**问题0**

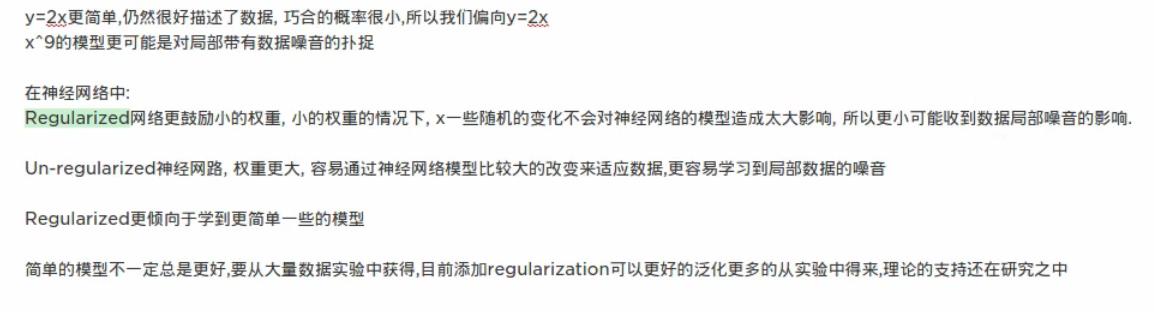
**平时调参的经验  
写出常见的正则化和防止过拟合的技术**

解决方法：

1）重新清洗数据，导致过拟合的一个原因也有可能是数据不纯导致的，如果出现了过拟合就需要我们重新清洗数据。

2）增[大数据](http://lib.csdn.net/base/hadoop" \o "Hadoop知识库" \t "http://blog.csdn.net/willduan1/article/details/_blank)的训练量，还有一个原因就是我们用于训练的数据量太小导致的，训练数据占总数据的比例过小。

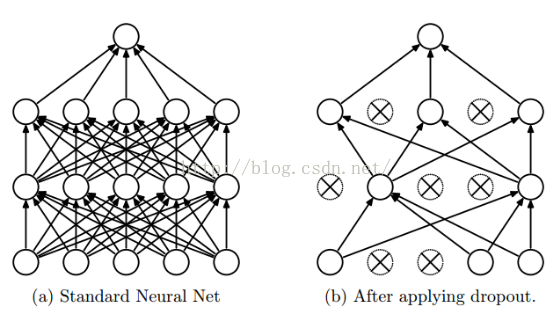
3）采用正则化方法。正则化方法包括L0正则、L1正则和L2正则，而正则一般是在目标函数之后加上对于的范数。但是在机器学习中一般使用L2正则，下面看具体的原因。



L0范数是指向量中非0的元素的个数。L1范数是指向量中各个元素绝对值之和，也叫“稀疏规则算子”（Lasso regularization）。两者都可以实现稀疏性，既然L0可以实现稀疏，为什么不用L0，而要用L1呢？个人理解一是因为L0范数很难优化求解（NP难问题），二是L1范数是L0范数的最优凸近似，而且它比L0范数要容易优化求解。所以大家才把目光和万千宠爱转于L1范数。

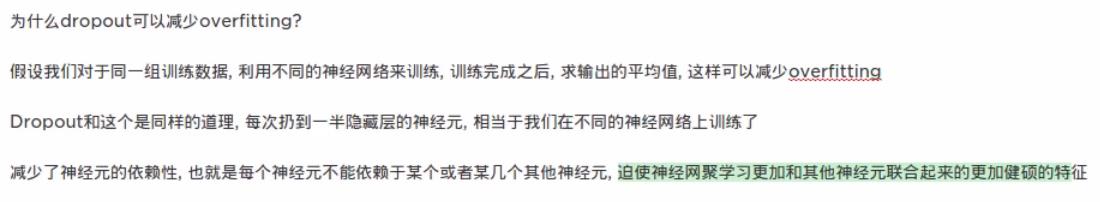
L2范数是指向量各元素的平方和然后求平方根。可以使得W的每个元素都很小，都接近于0，但与L1范数不同，它不会让它等于0，而是接近于0。L2正则项起到使得参数w变小加剧的效果，但是为什么可以防止过拟合呢？一个通俗的理解便是：更小的参数值w意味着模型的复杂度更低，对训练数据的拟合刚刚好（奥卡姆剃刀），不会过分拟合训练数据，从而使得不会过拟合，以提高模型的泛化能力。还有就是看到有人说L2范数有助于处理 condition number不好的情况下矩阵求逆很困难的问题。

