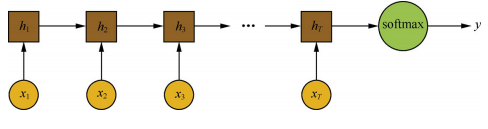
# 循环神经网络

**问题1 简述RNN和前馈神经网络(例如CNN)在处理文本数据时的过程和特点**

**答案** 首先，在传统的方法中，一般用TF-IDF来提取文本的特征向量，但是这显然丢失了位置信息。CNN在处理文本时通常将文本处理成等长，然后用卷积核池化的方式来提取特征向量，这样可以捕捉到文本中的局部信息，但是单词之间的长距离依赖还是很难学习到。循环神经网络之所以能很好的处理序列信息，因为在建模的过程中，文本数据是在不同的时间片依次输入到模型，从而使得历史信息通过隐层编码到当前的状态中，这样RNN就拥有了记忆功能（同样带来训练耗时的问题）。

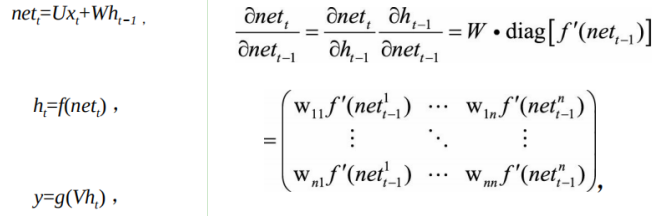


PS:激活函数一般使用tanh或者relu

**问题2 简述梯度消失和梯度爆炸，以及缓解办法。**

**1、**神经网络和梯度下降算法在上世纪就已经被提出，当时的激活函数一般使用sigmoid函数，但是这个函数具有饱和特性，当参数进入到激活函数的饱和区以后，梯度接近0，这使得参数难以继续更新-即梯度消失。

2、在RNN领域，一般采用BPTT算法反向传播，本质上是将T个时间步的模型等价的看做T层前馈神经网络，这样就和普通的反向传播算法一致。但是实践表明采用BPTT算法的RNN并不能成功的捕获长距离依赖问题，而这主要是由神经网络的梯度消失造成。分析如下：



其中的diag表示对角矩阵，由于f(net)是向量，与矩阵W相乘的过程中，可以等价的处理为对角矩阵，右图的计算结果称为雅克比矩阵，n为隐层单元数目。在反向传播的时候，如果雅克比矩阵的特征值大于1，则梯度值成指数增加导致梯度爆炸，反之导致梯度消失。

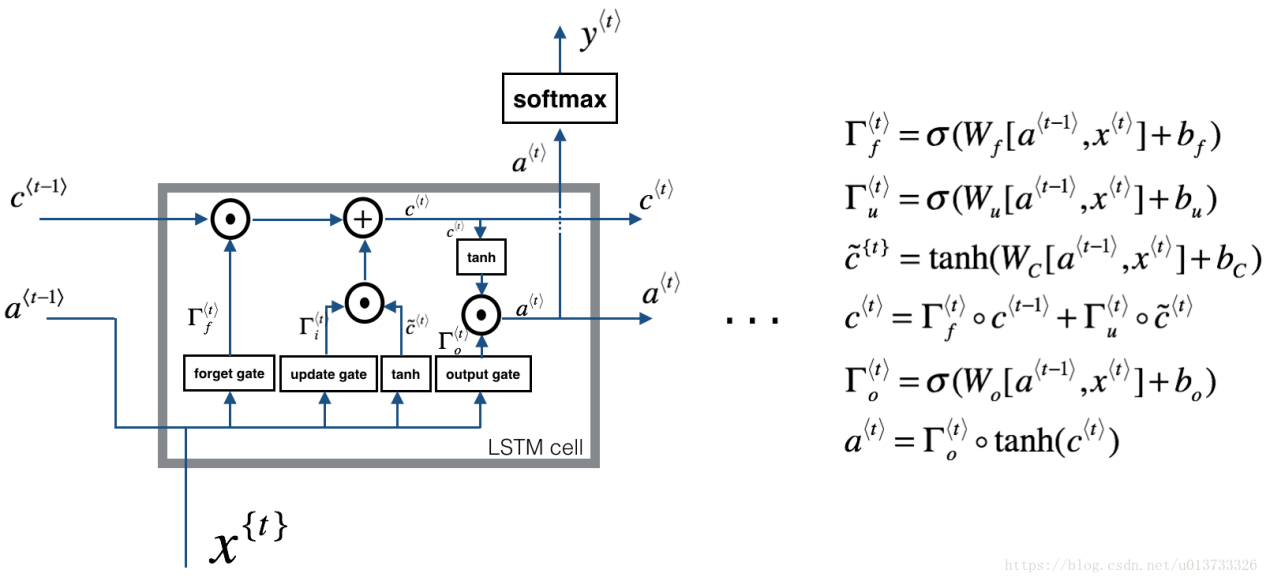
1. 梯度爆炸可以用梯度修剪缓解，梯度消失可以行三个方面考虑：初始化靠近单位矩阵、改进模型（LSTM、GRU）、残差网络

**问题3 在一般的前馈神经网络中，RELU可以有效的改善梯度消失问题，在RNN中可以使用RELU作为激活函数吗？**

1. 首先RELU之所以可以缓解梯度消失问题在于函数在x>0是没有饱和区，另外由于CNN的参数相互独立，可以相互抵消所以多层一般不会出现严重的数值问题，反观RNN，由于权重参数实际上是同一个矩阵在不同的时间步，所以多层会存在数值问题。
2. RNN梯度消失不仅受激活函数影响，还在于RNN参数的连乘。
3. 综上所述，只有将W初始化为单位矩阵时，RELU才能取得不错的效果。

**问题4 LSTM如何实现长短期记忆功能**

首先，LSTM的结构如下图所示：



其实LSTM的思路很清晰：2条记忆线、一个输入、一个输出。

**历史信号**：进来后，与遗忘门相乘决定多大程度采用历史信息。

**记忆单元**：用双曲激活当前输入后，和更新门相乘再和记忆相加→Ct。

**输出信号**：将选择的历史信息和当前记忆单元求和，然后和输出门相乘决定最后 的输出信号。

所有**门的数值**由历史信息和当前输入决定。

1. 不同的RNN之所以不能捕获长期记忆，是因为W连乘问题导致梯度消失。

2、LSTM和普通RNN的相比之所以能缓解梯度问题，本质上讲在于信号在不同的时间步之间传播时，其数值不在取决于一个单一的矩阵，而是将数值问题的压力分散到更多的矩阵，抽象一层来讲就是由于门控机制的引入，使得理论上而言当遗忘门数值接近1的时候可以顺利的将历史信息引入并传递下去。

PS：关于激活函数的选取：门控值采用sigmoid函数，候选记忆采用双曲正切。原因解释：为了符合门控的物理定义有必要采取“开关函数”，而选用tanh在于大多场景下特征分布以0为中心，而tanh在0附近有较大的梯度，有利于参数的快速收敛

Seq2seq

在解码时，输出由隐状态和前一词决定。

