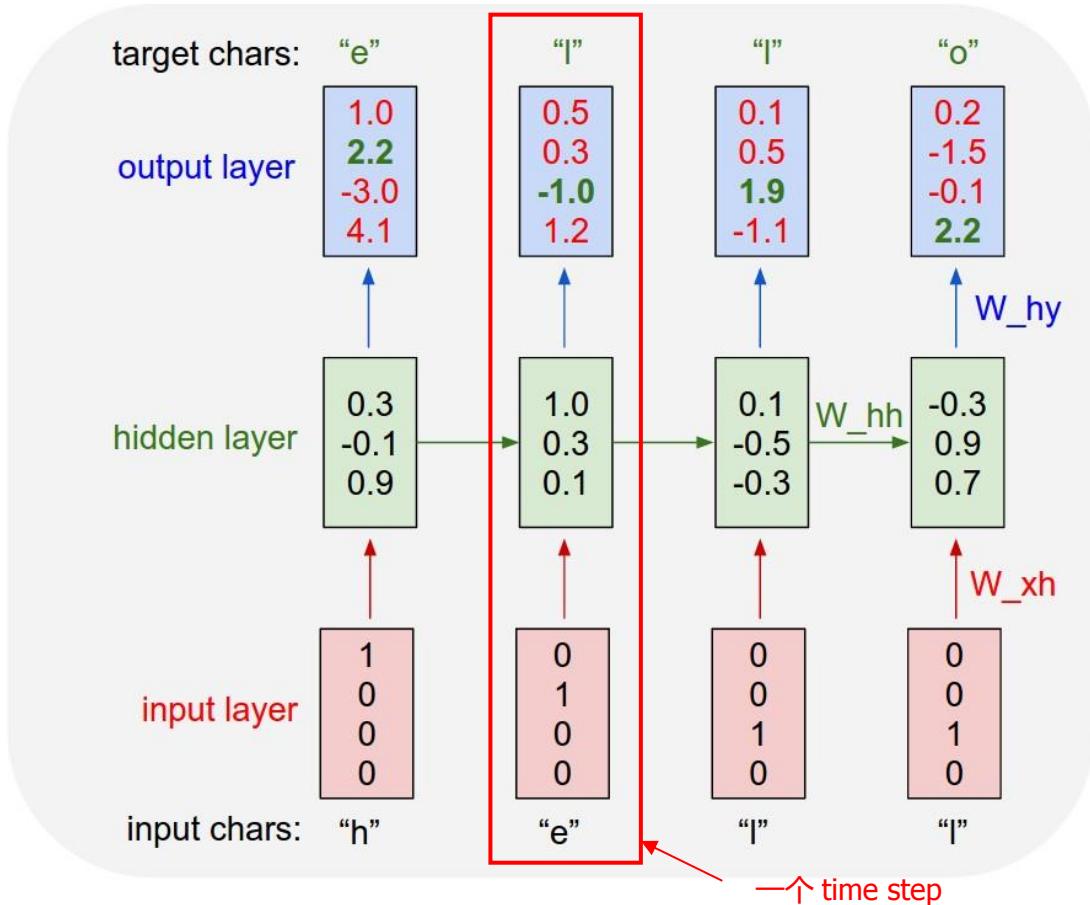


图 8.5 展开的 RNN 实例

以模型训练阶段为例。在第一步，当 RNN 看见了字符 “h”，在确定下一个字符时，它分配 1.0 的确信度到 “h”，2.2 的确信度到字符 “e”，-3.0 的确信度到字符 “l”，4.1 的确信度到字符 “o”。因为训练数据中，正确的下一个字符是 “e”，因此，训练算法将增加字符 “e”（绿色）的确信度，降低其他字符（红色）的确信度。

这个 RNN 的计算过程如下：

(1) 隐层的输出：(注意，与图 8.1, 8.2 不同，这里用符号 h 表示隐层的输出， y 表示 RNN 的输出) h_t 是时间步为 t 时的隐层输出。它是根据上一个状态 (时间步 $t-1$ 的隐层输出) 和当前的输入 x_t 来计算的。 $h_t = f_W(h_{t-1}, x_t)$

$$h_t = f_W(h_{t-1}, x_t)$$

new state | old state input vector at
 some function some time step
 with parameters W

注意在每个“时间步”使用的是同样的函数和参数。当隐层的激活函数是 tanh

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

因此计算时有个循环过程，循环 time steps 次数，每次用上一个时间步的状态进行计算。

(2) 输出层的输出：

$$y_t = W_{hy}h_t$$

Tips:

- (1) 基本 RNN 中输入的序列长度是可变的。之所以可变，关键就是每个时间步使用同样的函数和参数
- (2) 理论上 RNN 可以处理任意长度的序列，实践中还是设定了序列的长度。即规定了时间步的步数。实际上在训练的过程中，将模型上的一次训练的最后一个时间步的状态的输出，作为下一次训练时模型的初始状态，则等同于扩展了输入的序列。

设计 RNN 时几个重要的参数需要确定。(1) 输入序列的长度，即 RNN 时间步的步数。(2) 每个时间步的输入是一个向量(输入层)，向量的长度。(3) 隐层神经元的个数。(4) 输入层到隐层的权重 W_{xh} 是一个 tensor，它的 shape。(5) 从状态 h_{t-1} 到传统 h_t 传递时的权重 W_{hh} 的 shape。(6) 输出层的神经元数目。(7) 从隐层到输出层的权重 W_{hy} 的 shape。

图 8.5 中的一个时间步，我们可以按照一个前馈神经网络来理解。这里的隐层(即 cell)可以是多层。这就是我为什么说一个基本的 RNN 的 cell 是一个前馈神经网络。图 8.6 中隐层只是一层。如果是多层， $t-1$ 步隐层的每一层都会参与到 t 步隐层的对应每一层的计算。两个时间步之间的方块表示一个全连接层。它的神经元数和隐层的神经元数一致。对于图 8.5 的权重 W_{hh} 。

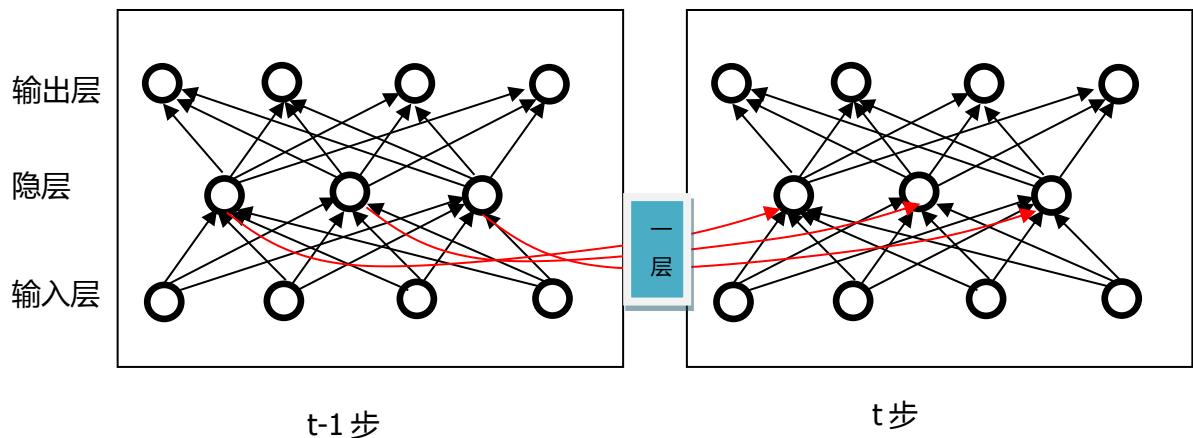


图 8.6 RNN 一个时间步

按照图 8.5 两个时间步之间的状态传递也经过了神经网络的一个层。不过在 tensorflow 是实施中，没有要求用户自己建立这一层，不知道是不是 RNN 函数内部建立了。另外每个时间步共享输入到隐层和隐层到输出层两个层的权重（共享权重）。

第二节：使用 TensorFlow 构建 RNN

8.2.1 介绍

用 Tensorflow 构建 RNN 模型时，先理解几个术语：

(1) - *num_Layers* - the number of RNN Layers

层数是指隐层中包含多少个子层，一个 cell 对应一层。例如，图 8.6 有一层，图 8.7 有两层。隐层有多个子层称为多层 RNN。

(2) - *num_steps* - the number of unrolled steps of RNN

时间步 time step 的步数，图 8.5 时间步为 4

(3) - *hidden_size* - the number of units

术语 *hidden size* 是指一个时间步的一个隐层的子层（一个 cell）包含多少个 unit。在基本的 RNN 结构，一个 unit 就是一个神经元。

RNN 实际上是对输入的序列中的一个时间步，如 x_t ，这个向量的每一位进行计算。每一位的计算都是由一个 unit 完成。每一个时间步就是一个神经网络，参考图 8.7 的描述。

注：我理解 cell 是指对一个向量进行计算的隐层单元。Cell 是由多个 unit 组成。对向量的每一位进行计算的。每个 unit 的结构在基本 RNN 中是一个神经元，在 LSTM 等复杂结构中，是如图 8.19 所示的结构。只不过该图应该是对标量进行计算，计算结果也是标量。

我们用图 8.7 来描述一个时间步的计算。输入 x^t 是一个向量。有一个全连接层和输入层相连。全连接层的神经元个数就是 $hidden_size$ 。全连接层的每个神经元的输出都和一个 unit 连接。这里的每个隐层是一个 cell，多个 cell 构成多层的隐层。时间步 t 每个 unit 的计算要使用时间步 $t-1$ 对应的 unit 的状态 $s_{L,n}^{t-1}$

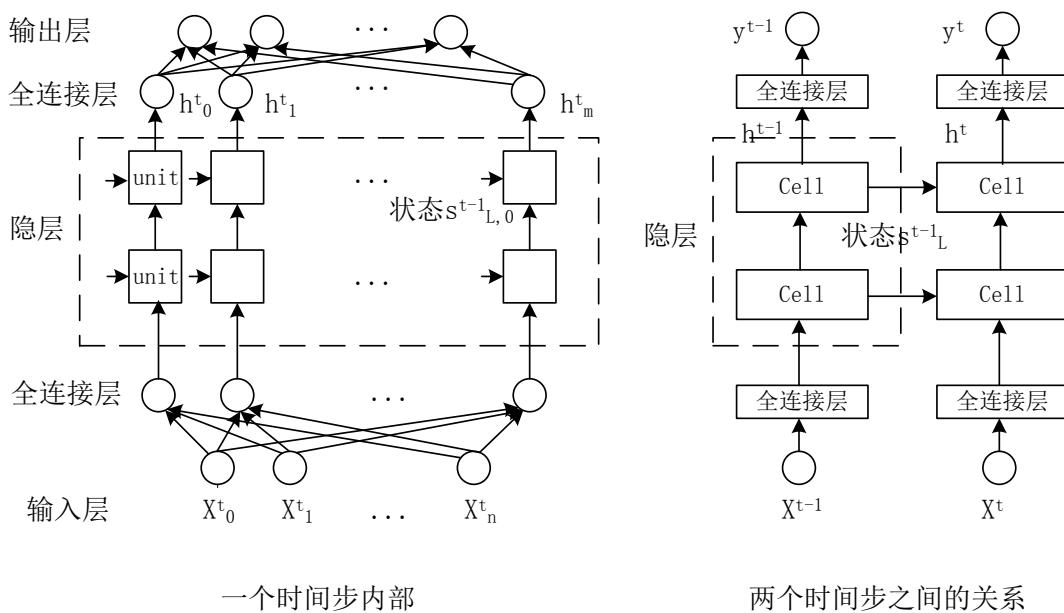


图 8.7 时间步

Tensorflow 提供的 cell 和 RNN 函数，只是实现了多个时间步的隐层部分。两个全连接层需要用户自己建立。每个时间步共享全连接层。

注：两个全连接层不是必须的。例如，在 `ptb_word_lm.py` 建立的模型中，它让词向量的长度等于 $hidden_size$ 。也就是说输入层全连接层的作用是变换向量的长度到 $hidden_size$ 。如果输入向量的长度等于 $hidden_size$ ，就没必要建立这个全连接层。

8.2.2 构建 RNN 的基本函数

Tensorflow 中构建 RNN 包括两个基本的元素：cell 和 rnn(参考
https://tensorflow.google.cn/versions/r1.9/api_guides/python/contrib.rnn)

1. Cell 的函数包括：

(1) tf.contrib.rnn.BasicRNNCell (或者 tf.nn.rnn.BasicRNNCell)

基本的 RNN Cell, 构建一个前馈神经网络的一层。参数如下：

- . num_units: 即 cell (隐层的一个子层) 的神经元数。
- . activation: 输出的激活函数，需要非线性激活函数。默认是 tanh.
- . reuse: (optional) Python boolean describing whether to reuse variables in an existing scope. If not True, and the existing scope already has the given variables, an error is raised.