4）采用dropout方法。这个方法在神经网络里面很常用。dropout方法是ImageNet中提出的一种方法，通俗一点讲就是dropout方法在训练的时候让神经元以一定的概率不工作。具体看下图：

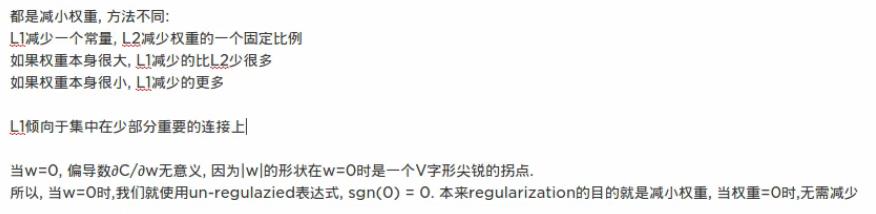


如上图所示，左边a图是没用用dropout方法的标准神经网络，右边b图是在训练过程中使用了dropout方法的神经网络，即在训练时候以一定的概率p来跳过一定的神经元。

**dropout为什么解决overfitting**



**为什么L1 regularization可以使参数优化到0**

****

**问题1****分析梯度弥散和梯度爆炸的原因，及其解决手段**

梯度消失：许多激活函数将输出值挤压在很小的区间内，在激活函数两端较大范围的定义域内梯度为0，导致权重更新的缓慢训练难度增加，造成学习停止。（前面层上的梯度是来自后面的层上项的乘积，当层数过多时，随着乘积的累积，将越来越小。）

梯度爆炸：前面层比后面层变化更快时，引起梯度爆炸。

解决手段：减少层数；增大学习率；用Relu代替sigmoid。

梯度爆炸的预防方案：L1或者L2正则化；

梯度爆炸的解决方案：梯度截断；

梯度消失的预防方案：LSTM结构；参数初始化；选用ReLU激活函数

**问题3****模型欠拟合的解决方法**

首先欠拟合就是模型没有很好地捕捉到数据特征，不能够很好地拟合数据。

1）添加其他特征项，有时候我们模型出现欠拟合的时候是因为特征项不够导致的，可以添加其他特征项来很好地解决。例如，“组合”、“泛化”、“相关性”三类特征是特征添加的重要手段，无论在什么场景，都可以照葫芦画瓢，总会得到意想不到的效果。除上面的特征之外，“上下文特征”、“平台特征”等等，都可以作为特征添加的首选项。

2）添加多项式特征，这个在机器学习算法里面用的很普遍，例如将线性模型通过添加二次项或者三次项使模型泛化能力更强。例如上面的图片的例子。

3）减少正则化参数，正则化的目的是用来防止过拟合的，但是现在模型出现了欠拟合，则需要减少正则化参数。

**问题2   
权重初始化（Weight Initialization）**

1）小的随机数

w= 0.01 \* np.random.randn(fan\_in,fan\_out)

2）神经元将饱和，梯度为0

w = 1.0 \* np.random.randn(fan\_in,fan\_out)

3）合理的初始化（Xavier init）

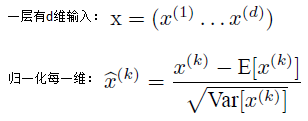
w = np.random.randn((fan\_in,fan\_out)/np.sqrt(fan\_in)

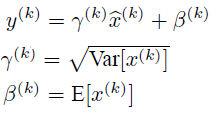
**batch normalizatin为什么可以防止梯度消失爆炸**

为了便于训练，我们经常归一化参数的初始值，通过mean=0, variance=1的高斯分布来初始化参数。在训练过程中，我们不同程度地更新参数，使用参数失去了归一化，这将降低训练速度且放大变化，网络越深问题越严重。

BN的本质原理：在网络的每一层输入的时候，又插入了一个归一化层，也就是先做一个归一化处理（归一化至：均值0、方差为1），然后再进入网络的下一层。BN是一个可学习、有参数（γ、β）的网络层。

