## Deep Residual Learning for Image Recognition

# 摘要：

更深的神经网络更难训练。我们提出了一种残差框架来使得深度超过以前的网络更易于训练。我们将层重新定义为学习关于输入的残差函数，而不是学习无关的函数。我们提供了多方面的经验性的证据表明这些残差网络更加易于优化，并且通过显著增加深度之后能获得准确度。在ImageNet上我们使用深度高达152层（8倍于VGG网络，但是复杂度却更低）的网络来验证。这些残差网络整体在ImageNet实现了3.57%的错误率。这个结果赢得了2015分类比赛的第一。我们也在CIFAR-10使用100和1000层的网络进行了分析。

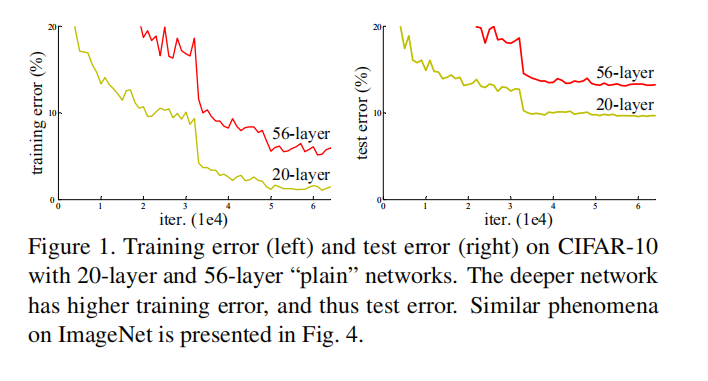
在许多视觉任务中，表征的深度是至关重要的。仅仅因为我们极大地深度，我们在COCO目标检测上面获得了28%的相对提升。深度残差网络是我们在2015 ILSVRC和COCO比赛上提交模型的基础，这些提交获得了多项第一....略

# 引言：

深度CNN引领了一系列图片检测任务上的突破。深度网络在一个端到端的多层的形式中，自然的集成了低/中/高水平的特征和分类器，并且特征的“级别”可以通过堆叠的层（深度）来丰富。**近期的证据揭示出网络深度是起决定性作用的，就是近期的在ImageNet上的领导性的结果都开发采用了非常深的模型，深度从16到30层。**许多其他的重要的视觉识别任务也很大的受益于非常深的模型。

由深度的重要性驱动，一个问题产生了：是不是学习更好的网络就像堆叠更多的层一样简单。（注：前面这句意思是，是否能简单的通过增加深度来增加模型能力呢？）回答这个问题的（达到这个目的）一个障碍是臭名昭著的梯度消失/爆炸问题。但是，这个问题很大程度上已经被批正则化和中间正则化层技术解决了，它们是的数十层的网络利用SDG在反向传播中收敛。

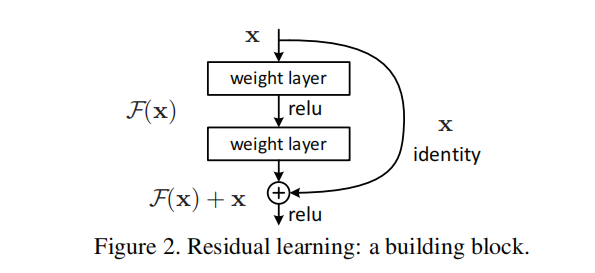
当更深的网络能够收敛，退化问题又出现了：随着网络深度的增加，准确度变得饱和（这不令人惊讶），然后迅速的下降。出乎意料的，这样的下降不是由于过拟合导致，冰洁添加更多的层给一个合适的深度模型导致更高的训练错误，正如11 12 中报告的以及我们的实验中彻底证明了的。图片1中展示了典型的例子：