(2) tf.contrib.rnn.BasicLSTMCell 详见 7.5 节

(3) tf.contrib.rnn.GRUCell (或者 tf.nn.rnn_cell.GRUCell) 。这个函数实现 Gated Recurrent Unit cell (参考论文 <http://arxiv.org/abs/1406.1078>) 。参数如下：

num_units: int, The number of units in the GRU cell.

activation: Nonlinearity to use. Default: tanh.

reuse: (optional) Python boolean describing whether to reuse variables in an existing scope. If not True, and the existing scope already has the given variables, an error is raised.

kernel_initializer: (optional) The initializer to use for the weight and projection matrices.

bias_initializer: (optional) The initializer to use for the bias.

name: String, the name of the layer. Layers with the same name will share weights, but to avoid mistakes we require reuse=True in such cases.

dtype: Default dtype of the layer (default of None means use the type of the first input). Required when build is called before call.

其他的 cell 函数有：

tf.contrib.rnn.MultiRNNCell	组合了多个基本 cell
tf.contrib.rnn.LayerRNNCell	
tf.contrib.rnn.FusedRNNCell	A fused RNN cell represents the entire RNN expanded over the time dimension. In effect, this represents an entire recurrent network.
tf.contrib.rnn.GRUBlockCellV2	The implementation is based on: http://arxiv.org/abs/1406.1078 Computes the GRU cell forward propagation for 1 time step.
tf.contrib.rnn.IndRNNCell	Independently Recurrent Neural Network (IndRNN) cell (cf. https://arxiv.org/abs/1803.04831).

<code>tf.contrib.rnn.IndyGRUCell</code>	Independently Gated Recurrent Unit cell. Based on IndRNNs (https://arxiv.org/abs/1803.04831)
<code>tf.contrib.rnn.IntersectionRNNCell</code>	Intersection Recurrent Neural Network (+RNN) cell. Architecture with coupled recurrent gate as well as coupled depth gate, designed to improve information flow through stacked RNNs.
<code>tf.contrib.rnn.SRUCell</code>	SRU, Simple Recurrent Unit. Implementation based on Training RNNs as Fast as CNNs (cf. https://arxiv.org/abs/1709.02755).

这些函数的使用细节请参见 TensorFlow 的 API。

2. RNN 的函数包括：

(1) `tf.nn.static_rnn(cell, inputs, initial_state=None, dtype=None, sequence_length=None, scope=None)`

该函数创建一个由参数 `cell` 设定的 RNN。用户可选择提供初始状态 `initial_state`。如果提供了序列长度向量 `sequence_length`，会进行动态计算。即，它不计算所有的 `steps`，而是计算完设定的长度后，传递计算结果到最终输出。将节约计算时间。

参数：

`cell`: 一个 RNNCell 实例

`inputs`: Inputs 是一个 list，list 的长度和 time steps 一样。list 中的每个元素是一个 tensor。Tensor 的 shape = [batch_size, input_size]。（注：input_size 应该是隐层的神经元个数。这里的输入是隐层的输入而不是模型的输入。）

`initial_state`: (可选) RNN 的初始状态。如果 `cell.state_size` 是一个整数，`initial_state` 必须是一个 Tensor 它的 shape=[batch_size, cell.state_size]。如果 `cell.state_size` 是 tuple，`initial_state` 应该是一个由 tensor 构成的 tuple 它的 shapes 是[batch_size, s]，`s = cell.state_size`。简单的说，`initial state` 的输入是一个 tensor，它的 shape=[batch_size, num]。num 值由使用的 RNN 类型 (`cell` 的类型) 确定，如，`BasicRNNCell` 或 `GRUCell`，num 值就是隐层的神经元数，如果是 `BasicLSTMCell`，`num=2*隐层的神经元数`。

`sequence_length`: 设定输入中每个序列的长度。是一个 int32 or int64 vector (tensor) size [batch_size], values in [0, T].

`scope`: VariableScope for the created subgraph; defaults to "RNN".

返回值：

返回一对 (outputs, state) 其中 Outputs 的结构和 inputs 是一样的，即一个长度为 time steps 的 list。List 的每个元素对应一个时间步的输出，是一个 tensor，它的 shape=[batch_size, 隐层的神经元数]。State 是最终的状态。

注：rnn 函数到底有多少个 time step 是由输入数据决定的。rnn 函数本身没有设置这一参数。

其他的 RNN 结构：

<code>tf.nn.dynamic_rnn</code>	Creates a recurrent neural network specified by RNNCell cell. Performs fully dynamic unrolling of inputs.
<code>tf.nn.raw_rnn</code>	This function is a more primitive version of dynamic_rnn that provides more direct access to the inputs each iteration. It also provides more control over when to start and finish reading the sequence, and what to emit for the output.
<code>tf.nn.bidirectional_dynamic_rnn</code>	Creates a dynamic version of bidirectional recurrent neural network.
<code>tf.nn.static_bidirectional_rnn</code>	Creates a bidirectional recurrent neural network.
<code>tf.nn.static_state_saving_rnn</code>	RNN that accepts a state saver for time-truncated RNN calculation.

8.2.3 实例 1：用 tensorflow 实现字符级语言模型

这一节用 tensorflow 实现一个字符级的语言模型。实现程序参见 min-char-rnn-tensorflow.py。对于字符采用 one-hot 编码。使用 TensorFlow 构建 RNN，实际上是使用一个 cell 函数和 rnn 函数把隐层给实现。图 8.5 的输出和输出部分还是要自己实现。模型如图 8.8 所示。

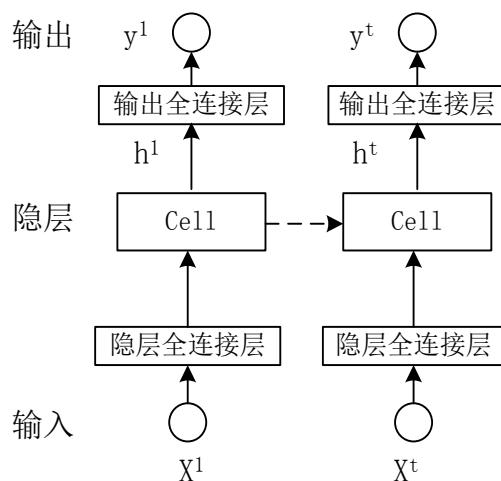


图 8.8：min-char-RNN 模型

输入数据经过全连接层转换成适合隐层处理的形式，隐层的输出经过全连接层转换成输出。每个时间步的全连接层是共享的。设计这个 RNN 模型时，我们需要考虑几个参数

```
hidden_size = 100
seq_length = 25
learning_rate = 0.01
epoch_size = 40
batch_size = 1000
```

hidden_size 是隐层的 unit 数目；Seq_length 是时间步数；Learning_rate 是训练模型时的学习率；epoch_size 是训练模型时的迭代次数；batch_size 是我们采用 minibatch GSD 训练算法时，batch 的大小。这里我们对每个字符采用的 one-hot 编码，因此 onehot 字符向量的大小是字符集的大小。

建立三个 placeholder，对应训练模型时输入数据

```
hstate = tf.placeholder(dtype=tf.float32, shape=[None,
hidden_size])
input = tf.placeholder(dtype=tf.float32, shape=[None,
seq_length, vocab_size])
target = tf.placeholder(dtype=tf.float32, shape=[None,
seq_length,vocab_size])
```

hstate 是初始化的隐状态。input 是输入的数据它的 shape 是[batch_size, seq_length, onehot 字符向量长度]，目标数据也是同样的 shape。

我们采用 BasicRNNCell 和 static_rnn 建立一个最基本的 RNN，即隐层。

```
cell = tf.contrib.rnn.BasicRNNCell(hidden_size)
outputs, state = tf.nn.static_rnn(cell, X,
initial_state=hstate, dtype=tf.float32)
```

我们再建立模型的损失函数。

```
loss =
tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits_v2(logits=Y, labels=T))
```

这里我们采用 softmax 交叉熵来计算损失函数。它是将 logit (神经网络输出层的输出) 转换成一个概率分布。如此，后面的所有工作就是围绕着，(1) 怎样将输入数据转化成 rnn 需要的数据形式，(2) 怎么转换 rnn 的输出和目标数据，以保证转换后的数据 (Y 和 T) 在损失函数中的 Y 和 T 是匹配的。

我们从 rnn 的输入开始来看前后的数据转换如何完成。 rnn 函数的输入 X 应该是一个 List 结构，每个元素对应一个时间步，每个元素的值是一个 tensor，它的 shape 是 [batch_size, onehot 字符向量长度]。Rnn 函数有两个输出，一个是隐层的输出数据

outputs，一个是最后一个时间步的状态值 state。Outputs 是个 tensor，它的 shape 和 input 一样。隐层的输出要和隐层的全连接层相乘。我们定义两个全连接层如下：

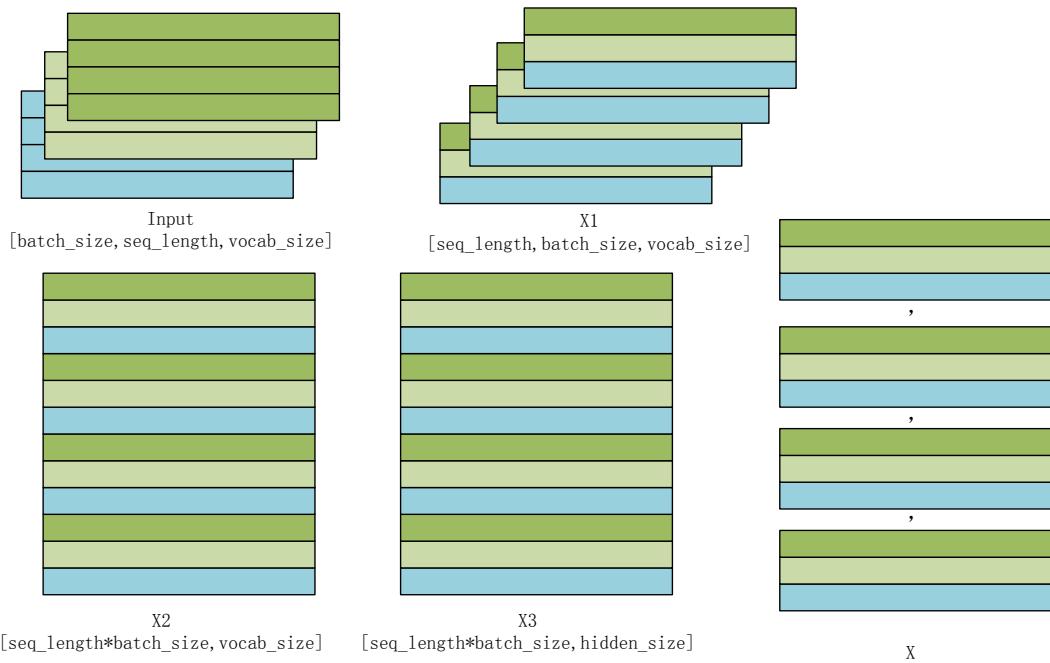
```
weights = {
    'hidden': tf.Variable(tf.random_normal([vocab_size,
hidden_size])*0.01), # Hidden layer weights
    'out': tf.Variable(tf.random_normal([hidden_size,
vocab_size])*0.01)
}
biases = {
    'hidden': tf.Variable(tf.zeros([hidden_size])),
    'out': tf.Variable(tf.zeros([vocab_size]))
}
```

隐层全连接层因为连接输入和隐层，因此权重向量是[onehot 字符向量大小, hidden_size]。输出全连接层因为连接模型的输出和隐层的输出，因此权重向量是[hidden_size, onehot 字符向量大小]。

我们先来看从 inputs 到 rnn 函数需要的输入的转变过程。因为 rnn 函数的输入 X 是一个 List 结构，每个元素对应一个时间步，每个元素的值是一个 tensor，它的 shape 是[batch_size, onehot 字符向量长度]。Placeholderinput 的 shape=[batch_size, seq_length, onehot 字符向量长度]。转变过程如下：

```
X1 = tf.transpose(input, [1, 0, 2])
X2 = tf.reshape(X1, [-1, vocab_size])
X3 = tf.nn.xw_plus_b(X2, weights['hidden'], biases['hidden'])
X = tf.split(value=X3, axis=0,
num_or_size_splits=seq_length)
```