归一化公式：



如果是仅仅使用上面的归一化公式，对网络某一层A的输出数据做归一化，然后送入网络下一层B，这样是会影响到本层网络A所学习到的特征的。比如我网络中间某一层学习到特征数据本身就分布在S型激活函数的两侧，你强制把它给我归一化处理、标准差也限制在了1，把数据变换成分布于s函数的中间部分，这样就相当于我这一层网络所学习到的特征分布被你搞坏了，这可怎么办？于是文献使出了一招惊天地泣鬼神的招式：变换重构，引入了可学习参数γ、β，这就是算法关键之处： 

上面的公式表明，通过学习到的重构参数γ、β，是可以恢复出原始的某一层所学到的特征的。

在使用BN前，减小学习率、小心的权重初始化的目的是：使其输出的数据分布不要发生太大的变化。

**batch normalizatin在CNN中卷积层怎么做**

BN层是对于每个神经元做归一化处理，甚至只需要对某一个神经元进行归一化，而不是对一整层网络的神经元进行归一化。既然BN是对单个神经元的运算，那么在CNN中卷积层上要怎么搞？假如某一层卷积层有6个特征图，每个特征图的大小是100\*100，这样就相当于这一层网络有6\*100\*100个神经元，如果采用BN，就会有6\*100\*100个参数γ、β，这样岂不是太恐怖了。因此卷积层上的BN使用，其实也是使用了类似权值共享的策略，**把一整张特征图当做一个神经元进行处理**。

卷积神经网络经过卷积后得到的是一系列的特征图，如果min-batch sizes为m，那么网络某一层输入数据可以表示为四维矩阵(m,f,w,h)，m为min-batch sizes，f为特征图个数，w、h分别为特征图的宽高。在CNN中我们可以把每个特征图看成是一个特征处理（一个神经元），因此在使用Batch Normalization，mini-batch size 的大小就是：m\*w\*h，于是对于每个特征图都只有一对可学习参数：γ、β。说白了吧，这就是相当于求取所有样本所对应的一个特征图的所有神经元的平均值、方差，然后对这个特征图神经元做归一化。

**batch normalizatin的作用**

1. 改善流经网络的梯度

2）允许更大的学习率，大幅提高训练速度：

你可以选择比较大的初始学习率，让你的训练速度飙涨。以前还需要慢慢调整学习率，甚至在网络训练到一半的时候，还需要想着学习率进一步调小的比例选择多少比较合适，现在我们可以采用初始很大的学习率，然后学习率的衰减速度也很大，因为这个算法收敛很快。当然这个算法即使你选择了较小的学习率，也比以前的收敛速度快，因为它具有快速训练收敛的特性；

3）减少对初始化的强烈依赖

4）改善正则化策略：作为正则化的一种形式，轻微减少了对dropout的需求

你再也不用去理会过拟合中dropout、L2正则项参数的选择问题，采用BN算法后，你可以移除这两项了参数，或者可以选择更小的L2正则约束参数了，因为BN具有提高网络泛化能力的特性；

5）再也不需要使用使用局部响应归一化层了（局部响应归一化是Alexnet网络用到的方法，搞视觉的估计比较熟悉），因为BN本身就是一个归一化网络层；

6）可以把训练数据彻底打乱（防止每批训练的时候，某一个样本都经常被挑选到，文献说这个可以提高1%的精度）。

注：以上为学习过程，在测试时，均值和方差(mean/std)不基于小批量进行计算， 可取训练过程中的激活值的均值。

**问题4   
写出三个损失函数，并分析特点和使用场景**

**问题5   
VGGNet,GoogleNet，Resnet的特点和对比**  
**问题7   
[深度学习](http://lib.csdn.net/base/deeplearning" \o "深度学习知识库" \t "http://blog.csdn.net/victoriaw/article/details/_blank)是否是万能的，深度学习未来发展的方向，你的观点**

**问题8   
CNN做卷积的运算时间复杂度**

****

## 0. 滤波器的大小选择

* 大部分卷积神经网络都会采用**逐层递增**（1⇒ 3 ⇒ 5 ⇒ 7）的方式。
* 每经过一次池化层，卷积层过滤器的深度都会乘以 2；

## 1. 权值共享：减轻过拟合 & 降低计算量

一个卷积层（Wx+b ⇒ ReLU ⇒ maxpooling）可以有多个不同的卷积核，而每一个卷积核都对应一个滤波后映射出的新图像，同一个新图像中的每一个像素都来自完全相同的卷积核，这就是卷积核的权值共享。