训练准确度的下降表明不是所有的系统都能类似的简单的进行优化。让我们考虑一个较浅的架构然后在它上面增加更多的层形成一个更深的版本。这里有一个通过构造更深的模型的解决问题的方法，添加的层是恒等映射。增加的层只是拷贝那个浅层网络学到的层。构造出更深的模型这个解决方案中的不会产生比浅层对应网络的更高的训练误差。但是实验显示当前我们手边没有这种效果等于或优于构造方案的解决方案（或者在可行的时间上没有）出现。（注：**巨大的成果来源于很朴素的想法，也就是更深层的网络的能取得更好效果的网络是存在的，但是目前我们还却没有训练出来，这是为什么呢？由此想到可能是我们哪里做得不正确导致的，**比如训练方法不对，或者当前的网络结构无法训练出来。**后面3.1中进一步的说明了，我们无法训练出的是什么呢，恒等映射我们就训练不出来，于是想到可以做某种改变，使得我们的层能够被训练出恒等映射。从训练模式上入手比较难，也不太可能，因为一般的各种训练模式大家都已经覆盖尝试过了没用，但是改变模型结构是容易的。我直接加上一条短连接，把堆叠栈的输入x加到输出y上面，然后训练原来的部分学习0，这个能做到的，那么就能训练出恒等映射了，于是残差网络的核心就出来了。后面更具体的结合相关知识的分析**）

本文中，我们提出一种**深度残差学习框架**来解决精度下降问题。不是希望每个堆叠层直接适配基础的期望的映射（注：一种形式下的真实函数），而是显式让它适配一个残差映射。形式上，将期望的底层映射记为H(x),我们让堆叠的非线性层适配另一个映射F(x) = H(x) - x. 此时原始的映射被转换为H(x) = F(x)+x.我们假设优化残差映射比优化原始的映射更容易。极端情况下，如果恒等映射是最优的，那么将残差推到0比起用非线性堆叠层去适配一个恒等映射是更容易的。

表达式F(x)+x可以用带有短连接的前馈神经网络实现，如图2。短连接就是图中跳过1层或多层的那个部分。在我们的情况中，短连接只是简单的做恒等映射，他们的输出和堆叠



层的输出相加。恒等的短连接既不增加额外的参数也不增加计算复杂度。整个网络仍然能够实用端到端的SGD优化的反向传播进行训练，用常见的框架就能轻松实现（比如caffe），

并且不用修改solver(注：解算器？）

我们在ImageNet上做了广详尽的实验来展示退化问题的应对效果和评价我们的模型。结果显示：1）我们的极深的残差网络易于优化，但是对应的一般网络会随着深度增加产生更高的训练错误。2）我们的深度残差网络能够轻易的从大幅增加的深度中获益，比起以前的网络产生好得多的结果。

在CIFAR-10上的结果也是类似的，这显示优化困难（注：退化问题）和我们的方法的（注：对这个问题的）有效性不是仅仅适用于特殊数据集。我们展示了在这个数据集的使用100层的训练成功的模型，甚至探索了1000层的模型。

在ImageNet分类数据及上，我们利用极深的残差网络取得了优秀的结果。我们的152层的残差网络是在ImageNet上出现过的最深的，同时相比VGG有更低的复杂度。我们总体前5错误率是3.57%，赢得了ILSVRC 2015 分类比赛的第一。**极深的表示（注：就是深度的特征图）在其它识别任务中也拥有优秀的泛化能力。**使得我们赢得了其它比赛的first place。xxx比赛略。 这个有力的证据表明，残差学习的原理是通用的，我们期望他可以适用于其它视觉和非视觉任务。

# 相关工作：

**残差表示：**在图片识别领域，VLAD是一种使用根据字典使用残差向量编码进行编码的表示。（注：深度学习之前的图像特征提取的一种方法）而Fisher向量可以被表示为VLAD的概率版本。他们两个都是图片检索和分类领域的有力的浅层表示。对于向量量化来说，编码残差向量比编码原始向量更有效。

在低层次视觉和计算机图像学领域，为了解决偏微分方差，广泛使用的多重网格法将系统重新阐述为一个多尺度的子问题，每个子问题复杂一个较粗和较细粒度之间的剩余解。多重网格法的一种替代是分层基础预处理法，这种方法依赖于代表两个尺度的残差向量的变量。已经证明这些解决方法收敛起来比没有利用残差性质的标准解决方法快。这些方法表明一个好的重新表述或者预处理能够简化优化过程。

**短连接：**实践和理论都引起短连接已经被研究了很长时间。一个早期的训练多层网络的实践是从输入到输出加上一个线性层。在【44，24】中，少量中间层直接连接到辅助的分类器中来解决梯度消失或爆炸问题。论文[39，38，31，47]的论文提出了用快捷连接实现的对中层响应、梯度和传播误差的方法。在【44】中，一个Inception层是由一个短的分支和少量更深的分支组成。