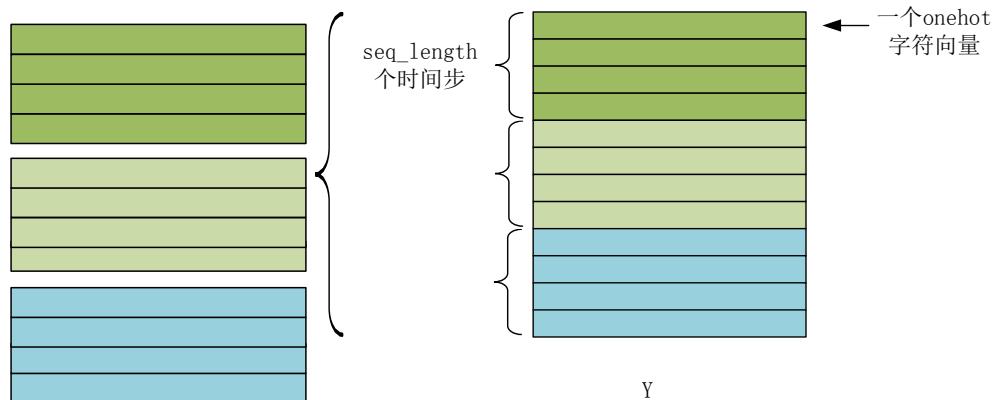
下图中的 Input 的 batch_size=3，时间步为 4



我们再来看怎样转变 rnn 函数的输出。Rnn 的函数的输出 `outputs` 与输入 `inputs` 的结构是一样的，即上图中的 `X` (有四个元素，每个元素对应一个时间步)。输出经过形式变换，然后和权重矩阵相乘。

```
outputs_ = tf.transpose(outputs, [1, 0, 2])
new_output = tf.reshape(tf.concat(outputs, 1), [-1,
hidden_size])
Y = tf.nn.xw_plus_b(new_output, weights[ 'out' ],
biases[ 'out' ])
```

`Outputs_` 的结构和 `Y` 的结构是



图中 `batch_size=3`。因此 `Y` 是一个 `shape=[batch_size*seq_length, vocab_size]` 的 Tensor。（一个 onehot 字符向量的长度等于字符集的大小）。

输入的 `target` 是一个 `shape=[batch_size, seq_length, vocab_size]` 的 tensor, 它需要转换成和 `Y` 一样的结构。

```
T = tf.reshape(target, [-1, vocab_size])
```

建立损失函数 loss 后，就可以建立优化器，然后使用优化器训练模型

```
optimizer = tf.train.AdamOptimizer(learning_rate)
grads_and_vars = optimizer.compute_gradients(loss)
train_op = optimizer.apply_gradients(grads_and_vars)
```

模型运行的代码如下：

```
with tf.Session() as sess:
    sess.run(init)
    p = 0
    k = 1
    while p + batch_size < len(data_ex)-seq_length:
        init_state = np.random.randn(batch_size, hidden_size)
        inputs, targets, p = getBatch(data_ex, batch_size,
                                      seq_length, vocab_size,
                                      p)
        feed_dict={hstate:init_state, input:inputs,
target:targets}
        _, _, loss_x, pred_s = sess.run(
            [train_op, state, loss, pred], feed_dict)

        if(k%10==0):
            print("%s,%s,%s"%(len(data_ex),p,loss_x))
            slist = np.argmax(inputs[0],1)
            rs =[ix_to_char[idx] for idx in slist]
            print(''.join(rs))
            rs =[ix_to_char[idx] for idx in
pred_s[0:seq_length]]
            print(''.join(rs))
        k += 1
```

其中的 getBatch 函数是从数据集建立 inputs 和 target 的函数。详细代码在我的 min-char-rnn-tensorflow.py

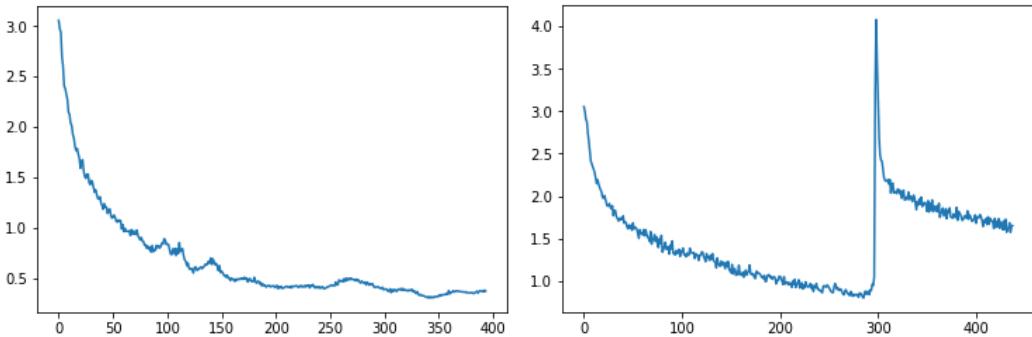
后记：

对于这个字符级的语言模型，目前的训练结果并不是很满意，因为不能做到较为准确的预测。于是我尝试寻找更低的损失函数，即对数据拟合的更好的模型。

关键是改变三个参数：

hidden_size = 400
epoch_size = 300
batch_size = 5000

当增大 batch_size 时，损失函数比较稳定地收敛，如下图所示。但再增大 batch_size，则出现了 resource exhausted error。是指 GPU 上的内存耗尽了。当增加 hidden_size，可以得到更低的损失函数，低至 0.37。此时对句子的预测，见下面的例子。可以看出，模型已经能很好地对原来的句子进行预测了。但继续增加 hidden_size 到 500（为了防止资源耗尽，降低了 batch_size），损失函数有较大的波动，且出现了突变的异常。此突变是由于 Adam 优化器的特性。换做其他优化器没有此异常



例句：

0.3835065

permit from dormont. a month later, i was frustrated and asked to speak with someone at the company

insitearom dormont abmonth later, i was frustrated and a ked to speak with someone at the company.

0.373651

ome cheap scotch. and that is exactly what i love. oh...but good food counts too. we each had a r neswhaep.sootch. and that is exactly srat i love. oh...but good food counts too we each had a re

8.2.4 实例 2：建立可以识别手写数字的 rnn 模型

本节我们将用 RNN 构建一个数字识别模型（rnn4mnist_basic.py）。不是说，RNN 构建的模型在识别手写数字时性能会更好。我们是想通过该模型理解怎样用 TensorFlow 构建 RNN。（大家可以比较使用不同类型的 RNN 获得的性能如何）

(1) 参数设置

```
learning_rate = 0.001
training_iters = 100000
batch_size = 128
display_step = 10
```

learning_rate 是优化函数的学习率；training_iters 是训练的迭代次数；batch_size 是 minibatch 学习中选取的批量数据的大小；display_step 是每隔这些学习步骤显示学习结果。

(2) RNN 网络的参数

```
n_input = 28
```

```
n_steps = 28  
n_hidden = 128  
n_classes = 10
```

mnist 数据集是灰度图片数据集。一张图片的大小是 28*28。当需要把一张图片带入 RNN 模型时，把一行作为输入的向量；而每行是一个 time step。**注：我们讨论的输入、隐层、输出层都是指在一个 time step 的范围。**因此 n_input 是一个输入向量的长度=28；n_hidden 是隐层的神经元个数；n_classes 是输出层的神经元个数；n_steps 是 rnn 的 time steps=28。图 7.5 中的 rnn 输入的一个输入向量如图 7.9 所示。

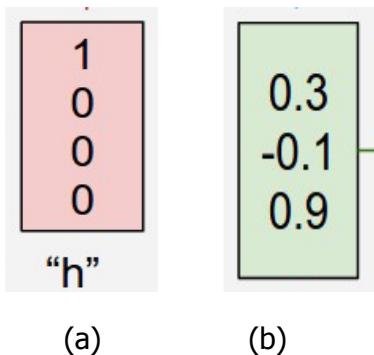


图 7.9 (a) rnn 的输入 ; (b) rnn 的隐层

(3) 定义 placeholder。

```
x = tf.placeholder("float", [None, n_steps, n_input])  
istate = tf.placeholder("float", [None, 2*n_hidden])  
y = tf.placeholder("float", [None, n_classes])
```

创建 x 和 y 占位符，在每趟训练时，将批量训练数据“喂给”模型训练数据和标签。rnn 隐层需要的“state”需要初始化。因为希望在每趟训练时使用不同的随机初始的值。因此也采用占位符的方式“喂给”模型。

x = tf.placeholder("float", [None, n_steps, n_input]) 表示创建的 x 占位符接受的训练数据的 shape=[None, n_steps, n_input]。None 表示训练数据的批量大小，不预设；训练数据是图片，图片的大小= n_steps* n_input。

istate = tf.placeholder("float", [None, n_hidden])。None 是 batch_size。第二个参数和 RNN 的类型 (Cell 的类型) 有关。如果是 BasicRNNCell 或 GNRCell，它等于隐层的神经元数目；如果是 LSTMCell，他是 2*隐层的神经元数目。

(4) 定义 Variable。前面已经讲过。TensorFlow 使用 tf.nn.rnn 和 tf.nn.rnn_cell 实施的 RNN，相当于把图 7.5 中隐层部分实施了。用户仍需要实施其他部分。即仍需要创建两个权重 W_xh 和 W_hy。

```
weights = {
```

```

'hiddens': tf.Variable(tf.random_normal([n_input, n_hidden])), # Hidden layer weights
'out': tf.Variable(tf.random_normal([n_hidden, n_classes]))
}
biases = {
    'hidden': tf.Variable(tf.random_normal([n_hidden])),
    'out': tf.Variable(tf.random_normal([n_classes]))
}

```

图 7.5 的 W_{xh} 权重这里定义为一个名为 `weights` 的 dictionary 结构中的一个“键：值”对，键的名称为“`hidden`”。 W_{hy} 也是一个“键：值”对，键的名称为“`out`”。

这些权重是待学习的参数，因此定义为 `Variable`。 W_{xh} 的 `shape`=[`n_input`, `n_hidden`]。这是因为，该矩阵参与运算 $h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$ 。 x_t 是输入的行向量， $x_t W_{xh}$ 的运算结果是一个列向量，长度等于隐层的神经元数。

W_{hy} 的权重是 [`n_hidden`, `n_classes`]。

注：所有的 time steps 共享 W_{xh} 和 W_{hy} 权重。

(5) 构建 RNN：创建一个函数完成创建 RNN。

```

def RNN(_X, _istate, _weights, _biases):
    _X = tf.transpose(_X, [1, 0, 2])
    _X = tf.reshape(_X, [-1, n_input])
    _X = tf.matmul(_X, _weights['hidden']) + _biases['hidden']

    cell = tf.nn.rnn_cell.BasicRNNCell(n_hidden)
    _X = tf.split(0, n_steps, _X)
    outputs, states = tf.nn.rnn(cell, _X, initial_state=_istate)
    return tf.matmul(outputs[-1], _weights['out']) + _biases['out']

```

该函数中将输入数据进行了变化。图 7.8 描述把 `batch_size=3` 的一个批量的图片的转换过程。`_X = tf.transpose(_X, [1, 0, 2])`，此时的参数 `_X` 是整个模型的输入。输出 `_X` 是经过转置后的 tensor。它的 `shape`=[`n_steps`, `batch_size`, `n_inputs`]。`_X = tf.reshape(_X, [-1, n_input])` 继续将 `_X` 转换成一个 tensor，它的 `shape`=[`n_steps*batch_size`, `n_inputs`]。

`_X = tf.matmul(_X, _weights['hidden']) + _biases['hidden']` 将转换后的 `_X` 和权重相乘。得到的 tensor 的 `shape`=[`n_steps*batch_size`, `n_classes`]。

因为 `tf.nn.rnn` 需要的 `input` 是一个 list，list 长度为 time steps，list 每个元素是一个 tensor，它的 `shape`=[`batch_size`, `n_hidden`]。`_X = tf.split(0, n_steps, _X)` 继续将 `_X`，进行分割。在维度 0 上按照 `n_steps` 进行分割。结果如图 7.10 所示。

