

梯度消失和梯度爆炸问题详解

1.为什么使用梯度下降来优化神经网络参数？

反向传播（用于优化神经网络参数）：根据损失函数计算的误差通过反向传播的方式，指导深度神经网络参数的更新优化。

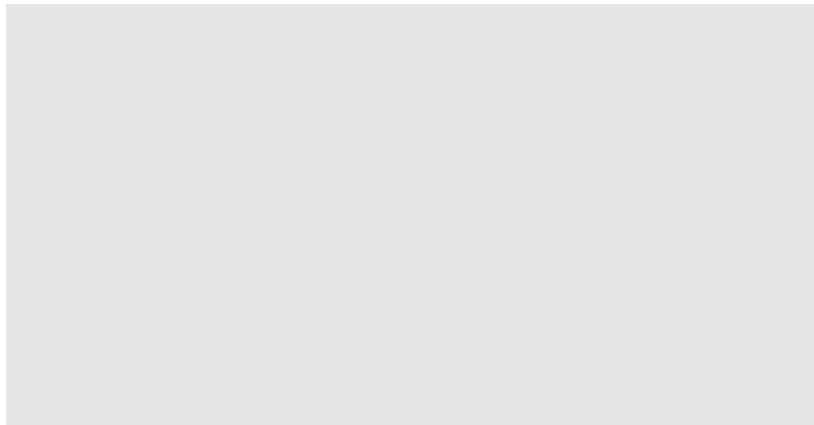
15赞

采取反向传播的原因：首先，深层网络由许多线性层和非线性层堆叠而来，每一层非线性层都可以视为是一个非线性函数（非线性来自于非线性激活函数），因此整个深度网络可以视为是一个复合的非线性多元函数。

我们最终的目的是希望这个非线性函数很好的完成输入到输出之间的映射，也就是找到让损失函数取得极小值。所以最终的问题就变成了一个寻找函数最小值的问题，在数学上，很自然的就会想到使用梯度下降来解决。

2.梯度消失、爆炸会带来哪些影响

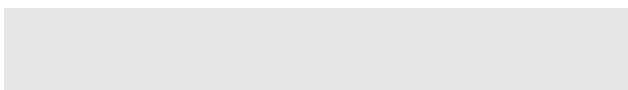
举个例子，对于一个含有三层隐藏层的简单神经网络来说，当梯度消失发生时，接近于输出层的隐藏层由于其梯度相对正常，所以权值更新时也就相对正常，但是当越靠近输入层时，由于梯度消失现象，会导致靠近输入层的隐藏层权值更新缓慢或者更新停滞。这就导致在训练时，只等价于后面几层的浅层网络的学习。



image

3.产生的原因

以最简单的网络结构为例，加入有三个隐藏层，每层的神经元个数都是1，且对应的非线性函数为 σ （其中 σ 为某个激活函数）如下图：



image

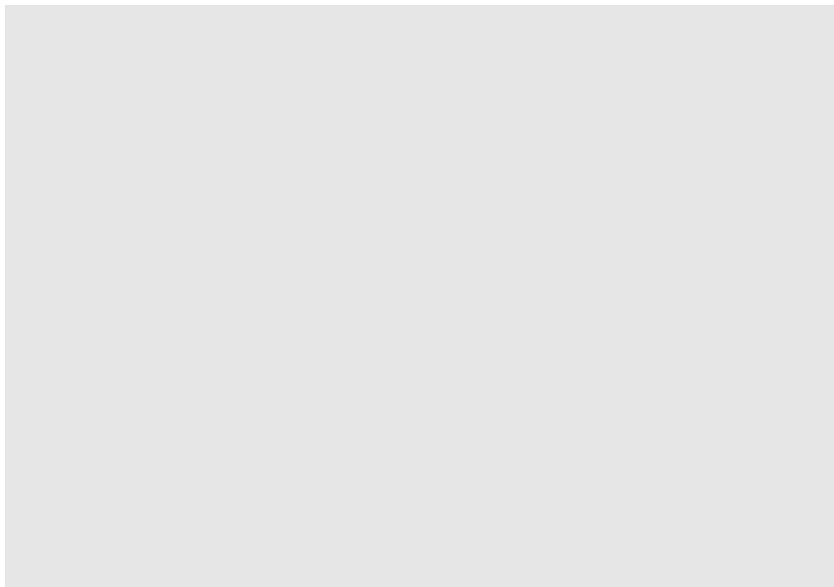
现在假设我们需要更新参数 w ，那么我们就要求出损失函数对参数 w 的导数，根据链式法则，可以写成下面这样：

写下你的评论...

评论1

赞15

...

