

梯度爆炸和梯度消失

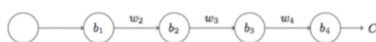
Created	@July 27, 2022 9:30 PM
Class	interview
Type	
Materials	
Reviewed	<input type="checkbox"/>

什么是梯度消失和梯度爆炸

梯度消失 (gradient vanishing) 和梯度爆炸 (gradient exploding) 一般出现在层数比较多的神经网络中。

以MLP为例，当梯度消失发生时，最后一个隐层梯度更新基本正常，但是越往前的隐层内更新越慢，甚至有可能出现停滞，此时，多层深度神经网络可能会退化为浅层的神经网络（只有后面几层在学习），因为浅层基本没有学习，对输入仅仅做了一个映射而已。

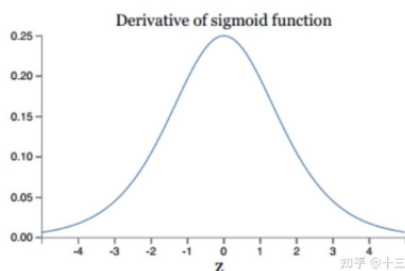
用一个每层只有一个节点的网络来描述：



根据链式法则，对b1进行求导可得：

$$\begin{aligned}\frac{\partial C}{\partial b_1} &= \frac{\partial C}{\partial y_4} \frac{\partial y_4}{\partial z_4} \frac{\partial z_4}{\partial y_3} \frac{\partial y_3}{\partial z_3} \frac{\partial z_3}{\partial y_2} \frac{\partial y_2}{\partial z_2} \frac{\partial z_2}{\partial y_1} \frac{\partial y_1}{\partial z_1} \frac{\partial z_1}{\partial b_1} \\ &= \frac{\partial C}{\partial y_4} \sigma(z_4)' w_4 \sigma(z_3)' w_3 \sigma(z_2)' w_2 \sigma(z_1)' 1\end{aligned}$$

激活函数 σ 的导数为 $\sigma' = \sigma(1 - \sigma)$ ，函数曲线如下图。函数的峰值为0.25。



如果参数在初始化的时候，初始值是小于1的，那么随着激活函数的连乘，就会导致当层数越多，梯度就会越小，因此就导致梯度消失现象相反，如果在初始化的时候，参数初始值特别大，当层数越多，就会造成梯度爆炸现象

如何解决？

梯度消失和梯度爆炸的原因是网络层次太深，网络权值更新不稳定造成的，本质是因为反向传播过程中，梯度的连乘效应造成的

另一个是使用sigmoid作为激活函数时，容易出现梯度消失问题

解决方法：

1. 压缩模型层数
2. BN
3. 改用其他的激活函数如ReLU
4. 采用ResNet中的残差连接结构