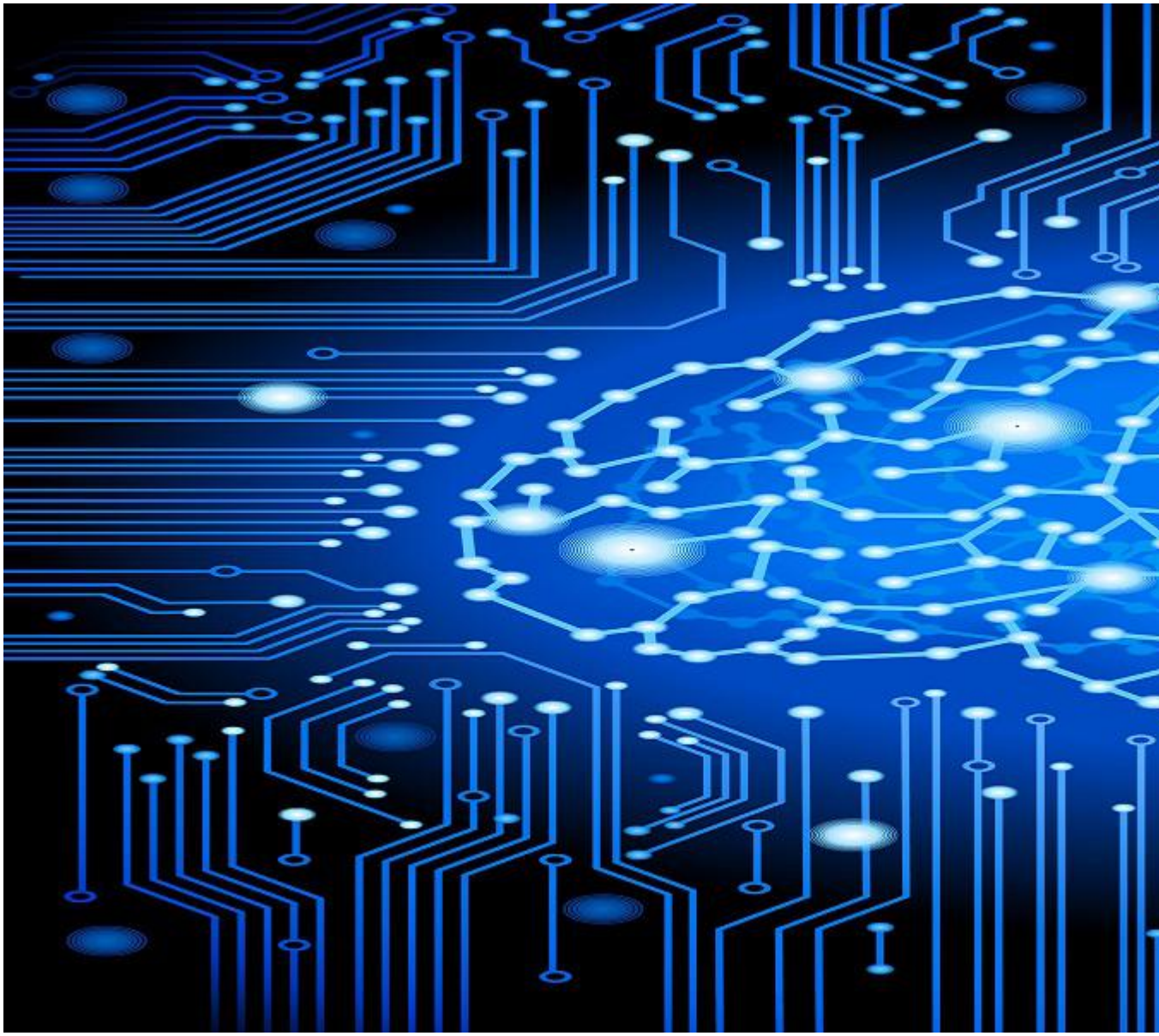
[关闭]

@hanbingtao 2017-08-28 19:54 字数 23656 阅读 221766

零基础入门深度学习(5) - 循环神经网络

机器学习 深度学习入门



无论即将到来的是大数据时代还是人工智能时代，亦或是传统行业使用人工智能在云上处理大数据的时代，作为一个有理想有追求的程序员，不懂深度学习（Deep Learning）这个超热的技术，会不会感觉马上就 out 了？现在救命稻草来了，《零基础入门深度学习》系列文章旨在讲帮助爱编程的你从零基础达到入门级水平。零基础意味着你不需要太多的数学知识，只要会写程序就行了，没错，这是专门为程序员写的文章。虽然文中会有很多公式你也许看不懂，但同时也会有更多的代码，程序员的你一定能看懂的（我周围是一群狂热的 Clean Code 程序员，所以我写的代码也不会很差）。

文章列表

零基础入门深度学习(1) – 感知器
零基础入门深度学习(2) – 线性单元和梯度下降
零基础入门深度学习(3) – 神经网络和反向传播算法
零基础入门深度学习(4) – 卷积神经网络
零基础入门深度学习(5) – 循环神经网络
零基础入门深度学习(6) – 长短时记忆网络(LSTM)
零基础入门深度学习(7) – 递归神经网络

往期回顾

在前面的文章系列文章中，我们介绍了全连接神经网络和卷积神经网络，以及它们的训练和使用。他们都只能单独的取处理一个个的输入，前一个输入和后一个输入是完全没有关系的。但是，某些任务需要能够更好的处理**序列**的信息，即前面的输入和后面的输入是有关系的。比如，当我们在理解一句话意思时，孤立的理解这句话的每个词是不够的，我们需要处理这些词连接起来的整个**序列**；当我们处理视频的时候，我们也不能只单独的去分析每一帧，而要分析这些帧连接起来的整个**序列**。这时，就需要用到深度学习领域中另一类非常重要神经网络：**循环神经网络(Recurrent Neural Network)**。RNN 种类很多，也比较绕脑子。不过读者不用担心，本文将一如既往的对复杂的东西剥茧抽丝，帮助您理解 RNNs 以及它的训练算法，并动手实现一个**循环神经网络**。

语言模型

RNN 是在**自然语言处理**领域中最先被用起来的，比如，RNN 可以为**语言模型**来建模。那么，什么是语言模型呢?

