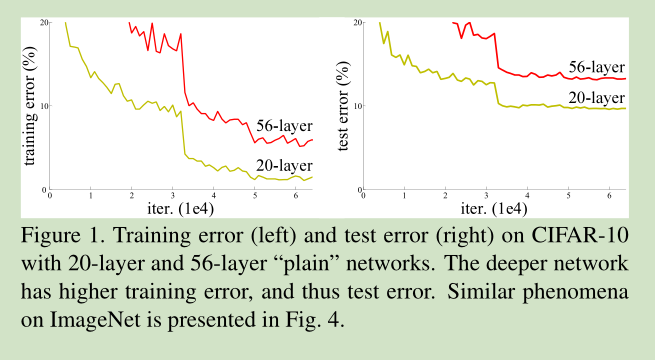
这是微软方面的最新研究成果， 在第六届ImageNet年度图像识别测试中，微软研究院的计算机图像识别系统在几个类别的测试中获得第一名。

本文是解决超深度CNN网络训练问题，152层及尝试了1000层。

随着CNN网络的发展，尤其的VGG网络的提出，大家发现网络的层数是一个关键因素，貌似越深的网络效果越好。但是随着网络层数的增加，问题也随之而来。

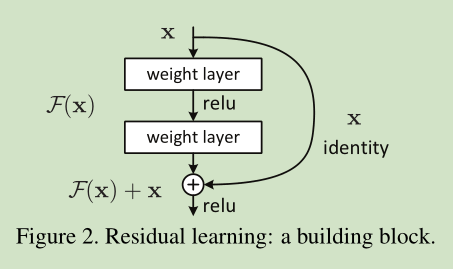
首先一个问题是 vanishing/exploding gradients，即梯度的消失或发散。这就导致训练难以收敛。但是随着 normalized initialization [23, 9, 37, 13] and intermediate normalization layers[16]的提出，解决了这个问题。

当收敛问题解决后，又一个问题暴露出来：随着网络深度的增加，系统精度得到饱和之后，迅速的下滑。让人意外的是这个性能下降不是过拟合导致的。如文献 [11, 42]指出，对一个合适深度的模型加入额外的层数导致训练误差变大。如下图所示：   


如果我们加入额外的 层只是一个 identity mapping，那么随着深度的增加，训练误差并没有随之增加。所以我们认为可能存在另一种构建方法，随着深度的增加，训练误差不会增加，只是我们没有找到该方法而已。

这里我们提出一个 deep residual learning 框架来解决这种因为深度增加而导致性能下降问题。 假设我们期望的网络层关系映射为 H(x), 我们让 the stacked nonlinear layers 拟合另一个映射， F(x):= H(x)-x , 那么原先的映射就是 F(x)+x。 这里我们假设优化残差映射F(x) 比优化原来的映射 H(x)容易。

F(x)+x 可以通过shortcut connections 来实现，如下图所示：



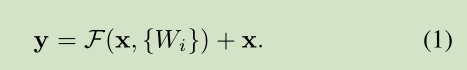
2 Related Work   
Residual Representations   
以前关于残差表示的文献表明，问题的重新表示或预处理会简化问题的优化。 These methods suggest that a good reformulation or preconditioning can simplify the optimization

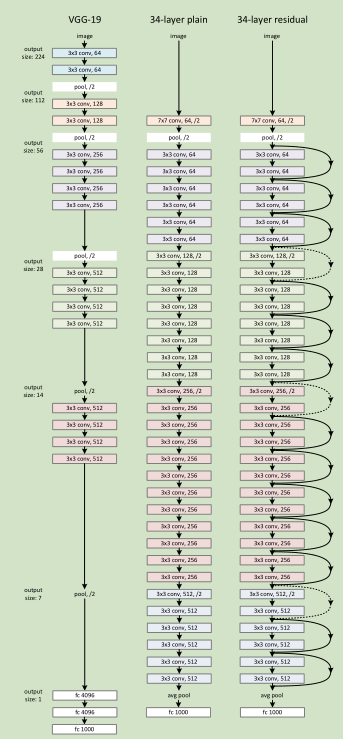
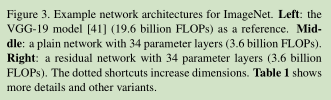
Shortcut Connections   
CNN网络以前对shortcut connections 也有所应用。

3 Deep Residual Learning   
3.1. Residual Learning   
这里我们首先求取残差映射 F(x):= H(x)-x，那么原先的映射就是 F(x)+x。尽管这两个映射应该都可以近似理论真值映射 the desired functions (as hypothesized)，但是它俩的学习难度是不一样的。

这种改写启发于 图1中性能退化问题违反直觉的现象。正如前言所说，如果增加的层数可以构建为一个 identity mappings，那么增加层数后的网络训练误差应该不会增加，与没增加之前相比较。性能退化问题暗示多个非线性网络层用于近似identity mappings 可能有困难。使用残差学习改写问题之后，如果identity mappings 是最优的，那么优化问题变得很简单，直接将多层非线性网络参数趋0。

实际中，identity mappings 不太可能是最优的，但是上述改写问题可能能帮助预处理问题。如果最优函数接近identity mappings，那么优化将会变得容易些。 实验证明该思路是对的。

3.2. Identity Mapping by Shortcuts   
图2为一个模块。A building block   
公式定义如下：   
  
这里假定输入输出维数一致，如果不一样，可以通过 linear projection 转成一样的。

3.3. Network Architectures   
  


Plain Network 主要是受 VGG 网络启发，主要采用3\*3滤波器，遵循两个设计原则：1）对于相同输出特征图尺寸，卷积层有相同个数的滤波器，2）如果特征图尺寸缩小一半，滤波器个数加 倍以保持每个层的计算复杂度。通过步长为2的卷积来进行降采样。一共34个权重层。   
需要指出，我们这个网络与VGG相比，滤波器要少，复杂度要小。

Residual Network 主要是在 上述的 plain network上加入 shortcut connections

3.4. Implementation   
针对 ImageNet网络的实现，我们遵循【21,41】的实践，图像以较小的边缩放至[256,480]，这样便于 scale augmentation，然后从中随机裁出 224\*224，采用【21,16】文献的方法。

4 Experiments   
