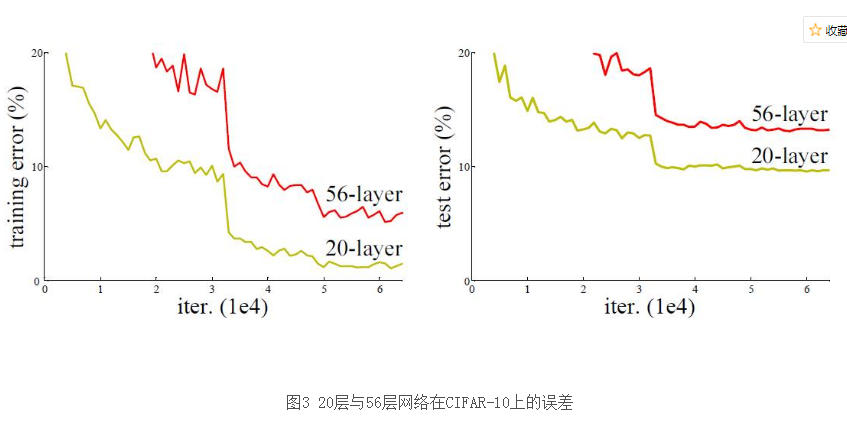
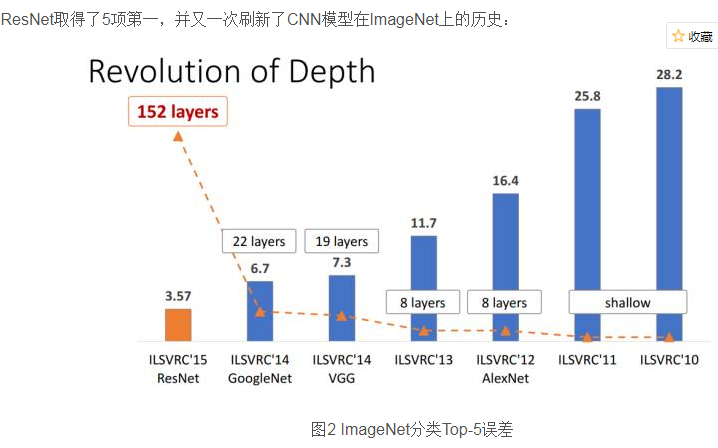
1.先看两张图片：分别是各个模型的误差和层数加大时的误差变化规律



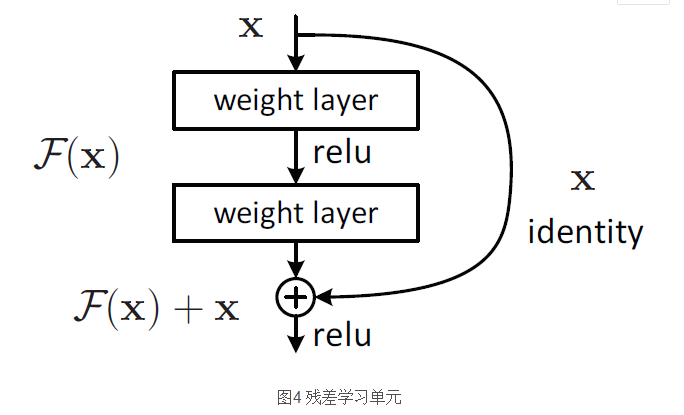
从经验来看，网络的深度对模型的性能至关重要，当增加网络层数后，网络可以进行更加复杂的特征模式的提取，所以当模型更深时理论上可以取得更好的结果，从图2中也可以看出网络越深而效果越好的一个实践证据。但是更深的网络其性能一定会更好吗？实验发现深度网络出现了退化问题（Degradation problem）：网络深度增加时，网络准确度出现饱和，甚至出现下降。这个现象可以在图3中直观看出来：56层的网络比20层网络效果还要差。这不会是过拟合问题，因为56层网络的训练误差同样高。我们知道深层网络存在着梯度消失或者爆炸的问题，这使得深度学习模型很难训练。但是现在已经存在一些技术手段如BatchNorm通过规范化输出数据来改变数据分布，是一个向前的过程来解决梯度弥散问题来缓解这个问题。因此，出现深度网络的退化问题是非常令人诧异的。

