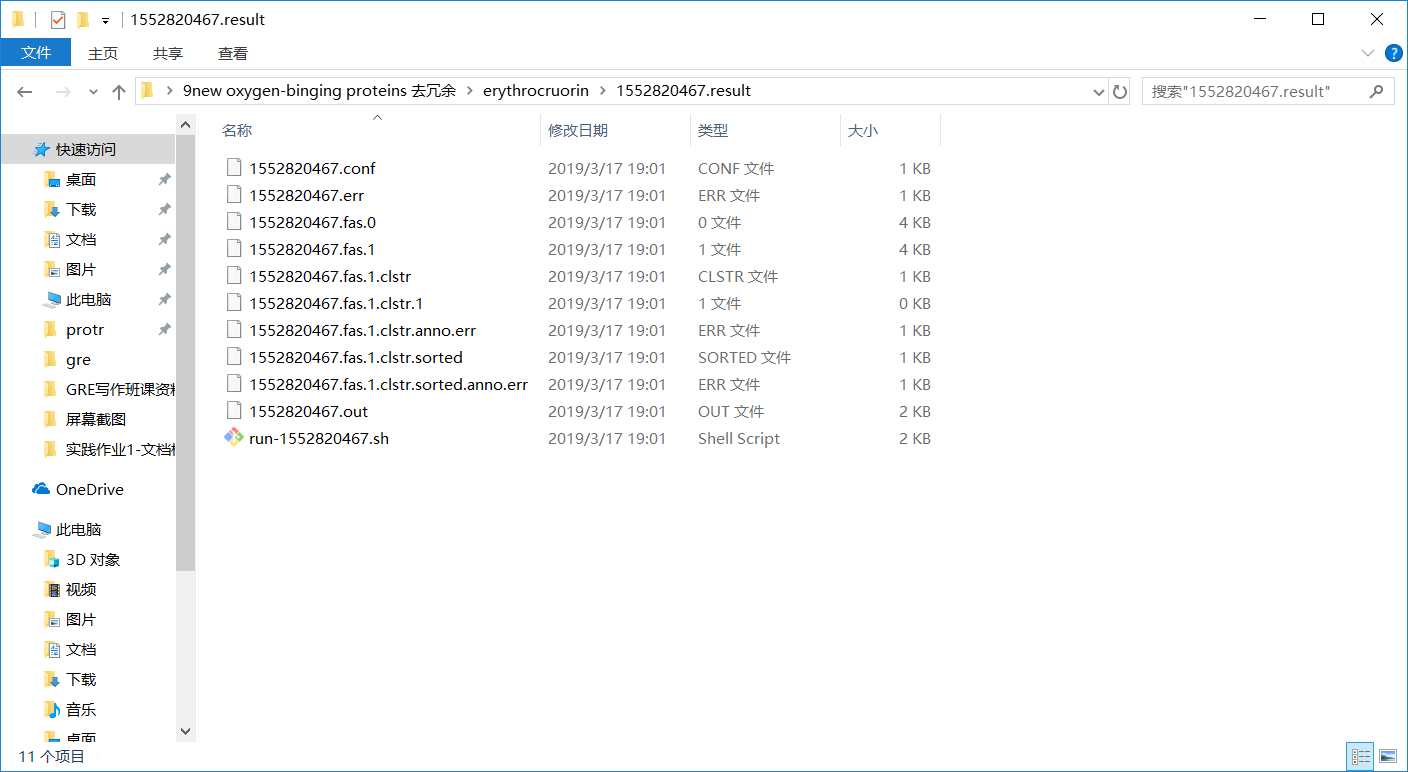