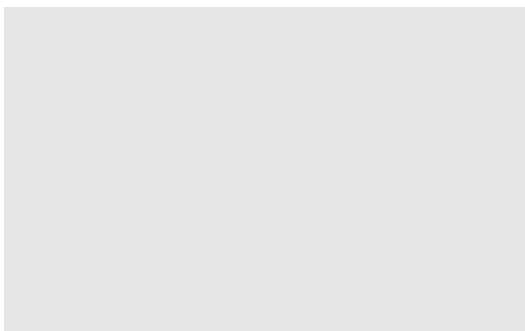


image

当我们对Sigmoid函数求导时，得到其结果如下：

由此可以得到它Sigmoid函数图像，呈现一个驼峰状（很像高斯函数），从求导结果可以看出，Sigmoid导数的取值范围在0~0.25之间，而我们初始化的网络权值通常都小于1，因此，当层数增多时，小于0的值不断相乘，最后就导致梯度消失的情况出现。同理，梯度爆炸的问题也就很明显了，就是当权值过大时，导致 ∞ ，最后大于1的值不断相乘，就会产生梯度爆炸。

Sigmoid函数求导图像



image

4.解决办法

梯度消失和梯度爆炸本质上是一样的，都是因为网络层数太深而引发的梯度反向传播中的连乘效应。

解决梯度消失、爆炸主要有以下几种方案：

4.1 换用Relu、LeakyRelu、Elu等激活函数

ReLU：让激活函数的导数为1

LeakyReLU：包含了ReLU的几乎所有优点，同时解决了ReLU中0区间带来的影响

ELU：和LeakyReLU一样，都是为了解决0区间问题，相对于来，elu计算更耗时一些（为什么）

具体可以看[关于各种激活函数的解析与讨论](#)

4.2 BatchNormalization

BN本质上是解决传播过程中的梯度问题，具体待补充完善，查看[BN](#)

4.3 ResNet残差结构

具体待补充完善，查看[ResNet](#)

4.4 LSTM结构

LSTM不太容易发生梯度消失，主要原因在于LSTM内部复杂的“门（gates）”，具体看[LSTM基本原理解析](#)

4.4 预训练加finetunning

此方法来自Hinton在06年发表的论文上，其基本思想是每次训练一层隐藏层节点，将上一层隐藏层的输出作为输入，而本层的输出作为下一层的输入，这就是逐层预训练。

训练完成后，再对整个网络进行“微调（fine-tunning）”。


此方法相当于是找全局最优，然后整合起来寻找全局最优，但是现在基本都是直接拿imagenet的预训练模型直接进行finetunning。

4.5 梯度剪切、正则

这个方案主要是针对梯度爆炸提出的，其思想是设值一个剪切阈值，如果更新梯度时，梯度超过了这个阈值，那么就将其强制限制在这个范围之内。这样可以防止梯度爆炸。

另一种防止梯度爆炸的手段是采用权重正则化，正则化主要是通过对网络权重做正则来限制过拟合，但是根据正则项在损失函数中的形式：

可以看出，如果发生梯度爆炸，那么权值的范数就会变的非常大，反过来，通过限制正则化项的大小，也可以在一定程度上限制梯度爆炸的发生。

 推荐阅读更多精彩内容 >

NLP三大特征提取器全梳理：RNN vs CNN vs Transformer

姓名：韩宜真 学号：17020120095 转载自：<https://zhuanlan.zhihu.com/p/18...>

17020120095 阅读 50 评论 0 赞 0

数据挖掘算法-逻辑回归模型

1.什么是逻辑回归模型？ 首先，我们不要被逻辑回归这个名字所误导，逻辑回归实际上是一个分类算法，它被用于将样本数据...

小飞的学习记录 阅读 19 评论 0 赞 1

基因组学中的深度学习

全文6,743字，阅读30分钟。 这一篇文章的主题是深度学习在基因组学中的应用情况的。文章较长，读完要花些时间，不...

黄树嘉 阅读 970 评论 0 赞 14

详解pytorch CNN操作

陈崇和 陈崇和之前写过两篇深度学习系列，其中第二篇介绍了如何编写CNN的顶层模块，本文的后续部分...

一堆卷积 一堆卷积土安用作降维或升维。以下所有例子都以语音/NLP的场景讲述，输入的特征为Data X 1 X...

习惯了千姿百态 阅读 110 评论 1 赞 2

机器学习启蒙—线性回归

前言 最近在网上一本亚马逊的技术大佬写的机器学习入门书籍，总共100多页，刚好最近在学习机器学习相关的知识，就...

shui水mo墨 阅读 167 评论 0 赞 2