# 面试-RNN的梯度消失有什么与众不同的地方

AINLP 1周前

以下文章来源于NLP从入门到放弃 ， 作者DASOU

**NLP从入门到放弃**

积累一些平时的工作经验和思考，主要是关于NLP，搜索和推荐，只写干货!



**如果面试官问【聊一下RNN中的梯度消失】**

盲猜很多同学的回答可以简化成这样形式【由于网络太深，梯度反向传播会出现连乘效应，从而出现梯度消失】

这样的回答，如果用在普通网络，类似MLP，是没有什么问题的，但是放在RNN中，是错误的。

**RNN的梯度是一个和，是近距离梯度和远距离梯度的和；**

RNN中的梯度消失的含义是远距离的梯度消失，而近距离梯度不会消失，从而导致总的梯度被近的梯度主导，同时总的梯度不会消失。

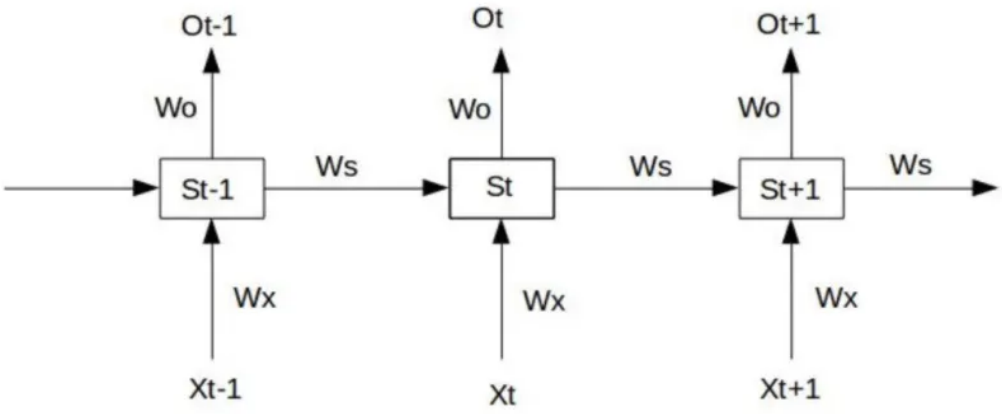
这也是为什么RNN模型能以学到远距离依赖关系。

简单的解释一下原因。

**首先，我们要明白一点，RNN是共享一套参数的（输入参数，输出参数，隐层参数），这一点非常的重要。**

当然，我们在理解RNN的时候，会把RNN按照时间序列展开多个模块，可能会认为是多套参数，这个是不对的哈。

如下所示：




然后，假设我们现在的时间序列为3，有如下公式存在：

$$S_1 = W_x X_1 + W_s S_0 + b_1 \; O_1 = W_o S_1 + b_2$$

$$S_2 = W_x X_2 + W_s S_1 + b_1 \; O_2 = W_o S_2 + b_2$$

$$S_3 = W_x X_3 + W_s S_2 + b_1 \; O_3 = W_o S_3 + b_2$$

现在假设我们只是使用t=3时刻的输出去训练模型，同时使用MSE作为损失函数，那么我们在t=3时刻，损失函数就是:

$$L_3 = \frac{1}{2}(Y_3 - O_3)^2$$

求偏导的时候，就是这样的情况：

$$\frac{\partial L_3}{\partial W_0} = \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial W_o}$$

$$\frac{\partial L_3}{\partial W_x} = \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial W_x} + \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial S_2}\frac{\partial S_2}{\partial W_x} + \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial S_2}\frac{\partial S_2}{\partial S_1}\frac{\partial S_1}{\partial W_x}$$

$$\frac{\partial L_3}{\partial W_s} = \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial W_s} + \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial S_2}\frac{\partial S_2}{\partial W_s} + \frac{\partial L_3}{\partial O_3}\frac{\partial O_3}{\partial S_3}\frac{\partial S_3}{\partial S_2}\frac{\partial S_2}{\partial S_1}\frac{\partial S_1}{\partial W_s}$$

其实看到这里，答案已经出来了。

我们以第二个公式为例，也就是对$w_x$ 求偏导，如果时间序列程度为t，我们简化一下成下面这个公式：

$$W_x = a_1 + a_2 + \ldots + a_t$$

时间序列越长，出现连乘的部分越集中出现在靠后面的公式上，比如$a_t$，但是前面的公式是不受影响的，比如$a_1$，也就是梯度是肯定存在的。

总结一下：RNN中的梯度消失和普通网络梯度消失含义不同，它的真实含义是远距离的梯度消失，而近距离梯度不会消失，同时总的梯度不会消失，从而导致总的梯度被近的梯度主导。

---



**推荐阅读**

这个NLP工具，玩得根本停不下来

征稿启示| 200元稿费+5000DBC（价值20个小时GPU算力）

完结撒花！李宏毅老师深度学习与人类语言处理课程视频及课件（附下载）

从数据到模型，你可能需要1篇详实的pytorch踩坑指南