与我们工作同时进行的，, “highway networks” [42, 43]提出使用门函数的短连接。这些们是依赖数据并且有参数，相等的，我们的是恒等捷径，没有参数。当一个门捷径关闭，高速路网络就是无残差的函数。相反的，我们的表达式残差函数是一直存在要学习的；我们么的恒等捷径不会关闭，所有的信息总是会传送，用增加的残差函数被学习到。另外，高速路网络没有演示出通过极大的深度增加获得了精度上的收益。

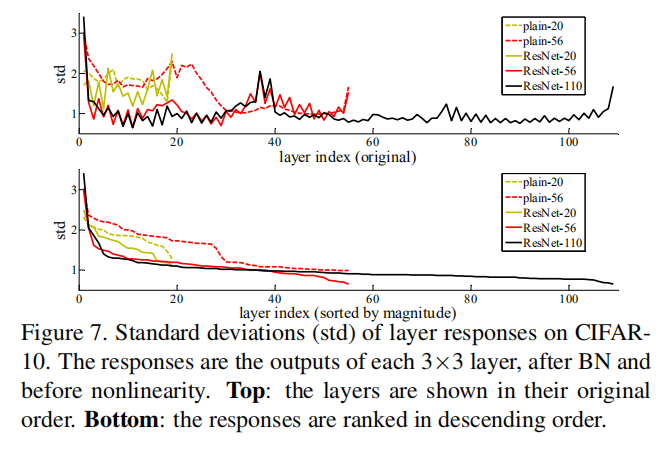
# 深度残差学习

## 3.1 残差学习

让我们考虑H(x)作为一个堆叠层要学习基础的映射（不一定是整个网络），X代表这些层（注：堆叠栈的那些层）的第一层的输入。如果假设多层非线性层能够渐进的估计复杂函数，那么它等价于假设这个多层非线性层可以渐进的估计残差函数，如H(x) - x (假设输入和输出尺寸相同）。所以我们现在用这些层去近似残差函数F(x)=H(x)-x，而不是去近似H(x)。原始的函数因此变成了F(x)+x。尽管两种形式都应该能渐进的近似到期望的函数（注：真实函数），但是他们的容易程度是不同的。

这种重构形式是受到退化问题的反直觉现象启发而想到的。在引言中我们讨论到，如果增加的层是恒等映射，那么更深的模型的训练错误不应该比它的浅层对应版本要大。**退化问题暗含了这样一个道理，前面的模型结构难以学到在多个非线性层间的恒等映射。**在残差学习的表达方式下，如果恒等映射是最优的，那么求解器（指残差函数部分？）可以简单的将权重参数推向0来接近恒等映射。

在实际中，不太可能恒等映射是最优的，但是我们的重构形式可以帮助预处理(我=解决？）这个问题。如果最优的函数更接近恒等映射而不是0映射，对于求解器来说找到关于恒等映射的扰动比找到一个形函数更容易。我们的实验（图片7）显示学习到的残差函数通常具有小的响应，表明恒等映射提供了合理的预处理（解决）。（注：？不太理解）



3.2 短连接进行恒等映射

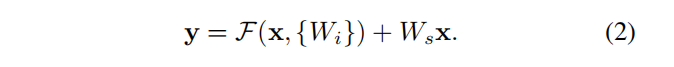
我们对没几层进行残差学习。一个构建块如图2所示。形式上，本文中，我们考虑将一个构建块定义为



x，y是被考虑的几个层的输入和输出。函数F(x,{Wi})代表学习到的残差映射。比如图2中的例子，残差函数有两层，F=W2δ(W1x),δ表示Relu函数，为了简化问题，忽略偏置。操作F+x通过短连接执行，它是元素加。我们将非线性变化用在相加之后。（比如图2对应的δ(y)）.

等式1中的短连接不引入额外的参数也不引入额外的计算复杂度。这不仅在实际中很有吸引力，并且对我们将残差网络和其它一般网络进行比较也是重要的。我们可以在具有相同的参数量，深度，宽度，和计算花费这样的公平的条件下比较残差网络和其它一般网络。

x和F的尺寸必须一样。如果不是的话（比如改变了输入/输出的的通道），我们可以对短连接执行线性投射Ws来匹配尺寸。



我们也可以在等式1中使用方形矩阵Ws。（注：不是上面那个Ws，意思应该是指做更复杂的变化）。但是我们后面的实验会显示恒等映射能够有效的解决退化问题并且是很经济的（注：节省参数，算力等资源），所以Ws仅仅用于匹配尺寸。