我们可以和电脑玩一个游戏，我们写出一个句子前面的一些词，然后，让电脑帮我们写下接下来的一个词。比如下面这句：

> 我昨天上学迟到了，老师批评了____。

我们给电脑展示了这句话前面这些词，然后，让电脑写下接下来的一个词。在这个例子中，接下来的这个词最有可能是『我』，而不太可能是『小明』，甚至是『吃饭』。

语言模型就是这样的东西：给定一个一句话前面的部分，预测接下来最有可能的一个词是什么。

语言模型是对一种语言的特征进行建模，它有很多很多用处。比如在语音转文本(STT)的应用中，声学模型输出的结果，往往是若干个可能的候选词，这时候就需要**语言模型**来从这些候选词中选择一个最可能的。当然，它同样也可以用在图像到文本的识别中(OCR)。

使用 RNN 之前，语言模型主要是采用 N-Gram。N 可以是一个自然数，比如 2 或者 3。它的含义是，假设一个词出现的概率只与前面 N 个词相关。我们以 2-Gram 为例。首先，对前面的一句话进行切词：

> 我 昨天 上学 迟到 了，老师 批评 了 ____。

如果用 2-Gram 进行建模，那么电脑在预测的时候，只会看到前面的『了』，然后，电脑会在语料库中，搜索『了』后面最可能的一个词。不管最后电脑选的是不是『我』，我们都知道这个模型是不靠谱的，因为『了』前面说了那么一大堆实际上是没有用到的。如果是 3-Gram 模型呢，会搜索『批评了』后面最可能的词，感觉上比 2-Gram 靠谱了不少，但还是远远不够的。因为这句话最关键的信息『我』，远在 9 个词之前！

现在读者可能会想，可以提升继续提升 N 的值呀，比如 4-Gram、5-Gram.....。实际上，这个想法是没有实用性的。因为我们想处理任意长度的句子，N 设为多少都不合适；另外，模型的大小和 N 的关系是指数级的，4-Gram 模型就会占用海量的存储空间。