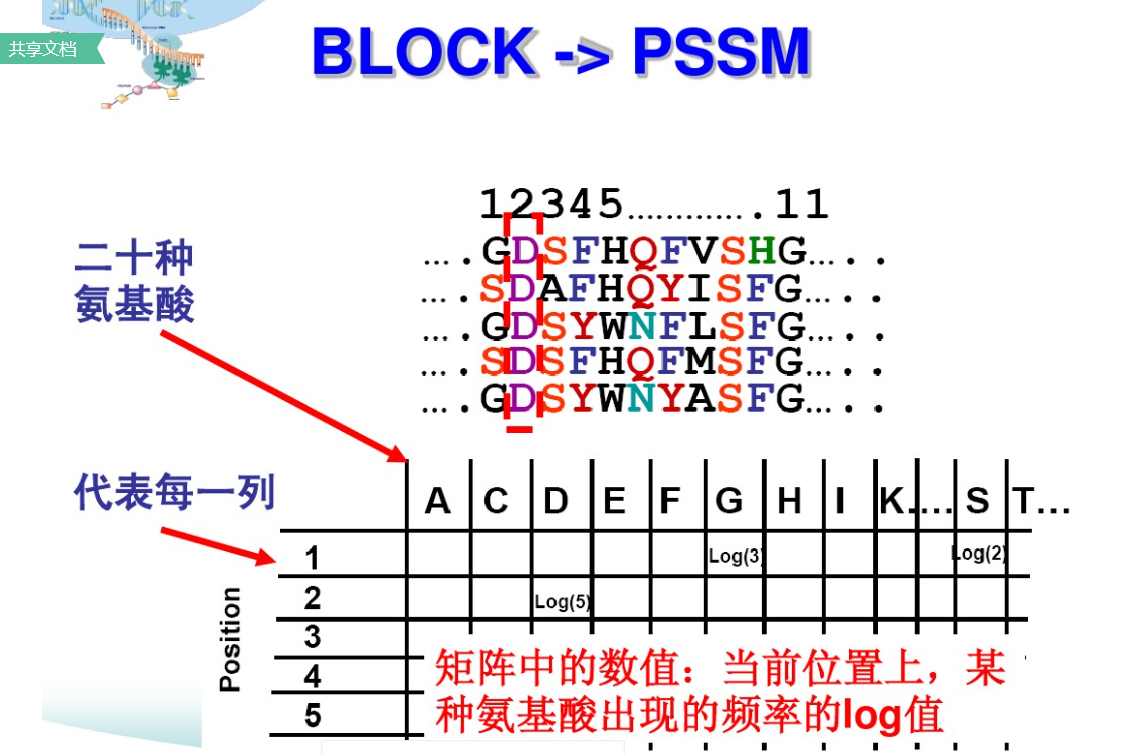
## 1.去冗余

