注：1）这篇文章主要参考Alex Graves的博士论文《Supervised Sequence Labelling with Recurrent Neural Networks》 详细介绍其关于BP算法的推导过程。

        2）论文没有提到过偏差的处理，不过如果能耐心推导出论文给出的公式，那么这将十分简单。

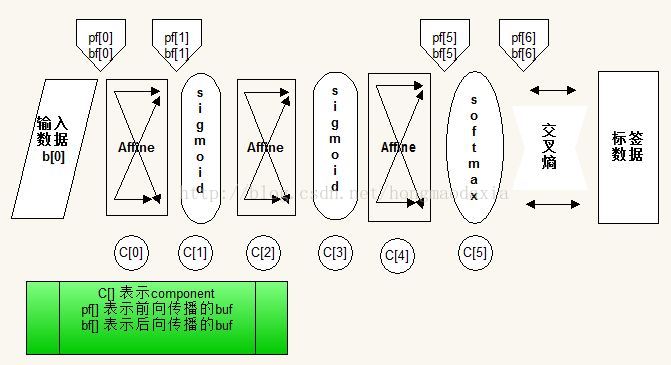
        3）由于是结合语音训练的，最后是softmax出来是获得有限结果的概率值，然后做交叉熵作为目标函数，所以可能与其他的网络不一致，但是反馈的推导应该是一样的，除了输出层和最后一个隐层之间的个推导。

  转载请保留出处 <http://write.blog.csdn.net/postedit/41809341>

**1.MLP 多层感知器**

        我觉的在看下面公式之前多少也要对BP有些了解吧，最起要明白，建立网络的目的是要拟合一个非线性函数，最终效果是整个网络就是一个函数。函数的参数便是里面的权重，所以主要目的就是如何训练得到这些参数。一般都是建立一个目标函数O，然后优化它，因为它和权重w有关，所以利用梯度下降法，来更新权重，所以文中最终的目的就是是在设法求dO/dw ……这些基本知识我还是不说了。

下面这张图是**我结合kaldi中dnn的代码画的图**不知对接下来讲的有没有帮助，我觉的kaldi的代码和论文这里的一抹一样，只是代码中把权重和层看作不同的component。



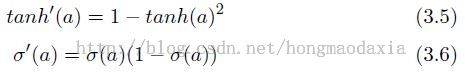
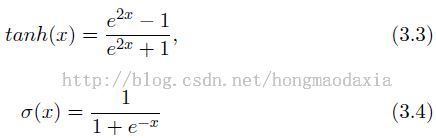
**1.1前向传播**

       普通的BP前向传播很简单，如果最后一层不是softmax层，那么就是 输入 ×  权重 然后通过一个激活函数（eg. sigmoid），加上偏差作为输出。为了简便以下直接通过截图，然后详细的介绍每一步的作用。



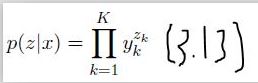
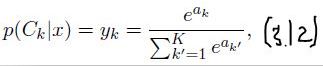
        其中I 代表输出单元个数，输入的数据是向量x，下表带有h的表示在隐层。左面两个式子就是求第一个隐层其中一个unit（单元）的输出。  3.1式，可以理解为一个过渡式子，但是这个式子很重要，**后面推导的时候就是用目标函数对a求导作为基本结构（我猜想这也是考虑到 后面对权重w求导方便）**。右面两个式子，是其余隐层单元的输出（因为这里要讲的是深度网络，所以有多个隐层）

        那个θ是一个非线性函数（激活函数），一般有下面两种选择：



      指所以选择非线性函数，因为其可以拟合线性的函数，发现非线性的分类边界（nonlinear classi cation boundaries），同时线性函数的组合还是线性的，而非线性则可以更好地”记忆“输入数据的特征。（算了这样下去就成翻译了，我还是注重推导吧，反正大家的ANN基本知识比我好）

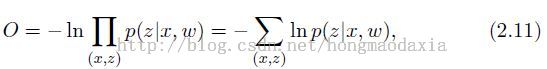
     前向传播到最后，当然要把输出数据和原来的数据进行比对然后更新权重（这里讲的是有监督学习）。对于语音识别网络结构CD-DNN-HMM熟悉的话，知道**网络最后的输出是关于每个phoneme或triphone的概率**，而概率的得来是通过最后一个softmax层求得并输出的（此层非常特殊，不像原来每个节点单独求输出b，而是等此层所有节点的a都求出后，一起求b【概率】）。式子如下

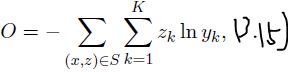
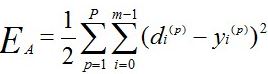