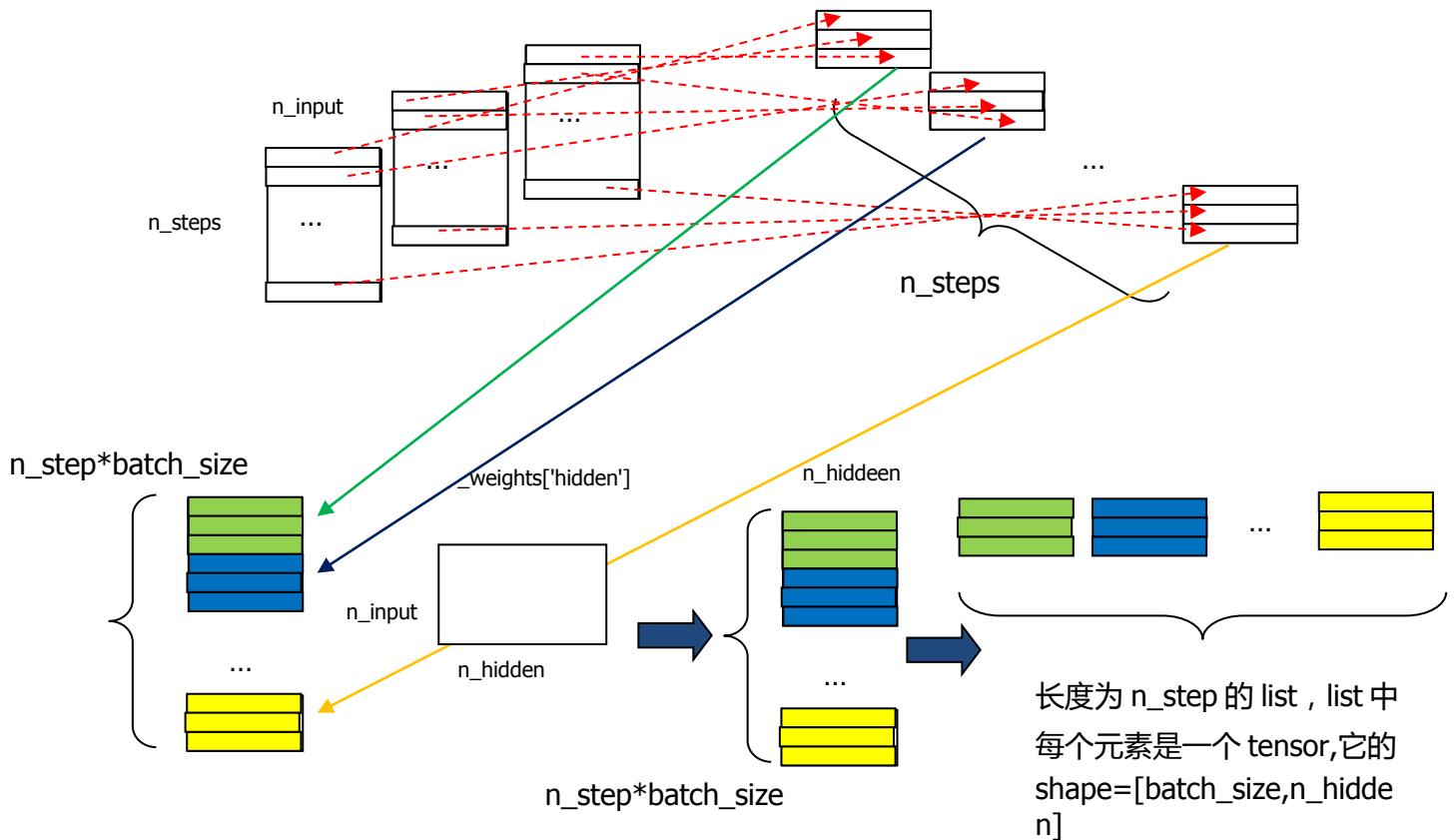


图 7.10 输入图片转换到 rnn 函数需要的格式的过程

输入图片 tensor 的 shape 是 $[batch_size, n_steps, n_input]$ 。权重 W_{xh} 的 shape = $[n_input, n_hidden]$ 。因此需要把输入的图片 tensor 进行转换 $_X = tf.reshape(_X, [-1, n_input])$ 。如此得到的 $_X$ 的 shape = $[n_steps * batch_size, n_input]$ 。再将输入和权重 W_{xh} 进行矩阵乘加上偏置进行计算

```
tf.matmul(outputs[-1], _weights['out']) + _biases['out']
```

得到的 tensor 的 shape = $[n_steps * batch_size, n_classes]$ 。

创建 RNN Cell

```
cell = tf.nn.rnn_cell.BasicRNNCell(n_hidden)
```

因为 rnn 函数的输入需要的是一个 list，list 中的每个元素是一个 tensor。Tensor 的 shape = $[batch_size, n_hidden]$ 。因此需要进行划分

```
_X = tf.split(0, n_steps, _X)
```

然后进行隐层的计算

```
outputs, states = tf.nn.rnn(cell, _X, initial_state=_istate)
```

outputs 是隐层的计算结果。Outputs 是一个 list , 长度为 time steps。每个元素是一个 tensor。一个 tensor 的 shape=[batch_size, 隐层的神经元数]。outputs[-1]是获取最后一个 time step。将它作为模型的输出。在输出层进行计算时将 outputs[-1]乘上权重 W_hy 得到输出 , 它是一个 tensor , 它的 shape=[batch_size, n_classes]

```
tf.matmul(outputs[-1], _weights['out']) + _biases['out']
```

调用定义的 RNN 函数得到预测结果 pred。

```
pred = RNN(x, istate, weights, biases)
```

(6) 迭代训练

```
while step * batch_size < training_iters:  
    batch_xs, batch_ys = mnist.train.next_batch(batch_size)  
    # Reshape data to get 28 seq of 28 elements  
    batch_xs = batch_xs.reshape((batch_size, n_steps, n_input))  
    # Fit training using batch data  
    sess.run(optimizer, feed_dict={x: batch_xs, y: batch_ys,  
                                   istate: np.zeros((batch_size, n_hidden))})  
  
    ....
```

可以看到在每次迭代中 , 产生一个批次的数据 , 带入模型。而每个批次中的状态是被初始化为 0。

比较一下 7.1 节对文本的处理。可以看到处理文本和处理图片时的差别。7.1 节是将一个文本文件作为训练集。将这一个文本转换成长串 , 然后切分成 time step 长度的子串但每次迭代时 , 使用上一次迭代的 RNN 最后的状态 , 作为下一次迭代时 RNN 的输入状态。因此实际上是整个长串作为了一个输入 , 即 RNN 的 time step 是整个长串的字符数。

而本节处理图片时 , 每个图片和下一个图片没有关联 , 因此将图片转换成 time step 个向量后 , RNN 的 time step 之间存在状态传递。

(7) 定义损失函数和优化器

```
cost = tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(pred, y))  
optimizer = tf.train.AdamOptimizer(learning_rate=learning_rate).minimize(cost)
```

练习 : 7.2 节用 Tensorflow 实现的字符级语言模型实验的是最基本的 BasicRNNCell 和 static_rnn。使用其他 Cell 和 RNN 模型构建该模型。

第三节：LSTM

RRN 的一个吸引人的地方在于它可以连接先前的信息到当前的任务，例如使用先前的视频帧帮助当前帧的理解。有时我们仅仅需要最近的信息，而不是太早以前的信息完成当前的任务。例如，一个语言模型试图根据前面的词预测下一个词。如果我们预测一个句子 “the clouds are in the sky” 中的最后一个词。根据句子中前面的词集合（前面的词是当前的词 context）“the clouds are in the”很明显这个词应该是 sky。这个例子中，我们需要的 context 相关信息可以不是很多。这个问题称为 Short Term Dependencies。

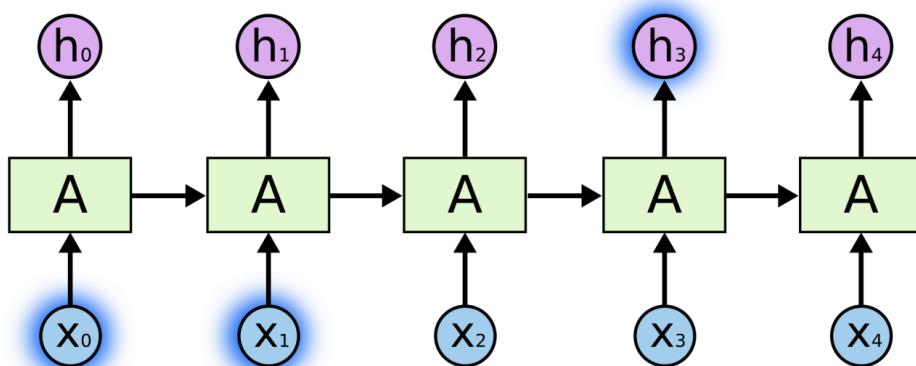


图 8.11. Short Term dependency

但有时我们需要更多的 context。再看一个例子，“I grew up in France... I speak fluent French.”（省略号表示还有很多句子）当预测最后一个词时，前面的信息 I speak fluent 给出暗示这个最后的词应该是一个语言名称。但如果想知道具体是哪一种语言，我们需要更多的 context。如此再往前寻找 context。“I grew up in France” 暗示是 French。可以看出从相关信息到待预测的词之间的 gap 很大。

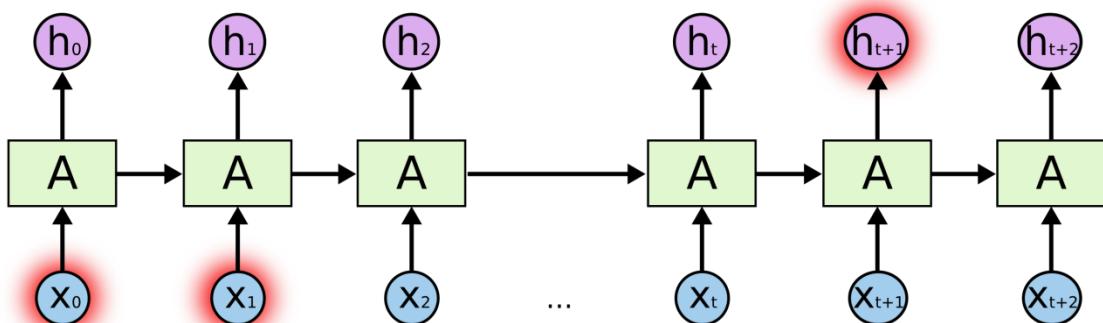


图 8.12. Long Term Dependency

当这个 gap 很大时前面讲述的基本 RNN 没有能力连接到 gap 前面的 context 去进行学习。这个问题称为 Long Term Dependencies.

8.3.1 LSTM 的结构

理论上 RNN 能够处理 long-term dependencies。但实践中有很多问题。[Hochreiter \(1991\) \[German\]](#) and [Bengio, et al. \(1994\)](#),指出了 RNN 在这个问题上的根本缺陷。但 LSTM 可以很好的解决这个问题。LSTM (Long Short Term Memory networks) 是一种特殊结构的 RNN，它可以学习 Long Term Dependencies。它由 Hochreiter & Schmidhuber 提出。在其后的研究中许多人的研究将 LSTM 改进使得它在很多任务上都非常成功。现在 LSTM 的应用非常广泛。所有的 RNN 都有一个链式结构，重复了神经网络的一个块（chunk 或 cell）。标准的 RNN 中重复的“块”有一个很简单的结构，例如一个单独的 tanh 层（**图中黄框表示一个层**）。（8.1 节 $h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$ ）。这里 x_t 是一个输入向量；tanh 层是神经网络的一层，即包含了多个 unit； h_t 是一个向量。

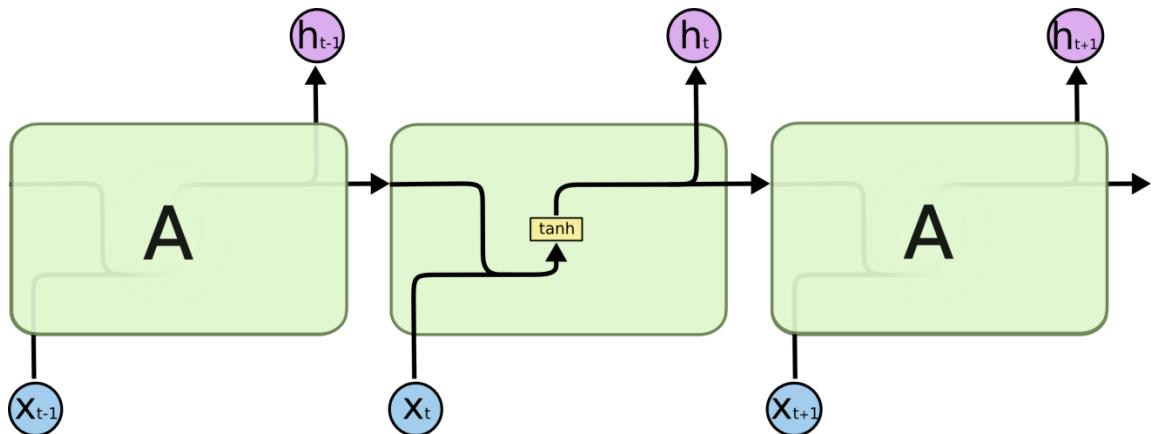


图 8.13. 在一个标准 RNN 中重复了一个单独的层

LSTM 也有这种链式结构，但是重复的块（chunk）有复杂的结构，例如有四个层，以一种特殊的形式交互。如图 8.14 所示。

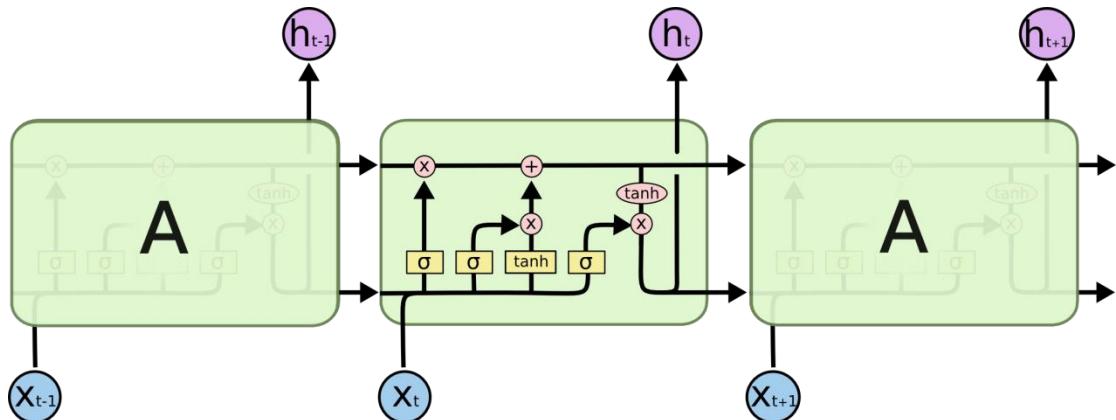


图 8.14 LSTM 中的块包含四个交互层

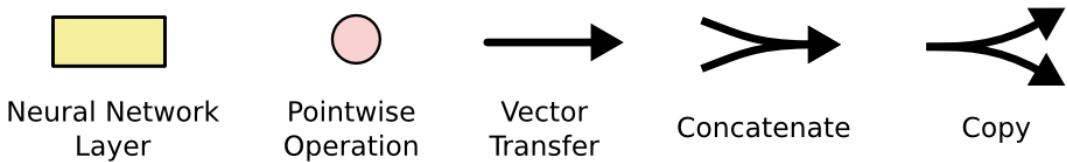


图 8.15 一些注释符号

图 8.15 给出描述 LSTM 需要的一些符号。图中每条线表示从一个节点的输出传递一个 Tensor 到一个节点的输入；粉色的圈表示逐点运算，例如向量相加；黄色的方框是待学习的神经网络的层；线的合并表示拼接操作；线段的分叉表示内容被复制，然后送到不同的节点。