残差函数F是灵活的。本文的实验中F包含两个或3个层，同时多个层是可能的。但是如果F只有一个层，等式1就会类似于一个线性层：y=W1x+x，这种形式我们没有看到改进。（注：因为缺少激活函数，就相当于y=W1x+x =（W1 + E）x，加了个单位tensor的权重，这样应该残差就没啥用了。）

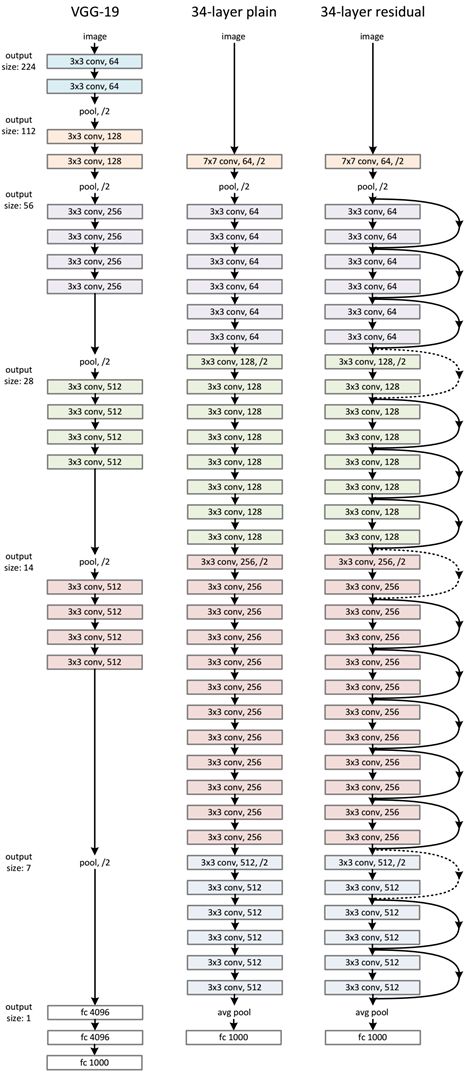
我们还注意到尽管为了简化上面的符号是用的全连接形式，但是它们对于卷积层是一样是用的。函数F(x, {Wi})能都代表多个卷积层。元素相加就是将两个特征图逐通道的相加。

## 3.3 网络架构

我们测试了多种一般网络和残差网络，观察到一致的现象。为了提供讨论是你，我们在ImageNet下描述两种模型，如下：

**一般网络** 我们的朴素网络的基线主要受VGG启发，卷积层主要是3\*3的核并且有以下两个简单的设计原则：i) 输出同样尺寸的特征图，那么这个层拥有同样数量的卷积核 ii）如果特征图尺寸减半，那么将卷积核的数量加倍，这样来让每层保持相同的时间复杂度。我们卷积层后面直接下采样，步长为2.网络的最后，采用一个全局平均池化层，以及1000条线的全连接层+softmax。权重层数量是34，如图3。