     左面那个就是标准的softmax函数，求Ck类的概率值。 右面那个函数**z是标签(或者说本应该的正确结果，但是完全由一个1其余为0组成比如[0,1,0,0,0]  ，1，0其实代表的是概率值，它们求和为1，可以理解为总共有5个分类结果，输入应该对应第二个分类)**

**1.2目标函数O**

结合2.11式（在下面，就是求交叉熵的）与3.13式可得3.15式——我们需要最小化的目标函数（交叉熵越小，说明模型越逼近真实的结果）



         目标函数以前用的是均方误差    

         最小化目标函数O就需要用梯度下降法【应该是随机梯度下降，因为基本上是训练一个数据（或mini-batch），更新一下权重，而不是把所有的都训练完再更新】

        关于梯度下降法，要理解清楚，要不然就不知道下面为什么要那么干，可参考博客 <http://www.cnblogs.com/iamccme/archive/2013/05/14/3078418.html> 。 简单说一下，如果要更新权重w，就要知道更新量,可以选取梯度最大的值(传说中的导数)，如果+则是最大化目标函数,如果减则是最小化目标函数.为了防止过快/慢则乘以一个系数.

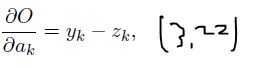
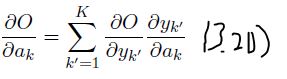
**1.3 后向传播**

这才是真正难的地方.大体说一下,由于最后一次层特殊是个softmax层,所以单独证明一下,后面的则完全一样了,通过找到递归的算法.使得公式十分简洁.

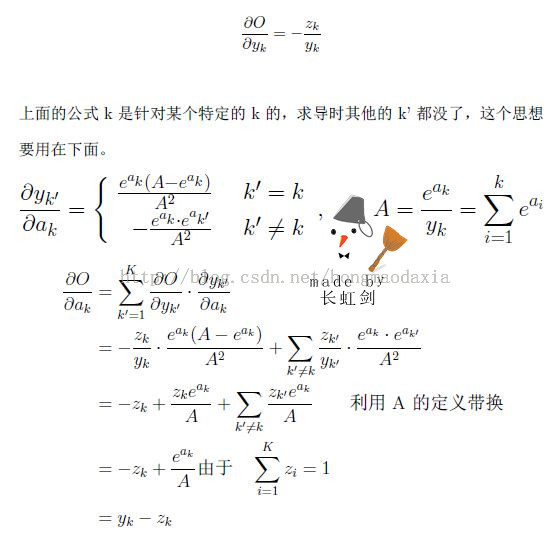
**再次强调,ah是指下面的 输入X 权重 求和的值,还没有通过节点,而 bh 是指ah 通过节点的激活函数θ的值,那么这个bh乘以权重whk与其他求和,就又得到上一层的ak,而论文的作者是以求dO/da 为核心，然后推导出递归公式。**

**先看softmax这层的推导**

根据3.15式和3.13式，不难想到3，20式 【牢记每个 单元/节点 的softmax值依赖于这层所有的a】

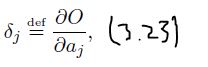


对于3.22式，有兴趣的同学可以自己推，下面是我的推导过程，截图如下，不知是否准确

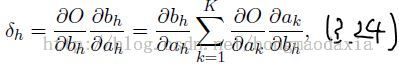
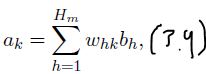


其他层的递归求法推导

之前我强调了很久的地方，是因为我在推导哦啊RNN和LSTM的时候，充分发现其巧妙之处，作者在接下来也直接给出了公式 3.23

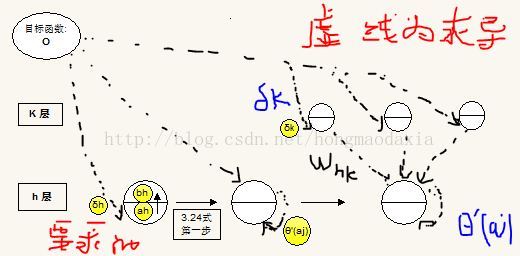


δk 在3.22式就已经得到了，接下来，论文通过简单的求导变换就推出出的递归公式，下面慢慢讲，首先是倒数第二个隐层

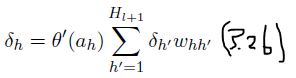
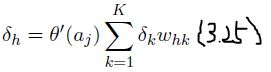
      类似3.1有 

