# Deep Residual Learning for Image Recognition

## 1. 思想

作者根据输入将层表示为学习**残差函数**。实验表明，残差网络更容易优化，并且能够通过增加相当的深度来提高准确率。
核心是解决了增加深度带来的副作用（梯度弥散或梯度爆炸），这样能够通过单纯地增加网络深度，来提高网络性能。

- 作者在ImageNet上实验了一个152层的残差网络，比VGG深8倍，取得了3.57%的错误率。
- 作者通过一系列实验证明了表示的深度（即网络的深度）对很多视觉识别任务都至关重要。仅仅由于使用了非常深的网络，作者就在COCO目标检测数据集上获得了**28%**的相对提升。

---

## 2. 笔记

**网络的深度为什么重要？**
因为CNN能够提取low/mid/high-level的特征，网络的层数越多，意味着能够提取到不同level的特征越丰富。并且，越深的网络提取的特征越抽象，越具有语义信息。

**为什么不能简单地增加网络层数？**

- 对于原来的网络，如果简单地增加深度，会导致梯度弥散或梯度爆炸。

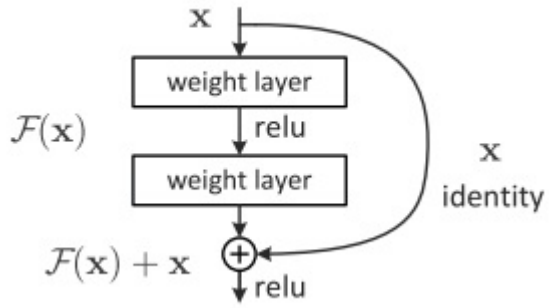
  对于该问题的解决方法是**正则化初始化**和**中间的正则化层（Batch Normalization）**，这样的话可以训练几十层的网络。

- 虽然通过上述方法能够训练了，但是又会出现另一个问题，就是**退化问题**，网络层数增加，但是在训练集上的准确率却饱和甚至下降了。这个不能解释为overfitting，因为overfit应该表现为在训练集上表现更好才对。
  退化问题说明了深度网络不能很简单地被很好地优化。

作者通过实验：通过浅层网络+ y=x 等同映射构造深层模型，结果深层模型并没有比浅层网络有等同或更低的错误率，推断退化问题可能是因为深层的网络并不是那么好训练，也就是求解器很难去利用多层网络拟合同等函数。

