

ResNet-史上最详细解读



科研盐酒员

东北大学->保送中科院直博

最近这一个星期阅读和学习了resnet文章，接下来我就用通俗的语言把残差解决的问题、原理、梯度推导过程、弊端给大家讲清楚。史上最详细、最基础的resnet的讲解。

梯度：

我先把梯度推导过程拿出来，毕竟很多人都会用resnet,但不知道梯度的推到过程，弄了半天也不知道怎么把图片旋转正过来，梯度推导过程主要有三点：

1.正向传播过程中，我把残差结构的batchnorm回归和激活函数relu的计算过程舍去，主要是方便理解，如要加上，无非就是在正向传播线性变化完后再进行一个非线性变化的计算公式，反向传播时进行复合函数求导即可，这里舍去方便理解。

2.我把普通CNN的梯度推到过程也写出来进行对比

正向传播

$$x_1 = x_0 + F(x_0, w_0)$$

$$x_2 = x_1 + F(x_1, w_1)$$

$$x_3 = x_2 + F(x_2, w_2)$$

$$x_4 = x_3 + F(x_3, w_3)$$

反向传播，用链式法则求梯度

$$\frac{\partial E}{\partial w_3} = \frac{\partial E}{\partial x_4} \cdot \frac{\partial x_4}{\partial w_3} = \frac{\partial E}{\partial x_4} \cdot F'(x_3, w_3)$$

$$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial x_3} \cdot \frac{\partial x_3}{\partial w_2} = \frac{\partial E}{\partial x_3} \cdot F'(x_2, w_2)$$

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial x_2} \cdot \frac{\partial x_2}{\partial w_1} = \frac{\partial E}{\partial x_2} \cdot F'(x_1, w_1)$$

标准 CNN

$$x_4 = F(x_3, w_3)$$

$$\frac{\partial E}{\partial w_3} = \frac{\partial E}{\partial x_4} \cdot \frac{\partial x_4}{\partial w_3} = \frac{\partial E}{\partial x_4} \cdot F'(x_3, w_3)$$

$$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial x_3} \cdot \frac{\partial x_3}{\partial w_2} = \frac{\partial E}{\partial x_3} \cdot F'(x_2, w_2)$$

反向传播

$$x_4 \rightarrow x_3 \rightarrow x_2 \rightarrow x_1 \rightarrow x_0$$

背景：

赞同 添加评论 分享 喜欢 收藏 申请转载



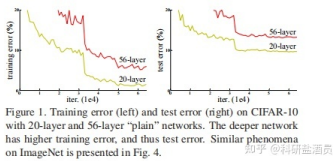
选择分享resnet也是因为它在deep learning领域有代表性，也对之后几年的很多研究打下了基础，包括文章读起来通俗易懂，对于学习神经网络也很有代表性，它的一作作者去了facebook工作。

一般的卷积神经网络都是卷积层、池化层、再加上一些防止过拟合的函数、最后就是全连接层得到最终的输出，有时加上softmax将输出转化为概率分布。

随着网络层数的加深，网络的表达能力会更强，这是因为卷积核的作用是提取图像的特征，然而一个卷积核是不够的，一个卷积核只能反应图像的某一个特征，所以我们需要多个卷积核，这些不同的卷积核可以提取到图像不同的特征，从而让我们的模型学习图像特征的能力更强。因此有足够的卷积核和足够的参数才可以更好表述原始图像的特征。因此深度网络有两个优势特点，1. 特征的等级随着网络深度的加深而变高；2. 越深的深度使网络的表达能力更强。

问题:

但是现在的深度网络却出现了两个明显的缺点，看到原论文中这两个图片，分别用的20层和56层的普通神经网络来跑cifar-10数据集，可以看到，越深的网络在训练集和测试集上的表现都不如浅的网络好。



原因:

主要两个原因导致，**第一个原因**就是层数加深，引起反向传播梯度消失或者梯度爆炸，**(反向传播是用来对网络的权重进行调整，包括卷积核的值，隐藏层的权重和偏置，这些都需要反向传播来调整；反向传播主要是计算变化因子来调整权重，而变化因子的计算首先得计算目标函数（预测值和真实值的差的平方和）对每层网络权重的偏导数）**，因此咱们在求反向传播求梯度时利用了链式法则，梯度值会进行一系列的连乘，也就会出现剧烈的缩减或者变大（这里也比较容易理解，假设每一层的误差梯度是一个小于1的数，在反向传播过程中，每向前传播一次，都要乘以一个小于1的误差梯度，当网络越来越深，所乘的小于1的系数越多，梯度就越趋近于零，反过来如果每一层的梯度是一个大于1的数，则会发生梯度爆炸），这种现象就阻碍了收敛。