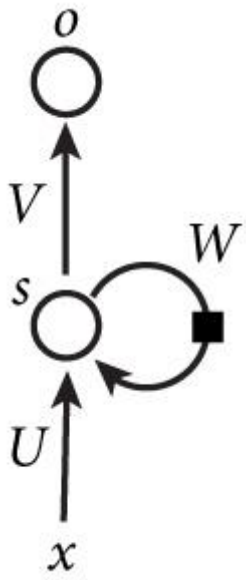
所以，该轮到 RNN 出场了，RNN 理论上可以往前看(往后看)任意多个词。

循环神经网络是啥

循环神经网络种类繁多，我们先从最简单的基本循环神经网络开始吧。

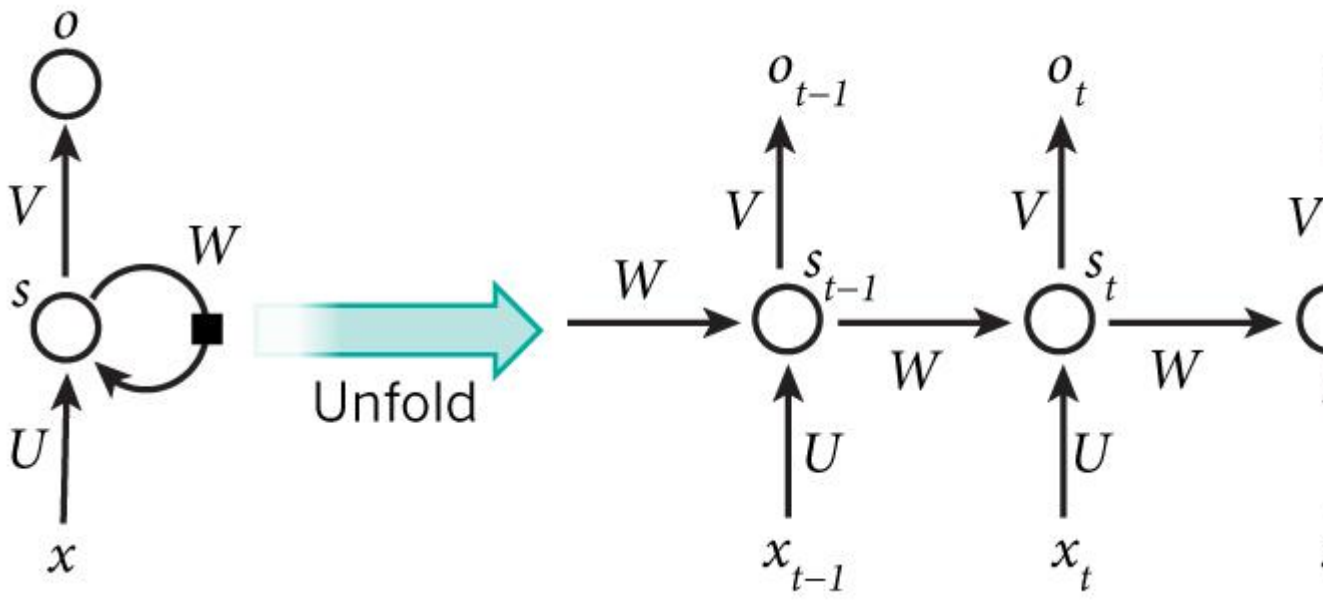
基本循环神经网络

下图是一个简单的循环神经网络如，它由输入层、一个隐藏层和一个输出层组成：



纳尼？！相信第一次看到这个玩意的读者内心和我一样是崩溃的。因为**循环神经网络**实在是太难画出来了，网上所有大神们都不得不用了这种抽象艺术手法。不过，静下心来仔细看看的话，其实也是很好理解的。如果把上面有 W 的那个带箭头的圈去掉，它就变成了最普通的**全连接神经网络**。x 是一个向量，它表示**输入层**的值（这里面没有画出来表示神经元节点的圆圈）；s 是一个向量，它表示**隐藏层**的值（这里隐藏层面画了一个节点，你也可以想象这一层其实是多个节点，节点数与向量 s 的维度相同）；U 是输入层到隐藏层的**权重矩阵**（读者可以回到第三篇文章零基础入门深度学习(3) – 神经网络和反向传播算法，看看我们是怎样用矩阵来表示**全连接神经网络**的计算的）；o 也是一个向量，它表示**输出层**的值；V 是隐藏层到输出层的**权重矩阵**。那么，现在我们来看看 W 是什么。**循环神经网络**的**隐藏层**的值 s 不仅仅取决于当前这次的输入 x，还取决于上一次**隐藏层**的值 s。**权重矩阵** W 就是**隐藏层**上一次的值作为这一次的输入的权重。

如果我们把上面的图展开，**循环神经网络**也可以画成下面这个样子：



现在看上去就比较清楚了，这个网络在 t 时刻接收到输入之后，隐藏层的值是，输出值是。关键一点是，的值不仅仅取决于，还取决于。我们可以用下面的公式来表示**循环神经网络**的计算方法：

式式

式 1 是**输出层**的计算公式，输出层是一个**全连接层**，也就是它的每个节点都和隐藏层的每个节点相连。V 是输出层的**权重矩阵**，g 是**激活函数**。式 2 是隐藏层的计算公式，它是**循环层**。U 是输入 x 的权重矩阵，W 是上一次的值作为这一次的输入的**权重矩阵**，f 是**激活函数**。