那么为什么要共享卷积核的权值参数呢？

* 降低模型复杂度以减轻过拟合；
* 降低计算量；

## 2. 待求参数数目的量化分析

考虑 103×103 的输入图像：

* 全连接，隐层神经元的数目为 106 时，则每一个输入像素与每一个隐层神经元之间都是待学习的参数，
  + 数目为 106×106=1012
* 卷积，卷积核的大小为 10×10 时，
  + 步长为 10，103×10310×10×(10×10)，103×10310×10表示的是输入图像可划分的块数，也即经卷机作用后的输出图像；
  + 步长为 1，(103−10+1)×(103−10+1)⋅(10×10)
  + 在不考虑步长的前提下，可近似将待学习的参数的数目视为 (103×103)⋅(10×10)

## 3. CNN 的卷积与信号与系统中的卷积

**CNN 的卷积并没有执行“翻转”操作，而是与输入图像做滑动窗口“相关”计算；**

如果 K 个输入通道（Xk,0≤k<K）的输入经若干卷积核的作用之后得到 L 个通道的输出（Yℓ,0≤ℓ<L），需要的卷积核的数目为 L×K：

Yℓ(m,n)==Xk(m,n)⋆Hkℓ(m,n)∑k=0K−1∑i=0I−1∑j=0J−1Xk(m+i,n+j)Hkℓ(i,j)

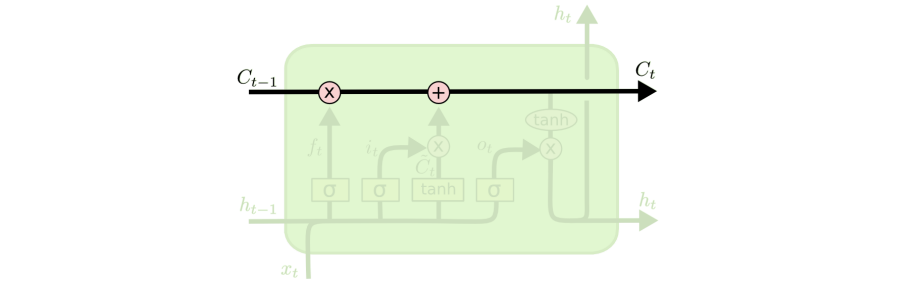
Hkℓ(i,j) 表示的是第 ℓ 列，第 k 行二维卷积核，卷积核的大小为 I⋅J。

**问题9****LSTM结构推导，为什么比RNN好？**

****

****

LSTM的关键就是state,就是对应上面的****主线数据****的传递

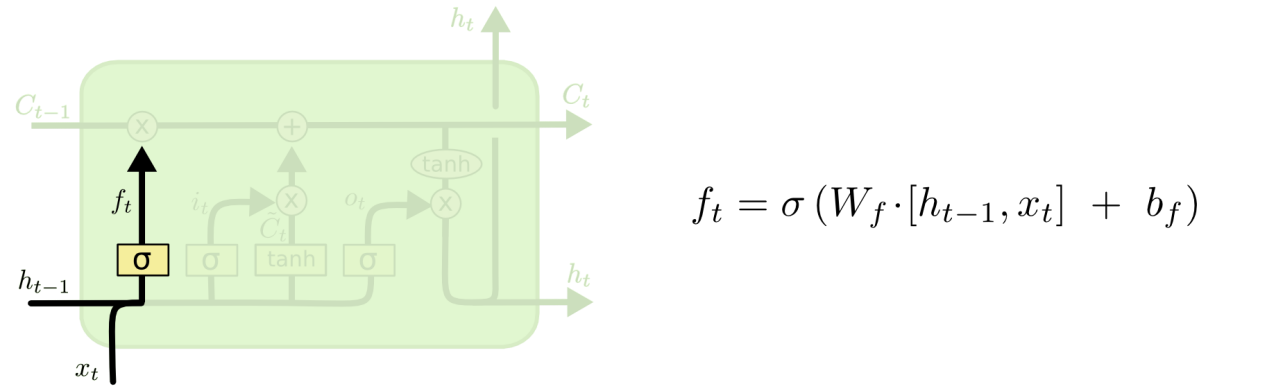


**前向计算step by step**

(1) 决定抛弃的信息

遗忘门 (forget gate layer)

