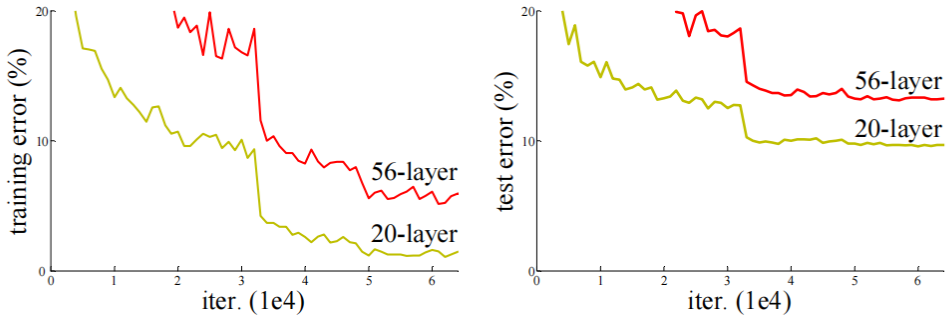
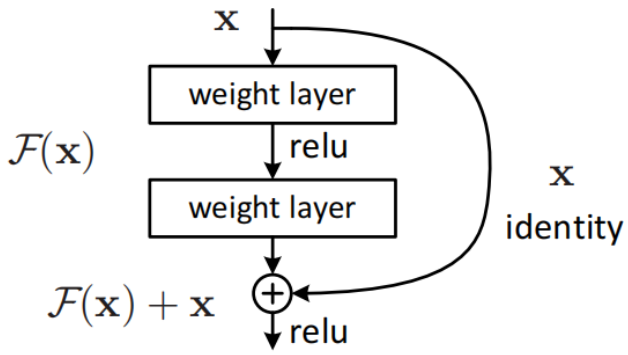
总结

ResNet 提出的背景是发现当神经网络很深的时候，其效果并不是期望中的那么好，模型的效果不升反降（模型退化），出现的原因是因为梯度消失或者梯度爆炸现象么？答案是否定的，因为梯度消失/爆炸现象已经被 Batch Normalization 很好地解决了。出现模型退化问题的原因是过拟合么？答案也是否定的，因为如果发生了过拟合，那么模型的训练误差会很小，测试误差很大。从下图可以看出，相比于 20 层的网络，56 层的网络训练误差和测试误差都很大，所以判断模型出现退化的原因并不是过拟合。



那么接下来思考，如果让深层网络较高的层学习恒等映射，那网络就相当于一个浅层网络，这样深层网络的效果是不是至少不会比浅层网络的效果差？于是 ResNet 最主要的学习思想就产生了，即残差思想。

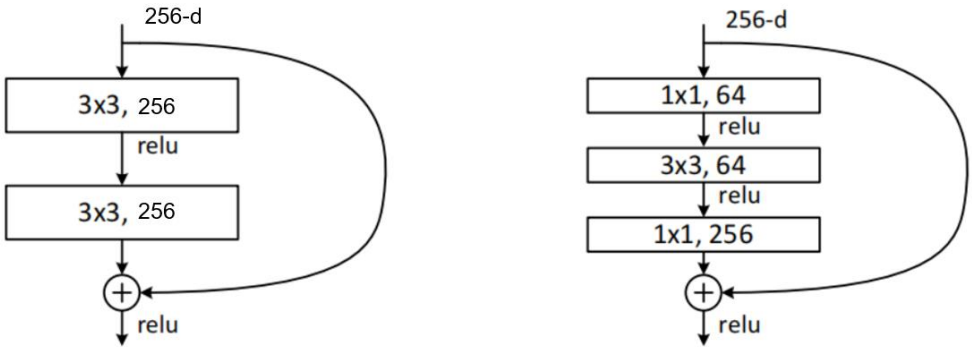


如上图所示，网络在没有加跳跃连接的时候，需要学习的是 $F(x)$，再加上跳跃连接后输出的是 $H(x)=F(x)+x$，x 是原本的输入，不需要网络学习，故现在网络需要学习的是 $F(x)=H(x)-x$。举一个更直观的例子：

假设网络输入 x=3，没有跳跃连接时网络输出 $F(3)=3.5$，引入跳跃连接后网络输出是 $H(3)=3.5=F'(3)+3$，那么 $F'(3)=0.5$，网络只需要学习输出和输入差值的那一部分，也就是例子中的 $F'(3)=0.5$ 这一部分，学习目标得以简化。

方法

①两种残差结构



关于残差结构作者设计了两种针对不同深度网络的残差块，叠加相应的残差块即可实现网络的搭建。首先图中左边这个结构是通过堆叠 2 个 3×3 卷积层实现残差函数，用于训练 ResNet18/34。

考虑到训练更深的网络所需付出的训练时间更多，又设计了一种瓶颈结构用于训练 ResNet50/101/512。在瓶颈结构中，残差函数通过堆叠 1×1、3×3、1×1 卷积层得以实现，其中 1×1 主要起到调整维数的作用。

对比左右两图的结构，我们可以发现，即使右图结构层数比左图多，但是这两种不同残差块有着相近的时间复杂度，且在输入维度相同的情况下，右图的参数量比左图小。

②网络结构

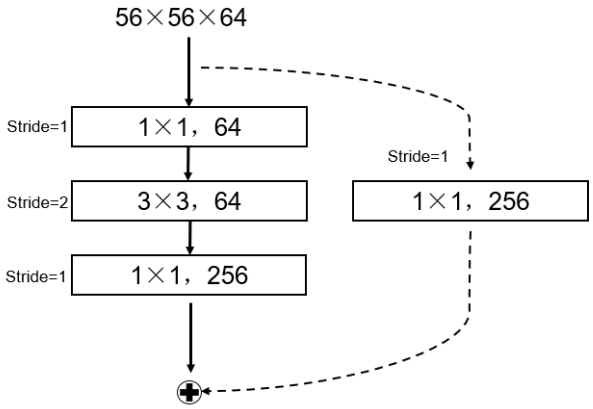
下图是不同层数的 ResNet 网络结构框架，可以看出是由残差块构建而成。

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

不同层数的 ResNet 都是先通过了 7×7 的卷积，再通过 3×3 的最大池化操作，然后都设计有四个残差块，不同的是残差块的参数设置以及残差块结构不同。

在代码实现中因为残差跳跃连接部分必须保证所叠加的两个特征图的形状相同，通过观察 ResNet50 结构可以发现在 conv2_1 部分，经过 3×3 的最大池化操作后，特征图的形状为 56×56×64 而此时该残差块的输出特征图的形状为 56×56×256，这两部分的通道数不同，如果使用实线的恒等映射连接则不能直接相加，为此还有一种虚线连接结构，也就是原论文提到的 OptionB，其结构示意

图如下图所示(以 Conv2_1 为例)。



在 50/101/152 层网络中的 conv2_1、conv3_1、conv4_1、conv5_1 都采用了虚线连接的结构，而在 18/34 层网络结构中 conv2_1 采用的还是实线恒等映射连接。

③跳跃连接部分

残差块中跳跃连接部分都设计为恒等映射，关于跳跃连接这部分的设计作者在《Identity Mapping in Deep Residual Networks》中进行了详细的阐述。

通过将恒等映射更换为"scaling"、"gating"、"1×1 convolution"、"dropout"，并做了对比分析，非恒等映射形式引入了更多的参数，这理应使得网络具有更强的拟合能力，但是网络的性能出现了退化，因此退化的原因不是拟合能力的不足，而是优化问题，最终得出结论：设计为恒等跳跃连接更有利于优化。