从上面的公式我们可以看出，**循环层**和**全连接层**的区别就是**循环层**多了一个**权重矩阵** W。

如果反复把**式 2** 带入到**式 1**，我们将得到：

从上面可以看出，**循环神经网络**的输出值，是受前面历次输入值、、、、...影响的，这就是为什么**循环神经网络**可以往前看任意多个**输入值**的原因。

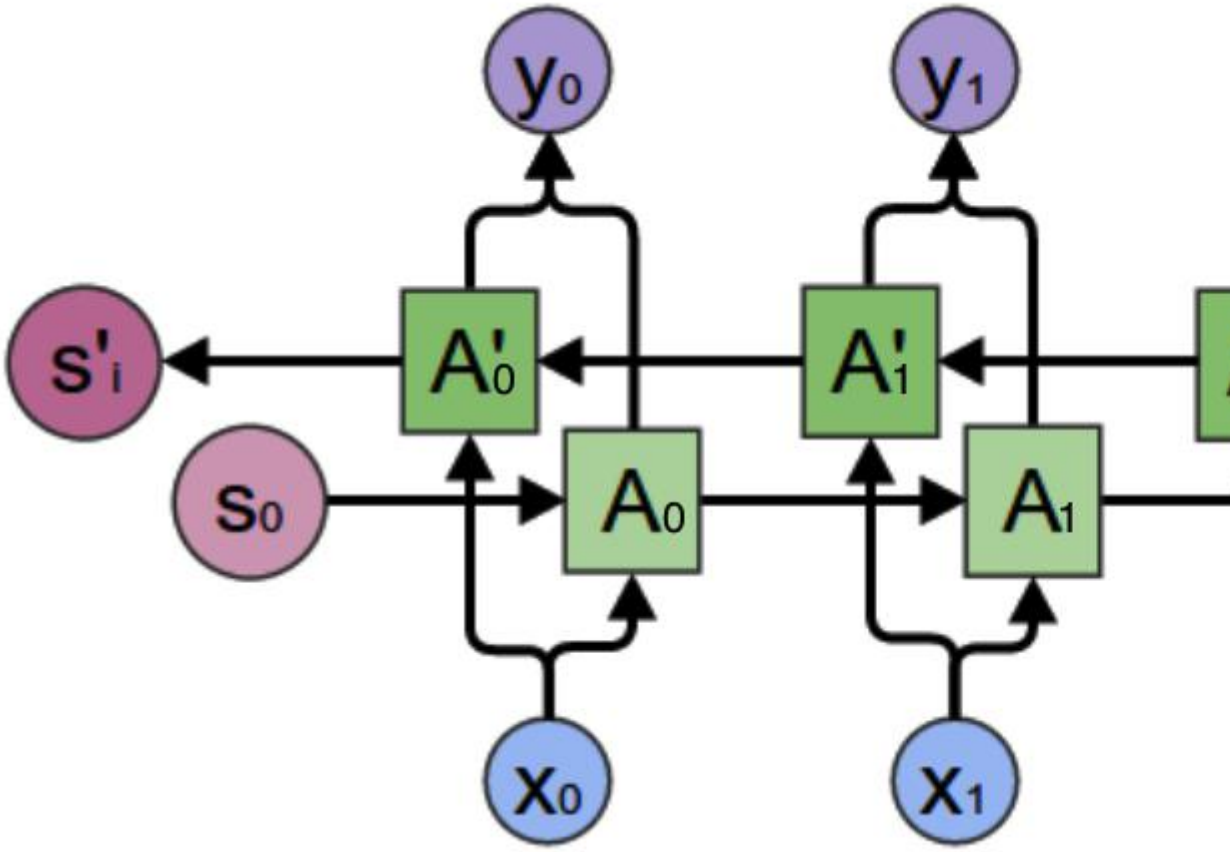
双向循环神经网络

对于**语言模型**来说，很多时候光看前面的词是不够的，比如下面这句话：

我的手机坏了，我打算____一部新手机。

可以想象，如果我们只看横线前面的词，手机坏了，那么我是打算修一修？换一部新的？还是大哭一场？这些都是无法确定的。但如果我们也看到了横线后面的词是『一部新手机』，那么，横线上的词填『买』的概率就大得多了。

在上一小节中的**基本循环神经网络**是无法对此进行建模的，因此，我们需要**双向循环神经网络**，如下图所示：



当遇到这种从未来穿越回来的场景时，难免处于懵逼的状态。不过我们还是可以用屡试不爽的老办法：先分析一个特殊场景，然后再总结一般规律。我们先考虑上图中，的计算。

从上图可以看出，**双向卷积神经网络**的隐藏层要保存两个值，一个 A 参与正向计算，另一个值 A'参与反向计算。最终的输出值取决于和。其计算方法为：

和则分别计算：

现在，我们已经可以看出一般的规律：正向计算时，隐藏层的值与有关；反向计算时，隐藏层的值与有关；最终的输出取决于正向和反向计算的**加和**。现在，我们仿照**式 1**和**式 2**，写出双向循环神经网络的计算方法：

从上面三个公式我们可以看到，正向计算和反向计算**不共享权重**，也就是说 U 和 U'、W 和 W'、V 和 V'都是不同的**权重矩阵**。

深度循环神经网络

前面我们介绍的**循环神经网络**只有一个隐藏层，我们当然也可以堆叠两个以上的隐藏层，这样就得到了**深度循环神经网络**。如下图所示：

我们把第 i 个隐藏层的值表示为、，则**深度循环神经网络**的计算方式可以表示为：

循环神经网络的训练

循环神经网络的训练算法：BPTT

BPTT 算法是针对**循环层**的训练算法，它的基本原理和 BP 算法是一样的，也包含同样的三个步骤：

1. 前向计算每个神经元的输出值；
2. 反向计算每个神经元的**误差项**值，它是误差函数 E 对神经元 j 的**加权输入**的偏导数；
3. 计算每个权重的梯度。

最后再用**随机梯度下降**算法更新权重。

循环层如下图所示：

前向计算

使用前面的**式 2** 对循环层进行前向计算：

注意，上面的、、都是向量，用**黑体字母**表示；而 U、V 是**矩阵**，用大写字母表示。**向量的下标**表示**时刻**，例如，表示在 t 时刻向量 s 的值。

我们假设输入向量 x 的维度是 m，输出向量 s 的维度是 n，则矩阵 U 的维度是，矩阵 W 的维度是。下面是上式展开成矩阵的样子，看起来更直观一些：

在这里我们用**手写体字母**表示向量的一个**元素**，它的下标表示它是这个向量的第几个元素，它的上标表示第几个**时刻**。例如，表示向量 s 的第 j 个元素在 t 时刻的值。表示**输入层**第 i 个神经元到**循环层**第 j 个神经元的权重。表示**循环层**第 t–1 时刻的第 i 个神经元到**循环层**第 t 个时刻的第 j 个神经元的权重。

误差项的计算

BTPP 算法将第 l 层 t 时刻的**误差项**值沿两个方向传播，一个方向是其传递到上一层网络，得到，这部分只和权重矩阵 U 有关；另一个是方向是将其沿时间线传递到初始时刻，得到，这部分只和权重矩阵 W 有关。

我们用向量表示神经元在 t 时刻的**加权输入**，因为：

因此：

我们用 a 表示列向量，用表示行向量。上式的第一项是向量函数对向量求导，其结果为 Jacobian 矩阵：

同理，上式第二项也是一个 Jacobian 矩阵：

其中，diag[a]表示根据向量 a 创建一个对角矩阵，即

最后，将两项合在一起，可得：

上式描述了将沿时间往前传递一个时刻的规律，有了这个规律，我们就可以求得任意时刻 k 的**误差项**：

式

式 3 就是将误差项沿时间反向传播的算法。

循环层将**误差项**反向传递到上一层网络，与普通的**全连接层**是完全一样的，这在前面的文章零基础入门深度学习(3) – 神经网络和反向传播算法中已经详细讲过了，在此仅简要描述一下。

循环层的**加权输入**与上一层的**加权输入**关系如下：

上式中是第 l 层神经元的**加权输入**(假设第 l 层是**循环层**)；是第 l–1 层神经元的**加权输入**；是第 l–1 层神经元的输出；是第 l–1 层的**激活函数**。