**解释**：上面这个式子，在不知不觉中牵扯到两个隐层 ，ah产生bh，bh产生ak，**记住a只是通过矩阵乘法还没通过节点，而b就是a通过后的值。**

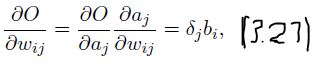
     给个我画的图片吧,



     通过3.24式与3.2，3.9式结合，便可得3.25式，这样推广一下除了最后的softmxt层其他层都可以按3.36这样递归来算

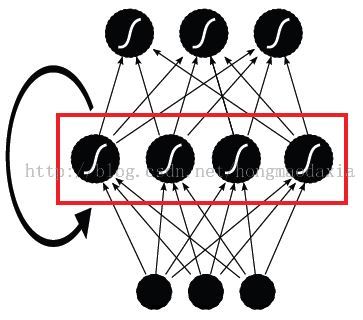


   虽然找到了递归关系，但是有朋友要问了，不是应该求权重w的倒数吗？怎么一直没看到？我想说的是如果一开始直接利用求w的倒数来推导，那么上面的公式将会复杂，还有论文后面的LSTM列出的公式（论文同样没给具体的推导，需要自己来推推）也利用dO/da 总结出公式……不说了，下面就看看这样之后求dw是多么简洁吧！



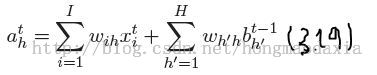
**2.RNN Recurrent Neural Networks(循环的神经网络)**

网络的结构就是每一个隐层,拥有上一时刻这个隐层输出的输入值.,原文示意图如下

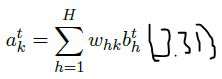


**2.1前向传播**

   直接看公式吧,很简单,注意上标是指时间t

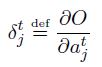
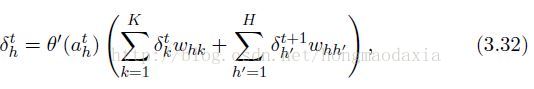
C:\Users\prm14\AppData\Local\YNote\data\353345737@qq.com\d7a129fba1804d0fb90977b3cc1318a1\center.jpeg

当然输出层的a则和以前的一样



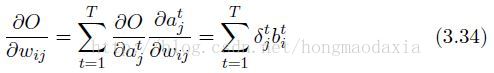
**2.2 后向传播**

先给公式

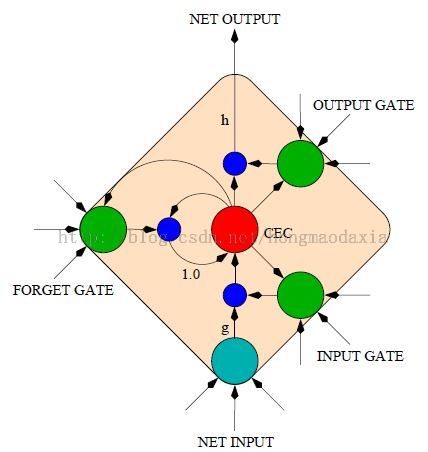


    其实也没什么难的,但是有个地方要仔细揣摩一下,仔细观察 公式3.23 用到了t+1,如果求导思路清晰则可以推,出来.但是也可以这样理解:前向传播时h层的输出bh影响了h层t+1时刻,所以按照多元函数求导的法则,自然要加上对t+1时刻的影响,或者干脆看作上面不仅有k层还有 h(t+1) 层,这样就好理解为什么有 δ(t+1)了吧,【**明白这个很重要，要不然LSTM给出的公式将更难理解**】

求对w的倒数和往常一样的简单



**3 LSTM Long Short-Term Memory （长短时间记忆）**

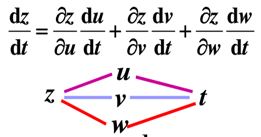


到这里，朋友们可能要失望了，因为接下来我也不贴结构图或者是公式。在这里给想推导出LSTM公式的同学一些经验。

1）先把结构图看明白，要对照论文给出的各个定义，结合Forward Pass ，注意图中C=1，其实这一个memory block 应该有多个输出。前万不要先看 Backward Pass

2)要时刻警惕脚标t ,t+1 ,t-1

3）开始推导后向算法的朋友，也不要害怕，其实就是下图这么回事



**推导过程中z到t可能间隔几层，而非上图中的一层，一定要明白谁是最后的求导对象。一般来说除了像3.24那种需要变换的，其余只要根据图中箭头的反像就可以慢慢链式的求出来**。 当然这是我的理解不一定对，大家数学好的可能就会说，这还用求吗？显然是这样，就像1+1=2一样的显然。

    这一块内容我看了很久，这三个推导看了大约5天，很多笔记在打印纸上，这里只能摘一点防止自己忘记，也为入门的朋友提供一些小的帮助。