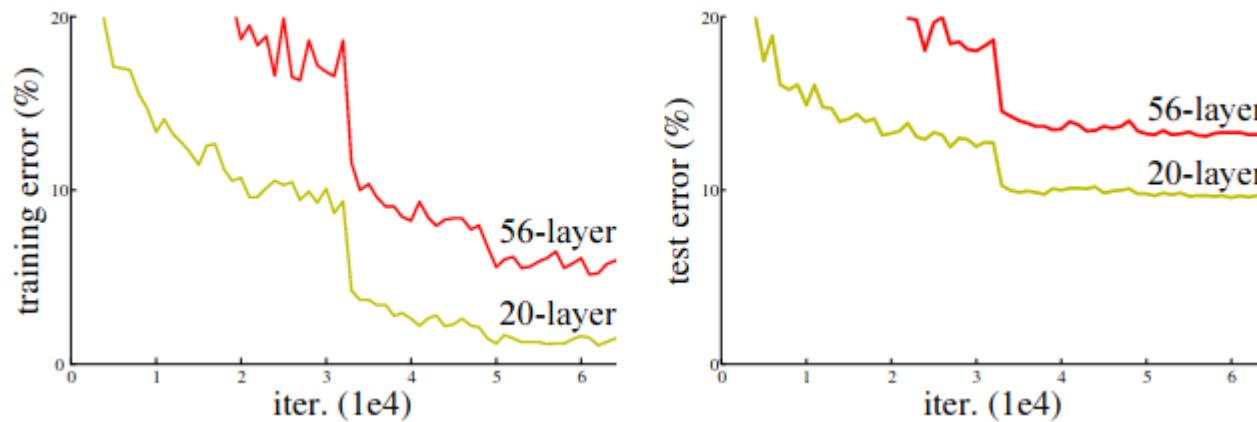


# ResNet

## 1. 要解决什么问题？

首先，ResNet提出之前，He发现了什么问题？根据经验我们知道，网络的深度对模型性能的提升很重要。从经验来看，网络越深，就可以进行更复杂的特征提取，能够拟合更复杂非线性关系，所以理论上来说，更深的网络肯定会表现出一个更好的性能。当然在一定程度上的加深网络结构确确实实会提升网络的性能，但是当网络的深度增加到一定的程度继续增加的时候，就会出现模型性能下降的现象，也就是**退化问题**。

针对该问题，首先想到的是过拟合，但是训练过程中的训练误差同样变差了，如下图。所以说也不是过拟合的问题。



其次是，梯度消失或者梯度爆炸？但是一些技术手段像BatchNorm和网络初始化（Kaiming初始化）可以缓解这个问题，所以为什么会出现这个问题呢？有点离谱。

ResNet论文对这个问题的解释是：

这个问题的核心在于**优化困难**，而不是表达能力不足。理论上，更深的网络应该至少能达到浅层网络的性能（通过将额外的层设置为恒等映射），但实际训练中，深层网络却表现更差。

**关键原因：**深层网络很难学习到恒等映射（identity mapping）。当网络层数增加时，优化器很难找到合适的参数让新增的层简单地传递输入（即学习恒等函数  $F(x) = x$ ）。

也就是说，假设我现在有一个网络A，我增加他的深度，在A的后面多加了几层得到网络B。理论上来说，如果我多加几层网络什么也不干（也就是恒等映射），A输出以后经过这几层之后也就是B的输出应该和A的输出一样，至少能保持性能不变，但是事实上不仅没有保持性能不变，反而出现了退化问题。说明，新增加的这几层难以学习到一个恒等映射，主要原因是这些层的输出经过非线性激活再想要一个恒等映射，简直是太难了。**什么都不做**恰好是当前神经网络最难做到的东西之一。

## 2. 解决方案

ResNet的解决方案：

ResNet引入了**残差连接**（residual connection）或**跳跃连接**（skip connection），将学习目标从  $H(x)$  改为学习残差  $F(x) = H(x) - x$ 。

这样做的好处是：

- 如果恒等映射是最优的，网络只需要将  $F(x)$  学习为0即可，这比直接学习恒等映射容易得多
- 梯度可以通过跳跃连接直接反向传播，缓解了深层网络的梯度问题
- 网络可以自适应地决定是否使用这一层的变换

用公式表示就是： $H(x) = F(x) + x$ ，那么网络要学习的函数就变为残差函数  $F(x) = H(x) - x$ ，也就是说不再去学习输入到输出的映射，而是学习输出减去输入这个差值。

使用数学语言描述：

残差块的数学公式：

$$y = F(x, Wi) + x$$

其中， $F = W_2\sigma(W_1x)$ 就是我们要学习的函数映射，他多个层的堆叠，论文中使用两个层W1和W2， $\sigma$ 代表激活函数Relu。

注意：一个Block块至少包含两个层，若只有一个层，则是：

$$y = F(x, Wi) + x = (W_1x) + x = (W_1 + 1)x$$

显然这样相当于只是对这一层的权重做了一些调整，没有跳跃连接。

不是寻找输入到输出的映射  
而是寻找到“输出减输入”

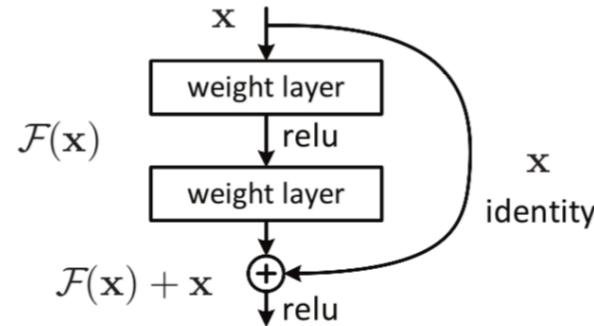


Figure 2. Residual learning: a building block.