2.在ResNet之前，Schmidhuber教授提出了Highway Network，原理与ResNet很相似。Highway Network 相当于修改了每一层的激活函数，此前的激活函数只是对输入做一个非线性变换https://img2018.cnblogs.com/blog/1305301/201809/1305301-20180929170751628-1672932219.png，Highway Network则允许保留一定比例的原始输入x,即https://img2018.cnblogs.com/blog/1305301/201809/1305301-20180929170947037-1305130811.png，其中T为变换系数，C为保留系数，论文中令C=1-T。这样前一层的信息，有一定比例可以不经过矩阵乘法和非线性变换，直接传输到下一层，仿佛一条信息高速公路，因此得名Highway Network。

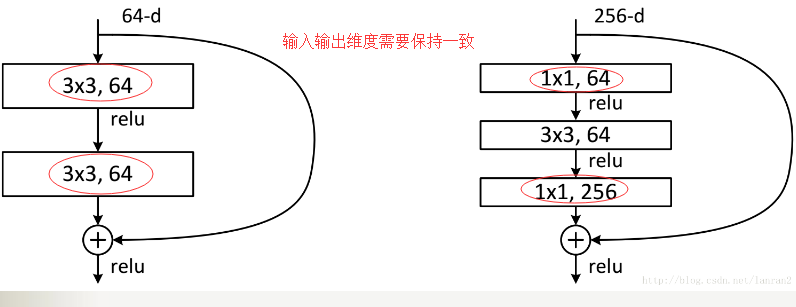
3.针对出现的“退化”问题，作者提出了残差网络的概念：

假设：如果多层非线性层叠加可以渐进逼近复杂的函数，那么也能渐进逼近残差函数：H(X)-X ，F(X):=H(X)−X，原来的网络结构变成了F(X)+X

skip connection（Identity Map）来直接连接浅层网络与深层网络

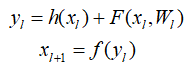


残差学习相比原始特征直接学习更容易。当残差为0时，此时堆积层仅仅做了恒等映射，至少网络性能不会下降，实际上残差不会为0，这也会使得堆积层在输入特征基础上学习到新的特征，从而拥有更好的性能

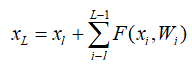


因为残差等于目标输出减去输入，即H(x)-x,因此输入、输出维度需保持一致

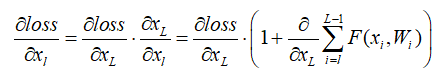
4. 为什么残差学习相对更容易？

首先残差单元可以表示为：

其中0?wx_fmt=png分别表示的是第0?wx_fmt=jpeg个残差单元的输入和输出，注意每个残差单元一般包含多层结构。 0?wx_fmt=png是残差函数，表示学习到的残差，而 0?wx_fmt=png表示恒等映射， 0?wx_fmt=png是ReLU激活函数。基于上式，我们求得从浅层 0?wx_fmt=png到深层0?wx_fmt=png的学习特征