所以，

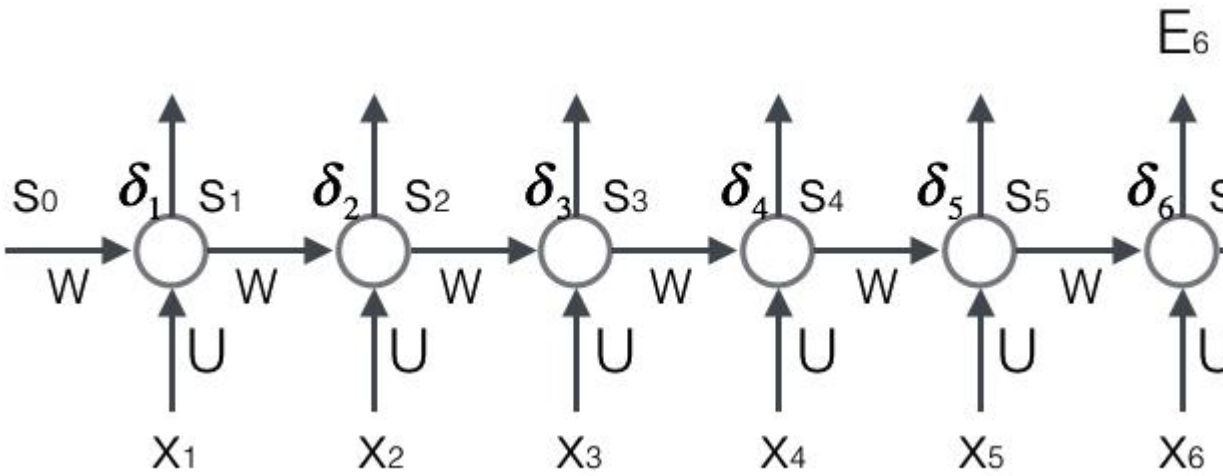
式

式 4 就是将误差项传递到上一层算法。

权重梯度的计算

现在，我们终于来到了 BPTT 算法的最后一步：计算每个权重的梯度。

首先，我们计算误差函数 E 对权重矩阵 W 的梯度。



上图展示了我们到目前为止，在前两步中已经计算得到的量，包括每个时刻 t **循环层**的输出值，以及误差项。

回忆一下我们在文章零基础入门深度学习(3) – 神经网络和反向传播算法介绍的全连接网络的权重梯度计算算法：只要知道了任意一个时刻的**误差项**，以及上一个时刻循环层的输出值，就可以按照下面的公式求出权重矩阵在 t 时刻的梯度：

式

在**式 5** 中，表示 t 时刻**误差项**向量的第 i 个分量；表示 t–1 时刻**循环层**第 i 个神经元的输出值。

我们下面可以简单推导一下**式 5**。

我们知道：

因为对 W 求导与无关，我们不再考虑。现在，我们考虑对权重项求导。通过观察上式我们可以看到只与有关，所以：

按照上面的规律就可以生成**式 5** 里面的矩阵。

我们已经求得了权重矩阵 W 在 t 时刻的梯度，最终的梯度是各个时刻的梯度**之和**：

式

式 6 就是计算**循环层**权重矩阵 W 的梯度的公式。

----------数学公式超高能预警----------

前面已经介绍了的计算方法，看上去还是比较直观的。然而，读者也许会困惑，为什么最终的梯度是各个时刻的梯度**之和**呢？我们前面只是直接用了这个结论，实际上这里面是有道理的，只是这个数学推导比较绕脑子。感兴趣的同学可以仔细阅读接下来这一段，它用到了矩阵对矩阵求导、张量与向量相乘运算的一些法则。

我们还是从这个式子开始：

因为与 W 完全无关，我们把它看做常量。现在，考虑第一个式子加号右边的部分，因为 W 和都是 W 的函数，因此我们要用到大学里面都学过的导数乘法运算：

因此，上面第一个式子写成：

我们最终需要计算的是：

式

我们先计算**式 7** 加号左边的部分。是**矩阵对矩阵求导**，其结果是一个四维**张量(tensor)**，如下所示：

接下来，我们知道，它是一个**列向量**。我们让上面的四维张量与这个向量相乘，得到了一个三维张量，再左乘行向量，最终得到一个矩阵：

接下来，我们计算**式 7** 加号右边的部分：

于是，我们得到了如下递推公式：

这样，我们就证明了：最终的梯度是各个时刻的梯度之和。

----------数学公式超高能预警解除----------

同权重矩阵 W 类似，我们可以得到权重矩阵 U 的计算方法。

式

式 8 是误差函数在 t 时刻对权重矩阵 U 的梯度。和权重矩阵 W 一样，最终的梯度也是各个时刻的梯度之和：

具体的证明这里就不再赘述了，感兴趣的读者可以练习推导一下。

RNN 的梯度爆炸和消失问题

不幸的是，实践中前面介绍的几种 RNNs 并不能很好的处理较长的序列。一个主要的原因是，RNN 在训练中很容易发生**梯度爆炸**和**梯度消失**，这导致训练时梯度不能在较长序列中一直传递下去，从而使 RNN 无法捕捉到长距离的影响。

为什么 RNN 会产生梯度爆炸和消失问题呢？我们接下来将详细分析一下原因。我们根据**式 3** 可得：

上式的定义为矩阵的模的上界。因为上式是一个指数函数，如果 t−k 很大的话（也就是向前看很远的时候），会导致对应的**误差项**的值增长或缩小的非常快，这样就会导致相应的**梯度爆炸**和**梯度消失**问题（取决于大于 1 还是小于 1）。