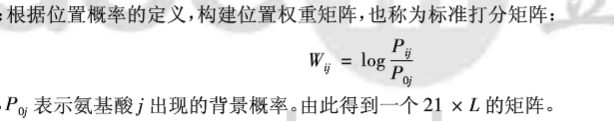
登陆web端cd-hit，把9大分类的氧结合蛋白的fasta文件上传，调整参数冗余度90%，获得去冗余后的数据。



## 2.pssm特征提取

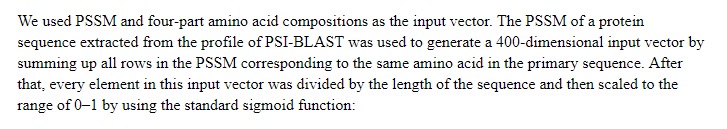
Pssm打分矩阵：使用nr库（蛋白质的非冗余数据库），使用PSI-BLAST，多次迭代，比对，得到每个氨基酸每个位置的打分。

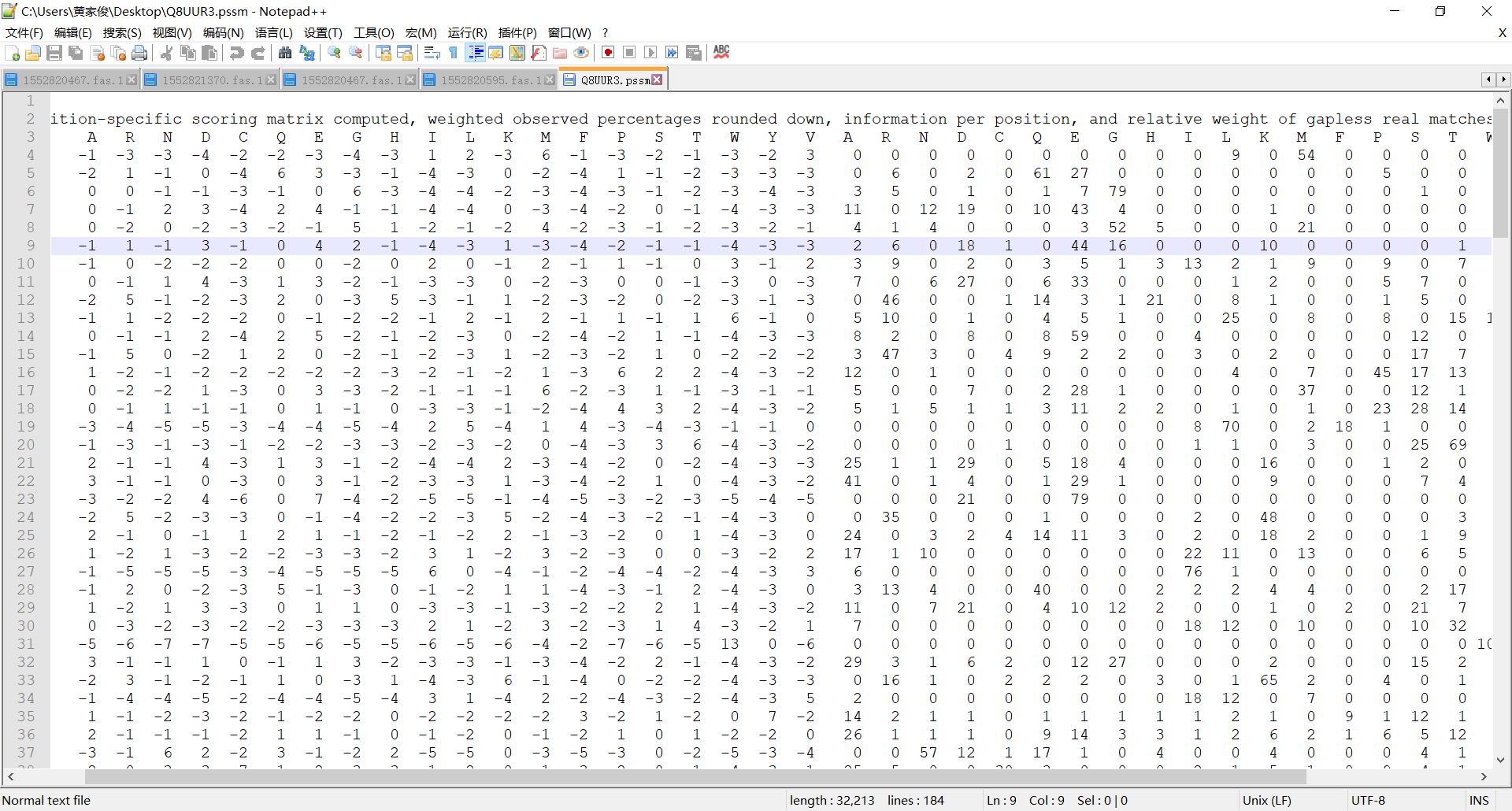










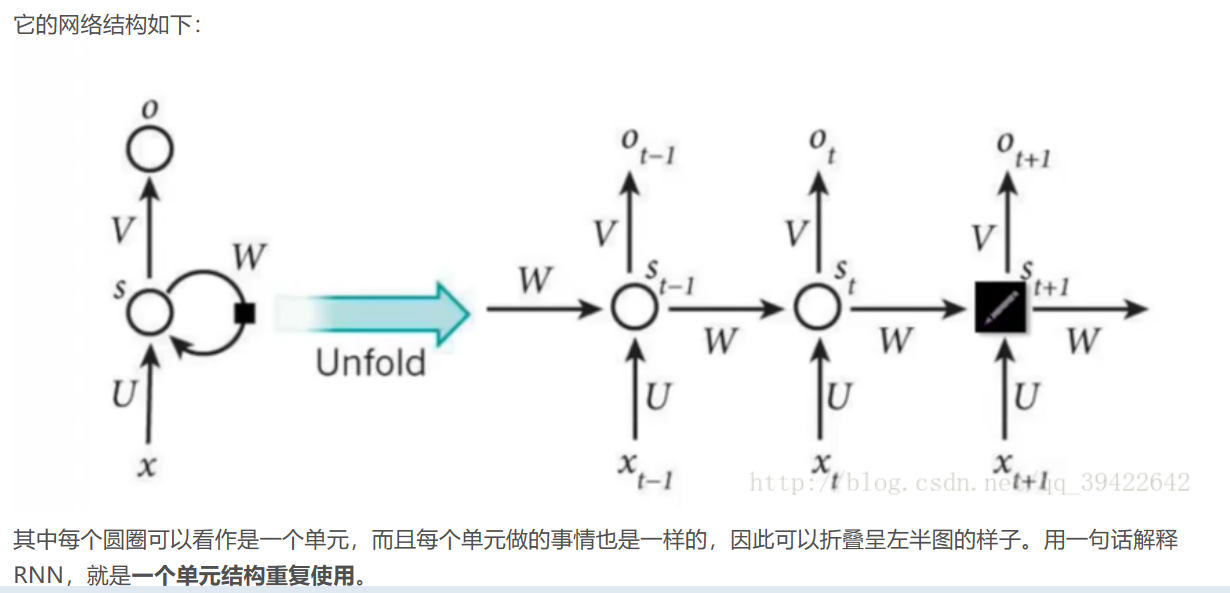


## 3.模型研究

CNN的基本结构包括两层，其一为特征提取层，每个神经元的输入与前一层的局部接受域相连，并提取该局部的特征。一旦该局部特征被提取后，它与其它特征间的位置关系也随之确定下来；其二是特征映射层，网络的每个计算层由多个特征映射组成，每个特征映射是一个平面，平面上所有神经元的权值相等。

* **卷积层（Convolutional layer）**，卷积神经网路中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从低级特征中迭代提取更复杂的特征。（filter数量自己设置）