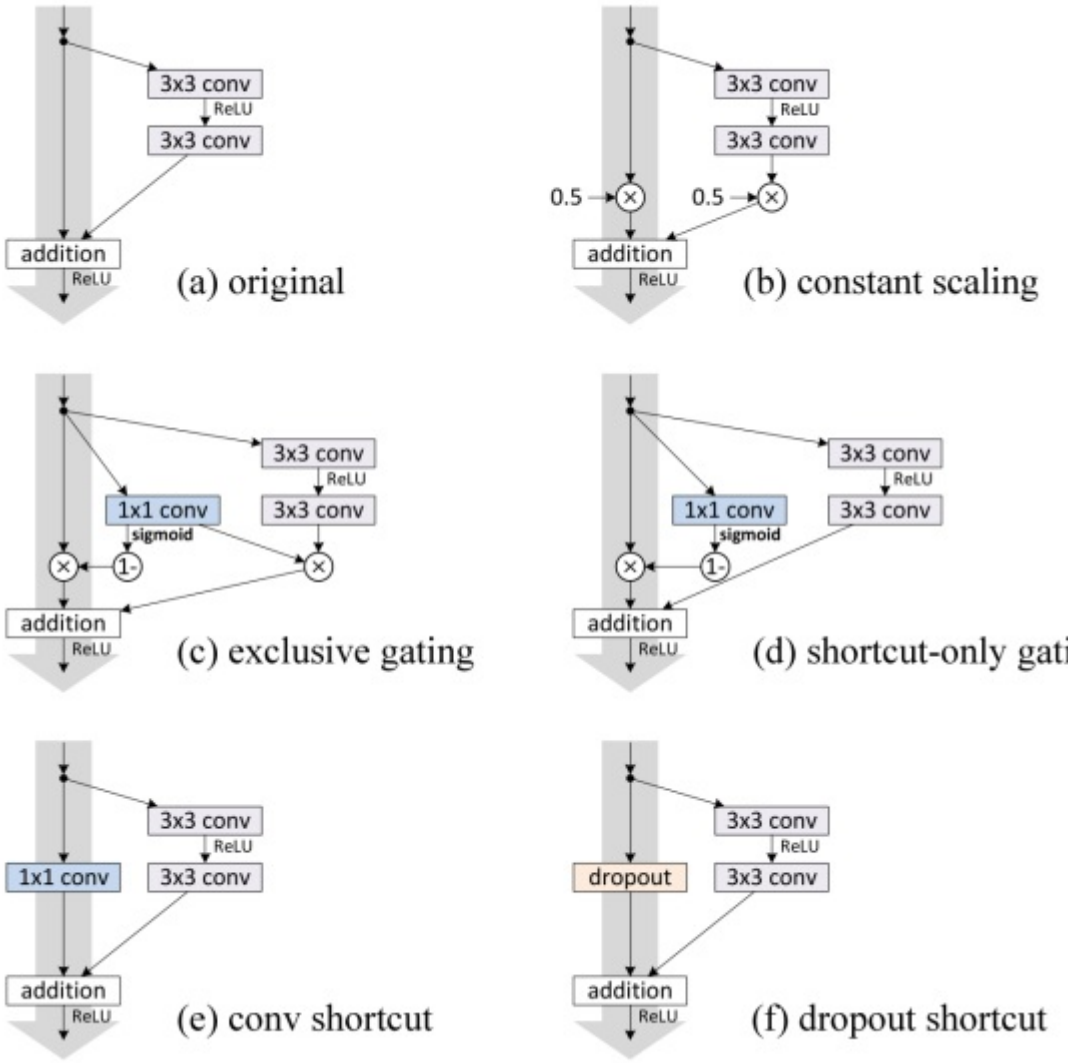
### 怎么解决退化问题？

深度残差网络。如果深层网络的后面那些层是恒等映射，那么模型就退化为一个浅层网络。那现在要解决的就是学习恒等映射函数了。 但是直接让一些层去拟合一个潜在的恒等映射函数H(x) = x，比较困难，这可能就是深层网络难以训练的原因。但是，如果把网络设计为H(x) = F(x) + x,如下图。我们可以转换为学习一个残差函数F(x) = H(x) - x. 只要F(x)=0，就构成了一个恒等映射H(x) = x. 而且，拟合残差肯定更加容易。




**其他的参考解释**

- F是求和前网络映射，H是从输入到求和后的网络映射。比如把5映射到5.1，那么引入残差前是F'(5)=5.1，引入残差后是H(5)=5.1, H(5)=F(5)+5, F(5)=0.1。这里的F'和F都表示网络参数映射，**引入残差后的映射对输出的变化更敏感**。比如s输出从5.1变到5.2，映射F'的输出增加了1/51=2%，而对于残差结构输出从5.1到5.2，映射F是从0.1到0.2，增加了100%。明显后者输出变化对权重的调整作用更大，所以效果更好。残差的思想都是去掉相同的主体部分，从而突出微小的变化，看到残差网络我第一反应就是差分放大器...地址
- 至于为何shortcut的输入时X，而不是X/2或是其他形式。kaiming大神的另一篇文章[2]中探讨了这个问题，对以下6种结构的残差结构进行实验比较，shortcut是X/2的就是第二种，结果发现还是第一种效果好啊（摊手）。




这种残差学习结构可以通过前向神经网络+shortcut连接实现，如结构图所示。而且shortcut连接相当于简单执行了同等映射，不会产生额外的参数，也不会增加计算复杂度。 而且，整个网络可以依旧通过端到端的反向传播训练。

ImageNet上的实验证明了作者提出的加深的残差网络能够比简单叠加层生产的深度网络更容易优化，而且，因为深度的增加，结果得到了明显提升。另外在CIFAR-10数据集上相似的结果以及一系列大赛的第一名结果表明ResNet是一个通用的方法。

---

# 相关的工作

- **残差表示**
  **VALD，Fisher Vector**都是是对残差向量编码来表示图像，在图像分类，检索表现出优于编码原始向量的性能。
  在low-level的视觉和计算机图形学中，为了求解偏微分方程，广泛使用的**Multigrid**方法将系统看成是不同尺度上的子问题。每个子问题负责一种更粗糙与更精细尺度的残差分辨率。**Multigrid**的一种替换方法是层次化的预处理，层次化的预处理依赖于两种尺度的残差向量表示。实验表明，这些求解器要比对残差不敏感的求解器收敛更快。

- **shortcut连接**
  shortcut连接被实验和研究了很久。**Highway networks**也使用了带有门函数的shortcut。但是这些门函数需要参数，而ResNet的shortcut不需要参数。而且当**Highway networks**的门函数的shortcut关闭时，相当于没有了残差函数，但是ResNet的shortcut一直保证学习残差函数。而且，当**Highway networks**的层数急剧增加时，没有表现出准确率的上升了。总之，ResNet可以看成是**Highway networks**的特例，但是从效果上来看，要比**Highway networks**好。