通常来说，**梯度爆炸**更容易处理一些。因为梯度爆炸的时候，我们的程序会收到 NaN 错误。我们也可以设置一个梯度阈值，当梯度超过这个阈值的时候可以直接截取。

梯度消失更难检测，而且也更难处理一些。总的来说，我们有三种方法应对梯度消失问题：

1. 合理的初始化权重值。初始化权重，使每个神经元尽可能不要取极大或极小值，以躲开梯度消失的区域。
2. 使用 relu 代替 sigmoid 和 tanh 作为激活函数。原理请参考上一篇文章零基础入门深度学习(4) - 卷积神经网络的**激活函数**一节。
3. 使用其他结构的 RNNs，比如长短时记忆网络（LTSM）和 Gated Recurrent Unit（GRU），这是最流行的做法。我们将在以后的文章中介绍这两种网络。

RNN 的应用举例——基于 RNN 的语言模型

现在，我们介绍一下基于 RNN 语言模型。我们首先把词依次输入到循环神经网络中，每输入一个词，循环神经网络就输出截止到目前为止，下一个最可能的词。例如，当我们依次输入：

我 昨天 上学 迟到 了

神经网络的输出如下图所示：

其中，s 和 e 是两个特殊的词，分别表示一个序列的开始和结束。

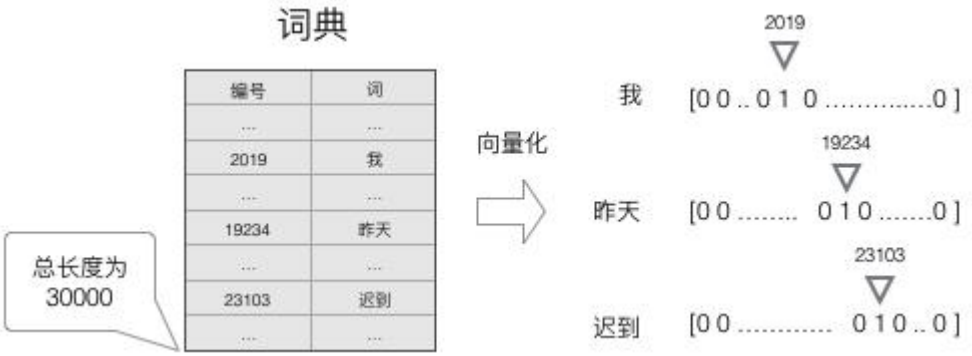
向量化

我们知道，神经网络的输入和输出都是**向量**，为了让语言模型能够被神经网络处理，我们必须把词表达为向量的形式，这样神经网络才能处理它。

神经网络的输入是**词**，我们可以用下面的步骤对输入进行**向量化**：

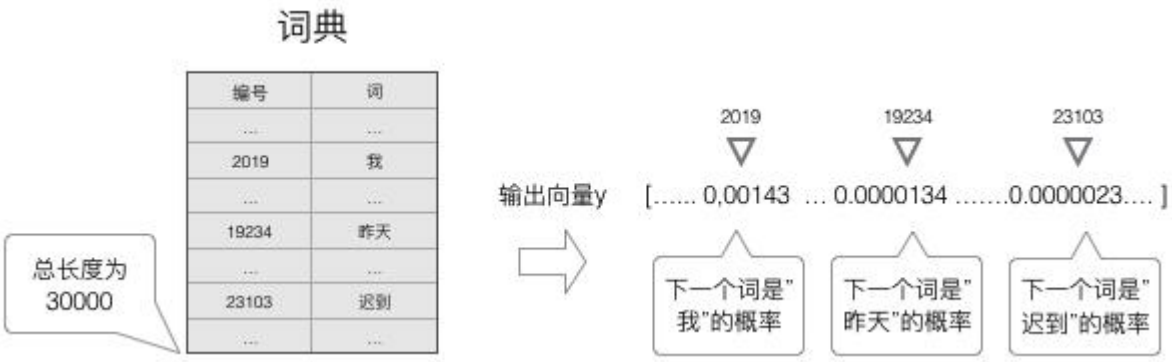
1. 建立一个包含所有词的词典，每个词在词典里面有一个唯一的编号。
2. 任意一个词都可以用一个 N 维的 one-hot 向量来表示。其中，N 是词典中包含的词的个数。假设一个词在词典中的编号是 i，v 是表示这个词的向量，是向量的第 j 个元素，则：

上面这个公式的含义，可以用下面的图来直观的表示：



使用这种向量化方法，我们就得到了一个高维、**稀疏**的向量（稀疏是指绝大部分元素的值都是 0）。处理这样的向量会导致我们的神经网络有很多的参数，带来庞大的计算量。因此，往往会需要使用一些降维方法，将高维的稀疏向量转变为低维的稠密向量。不过这个话题我们就不再这篇文章中讨论了。

语言模型要求的输出是下一个最可能的词，我们可以让循环神经网络计算计算词典中每个词是下一个词的概率，这样，概率最大的词就是下一个最可能的词。因此，神经网络的输出向量也是一个 N 维向量，向量中的每个元素对应着词典中相应的词是下一个词的概率。如下图所示：