σ是Sigmoid激励函数，因为它的值域是(0,1)，0代表遗忘所有信息，1代表保留所有信息

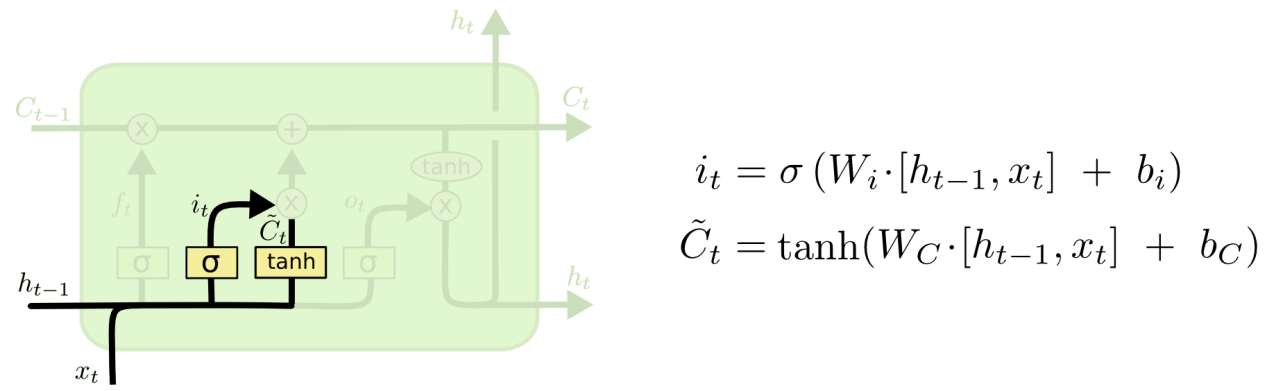


(2) 决定存储的新信息

包括两个部分

第一个是输入门 (input gate layer)，对应的是Sigmoid函数

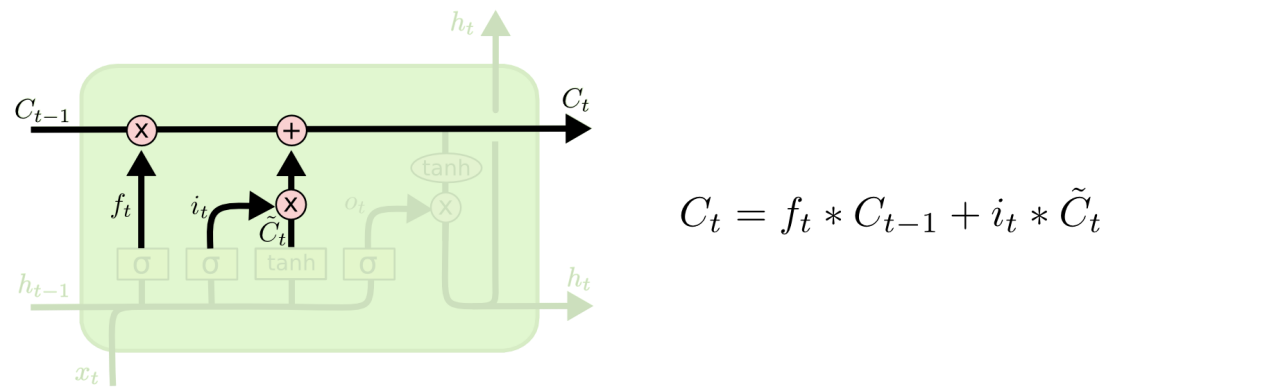
第二个是经过tanh激励函数



(3) 更新state Ct−1成Ct

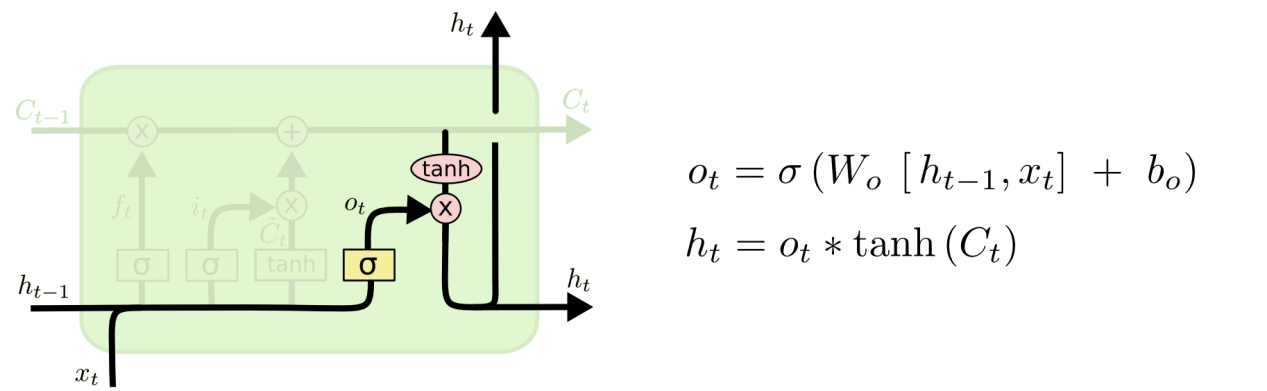
ft是经过Sigmoid函数的，所以值域在(0,1)之间，Ct−1点乘0-1之间的数实际就是对Ct−1的一种缩放，（可以认为是记住之前信息的程度）

然后加入进来的新的信息

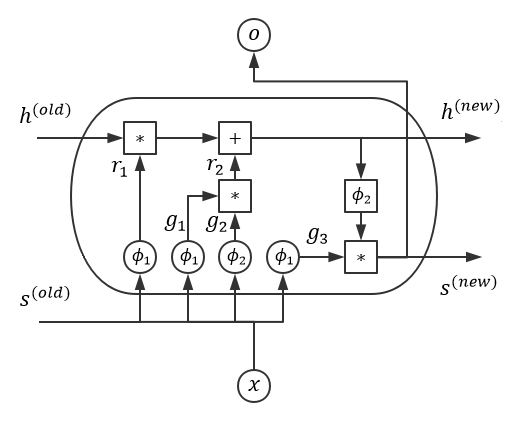


(4) 最后计算输出

输出门(output gate layer)



最后再放一下之前的图, 数据流向可能更清晰



后向过程

后想过程使用的算法这里是BPTT,跟一般的BP过程的区别就是在于Through Time部分,这里你可以展开来看整个网络,但简单说就是多了处理不同时间步间数据传输时产生的梯度的问题.

跟前向过程正好相反,这里先从输出开始进行.

优化

参数上主要的一个优化方法是在初始化时,设置大的遗忘门偏置,能够提高初始的记忆能力.这是在遗忘门提出来的一个老的技巧,但还是很有效的.

**问题10   
类别不平衡时怎么办？**

**CNN最成功的应用是在CV，那为什么NLP和Speech的很多问题也可以用CNN解出来？为什么AlphaGo里也用了CNN？这几个不相关的问题的相似性在哪里？CNN通过什么手段抓住了这个共性？**

**什么时候用local-conv？什么时候用全卷积（每一个点用同一个filter）？**

**为什么引入非线性激励函数？**

**常用的激励函数？**

问题11   
Random forest和GBT描述；

问题12   
random forest的**[算法](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库" \t "http://blog.csdn.net/victoriaw/article/details/_blank)**描述+bias和variance的分解公式；

问题13   
用EM理论推导GMM

问题14   
SVM的dual problem推导

问题15   
HMM原理

问题16   
HMM和CRF的本质区别；

问题17   
频率学派和贝叶斯派的本质区别；

问题18   
常用的优化方法；

问题19   
矩阵行列式的物理意义（行列式就是矩阵对应的线性变换对空间的拉伸程度的度量，或者说物体经过变换前后的体积比）

问题20   
**[Python](http://lib.csdn.net/base/python" \o "Python知识库" \t "http://blog.csdn.net/victoriaw/article/details/_blank)**写k-means

问题21   
动态预测每个区域的用车需求量；

问题22   
怎么样识别文本垃圾信息