深度残差网络和Highway网络

转载 xiaocong1990 于 2017-05-18 20:37:41 发布 2022 收藏 2
分类专栏： 深度学习

今天讲的这两种网络结构都是最新被业界针对 图像处理 问题提出的最新的结构，主要解决就是超深层的网络在训练优化中遇到的问题。说实话这两种模型就本身来说数学公式都不复杂，但是确实在实战中取得了非常好的效果（深度残差网络帮助微软的团队以绝对优势获得了2015 Image Cup的冠军），这也从侧面说明了**深度学习**是一门以实践为主导的学科，在这个领域里实践才是检验真理的唯一标准。（很多新的结构都是因为在实践中取得了不错的效果，然后被一些大牛通过一些高大上概念进行包装，最后再以一种很牛逼的姿态传递到我们的面前，令我们膜拜）。

首先来说一下深度 残差网络 ，下面是深度残差网络的**架构**图
（来自论文《Deep Residual Learning for Image Recognition》）

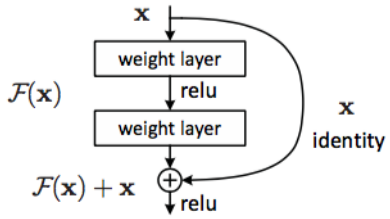


Figure 2. Residual learning: a building block.

之所以说起名“残差”网络，是因为假设网络要学习的假说是H(x)，那么由于图中identity x之间跨过了2层，那么其实相当于拟合的是F(x)=H(x)-x，这就是残差概念的来源，这是论文里的说法。其实我感觉作者在提出这个结构的时候，打破了传统的神经网络n-1层的输出只能给n层作为输入的惯例，使某一层的输出可以直接跨过几层作为后面某一层的输入。乍一看这样的结构没啥的，貌似没有什么特别厉害的地方，其实不然

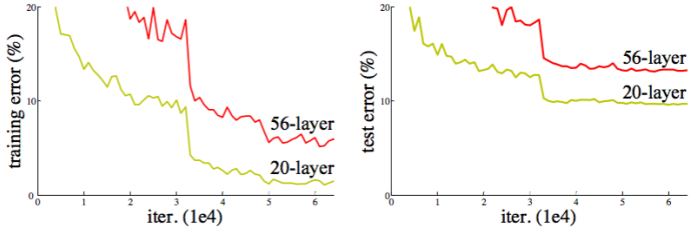


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

上图就是其构造深度残差网络的构思来源图，一个是56层的网络一个是20层的网络，从原理上来说其实56层网络的解空间是包括了20层网络的解空间的，换而言之也就是说，56层网络取得的性能应该大于等于20层网络的性能的。但是从训练的迭代过程来看，56层的网络无论从训练误差来看还是**测试**误差来看，误差都大于20层的网络（这也说明了为什么这不是过拟合现象，因为56层网络本身的训练误差都没有降下去）。导致这个原因就是虽然56层网络的解空间包含了20层网络的解空间，但是我们在训练网络用的是随机梯度下降策略，往往解到的不是全局最优解，而是局部的最优解，显而易见56层网络的解空间更加的复杂，所以导致使用随机梯度下降**算法**无法解到最优解。

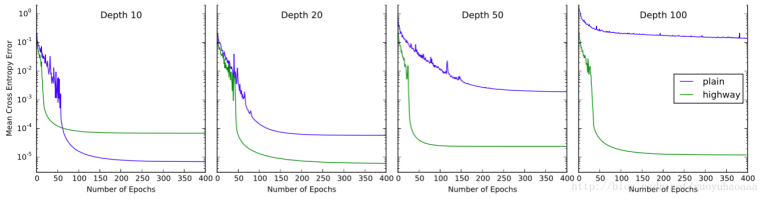
其实在构造这个网络的时候，我们完全可
常好的结果了，我在构造56层网络的时候

层只做inentity map至少效果不会差于20层的网络。于是乎深度残差的网络就提出了，这个思想其实不复杂，说白了打破了每一层网络输入只能来自于上一层网络输出的规律，可以让一些网络的输出直接跳过几层到达后面的输入。这样的网络确实也取得了非常好的效果。另外要注意的是，在真正训练的时候，有几点trick要注意：1、注意层与层之间使用batch-normalization技术，否则由于网络过深会导致梯度消失的问题，导致网络训练无法收敛；2、论文里说了，为了保持每一层网络的参数差不多，每当经过了pooling层输入的维度减少了一般，那么filter的个数就要增加一倍。

说完了深度残差网络，我们再来说说Highway网络。这篇网络来源于论文《Highway Networks》
所谓Highway网络，无非就是输入某一层网络的数据一部分经过非线性变换，另一部分直接从该网络跨过去不做任何转换，就想走在高速公路上一样，而多少的数据需要非线性变换，多少的数据可以直接跨过去，是由一个权值矩阵和输入数据共同决定的。下面是Highway网络的构造公式：

向量由两项组成。叫做transform gate ，叫做carry gate。和的激活函数都是函数。
T算出来的是一个向量,其中每个数字都是（0，1）之间的浮点数，代表y中由x变化后的内容所占的比例；
C算出来的也是一个向量,其中每个数字也都是（0，1）之间的浮点数，代表y中由x本身内容所占的比例；
（为了简便起见，有时候令， 代表了维度和 一样长的向量）从公式中我们需要注意的是，由于是点乘，当取了 之后那么必须是同样的维度。如果我们想更改x的维度从A变成B的话，一种方法是采用zero-padding和下采样的方法，或者是引入一个维度为A*B的变换矩阵，使每次都乘上这个矩阵。

主要解决的是多层深度神经网络的训练收敛问题，即使层数很多也可以使用简单的方法比方说 backpropogation来进行训练，保证合理的迭代范围内收敛，而传统的网络是很难保证收敛的。如下图所示：



当网络很深的时候，使用了Highway的网络更容易收敛。

原文里说道：
A highway layer can smoothly vary its behavior between that of a plain layer and that of a layer which simply passes its inputs through.
也就是说Highway也就是让输入数据的一部分变换，另一部分直接通过，相当于整体上来讲在这两者的效果中选了一个均衡。

从广义的角度来说，Highway更像是一种思想，它不但可以用在全连接网络中，也可以用在卷积神经网络中，原文里说：“Convolutional highway layers are constructed similar to fully connected layers. Weight-sharing and local receptive fields are utilized for both H and T transforms. We use zero-padding to ensure that the block state and transform gate feature maps are the same size as the input.”。

其实深度残差网络和Highway网络这两种网络结构都能够让一部分的数据可以跳过某些变换层，而直接到后面的层中去，只不过Highway网络需要一个权值来控制每次直接通过的数据量，而深度残差网络就直接让一部分数据通到了后面。从大量的实验中，我感觉这两种网络只有在很深的场景中才能发挥出“威力”，如果本身网络层数较浅，勉强使用这两种结构是很难得到好的结果的。