Softmax 层

前面提到，**语言模型**是对下一个词出现的**概率**进行建模。那么，怎样让神经网络输出概率呢？方法就是用 softmax 层作为神经网络的输出层。

我们先来看一下 softmax 函数的定义：

这个公式看起来可能很晕，我们举一个例子。Softmax 层如下图所示：

从上图我们可以看到，softmax layer 的输入是一个向量，输出也是一个向量，两个向量的维度是一样的（在这个例子里面是 4）。输入向量 x=[1 2 3 4]经过 softmax 层之后，经过上面的 softmax 函数计算，转变为输出向量 y=[0.03 0.09 0.24 0.64]。计算过程为：

我们来看看输出向量 y 的特征：

1. 每一项为取值为 0–1 之间的正数；
2. 所有项的总和是 1。

我们不难发现，这些特征和**概率**的特征是一样的，因此我们可以把它们看做是概率。对于**语言模型**来说，我们可以认为模型预测下一个词是词典中第一个词的概率是 0.03，是词典中第二个词的概率是 0.09，以此类推。

语言模型的训练

可以使用**监督学习**的方法对语言模型进行训练，首先，需要准备训练数据集。接下来，我们介绍怎样把语料

我 昨天 上学 迟到 了

转换成语言模型的训练数据集。

首先，我们获取**输入–标签**对：

输入	标签
s	我
我	昨天
昨天	上学
上学	迟到
迟到	了
了	e

然后，使用前面介绍过的**向量化**方法，对输入 x 和标签 y 进行**向量化**。这里面有意思的是，对标签 y 进行向量化，其结果也是一个 one–hot 向量。例如，我们对标签『我』进行向量化，得到的向量中，只有第 2019 个元素的值是 1，其他位置的元素的值都是 0。它的含义就是下一个词是『我』的概率是 1，是其它词的概率都是 0。

最后，我们使用**交叉熵误差函数**作为优化目标，对模型进行优化。

在实际工程中，我们可以使用大量的语料来对模型进行训练，获取训练数据和训练的方法都是相同的。

交叉熵误差

一般来说，当神经网络的输出层是 softmax 层时，对应的误差函数 E 通常选择交叉熵误差函数，其定义如下：

在上式中，N 是训练样本的个数，向量是样本的标记，向量是网络的输出。标记是一个 one-hot 向量，例如，如果网络的输出，那么，交叉熵误差是（假设只有一个训练样本，即 N=1）：

我们当然可以选择其他函数作为我们的误差函数，比如最小平方误差函数(MSE)。不过对概率进行建模时，选择交叉熵误差函数更 make sense。具体原因，感兴趣的读者请阅读参考文献 7。

RNN 的实现

> 完整代码请参考 GitHub: https://github.com/hanbt/learn_dl/blob/master/rnn.py (python2.7)

为了加深我们对前面介绍的知识的理解，我们来动手实现一个 RNN 层。我们复用了上一篇文章零基础入门深度学习(4) - 卷积神经网络中的一些代码，所以先把它们导入进来。

```
1. import numpy as np
2. from cnn import ReluActivator, IdentityActivator, element_wise_op
```

我们用 RecurrentLayer 类来实现一个**循环层**。下面的代码是初始化一个循环层，可以在构造函数中设置卷积层的超参数。我们注意到，循环层有两个权重数组，U 和 W。

```
1. class RecurrentLayer(object):
2.     def __init__(self, input_width, state_width,
3.                  activator, learning_rate):
4.
5.         self.input_width = input_width
6.         self.state_width = state_width
7.         self.activator = activator
8.         self.learning_rate = learning_rate
9.         self.times = 0       # 当前时刻初始化为 t0
10.        self.state_list = [] # 保存各个时刻的 state
```

```
        self.state_list.append(np.zeros(
            (state_width, 1)))          # 初始化 s0
        self.U = np.random.uniform(-1e-4, 1e-4,
            (state_width, input_width)) # 初始化 U
        self.W = np.random.uniform(-1e-4, 1e-4,
            (state_width, state_width)) # 初始化 W
```

在 forward 方法中，实现循环层的前向计算，这部分比较简单。

```
    def forward(self, input_array):
        '''
        根据『式 2』进行前向计算
        '''
        self.times += 1
        state = (np.dot(self.U, input_array) +
                 np.dot(self.W, self.state_list[-1]))
        element_wise_op(state, self.activator.forward)
        self.state_list.append(state)
```

在 backword 方法中，实现 BPTT 算法。

```
    def backward(self, sensitivity_array,
                 activator):
        '''
        实现 BPTT 算法
        '''
        self.calc_delta(sensitivity_array, activator)
        self.calc_gradient()

    def calc_delta(self, sensitivity_array, activator):
        self.delta_list = [] # 用来保存各个时刻的误差项
        for i in range(self.times):
            self.delta_list.append(np.zeros(
                (self.state_width, 1)))
        self.delta_list.append(sensitivity_array)
```

```
        # 迭代计算每个时刻的误差项
        for k in range(self.times - 1, 0, -1):
            self.calc_delta_k(k, activator)

    def calc_delta_k(self, k, activator):
        '''
        根据 k+1 时刻的 delta 计算 k 时刻的 delta
        '''
        state = self.state_list[k+1].copy()
        element_wise_op(self.state_list[k+1],
                        activator.backward)
        self.delta_list[k] = np.dot(
            np.dot(self.delta_list[k+1].T, self.W),
            np.diag(state[:,0])).T

    def calc_gradient(self):
        self.gradient_list = [] # 保存各个时刻的权重梯度
        for t in range(self.times + 1):
            self.gradient_list.append(np.zeros(
                (self.state_width, self.state_width)))
        for t in range(self.times, 0, -1):
            self.calc_gradient_t(t)
        # 实际的梯度是各个时刻梯度之和
        self.gradient = reduce(
            lambda a, b: a + b, self.gradient_list,
            self.gradient_list[0]) # [0]被初始化为 0 且没有被修改过

    def calc_gradient_t(self, t):
        '''
        计算每个时刻 t 权重的梯度
        '''
        gradient = np.dot(self.delta_list[t],
            self.state_list[t-1].T)
        self.gradient_list[t] = gradient
```