8.3.2 LSTM 的核心思想

与基本 RNN 相比，LSTM 的关键是增加了块（chunk 或 cell）状态，即贯穿图的那条水平线。块的状态（cell state）可以理解为是一种传送带。信息沿着它传递。

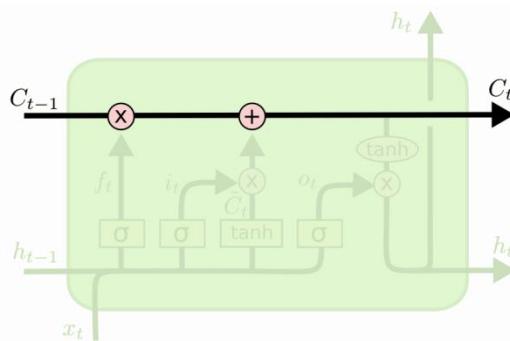


图 8.16 RNN 的信息传递

通过调整称为 gate 的结构，LSTM 有能力移除或添加信息给单元状态（cell state）。Gate 是一种方式或通道，选择性的让信息通过。它由一个 sigmoid 层和一个逐点相乘的运算操作组成。如图 7.17 所示。

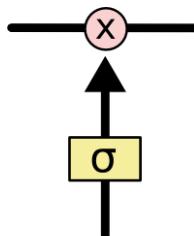


图 8.17 Gate 结构

Sigmoid 层输出 0~1 之间的数值，描述了每个构件（ component ）有多少部分允许通过，例如，0 表示不让通过，1 表示全部通过。一个 LSTM 有三种 gate（ forget gate, input gate, output gate ）来包含和控制 cell state。（图 8.14 中一个块里面的三个 σ ）

8.3.3 LSTM 的工作过程

LSTM 的第一步是决定什么样的信息应该通过 cell state 传递。这个决策由第一个 sigmoid 层决定（图 8.18），称为 **forget gate**。Sigmoid 层的输入是 h_{t-1} 和 x_t 。对于 cell state C_{t-1} 的每个值，forget gate 输出一个 0~1 之间的一个值。1 表示完全通过，0 表示阻止。

我们回到语言模型的例子。该语言模型试图基于前面的词预测下一个词。该任务中，cell state 可以包括当前主语的性别这样的信息，如此可以使用正确的介词。当看见一个新的主语，我们应该忘记上一个主语对应的性别。（我理解上面这个例子的意思是一个句子包含多个主语）

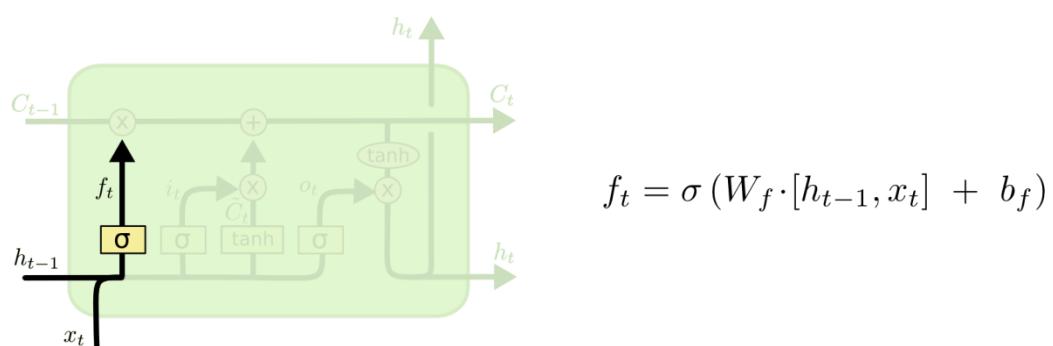


图 8.18 LSTM 块的第一层：一个 Sigmoid 层

$[h_{t-1}, x_t]$ 表示的是将 h_{t-1} 和 x_t 拼接操作到一个矩阵。权重矩阵 W_f 实际上包含 h_{t-1} 的权重 W_{fh} 和 x_t 的权重 W_{fx} 。上面的操作 $W_f.[h_{t-1}, x_t]$ 可以分解成 $W_{fh}.h_{t-1} + W_{fx}.x_t$ 。

下一步是要确定我们将要存储什么新信息在 cell state 中，见图 8.19。它包含两部分。首先一个 sigmoid 层称作 “**input gate**” 决定我们应该更新哪个值。下一步，一个 tanh 层创建一个新的候选值的向量 \tilde{C}_t 。它能被加到这个状态中。紧接着，联合这两个状态来创建一个新的状态。在语言模型的例子中，我们想加新主语的性别到 cell state，来替换旧的状态。

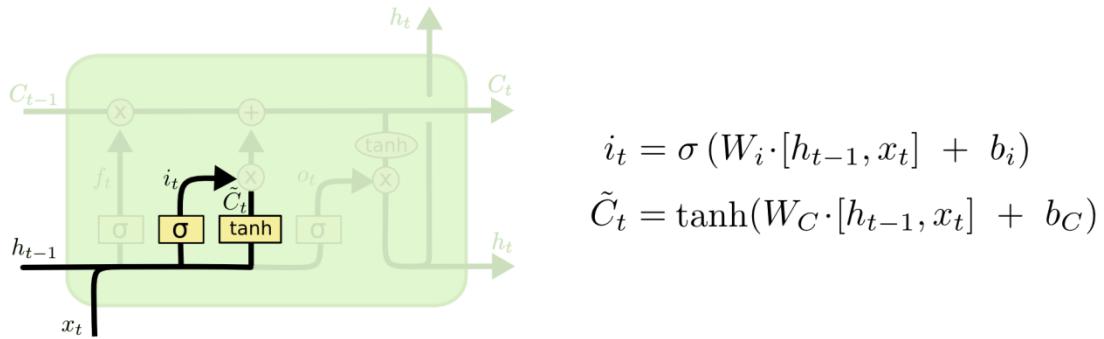


图 8.19 LSTM 块的第二，三层

下面的计算将旧的状态 C_{t-1} 更新到新的 cell state C_t 。旧状态乘上 f_t 于是忘记早先决定忘记的。然后加上 $i_t * \tilde{C}_t$ 。这是新的候选值。在语言模型的例子中，相当于我们实际上放弃了关于旧的主语的性别信息，加上了在上一步骤（图 7.19）中确定的新的信息。

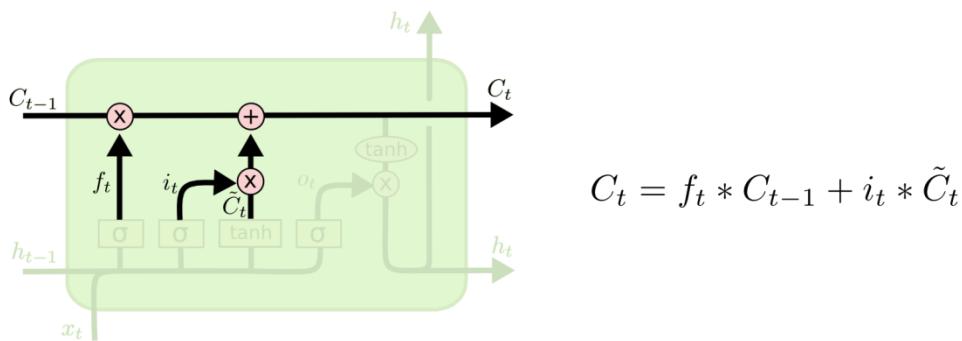


图 8.20 LSTM 块的第二，三层的输出

最后，需要确定输出值。输出应该基于 cell state。我们运行一个 sigmoid 层。它担负 gate 功能，即 output gate，它决定 cell state 的什么部分应该输出。让 cell state 通过 tanh（把值规范化到-1 和 1 之间），再乘上 sigmoid gate 的输出。如此我们仅仅输出我们确定想输出的部分。

再以语言模型为例。语言模型看见了一个主语，它可以想输出与动词相关的信息，因为动词是紧接着要到来的词。例如，它可以输出是否主语是单数还是复数。如此我们知道下一步形成一个动词时应该配合这个信息。

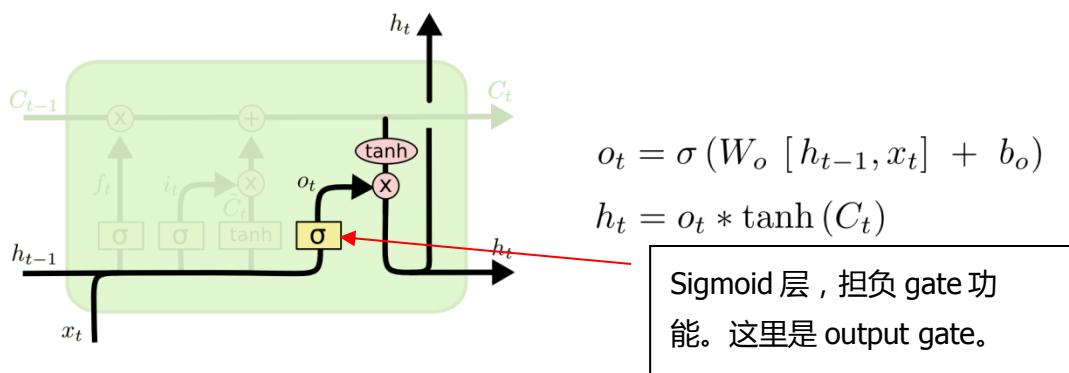


图 8.21 LSTM 块的第四层

第四节：LSTM 的变体

迄今我们描述的 LSTM 是一个标准的 LSTM。但不是所有 LSTM 都与上面的结构相同。事实上，每篇涉及 LSTM 的论文都有自己的结构，和标准结构会有些差异。

一个受欢迎的 LSTM 变体 Gers & Schmidhuber 加入了 peephole connection。这意味着让 gate layer 看见 cell state。见图 7.22 Cell state C_{t-1} 参与了 forget gate f_t 的计算。

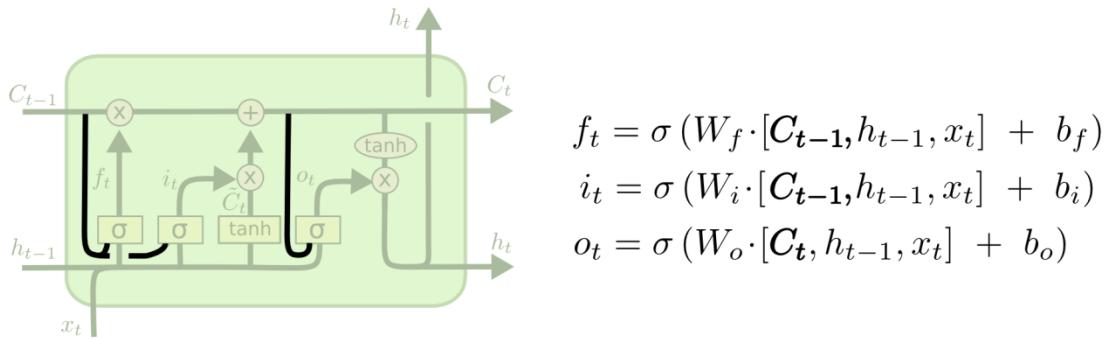


图 8.22 peephole connection LSTM

上图中所有的 gate 都添加了 peephole。有些论文并不是所有的 gate 都应用 peephole。

另一个 LSTM 变体将 forget gate 和 input gate 结合使用。它不像基本 LSTM 中单独决定什么应该被忘记 (forget gate)，应该加什么新信息 (input gate)。该变体将两个决策一起决定。仅仅当要输入一些信息，那么相应位置的旧信息把它忘记，其他位置的信息不变。仅仅在旧的 cell state 中被忘记的部分输入新值。