形象的说：好比，我现在写完了一篇会议文章，现在导师让我把会议修改成一篇期刊。两种方案：

- 我完全重新开始写，就是学习  $H(x)$ ，学习输入到输入的完整映射关系。
- 在原先会议文章  $x$  的基础之上稍作修改，得到最终的  $H(x)$ ，那我需要学习的就是如何在  $X$  上做的修改，也就是  $H(x)-X$ ，学习要修改什么，也就是这个偏差。肯定这个更简单。

另一方面，ResNet基本上解决了梯度消失的问题：

下图为残差连接的梯度推导：

$$\frac{\partial(f(x)+x)}{\partial x} = \frac{\partial f(x)}{\partial x} + 1$$

$$h(x) = f(x) + x$$

$$z(x) = g(h(x)) + h(x)$$

$$\frac{dz}{dx} = \frac{dz}{dh} \cdot \frac{dh}{dx}$$

$$= \left[ \frac{d}{dh} (g(h) + h) \right] \cdot \left[ \frac{d}{dx} (f(x) + x) \right]$$

$$= [g'(h) + 1] \cdot [f'(x) + 1]$$

$$= g'(h) \cdot f'(x) + f'(x) + g'(h) + 1$$

可以看到残差连接不论多少个残差块的堆叠都会存在一个梯度“1”，从而避免反向传播中梯度消失的问题。

思考：假如残差连接不是  $X$ ，而是  $\lambda X$ ，会发生什么？

单个:

$$\frac{\partial f(x) + \lambda x}{\partial x} = \frac{\partial f(x)}{\partial x} + \lambda.$$

两个:

$$h(x) = f(x) + \lambda x$$

$$z(x) = g(h(x)) + \lambda h(x)$$

$$\frac{dz}{dx} = \frac{dz}{dh} \cdot \frac{dh}{dx}$$

$$= \left[ \frac{d}{dh} (g(h) + \lambda h) \right] \cdot \left[ \frac{d}{dx} (f(x) + \lambda x) \right]$$

$$= [g'(h) + \lambda] \cdot [f'(x) + \lambda]$$

$$= g'(h) \cdot f'(x) + g'(h) \cdot \lambda + f'(x) \cdot \lambda + \lambda^2.$$

可以看到，叠加k个残差块，会存在一个梯度  $\lambda^k$ ，随着叠加块的数量k的增加。若  $\lambda > 1$ ，则会出现梯度爆炸，若  $\lambda < 1$ ，则会出现梯度消失的问题。

### 3. Why it works?

- 首先是让网络学习恒等映射很困难，而残差连接让网络的学习目标学习残差，也就是让多个层的堆叠去学习输出0，更简单。
- 另一个原因那就是跳连接相加可以实现不同分辨率特征的组合，因为浅层容易有高分辨率但是低级语义的特征，而深层的特征有高级语义，但分辨率就很低了。[\[2\]](#)
- 让模型自身有了更加“灵活”的结构，即在训练过程本身，模型可以选择在每一个部分是“更多进行卷积与非线性变换”还是“更多倾向于什么都不做”，抑或是将两者结合。
- 即使BN过后梯度的模稳定在了正常范围内，但梯度的相关性实际上是随着层数增加持续衰减的。而经过证明，ResNet 可以有效减少这种相关性的衰减。[\[1\]](#)

**总结：ResNet通过引入残差连接的方式解决了深层网络的退化问题。**

参考阅读：

- <https://www.zhihu.com/question/64494691> (较详细)
- <https://zhuanlan.zhihu.com/p/556521788>
- [https://zyc.ai/sketch/career/interview\\_resnet/](https://zyc.ai/sketch/career/interview_resnet/)
- <https://www.cnblogs.com/boligongzhu/p/15085678.html>
- [1] The Shattered Gradients Problem: If resnets are the answer, then what is the question?
- [2] Lin T Y , Dollár, Piotr, Girshick R , et al. Feature Pyramid Networks for Object Detection[J]. 2016.

关于一些面试中可能会问到的问题：

- ResNet主要解决什么问题？为什么会有 ResNet？

ResNet 主要解决的是深度神经网络的“退化问题”。当网络加深到一定程度后，训练误差不仅不下降，反而会上升。ResNet 通过**残差学习（Residual Learning）** 的方式，让深层网络更容易优化，从而使非常深的网络可以被成功训练。

- 深度网络退化的原因

退化不是因为过拟合，而是优化困难。深层网络要拟合一个复杂映射  $H(x)$ ，难度很高；而且反向传播的梯度容易消失，使得前面的层无法更新。在这种条件下，越深的网络越难训练，就出现了退化现象。

- ResNet 针对退化问题提出的残差网络思想

ResNet 把原来的映射  $H(x)$  换成了残差映射  $F(x)=H(x)-x$ 。如果最优映射接近恒等映射，网络只需要学习一个“微调”，即  $F(x)=0$ 。同时 shortcut 结构让梯度能够直接传到前面层，避免梯度消失。这样就解决了深度网络难训练和退化的问题。

- ResNet 中下采样如何实现？

ResNet 的下采样发生在每个 stage 的第一个 block 中，通过**stride=2 的卷积**实现。同时，为了让 shortcut 分支尺寸匹配，会用一个**stride=2 的 1×1 卷积**做下采样。整个 network 的空间尺度会从  $224 \rightarrow 112 \rightarrow 56 \rightarrow 28 \rightarrow 14 \rightarrow 7$ 。

ResNet 主要解决的是

**深度神经网络的“退化问题”**

◦

当网络加深到一定程度后，训练误差不仅不下降，反而会上升。

ResNet 通过

**残差学习 (Residual Learning)**

的方式，让深层网络更容易优化，从而使非常深的网络可以被成功训练。