有意思的是，BPTT 算法虽然数学推导的过程很麻烦，但是写成代码却并不复杂。

在 update 方法中，实现梯度下降算法。

```
def update(self):
    '''
    按照梯度下降，更新权重
    '''
    self.W -= self.learning_rate * self.gradient
```

上面的代码不包含权重 U 的更新。这部分实际上和全连接神经网络是一样的，留给感兴趣的读者自己来完成吧。

循环层是一个**带状态**的层，每次 forword 都会改变循环层的内部状态，这给梯度检查带来了麻烦。因此，我们需要一个 reset_state 方法，来重置循环层的内部状态。

```
def reset_state(self):
    self.times = 0       # 当前时刻初始化为 t0
    self.state_list = [] # 保存各个时刻的 state
    self.state_list.append(np.zeros(
        (self.state_width, 1)))     # 初始化 s0
```

最后，是梯度检查的代码。

```
def gradient_check():
    '''
    梯度检查
    '''
    # 设计一个误差函数，取所有节点输出项之和
    error_function = lambda o: o.sum()

    rl = RecurrentLayer(3, 2, IdentityActivator(), 1e-3)

    # 计算 forward 值
```

```
    x, d = data_set()
    rl.forward(x[0])
    rl.forward(x[1])

    # 求取 sensitivity map
    sensitivity_array = np.ones(rl.state_list[-1].shape,
                                dtype=np.float64)
    # 计算梯度
    rl.backward(sensitivity_array, IdentityActivator())

    # 检查梯度
    epsilon = 10e-4
    for i in range(rl.W.shape[0]):
        for j in range(rl.W.shape[1]):
            rl.W[i,j] += epsilon
            rl.reset_state()
            rl.forward(x[0])
            rl.forward(x[1])
            err1 = error_function(rl.state_list[-1])
            rl.W[i,j] -= 2*epsilon
            rl.reset_state()
            rl.forward(x[0])
            rl.forward(x[1])
            err2 = error_function(rl.state_list[-1])
            expect_grad = (err1 - err2) / (2 * epsilon)
            rl.W[i,j] += epsilon
            print 'weights(%d,%d): expected - actural %f - %f' % (
                i, j, expect_grad, rl.gradient[i,j])
```

需要注意，每次计算 error 之前，都要调用 reset_state 方法重置循环层的内部状态。下面是梯度检查的结果，没问题！

小节

至此，我们讲完了基本的**循环神经网络**、它的训练算法：BPTT，以及在语言模型上的应用。RNN 比较烧脑，相信拿下前几篇文章的读者们搞定这篇文章也不在话下吧！然而，**循环神经网络**这个话题并没有完结。我们在前面说到过，基本的循环神经网络存在梯度爆炸和梯度消失问题，并不能真正的处理好长距离的依赖（虽然有一些技巧可以减轻这些问题）。事实上，真正得到广泛的应用的是循环神经网络的一个变体：**长短时记忆网络**。它内部有一些特殊的结构，可以很好的处理长距离的依赖，我们将在下一篇文章中详细的介绍它。现在，让我们稍事休息，准备挑战更为烧脑的**长短时记忆网络**吧。

参考资料

1. RECURRENT NEURAL NETWORKS TUTORIAL

2. Understanding LSTM Networks
3. The Unreasonable Effectiveness of Recurrent Neural Networks
4. Attention and Augmented Recurrent Neural Networks
5. On the difficulty of training recurrent neural networks, Bengio et al.
6. Recurrent neural network based language model, Mikolov et al.
7. Neural Network Classification, Categorical Data, Softmax Activation, and Cross Entropy Error, McCaffrey

内容目录

- 机器学习 7
  - 零基础入门深度学习(7) - 递归神经网络
  - 零基础入门深度学习(6) - 长短时记忆网络(LSTM)
  - 零基础入门深度学习(5) - 循环神经网络
  - 零基础入门深度学习(4) - 卷积神经网络
  - 零基础入门深度学习(3) - 神经网络和反向传播算法
  - 零基础入门深度学习(2) - 线性单元和梯度下降
  - 零基础入门深度学习(1) - 感知器
- 深度学习入门 7
  - 零基础入门深度学习(7) - 递归神经网络
  - 零基础入门深度学习(6) - 长短时记忆网络(LSTM)
  - 零基础入门深度学习(5) - 循环神经网络
  - 零基础入门深度学习(4) - 卷积神经网络
  - 零基础入门深度学习(3) - 神经网络和反向传播算法
  - 零基础入门深度学习(2) - 线性单元和梯度下降
  - 零基础入门深度学习(1) - 感知器

-
-
-
    - 下载客户端
    - 关注开发者
    - 报告问题，建议
    - 联系我们
    -

添加新批注

保存 取消

在作者公开此批注前，只有你和作者可见。

保存 取消

修改 保存 取消 删除

- 私有
- 公开
- 删除

查看更早的 5 条回复

回复批注

×

通知

取消 确认

-
-