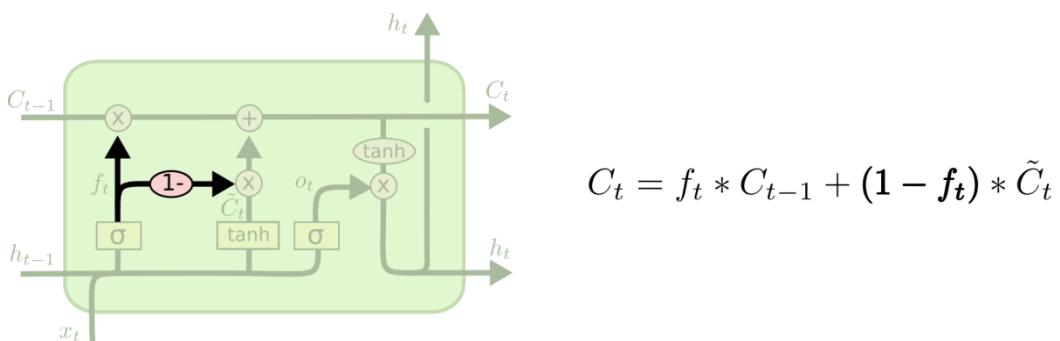


图 8.23 LSTM 的一个变体

一个更动态的变体是 GRU (Gated Recurrent Unit) [Cho, et al. \(2014\)](#)。它结合 input gate 和 forget gate 为一个新的 gate，称作 update gate。它也合并了 cell state 和隐

层，并做了一些其他的改变。其模型比 LSTM 更简单，少了一个传递状态 C_t ，也非常受欢迎。

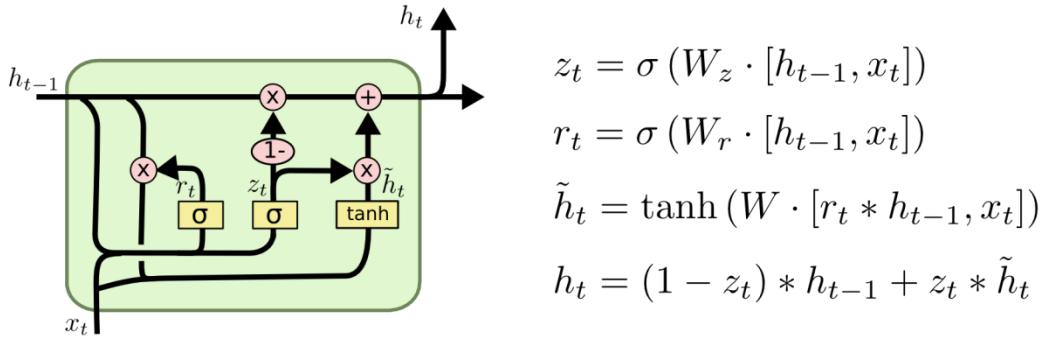


图 8.24 GRU

上面仅介绍了部分 LSTM 变体。有人对这些变体做了比较 [Greff, et al. \(2015\)](#)，发现它们其实都差不多。也有人测试了超过 1 万个 RNN 结构 [Jozefowicz, et al. \(2015\)](#)，发现有些变体在确定任务中比 LSTM 更好。

8.2.4 RNN 的发展方向

LSTM 在大部分任务上都工作的很好。LSTM 使得 RNN 向前发展了一大步。而下一个将使得 RNN 发展一大步的是 attention (翻译做注意力机制，将在第十章介绍)。

其思想是让 RNN 的每一个 time step 挑选信息，以看得到其他的信息集合。例如，在使用 RNN 创建一个描述图片内容的短文 (caption) 任务中，让输出的每个 word 可以看见挑选的一部分图片。关于 attention 可以参考论文 [Xu, et al. \(2015\)](#)。

RNN 发展的另外的方向还包括 Grid LSTMs by [Kalchbrenner, et al. \(2015\)](#)。

还有一些工作在 generative models 中使用 RNN，例如 [Gregor, et al. \(2015\)](#), [Chung, et al. \(2015\)](#), or [Bayer & Osendorfer \(2015\)](#) 也是一个 RNN 未来的方向。

第五节：用 Tensorflow 构建词级 LSTM 语言模型

8.5.1 介绍

Wojciech Zaremba 的论文 RECURRENT NEURAL NETWORK REGULARIZATION 实施了一个 word 级的 LSTM 语言模型（见 `ptb_word_lm.py`）。它在 Penn Tree Bank 数据集上进行 words 级别的预测，即输入是一个 words 的序列，输出是每个输入 word 的下一个邻近词的预测。Tensorflow 的 tutorial 讲解了该模型

<https://tensorflow.google.cn/tutorials/sequences/recurrent>。为了方便理解和讲解，

本文将原程序改写成了一个简单版本，`ptb_word_lm_simple.py`。这个模型有几个特点：（1）使用词向量的方法；（2）词向量的长度和隐层的 `hidden_size` 一致，则不需要使用一个全连接层连接输入和隐层；（3）使用序列损失函数。

构建 LSTM 模型和的步骤和 8.2 节的步骤一样。包括使用 LSTM Cell 函数构建 Cell 和使用 `rnn` 函数构建图 8.5 的隐层。构建 `LSTMcell` 的函数有很多

- (1) `tf.contrib.rnn.BasicLSTMCell` (或者 `tf.nn.rnn_cell.BasicLSTMCell`) 构建最基本的 LSTM Cell
- (2) `tf.contrib.rnn.LSTMBlockCell` 实现了论文 <http://arxiv.org/abs/1409.2329> 的 LSTM
- (3) `tf.contrib.rnn.LSTMCell` (或 `tf.nn.rnn_cell.LSTMCell`) 可以构建更高级、复杂的 LSTM Cell。

这一节详细讲解 `BasicLSTMCell`。基本的 LSTM RNN Cell (实现了论文

<https://arxiv.org/pdf/1409.2329.pdf>) tensorflow 中的函数

`tf.contrib.rnn.BasicLSTMCell` 参数如下：

num_units: int, nj 见 8.5.1 节的解释。

forget_bias: float, 加到 forget gates 的偏置。Must set to 0.0 manually when restoring from CudnnLSTM-trained checkpoints.

state_is_tuple: 如果为 True , 接收和返回的 states 是一个二元组{ `c_state` , `m_state` }。如果为 False, 两个状态沿着列拼接到一起。

activation: 内部状态激活函数，默认是 `tanh`。

8.5.2 数据集

该模型需要的 PTB 数据集 <http://www.fit.vutbr.cz/~imikolov/rnnlm/simple-examples.tgz>。该数据集已经被预处理，包含 10000 个 word，包括句子结束标记和一个特殊的符号`<unk>`指代很少出现的词。Tensorflow tutorial 提供的 `reader.py` 转换数据集，分配词编号。我们使用 `reader.py` 获取数据

```
raw_data = reader.ptb_raw_data(data_path)
train_data, valid_data, test_data, _ = raw_data
```

注：深度学习处理文本时首先确定一个词汇表，且为一个特殊字符 UNK 分配编号。然后把原始文本中的词用编号替换。未出现在词汇表中的词用 UNK 的编号替换。

8.5.3 模型结构

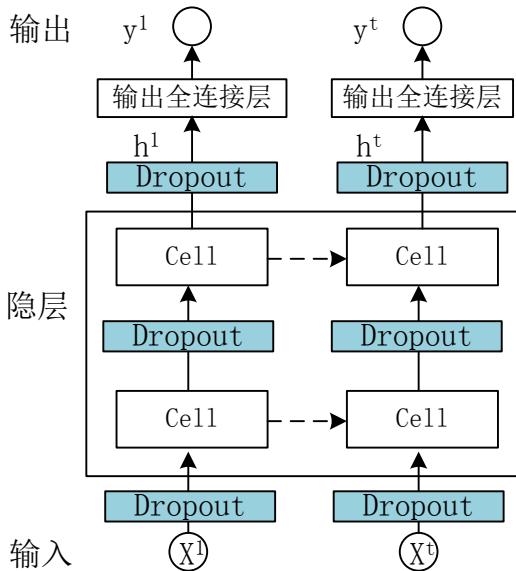


图 8.25 多层 LSTM

图中的 X^t 是一个词经过查找表映射到的一个词向量。 y^t 是一个长度为词汇表大小的向量。

1. 输入

在将输入数据“喂给”LSTM模型前，输入的词是以编号提供的，应该被转换成词向量。在该例子中，查找表被随机初始化，也作为模型的参数来学习。

```
lookup_table = tf.Variable(tf.random_uniform([vocab_size, hidden_size], -1.0, 1.0), name="Lookuptable")
inputs = tf.nn.embedding_lookup(lookup_table, input_x)
```

创建一个 variable，作为 embeddings。然后查找表函数 `tf.nn.embedding_lookup` 会将输入转换成词向量。

2. 多层 RNN

要想使得模型有更强的表达能力，可以加多个 LSTM 层来处理数据。第一层的输出将成为第二层的输入，以此类推。MultiRNNCell 类可以实施多层 RNN。

```
if is_training and keep_prob < 1:
    inputs = tf.nn.dropout(inputs, keep_prob)

cell1 = tf.contrib.rnn.BasicLSTMCell(
    hidden_size, forget_bias=0.0, state_is_tuple=True,
```

```

reuse=False)

cell1 = tf.contrib.rnn.DropoutWrapper(cell1, output_keep_prob
= keep_prob)
cell2 = tf.contrib.rnn.BasicLSTMCell(
    hidden_size, forget_bias=0.0, state_is_tuple=True,
reuse=False)

cell2 = tf.contrib.rnn.DropoutWrapper(cell2,
output_keep_prob=keep_prob)
cell = tf.contrib.rnn.MultiRNNCell([cell1, cell2],
state_is_tuple=True)

initial_state = cell.zero_state(batch_size, tf.float32)
inputs = tf.unstack(inputs, num=num_steps, axis=1)

outputs, state = tf.nn.static_rnn(cell, inputs,
initial_state=initial_state)

```

代码中还使用了几个技巧以得到更好性能的模型：(1)有计划的减小学习率；(2)在LSTM层之间运用dropout。

3. 输出层

隐层的输出经过一个全连接层的变换，转换成概率分布

```

output = tf.reshape(tf.concat(outputs, 1), [-1, hidden_size])

softmax_w = tf.get_variable(
    "softmax_w", [hidden_size, vocab_size],
dtype=tf.float32)
softmax_b = tf.get_variable("softmax_b", [vocab_size],
dtype=tf.float32)

# tf.nn.xw_plus computes matmul(x, weights) + biases. add
softmax on outputs
logits = tf.nn.xw_plus_b(output, softmax_w, softmax_b)
# Reshape logits to be a 3-D tensor for sequence loss
logits = tf.reshape(logits, [batch_size, num_steps,
vocab_size])

```

4. 损失函数

损失函数采用 `tf.contrib.seq2seq.sequence_loss`。它是一个序列的 logit
(logit 就是神经网络的输出层，循环神经网络中，每个时间步的输出就是一个 logit)
的交叉熵损失函数。

```
loss = tf.contrib.seq2seq.sequence_loss(  
    logits,  
    targets,  
    tf.ones([batch_size, num_steps], dtype=tf.float32),  
    average_across_timesteps=False,  
    average_across_batch=True)  
  
# Update the cost  
cost = tf.reduce_sum(loss)
```

其参数描述如下：

Logits，是一个 tensor，它的 `shape=[batch_size, sequence_length, num_decoder_symbols]`，float 类型。`sequence_length` 即时间步的步数，`num_decoder_symbols` 即一个时间步输出的向量长度。

target，一个 tensor，`shape= [batch_size, sequence_length]`，float 类型。即每个时间步应该的输出。每个元素是一个词在词汇表中的编号。

weight，一个 tensor，`shape= [batch_size, sequence_length]`，float 类型，构造在每个预测上的权重。也可以使用 weight 当作 mask。有些的时间步如果设置为 1，不需要的时间步（padded 时间步）设为 0。

average_across_timesteps，如果为 True，计算所有时间步上损失值的均值。

average_across_batch，如果为 True，计算所有 batch 上损失值的均值。

5. 优化器

建立梯度下降优化器

```
tvars = tf.trainable_variables()  
grads, _ = tf.clip_by_global_norm(tf.gradients(cost,  
tvars),max_grad_norm)  
optimizer = tf.train.GradientDescentOptimizer(learning_rate)  
train_op = optimizer.apply_gradients(zip(grads, tvars),  
global_step=tf.train.get_or_create_global_step())
```