---

# 深度残差学习

- **残差学习**
  根据多层的神经网络理论上可以拟合任意函数，那么可以利用一些层来拟合函数。问题是直接拟合H(x)还是残差函数，由前文，拟合残差函数F(x) = H(x) - x更简单。虽然理论上两者都能得到近似拟合，但是后者学习起来显然更容易。
  作者说，这种残差形式是由**退化问题**激发的。根据前文，如果增加的层被构建为同等函数，那么理论上，更深的模型的训练误差不应当大于浅层模型，但是出现的**退化问题**表面，求解器很难去利用多层网络拟合同等函数。但是，残差的表示形式使得多层网络近似起来要容易的多，如果同等函数可被优化近似，那么多层网络的权重就会简单地逼近0来实现同等映射，即F(x) = 0。
  实际情况中，同等映射函数可能不会那么好优化，但是对于残差学习，求解器根据输入的同等映射，也会更容易发现扰动，总之比直接学习一个同等映射函数要容易的多。根据实验，可以发现学习到的残差函数通常响应值比较小，同等映射（shortcut）提供了合理的前提条件。

- **通过shortcut同等映射**

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}.$$

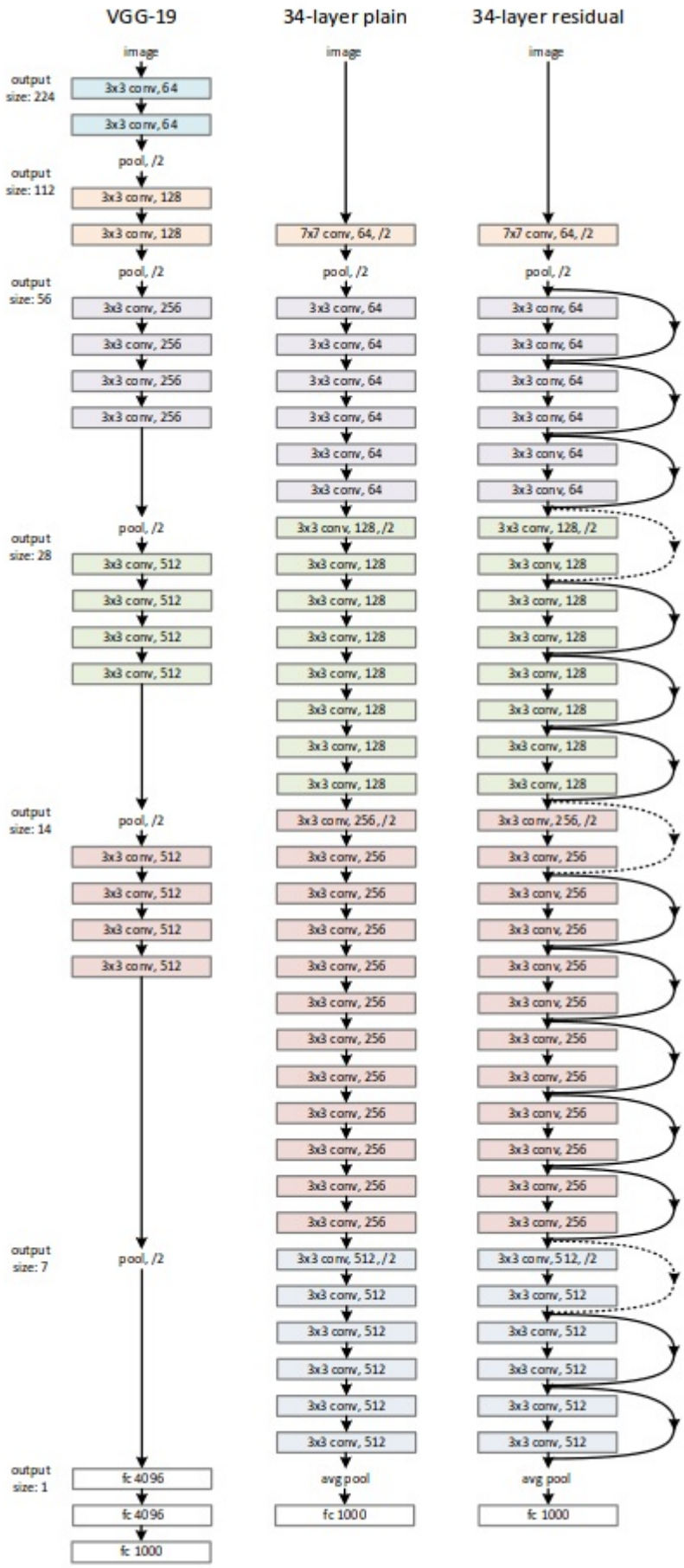
$$\mathcal{F} = W_2\sigma(W_1\mathbf{x})$$