* **线性**[**整流**](https://www.baidu.com/s?wd=%E6%95%B4%E6%B5%81&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd)**层（Rectified Linear Units layer, ReLU layer）**，这一层神经的活性化函数（Activation function）使用线性整流（Rectified Linear Units, ReLU）。
* **池化层（Pooling layer）**，通常在卷积层之后会得到维度很大的特征，将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。
* **全连接层（ Fully-Connected layer）**, 把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

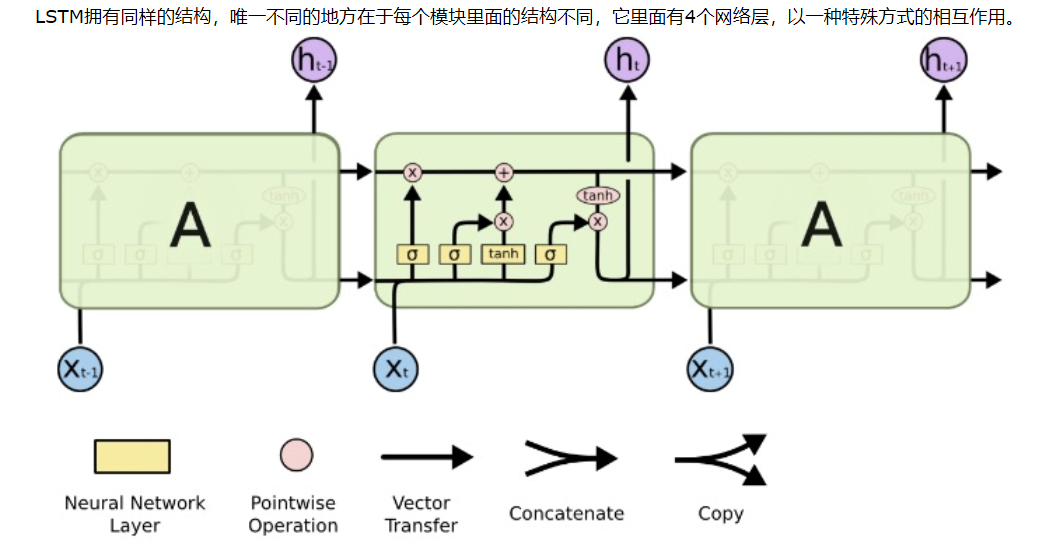
RNN：递归神经网络和传统神经网络不同的一个点在于，每次的训练，神经元和神经元之间需要传递一些信息。本次的训练，神经元需要使用上一次神经元作用之后的状态信息。类似递归函数一样。