但是现在使用标准初始化或者在网络中间层进行归一化，解决了反向传播的梯度消失或者爆炸问题。

第二个原因就是如果解决了梯度消失的问题后，仍然会存在层数深的网络没有浅的网络效果好，这就是退化问题。

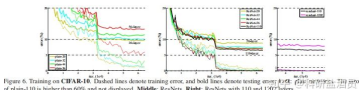
退化问题的产生

原始层：由一个已经学会的较浅模型复制而来

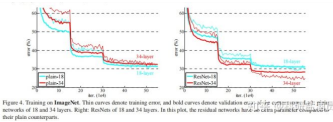
在原始层基础上加上附件层：设置为恒等，但是实际情况中附加层不能被训练为恒等，导致浅层和深层之间就有了不同的误差。

解决方式:

残差结构可以很好的解决梯度消失或者爆炸、退化问题，具体的效果图可以看这，下面这两张图一个在cifar-10上做的实验，一个在ImageNet数据集上做的实验，可以看到有残差的网络极大的解决了梯度消失和爆炸、退化的问题。



实现代表的是验证集的错误率，虚线代表的是训练集的错误率



再说一下这个残差结构，原文中是有两种不同的残差结构，左边针对层数较少的网络所使用，右边是针对层数较深的网络如ResNet-50/101/152。

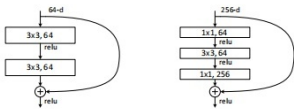


Figure 5. A deeper residual function \mathcal{F} for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a “bottleneck” building block for ResNet-50/101/152.

对于一个堆积层结构（几层堆积而成）当输入为 x 时其学习到的特征记为 $H(x)$ ，普通神经网络拟合的是 $H(x)$ ，现在我们把 $H(x)$ 可以理解为层输入 x 和残差函数 $F(x)$ 的加和，因为残差的定义就是预测值和观察值的差，所以 $H(x)-x$ 也就是残差 $F(x)$ ，这两种表达的效果相同，但是优化的难度却并不相同。

现在残差网络直接用来拟合残差 $f(x)$ 即可，相比于浅层网络出现的退化问题，我们只需要将附加层变成理论上的恒等变化即可，所以我们让残差中的卷积层和隐藏层的权重 W 尽可能区域 0，使得拟合目标 $f(x)$ 趋于零。这样就能解决附加层误差大的问题。

当残差为0时，此时堆积层仅仅做了恒等映射，至少网络性能不会下降，实际上残差不会为0，因此这个残差既极大的缓解了退化问题，也能使得堆积层在输入特征基础上学习到新的特征，从而拥有更好的性能。

残差的输入输出问题：

同等维度shape

当输入、输出通道数相同时，我们自然可以如此直接使用 x 进行相加。而当它们之间的通道数目不同时，我们就需要考虑建立一种有效的 identity mapping 函数从而可以使得处理后的输入 x 与输出 $F(x)$ 的通道数目相同即可

Shape不相同的情况下

因为我们输入的 x 经过卷积等等的操作的，输出维度可能会发生改变，此时输出的残差 $F(x)$ 跟额外层的 x 维度不匹配，因此我们采用 1×1 的卷积核来对额外层 x 进行维度的改变，使得单独变化后的 x 能与残差 $F(x)$ 相加， 1×1 的卷积核学习能力小，根本学不到相邻像素的空间信息，无法提取出轮廓特征，这里 1×1 的目的是为了压缩，减少其他尺寸卷积核的运算量，尽量不改变额外层 x 的特征信息。

梯度解决：

这样反向求偏导过程中也解决了梯度消失和梯度爆炸的问题，对任何一个残差求偏导都会有一个 1。解决退化问题是由于附加层的恒等变换，误差过大， $F(x)$ 趋于零时恰恰解决了这个问题。

具体原因参考文章开头

优缺点：

- 1、既利用了深层次的神经网络又避免了梯度消散和退化的问题。
- 2、resnet 看起来很深但实际起作用的网络层数不是很深，大部分网络层都在防止模型退化，误差过大。而且残差不能完全解决梯度消失或者爆炸、网络退化的问题，只能是缓解！

留一个小问题：

用普通网络和残差网络分别对 cifar-10 和 ImageNet 数据集做分类实验，cifar-10 数据在深层网络和浅层网络上的误差差距很大，反而数据大，样本大的 ImageNet 数据无论是在普通 CNN，还是 ResNet 网络上，其深层网络和浅层网络的误差差距不大？

▲ 赞同 ▼

● 添加评论

🔗 分享

❤ 喜欢

★ 收藏

📄 申请转载

...



推荐阅读



深入理解ResNet 到 IRNet

玖零猴 发表于Deep ...

Resnet50详解与实践（基于mindspore）

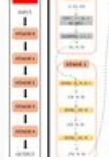
1. 简述Resnet是残差网络(Residual Network)的缩写,该系列网络广泛用于目标分类等领域以及作为计算机视觉任务主干经典神经网络的一部分，典型的网络有resnet50, resnet101等。Resnet网络的...

micro... 发表于AI框架及...



ResNet及其变种的结构梳理、有效性分析与代码解读

Pascal



ResNet50 解

臭咸鱼

还没有评论

写下你的评论...