  F(x)与x相加就是就是逐元素相加，但是如果两者维度不同，需要给x执行一个线性映射来匹配维度：

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}.$$

  用来学习残差的网络层数应当大于1，否则退化为线性。文章实验了layers = 2或3，更多的层也是可行的。
  **用卷积层进行残差学习：**以上的公式表示为了简化，都是基于全连接层的，实际上当然可以用于卷积层。加法随之变为对应channel间的两个feature map逐元素相加。

- **网络结构**
  作者由VGG19设计出了plain 网络和残差网络，如下图中部和右侧网络。然后利用这两种网络进行实验对比。


VGG-19
34-layer plain
34-layer residual
image
output size: 224
3x3 conv, 64
3x3 conv, 64
pool, /2
output size: 112
3x3 conv, 128
3x3 conv, 128
pool, /2
output size: 56
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
output size: 28
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 14
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 7
fc 4096
fc 4096
fc 1000
output size: 1
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 256, /2
3x3 conv, 512, /2
avg pool
fc 1000

**key point：**
设计网络的规则：1.对于输出feature map大小相同的层，有相同数量的filters，即channel数相同；2. 当feature map大小减半时（池化），filters数量翻倍。
对于残差网络，维度匹配的shortcut连接为实线，反之为虚线。维度不匹配时，同等映射有两种可选方案：

1. 直接通过zero padding 来增加维度（channel）。
2. 乘以W矩阵投影到新的空间。实现是用1x1卷积实现的，直接改变1x1卷积的filters数目。这种会增加参数。

- **实施**
  **key point：**

  训练测试的multi-scale，BN，color augmentation. 测试时的10-cut.

---

# 实验

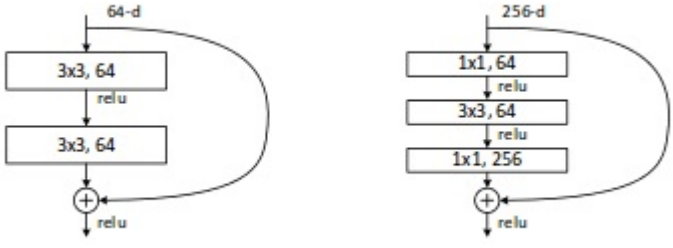
**key point：**

1. 实验了plain-18和plain-34，展示了退化问题。说明了退化问题不是因为梯度弥散，因为加入了BN。另外也不能简单地增加迭代次数来使其收敛，增加迭代次数仍然会出现退化问题。

2. 实验了ResNet-18和ResNet-34不会出现退化问题，ResNet-34明显表现的比ResNet-18和plain-34好，证明了残差学习解决了随网络深度增加带来的退化问题。 而且同等深度的plain-18和ResNet-18，残差网络更容易优化，收敛更快。

3. 对于同等映射维度不匹配时，匹配维度的两种方法，zero padding是参数free的，投影法会带来参数。作者比较了这两种方法的优劣。实验证明，投影法会比zero padding表现稍好一些。因为zero padding的部分没有参与残差学习。实验表明，将维度匹配或不匹配的同等映射全用投影法会取得更稍好的结果，但是考虑到不增加复杂度和参数free，不采用这种方法。

4. **更深的瓶颈结构:**




作者探索的更深的网络。 考虑到时间花费，将原来的building block(残差学习结构)改为瓶颈结构，如上图。首端和末端的1x1卷积用来削减和恢复维度，相比于原本结构，只有中间3x3成为瓶颈部分。这两种结构的时间复杂度相似。此时投影法映射带来的参数成为不可忽略的部分（以为输入维度的增大），所以要使用zero padding的同等映射。
替换原本ResNet的残差学习结构，同时也可以增加结构的数量，网络深度得以增加。生成了ResNet-50，ResNet-101，ResNet-152. 随着深度增加，因为解决了退化问题，性能不断提升。

作者最后在Cifar-10上尝试了1202层的网络，结果在训练误差上与一个较浅的110层的相近，但是测试误差要比110层大1.5%。作者认为是采用了太深的网络，发生了过拟合。

5. 最后作者把ResNet用到了其他比赛上，拿了很多冠军...