周报二

**一、参考文献：《基于RNN的语言模型》**

**二、网址：<https://www.zybuluo.com/hanbingtao/note/541458>**

**三、关键词：机器学习；深度学习入门；RNN；语言模型**

**四、收获：**

**（1）RNN被应用于自然语言处理的原因**



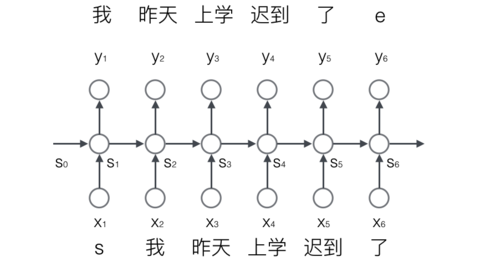
循环神经网络（RNN）既有前馈通路，又有反馈通路，因此在语音识别中可以有效的处理时间序列的上下文信息，可以做到联系上下文，因而RNN最先是在自然语言处理领域被运用起来，RNN可以为语言模型来建模，何为语言模型？通俗的说就是给定一句话的前半部分，预测接下来的词是什么，也就是说对一种语言的特征进行建模。

在使用RNN前，采用的是N-Gram：是一种语音识别中常用的语言模型，该模型的特点是第n个词的出现只与前面N-1个词相关，而与其它任何词都不相关，例子：我今天吃饭\_\_\_\_，如果采用2-Gram进行建模，在电脑预测时只会看见前面的【了】字，电脑只会在语料库中找出在【了】后面与之匹配度最大的概率，显然错误率会很大，而提升N值的大小是没有实用性的，N越大，所占用的内存空间就会越大。因此才用RNN具有反馈的特性，理论上可往前或往后看任意多个词来用于语言模型

1. **语言模型建模过程**

语料：“我 昨天 上学 迟到 了”

1. **将词输入到神经网络**



一个词代表的是一个神经元，其中s，e表示的是序列的开始和结束，未输入一个词，循环神经网络就输出截止到下一个最有可能的词出现

1. **输入与输出的向量化**

神经网络的输入与输出都是向量，因而要把词变成词向量，这就要用到one-hot编码方式

one-hot编码方式：N位状态寄存器来对N个状态进行编码，且其中只有一位有效

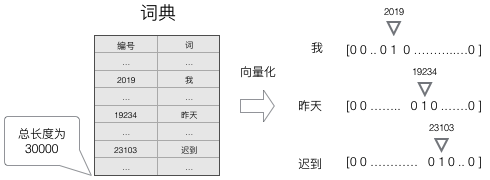
例如对六个状态进行编码：

自然顺序码为 000,001,010,011,100,101

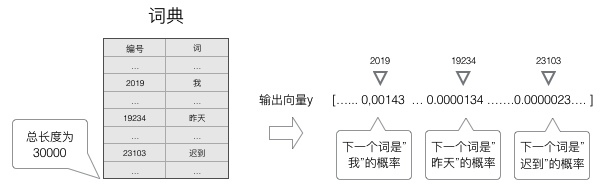
独热编码则是 000001,000010,000100,001000,010000,100000

像二进制编码一样，建立一个包含所有词的词典，每个词在词典中有唯一有效的编号，且每个词Vj都可以用N维的one-hot向量表示,N是词典中所有词的个数,j是词典中的第j个元素

输入



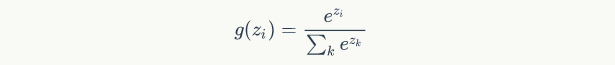
输出

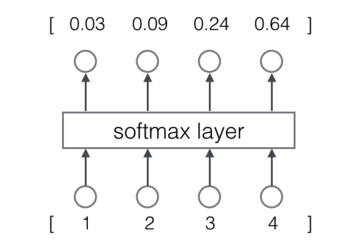


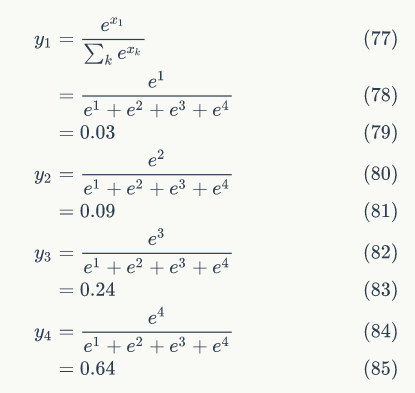
1. **使用Softmax层作为神经网络输出层输出概率**

语言模型就是对下一个词的概率进行建模，因此采用使用Softmax层作为神经网络输出层输出概率

softmax函数







输入x=[1 2 3 4]经过softmax层后输出y=[0.03 0.09 0.24 0.64],输出y最后加在一起和为1，也就是说，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。

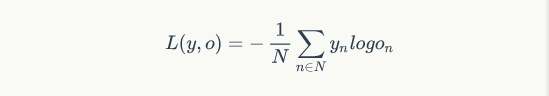
1. **语言模型的训练**

采用监督学习方法对语言模型进行训练

监督学习方法：就是对于一个训练样本，知道训练样本的特征x，训练样本的实际值y，就可以根据模型h(x)计算得到输出预测值，使模型计算出来的预测值与实际值越接近越好，个人理解，比如在微信上经常会出现的根据一些问题的答题然后来预测出你的性格或者应该从事什么专业，前期一些人的测试的答案作为样本值去训练模型，模型看到足够多的样本就能总结相应的规律，从而得出我们的测试结果。

采用交叉熵误差函数来将对模型进行优化，使预测值更加的准确

交叉熵误差函数



N是训练样本个数，yn是样本的标记，也就是实际值，on是网络的输出，假设y1=[1,0,0,0]，o=[0.03,0.09,0.24,0.64],N=1，计算方法如下：