6. 训练模型

从数据集中抽取一个 batch 数据进行训练。

```
fetches = {
    "cost": cost,
    "final_state": final_state,
    "train_op": train_op
}
init = tf.global_variables_initializer()
session = tf.Session()
session.run(init)
p = 0
while p<len(train_data)-batch_size:
    i_inputs, i_targets, p = getBatch(train_data, batch_size,
num_steps, p)
    feed_dict = {input_x:i_inputs, targets:i_targets}
    vals = session.run(fetches, feed_dict)
    cost = vals["cost"]
    state = vals["final_state"]
    print(cost)
```

8.5.4 高维输出的问题

在上述的语言模型中，输出结果是词。最简单的方法是如上用一个向量来描述词的分布。如果词汇表很大，例如有 10 万个词，因此输出结果是一个长度为 10 万的向量。这需要非常大的计算代价。

传统的做法是限制词汇表的大小，例如限制到 1 万到 2 万。D-softmax(参见论文“Strategies for Training Large Vocabulary Neural Language Models”)是一种可行的方法。但目前没有 Tensorflow 的官方实施。

第六节: 训练 RNN 的一些技巧

<https://danijar.com/tips-for-training-recurrent-neural-networks/>

第九章：基于 LSTM 的文本情感分析

第一节：介绍

<http://deeplearning.net/tutorial/lstm.html>

这篇文章实施了一个电影评论数据集

(<http://ai.stanford.edu/~amaas/data/sentiment/>) 的情感分类模型。它完成一个正向、负向情感分类的任务。

该文建立的 LSTM 模型是传统 LSTM 模型的一个简化版。Output gate 不依赖 cell 的状态 C_t 。

注：按照该文的说法，第 7 章我们介绍的 LSTM 就是该简化版，如图 8.1 所示。而传统版计算 $o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o)$ 。 V_o 是一个权重矩阵。

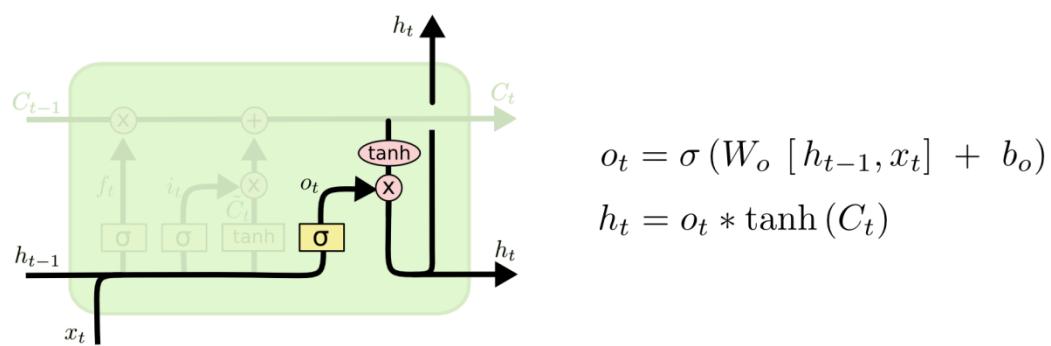
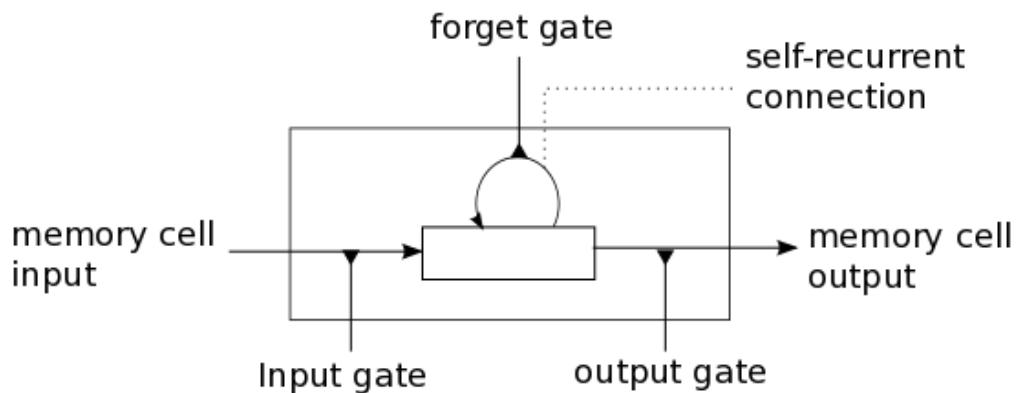


图 8.1 简化版的 LSTM

该文的模型如图 8.2 所示。和传统的 RNN 不同的是在 LSTM 层上加一个 average pooling 层，再加上一个 logistics 回归层。因此从输入序列 $x_0, x_1, x_2, \dots, x_n$ ，LSTM 层的 cell 将产生一个输出序列 h_0, h_1, \dots, h_n 。这个序列被求平均，得到一个输出 h 。最后 logistics 回归层产生一个类标签。

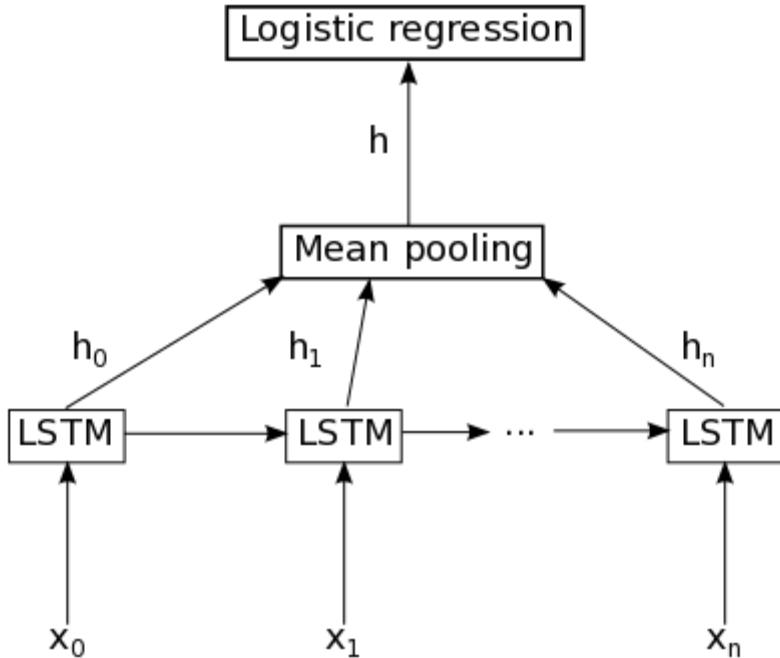


图 8.2 一个评论数据情感分类的 RNN 模型

输入的文本长度做成了固定长度。或者用户自己规定最大长度，或者以训练集中最长文本作为最大长度。不足的补零（词的编号 0，代表没有词）。查找表采用的是随机初始词向量。

当前在用 TensorFlow 实施该模型时，一开始用训练集最长文本的长度作为 time step 的值，即每个文档被补齐到相同的长度。但这时模型在处理文本时效率非常之低。长时间的未能处理完文本。我决定用动态 time step 方式来建立模型，即每篇文本的长度不固定，time step 不固定。

第二节：TensorFlow 实施

为了简化问题，我们没有使用这篇论文提及的电影评论数据集，而是使用 7.2 节的电影评论数据集和它的数据集处理程序。另外，我们也没有去开房该文设计的 cell，而是使用了 `tf.contrib.rnn.BasicLSTMCell`。下面，我只讲述核心的代码

```
tf.reset_default_graph()
time_step=x_train.shape[1]

input_x = tf.placeholder(tf.int32, [None, time_step], name="input_x")
input_y = tf.placeholder(tf.float32, [None, num_classes], name="input_y")
```

`tf.reset_default_graph()` 这条语句重新设置计算图

为输入数据设置 `placeholder()`

```
vocab_size=len(vocab_processor.vocabulary_)
C = tf.Variable(tf.random_uniform([vocab_size, embedding_dim], -1.0, 1.0),
name="lookuptable")
inputs = tf.nn.embedding_lookup(C, input_x)
```

建立查找表，将输入数据转换成词向量，并拼接

```
cell = tf.contrib.rnn.BasicLSTMCell(
    hidden_size, forget_bias=0.0, state_is_tuple=True, reuse=False)
inputs = tf.unstack(inputs, num=time_step, axis=1)
outputs, state = tf.nn.static_rnn(cell, inputs, dtype=tf.float32)
```

建立基本的 LSTM 模型。LSTM 和传统 rnn 的区别只是体现在 `cell` 上。此处，我们没有在 `static_rnn` 函数中给 `state` 参数赋值，则此时使用默认的值初始化 `state`。

```
#mean pooling
np_output=np.array(outputs)
mean_output=sum(np_output)/time_step
```

这段代码实现图 8.2 的 mean pooling 功能。为了操作的方便，我们将数据转换成 numpy 的 ndarray 数据结构

```
# logistic regression
W = tf.Variable(tf.random_uniform([hidden_size, num_classes], -1.0, 1.0))
b = tf.Variable(tf.zeros([num_classes]))
scores = tf.matmul(mean_output, W) + b
loss = tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(logits=scores,
labels=input_y))
```

这一段代码实现图 8.2 的 logistic 回归的功能。其实就是神经网络的一个隐层加上一个 softmax 激活函数。隐层的计算结果 (scores) 使用 tf.nn.softmax_cross_entropy_with_logits_v2 函数可以直接计算 logistic 回归的输出和目标值的损失函数。

```
optimizer = tf.train.AdamOptimizer(0.001)
train_op = optimizer.minimize(loss)
```

然后，建立优化器。下面的代码是度量模型的精确度

```
preds=tf.argmax(tf.nn.softmax(scores),1)
correct_predictions = tf.equal(preds, tf.argmax(input_y, 1))
accuracy = tf.reduce_mean(tf.cast(correct_predictions, "float"), name="accuracy")
```

全部源码参见 [github](#) 的 lstm-sentiment.py。

第十章：注意力机制

Attention Mechanism (翻译做注意力机制) 是 Dzmitry Bahdanau 在论文“Neural machine translation by jointly learning to align and translate” 中提出的。两个很常用的两个 attention function: MLP Attention [1] and Bilinear Attention [2]

Reference

- [1] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved Representation Learning for Question Answer Matching. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 464–473
- [2] Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2358–2367

第十一章：RNN Encoder-Decoder 和生成新闻标题

RNN Encoder-Decoder 模型是一个 sequence to sequence 模型。这种模型最常见的任务是机器翻译。从一种语言的句子（sequence）翻译成另一种语言的句子（sequence）。本章学习 RNN Encoder-Decoder 模型，找的应用场景是关键词抽取。参考了 ACL2017 论文《Deep Keyphrase Generation》，不过我对该论文存疑，因为产生的关键词不是序列（序列中前后词具有相关性），需要采用 RNN Encoder-Decoder 模型吗？（我觉得一个 RNN 模型就可以了）下面的内容结合了该论文和其他的 RNN Encoder-Decoder 相关论文。

Keywords 和 keyphrase 是指词语或词组的集合，用于对文档的内容做主要语义描述。传统的关键短语抽取（非深度学习的方法）通常包含两个步骤：（1）从文本内容里抽取潜在的候选短语；（2）对候选短语排序。传统的方法有个缺点，它们都是使用文本内容里出现的短语（present keyphrase），而不能基于文本的语义，即总结了文本的内容，但没有出现在文本里的词（absent keyphrase）不会被选择。人类在从一段文本内容中抽取关键词时，往往产生的词不一定是出现在文本中的，但是和文本内容语义相关的。该文建立的 RNN 模型试图捕捉文本的语义和语法特征。

1. 工作原理

基于 RNN 的 Encoder-decoder 模型。其思想是，使用 Encoder 压缩文本内容到一个隐描述空间，然后用一个 Decoder 产生对应的 keyphrase。Encoder 和 Decoder 都是用 RNN 来实施。

Encoder 通过迭代执行下面的公式

$$h_t = f(x_t, h_{t-1})$$

转换长度可变的输入序列（文本） $x=(x_1, x_2, \dots, x_T)$ 到一个隐描述序列
 $h=(h_1, h_2, \dots, h_T)$ 。这里， f 是一个非线性函数。我们可以获得一个 context 向量 c
 $c=q(h_1, h_2, \dots, h_T)$