值得一提的是我们的模型箱体VGG拥有更少的卷积核和更低的复杂度。我们的34层只有36亿FLOPs，只是VGG的18%。



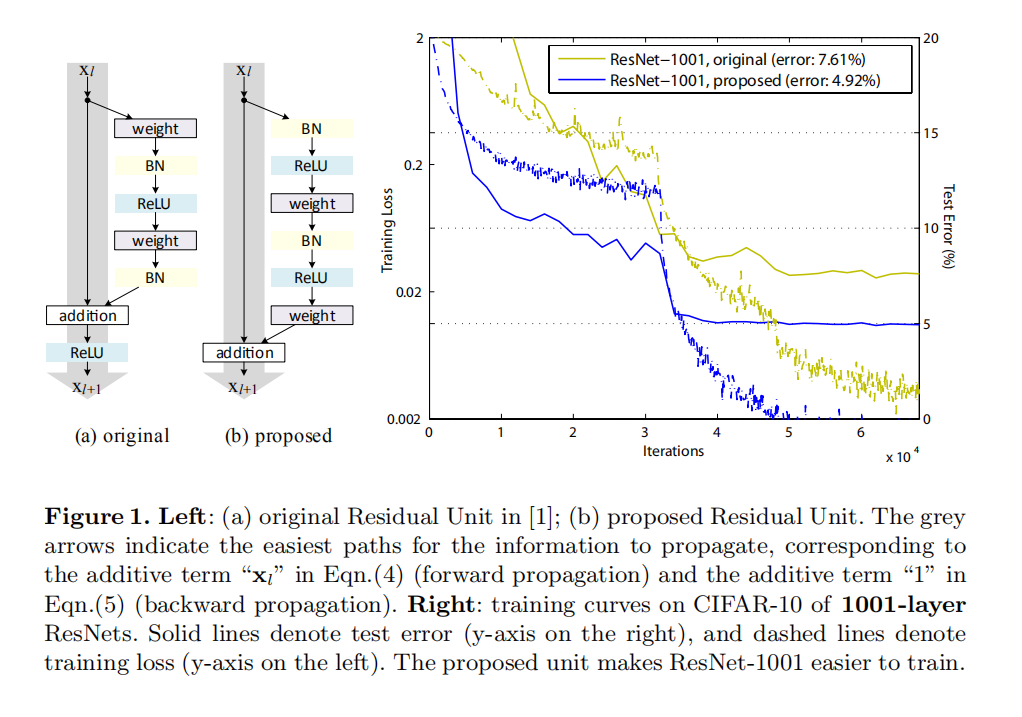
**残差网络：**基于上面的一般网络，我们插入短连接，这样就将一般网络变成了它的残差版本。当输出和输出尺寸相同时，这个恒等的短链接能够直接用，当尺寸增加，我们考虑两种操作A）短连接仍然执行恒等变化，为额外的维度添加0填充；B）做短连接的映射，如等式2中所示，使用1\*1的卷积完成。对于特征图尺寸不同的情况，A和B都采用步长为2的方式处理。（注：这里什么意思，似乎没太讲明白）。

## 3.4 实现

我们在ImageNet上的实现遵循了实践[21, 41].按照图片短边缩放到[256-480]之间的随机数的倍数来缩放图片，以此进行尺寸增强。用224\*224的区域在图片或者它的水平翻转上做随机采样，将每个像素减去均值。使用21的方式进行标准的颜色增强。我们使用在卷积之后，激活之前使用BN，遵循【16】。 遵循13的方式进行权重初始化并且从头开始训练网络。SGD配合256的mini批进行优化。学习率从0.1开始，当错误停滞时就除以10，模型训练达到60\*10^4个迭代。我们使用权重衰减0.0001和动量0.9。遵循16，我们没有使用dropout。

在测试中，为了对比研究，我们使用了标准的10 -crop 测试。为了最好的结果，我们应用了全卷积形式，就像[41,13]一样，然后使用了5种图片尺寸小的平均分数。短边长分别为 {224, 256, 384, 480, 640}).

残差块的结构，作者后续论文给出的改进版的



# 4.实验

暂略...

进一步的分析：

为什么叫残差网络：

残差是统计学中的术语，表示的是预测值和观测值（测量值）之间的差值，注意残差与误差的区别，误差是指观测值（测量值）与真实值之间的差值（就是比如用直尺测量长度，通常是各种原因导致结果有误差的，比如手抖，比如尺子不精确，比如眼睛看斜了）。

根据文中的公式，现在的H(x)变成了H(x)=F(x)+x，在单位映射中，y= x，x便是当成观测值，那么此时预测值H(x)减去观察值x，得到的就是残差F(x)=H(x)-x，因此叫做残差网络。

知乎的一个总结

二、ResNet想法来源

为什么想到使用残差？

-对于VLAD和Fisher Vector来说，对残差向量编码比对原始向量编码效率更高

-用Multigrid解偏微分方程（Partial Differential Equations，PDE）时，使用残差向量对于优化更好，收敛速度更快

为什么想到使用跨层连接？

-在多层感知机（multi-layer perceptrons，MLP）中加一层从输入到输出的线性层

-GoogLeNet中使用辅助分类器防止梯度爆炸 / 消失

-在此之前已有研究者使用跨层连接对响应和梯度中心化（center）处理

-inception结构本质也是跨层连接

-highway网络也使用到了跨层连接

模型集成：整个ResNet类似于多个网络的集成，原因是删除ResNet的部分网络结点不影响整个网络的性能，但VGGNet会崩溃，具体可以看这篇NIPS论文：

[Residual Networks Behave Like Ensembles of Relatively Shallow Networks](https://link.zhihu.com/?target=http://papers.nips.cc/paper/6556-residual-networks-behave-like-ensembles-of-relatively-shallow-networks" \t "https://zhuanlan.zhihu.com/p/_blank)