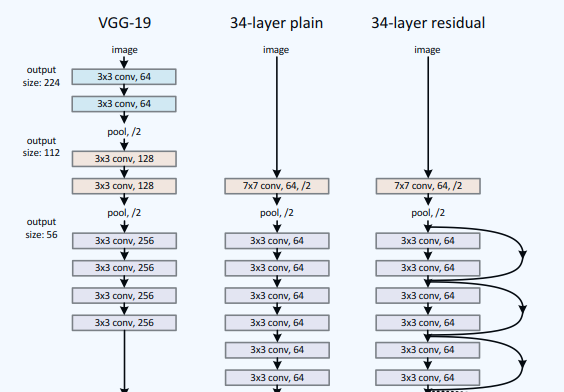


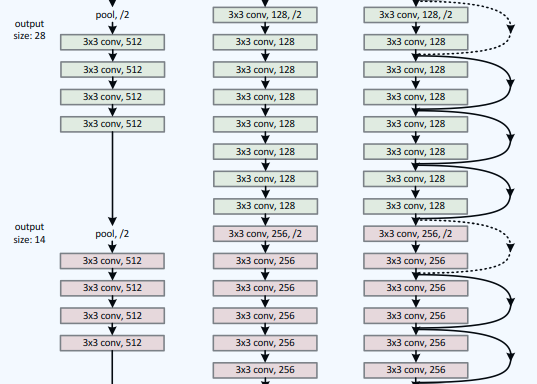
链式求导法则

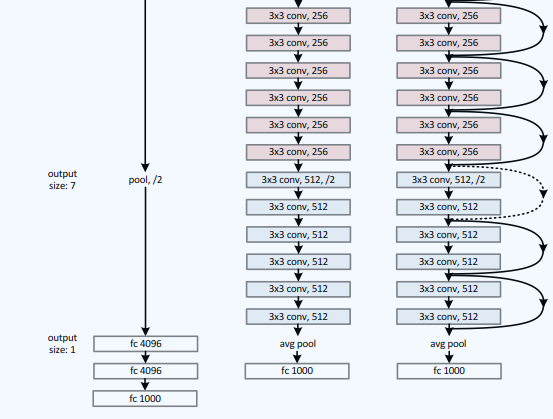


式子的第一个因子（δloss/δxl）表示的损失函数到达的梯度，小括号中的1表明短路机制可以无损地传播梯度，而另外一项残差梯度则需要经过带有weights的层，梯度不是直接传递过来的。残差梯度不会那么巧全为-1，而且就算其比较小，有1的存在也不会导致梯度消失。所以残差学习会更容易。要注意上面的推导并不是严格的证明。

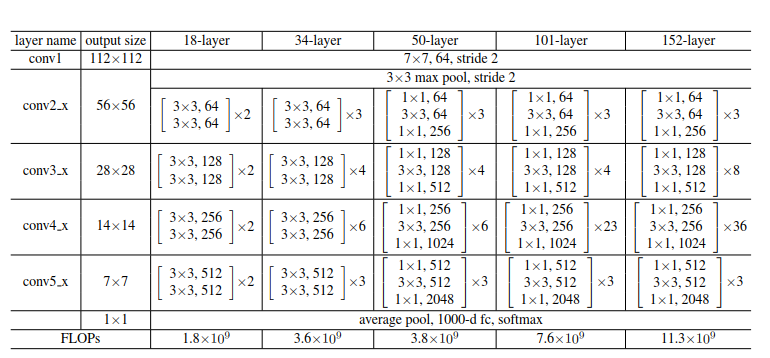
5.ResNet的模型



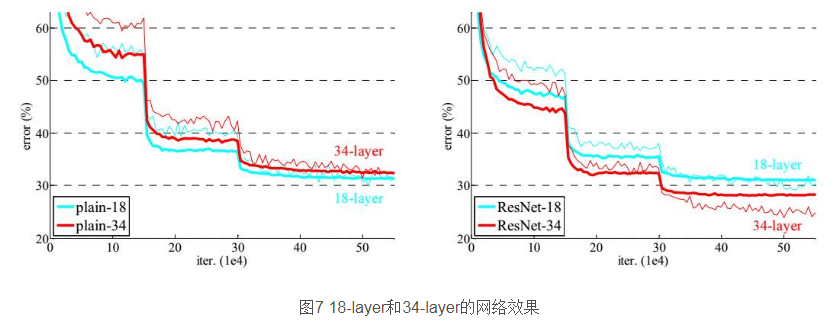




6.ResNet模型的各层参数



7，效果比较



说明：

①两折的基础网络是相同的，resNet同plainNet

②不同的是ResNet使用了shortcut connection：使得每两个卷积增加了一个shortcut connnection

③所有的shortcut 采用Identity Mapping，如果维度增加了，则采用0补足，从而相对于plainNet不会带来任何额外参数

④结果ResNet效果都要比plainNet要好:

1.通过table2，可以看出resNet可以比较好的控制“退化”问题，

2.top-1 error的对比，更能体现残差在提高网络深度方面更加有效

3.收敛速度：通过对比18-layer发现resNet和plainNet的最终精度差不多，说明网络深度不是特别深的时候，SGD能然可以对plainNet找到一个不错的最优解，但是Resnet提供了一种收敛速度更快的结构

8.训练过程中，也采取了类似VGG中的图像翻转、裁剪技术