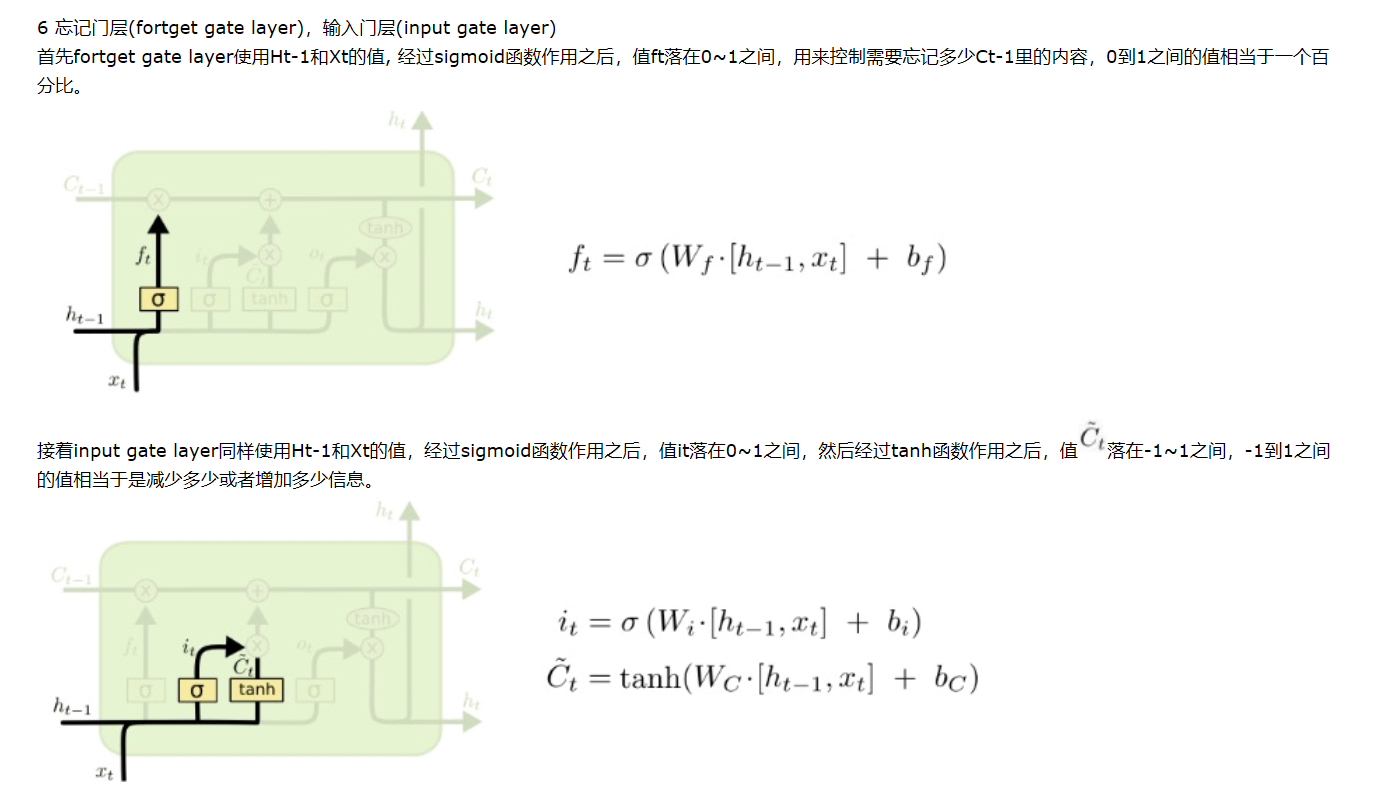
Xt:表示t时刻的输入，ot:表示t时刻的输出，St:表示t时刻的记忆

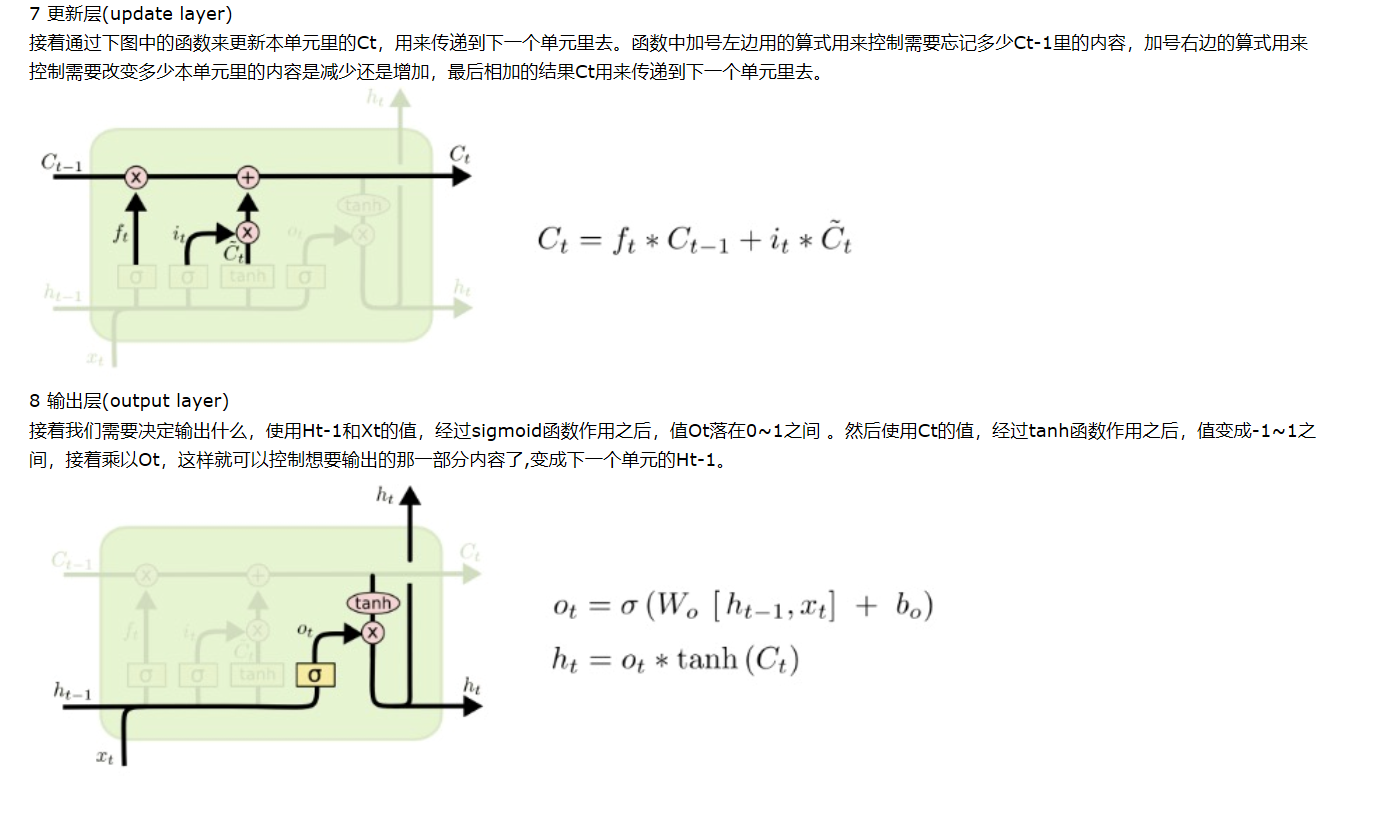
St=f(U∗Xt+W∗St−1)St=f(U∗Xt+W∗St−1)

LSTM：是一种特殊的RNN, 用来解决长期依赖问题。和传统的RNN一样，网络结构是重复的，每次的训练，神经元和神经元之间需要传递一些信息。传统的RNN，每个重复的模块里都有一个简单tanh层。



状态单元(cell state) 可以长期保存某些状态，cell state的值通过忘记门层（forget gate），输入门层（input gate layer）, 更新门层来控制实现保留多少旧状态，更新多少新的状态。





暂定模型：pssm+一维cnn+LSTM