Decoder 是另外一个 RNN。它使用一个条件语言模型

$$s_t = f(y_{t-1}, s_{t-1}, c)$$

$$p(y_t | y_1, \dots, y_{t-1}, x) = g(s_t, c)$$

将 context 向量分解成一个长度可变的序列 $y = (y_1, y_2, \dots, y_T)$ 。 s_t 是 decoder RNN 在时刻 t 的隐状态。非线性函数 g 是一个 softmax 分类器，它的输出是词汇表中所有词的概率。 y_t 是在 t 时刻被预测的词项，即选取 $g(\cdot)$ 中对应最大概率的词项。

Encoder 和 Decoder 网络联合训练，以达到给定源序列，最大化目标序列的条件概率。训练完成后使用 beam search 来产生 phrase。一个 max heap 被维护用于从预测的概率值中，选择预测的 word 序列。

2. 实施细节

Encoder 应用一个双向 gated recurrent unit (GRU)。有研究已经证实，GRU 比简单的 RNN 和 LSTM 中的简单结构在语言模型上具有更好的性能。因此，上面的非线性函数用 GRU 替换。

Decoder 中使用一个前向 GRU。另外一个 attention mechanism 被采纳，用于提高性能。

第十二章：基于 BiLSTM 的问答系统

第十三章：最新发展

Appendix A: 常用 TensorFlow 函数

1. tf.split

```
tf.split(  
    value,  
    num_or_size_splits,  
    axis=0,  
    num=None,  
    name='split'  
)
```

该函数将一个 tensor 划分成多个子 tensor。Value 是待划分的 tensor, axis 是沿着哪个轴划分 ; num_or_size_splits 是划分成多少个部分。原始 tensor 的长度 , 必须能被 num_or_size_splits 整除。

```
x =  
tf.Variable([[1,2],[3,4],[5,6],[7,8],[9,10],[11,12],[13,14],[  
15,16]])  
data = tf.split(axis=0, num_or_size_splits=2, value=x)  
  
init = tf.global_variables_initializer()  
sess = tf.Session()  
sess.run(init)  
d = sess.run(data)  
print(d)  
  
划分成两个部分  
[array([[1, 2], [3, 4], [5, 6], [7, 8]]),  
 array([[ 9, 10], [11, 12], [13, 14], [15, 16]])]
```

2. tf.reshape

```
tf.reshape(
```

```
    tensor,  
    shape,  
    name=None  
)
```

重新安排一个 tensor 的 shape. 该函数的 shape 参数是重新设定的 shape。如果 shape 中一个维度的值给的是-1，表示该维度的 size 需要计算，以保证其他维度上给的 size

```
x1 =  
tf.Variable([[[1,2],[3,4],[5,6]],[[7,8],[9,10],[11,12]]])  
x2 = tf.reshape(tensor=x1,shape=[-1,2,2])  
init = tf.global_variables_initializer()  
sess = tf.Session()  
sess.run(init)  
d = sess.run(x2)  
print(d)
```

X1 的形状如下：

1,2	7,8
3,4	9,10
5,6	11,12

X2 的形状

1,2	5,6	9,10
3,4	7,8	11,12

3. tf.concat

```
tf.concat(  
    values,  
    axis,  
    name='concat'  
)
```

将多个 tensor 沿着 tensor 的一个维度，重新连接 tensor

```
x1 = tf.Variable([[1,2],[3,4],[5,6]])
x2 = tf.Variable([[7,8],[9,10],[11,12]])
op = tf.concat(axis=1, values=[x1,x2])
init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init)
d = sess.run(op)
print(d)
```

结果是：

```
[[ 1  2  7  8]
 [ 3  4  9 10]
 [ 5  6 11 12]]
```

4. tf.nn.xw_plus_b

tf.nn.xw_plus_b(

 x,

 weights,

 biases,

 name=None

)

计算 matmul(x, weights) + biases.

X 是 2-D tensor [batch_size, 输入数据的维度]

Weights 是 2-D tensor[输入数据维度，神经元数]

Biases 是 1-D tensor [神经元数]

```
W = tf.Variable([[1,1,1],[2,2,2]])
X = tf.Variable([[1,2],[3,4],[5,6]])
b = tf.Variable([1,2,3])
scores = tf.nn.xw_plus_b(x=X, weights=W, biases=b)

init = tf.global_variables_initializer()
sess = tf.Session()
sess.run(init)
r=sess.run(scores)
print(r)
```

结果是

```
[[ 6  7  8]
 [12 13 14]
 [18 19 20]]
```

5. tf.nn.softmax_cross_entropy_with_logits_v2

```
tf.nn.softmax_cross_entropy_with_logits_v2(
    _sentinel=None,
    labels=None,
    logits=None,
    dim=-1,
    name=None
)
```

注：tf.nn.softmax_cross_entropy_with_logits 废弃了，推荐使用 v2 这个新版本。

计算 logits 和 labels 之间的 softmax 交叉熵。一个最通常的使用方法是 logits 和 labels 都有一个相同的 shape=[batch_size, num_classes]。Logit 和 labels 必须有相同的 dtype。Labels 的每一行应该是有效的概率分布。Logits 的每一行是一条数据在所有类别上计算的分值，不需要规范化。该函数有进行规范化的步骤。参看一下代码了解该函数的工作原理。

```
y_true = tf.convert_to_tensor(np.array([[0.0, 1.0, 0.0],[0.0, 0.0, 1.0]]))
y_hat = tf.convert_to_tensor(np.array([[0.5, 1.5, 0.1],[2.2, 1.3, 1.7]]))

# first step
y_hat_softmax = tf.nn.softmax(y_hat)

# second step
y_cross = y_true * tf.log(y_hat_softmax)

# third step
result = - tf.reduce_sum(y_cross, 1)
result_tf = tf.nn.softmax_cross_entropy_with_logits_v2(labels = y_true, logits = y_hat)
```

```

with tf.Session() as sess:
    sess.run(result)
    sess.run(result_tf)

    print('y_hat_softmax:\n{}'.format(y_hat_softmax.eval()))
    print('y_true: \n{}'.format(y_true.eval()))
    print('y_cross: \n{}'.format(y_cross.eval()))
    print('result: \n{}'.format(result.eval()))
    print('result_tf: \n{}'.format(result_tf.eval()))

```

上面的程序，我们可以把 y_{hat} 的一行理解为一条预测结果。每一列是在每个类别上（该程序有三个类别）计算的分数。 y_{true} 的一行理解为真实的标签，即为 1 的列表示分类为该类别。调用该函数时

`tf.nn.softmax_cross_entropy_with_logits_v2(labels = y_true, logits = y_hat)`，该函数的工作包括三个步骤：

(1) 应用 softmax 到参数 logits (这里是 y_{hat}) 将它们的值规范化到[0-1]
 $y_{hat_softmax} = softmax(y_{hat})$.

(2) 计算交叉熵损失函数: $y_{cross} = y_{true} * tf.log(y_{hat_softmax})$

(3) 在一条实例的不同类别上求和: $-tf.reduce_sum(y_{cross}, reduction_indices=[1])$

6. tf.contrib.seq2seq.sequence_loss

```

tf.contrib.seq2seq.sequence_loss(
    logits,
    targets,
    weights,
    average_across_timesteps=True,
    average_across_batch=True,
    softmax_loss_function=None,
    name=None
)

```

对一个序列计算加权的交叉熵损失函数。

logits: A Tensor of shape [batch_size, sequence_length, num_decoder_symbols] and dtype float. The logits correspond to the prediction across all classes at each timestep.

targets: A Tensor of shape [batch_size, sequence_length] and dtype int. The target represents the true class at each timestep.

weights: A Tensor of shape [batch_size, sequence_length] and dtype float. weights constitutes the weighting of each prediction in the sequence. When using weights as masking, set all valid timesteps to 1 and all padded timesteps to 0, e.g. a mask returned by tf.sequence_mask.

average_across_timesteps: If set, sum the cost across the sequence dimension and divide the cost by the total label weight across timesteps.

average_across_batch: If set, sum the cost across the batch dimension and divide the returned cost by the batch size.

softmax_loss_function: Function (labels, logits) -> loss-batch to be used instead of the standard softmax (the default if this is None). Note that to avoid confusion, it is required for the function to accept named arguments.

7. tf.argmax

```
tf.argmax(  
    input,  
    axis=None,  
    name=None,  
    dimension=None,  
    output_type=tf.int64  
)
```

该函数查找当前 tensor 中在某个维度上的最大数据所对应的下标。该函数经常被用于确定模型分类的类别。

Input 是 tensor, axis 表示沿着哪个轴去确定

```
x = [[1,2,3],[6,5,4],[1,2,1]]  
a = tf.argmax(x, 1)  
  
sess = tf.Session()  
r = sess.run(a)  
print(r)
```

结果是[2,0,1]

8. tf.nn.embedding_lookup

```
tf.nn.embedding_lookup(  
    params,
```

```
ids,  
partition_strategy='mod',  
name=None,  
validate_indices=True,  
max_norm=None  
)
```

进行表示向量的查找。Params 可以是一个 tensor , 它描述了完整的表示向量。或者可以是一个 list 结构 , 每个元素是一个 tensor , 所有的元素有相同的 shpae。

ids 是一个 tensor。包含要从 params 中查找到的 ids。

查找返回的结果是一个 tensor。它的 shape 是 shape(ids) 再增加一个维度 , 表示向量。

```
a = tf.constant([[0,0,0,0,1],[0,0,0,1,0],[0,0,1,0,0],[0,1,0,0,0],[1,0,0,0,0]])  
input = [[1,2],[2,3]]  
b = tf.nn.embedding_lookup(params=a, ids=input)  
  
sess = tf.Session()  
r=sess.run(b)
```

上面程序的运行后 , r 的结果是 :

```
array([[[0, 0, 0, 1, 0],  
       [0, 0, 1, 0, 0]],
```

```
       [[0, 0, 1, 0, 0],  
        [0, 1, 0, 0, 0]]])
```

9. **tf.reduce_mean**

```
tf.reduce_mean(  
    input_tensor,  
    axis=None,  
    keepdims=None,  
    name=None,  
    reduction_indices=None,
```

```
    keep_dims=None  
)  
在 tensor 的某个维度上计算均值 ( 约减操作 )。axis: 在哪个维度上进行均值计算约  
减。如果 None (默认), 所有的维度上计算均值 , 进行约减。必须是在范围 [-  
rank(input_tensor), rank(input_tensor)).
```

例子

```
x = tf.constant([[1., 1.], [2., 2.]])  
tf.reduce_mean(x) # 1.5  
tf.reduce_mean(x, 0) # [1.5, 1.5]  
tf.reduce_mean(x, 1) # [1., 2.]
```

10. tf.truncated_normal

```
tf.truncated_normal(  
    shape,  
    mean=0.0,  
    stddev=1.0,  
    dtype=tf.float32,  
    seed=None,  
    name=None  
)
```

产生正太分布的随机数 , 但是如果产生的随机数大于从均值开始的 2 倍标准差 , 这放
弃重新产生。

11. tf.random_uniform

```
tf.random_uniform(  
    shape,  
    minval=0,  
    maxval=None,  
    dtype=tf.float32,  
    seed=None,
```

```
    name=None  
)  
tf.nn.dropout(  
    x,  
    keep_prob,  
    noise_shape=None,  
    seed=None,  
    name=None  
)
```

13. tf.nn.max_pool

```
tf.nn.max_pool(value, ksize, strides, padding, data_format='NHWC', name=None)
```

Performs the max pooling on the input.

参数：

value: A 4-D Tensor with shape [batch, height, width, channels] and type tf.float32.

ksize: A list of ints that has length ≥ 4 . The size of the window for each dimension of the input tensor.

strides: A list of ints that has length ≥ 4 . The stride of the sliding window for each dimension of the input tensor.

padding: A string, either 'VALID' or 'SAME'. The padding algorithm. See the comment here

data_format: A string. 'NHWC' and 'NCHW' are supported.

name: Optional name for the operation.

Returns: A Tensor with type tf.float32. The max pooled output tensor.

14. tf.slice

```
tf.slice(  
    input_,
```

```
begin,  
size,  
name=None  
)
```

从一个 tensor 中抽取一个 slice。Input 是输入的 tensor ; begin 是对应着 input 的一个 shape , 只不过里面存放着从哪个维度的那个下标开始抽取。Size 也是对应着 input 的一个 shape , 里面存放着在哪个维度上抽取多少个。

```
t = tf.constant([[[1, 1, 1], [2, 2, 2]],  
                 [[3, 3, 3], [4, 4, 4]],  
                 [[5, 5, 5], [6, 6, 6]]])  
tf.slice(t, [1, 0, 0], [1, 1, 3])      # [[[3, 3, 3]]]  
tf.slice(t, [1, 0, 0], [1, 2, 3])      # [[[3, 3, 3],  
                                         # [4, 4, 4]]]  
tf.slice(t, [1, 0, 0], [2, 1, 3])      # [[[3, 3, 3]],  
                                         # [5, 5, 5]]]
```

Appendix B: TensorFlow 的多线程

Tf.train 可以创建多个线程，完成多个队列中的 op。