# 用于图像识别的深度残差学习

## 摘要

更深的网络更难训练，一般有梯度消失（爆炸）、过拟合、网络退化等问题。在本文中，作者提出了残差学习网络去训练比以往网络更深的网络。网络将学习相较于输入的残差函数而不是基于整个网络的误差函数。这样的残差网络更容易优化，并且随着深度的大幅提高可以获得更高的准确率。在ImageNet数据集上，作者提出的ResNet虽然达到了152层，是VGG网络的8倍，但网络的复杂度仍比VGG低。同时，作者的ResNet也获得了多项大赛的第一名。

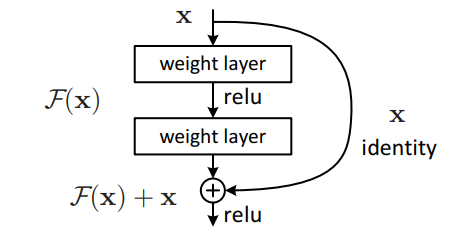
|  |
| --- |
| 区分梯度消失（爆炸）、过拟合和网络退化：  梯度消失（爆炸）：网络学不到东西；  过拟合：网络在训练集上表现良好，在测试集上表现糟糕；  网络退化：当网络深度增加时，网络的性能不但没有提升反而下降。 |

## 导论

深度网络用端对端的多层方式自然地融合了低、中、高层次的特征，并且可以通过增加网络的深度来增加特征的层级。最近的研究表明（如VGG和GoogLeNet）网络的深度非常重要，在ImageNet数据集上表现良好的网络都有较深的深度。网络堆叠深度的增加不是一件容易的事情，遇到的问题有：

* 第一个问题是梯度消失或梯度爆炸问题：这个问题已经在很大程度上通过各种归一化和初始化技巧来解决（泽威尔初始化、MSRA初始化），这使得具有数十层网络可以被训练。
* 第二个问题是网络退化问题：当网络的层数增加后，网络在训练集和测试集上的表现都下降。

作者提出了一种方法用来解决上述问题：在原来网络的基础上增加了一个恒等映射。这样就保证了更深的网络不会产生比原来网络更大的误差。但是实验表明这样的网络（由于包含了递归结构很难被优化）不能在可行的时间内求出网络的解。在本篇论文中，作者提出了深度残差学习框架来解决网络退化问题。不是希望网络直接拟合底层映射，而是去拟合残差映射。



设需要拟合的底层映射为H(x)，作者让堆叠的非线性层去拟合另一个映射F(x) = H(x) – x，这样原来需要拟合的映射被转换为F(x) + x。作者假设优化残差映射比优化原来的映射更容易。极端情况下，如果恒等映射是最优的，残差网络比原来的非线性网络更容易拟合。

式子F(x) + x可以通过加了“快捷连接”的前馈神经网络来实现，“快捷连接”是指跳过了一个或多个层。在本篇论文中，“快捷连接”仅仅被用来做恒等映射，并将其输出添加到堆叠层的输出。恒等连接既没有增加额外的参数，也没有提升计算的复杂性。整个网络依然可以通过SGD来进行反向传播来进行优化。

|  |
| --- |
| “快捷连接”是指跳过了一个或多个层 |

在ImageNet测试中，作者的深度残差网络很容易优化，而普通的网络在深度叠加时出现了网络退化现象。而且作者的深度残差网络因为深度的增加网络的性能大幅提升。在CIFAR-10的测试集上也表现出类似的效果。

作者的152层的ResNet虽然是参加ImageNet的网络中最深的网络，但是它的复杂度仍然低于VGG，在当年的大赛中获得了第一名，取得了比人类更优秀的成绩。该网络也可以被泛化迁移到其他识别任务上，并且让作者在其他比赛中也获得了第一名。这表明残差学习的原理是通用的，可以适用于其他视觉或非视觉问题。

## 相关工作

残差表示

在图像识别领域，VLAD是通过相对于字典的残差向量进行编码的典型代表。Fisher Vector可以被公式化为VLAD的概率版本。它们都是用于图像检索和分类的浅层网络代表。对于向量量化，残差向量编码展现出比原始向量编码更有效的效果。

|  |
| --- |
| VLAD是一种图像检索方法 |

在低级视觉（即对像素进行操作）和计算机图形领域，为了求解偏微分方程（PDEs），多重网格方法将整个问题分解为不同层次上的子问题，每个子问题都在粗糙和精细之间的层次上求解残差解。多重网格的替代方案是分层基预处理，它依赖代表两个维度的残差向量的变量。这些解决方法比不知道解的残差特性的标准解决方法收敛快得多，这些表明一个好的重构或预处理可以简化优化过程。

快捷连接

快捷连接已经被研究了很长时间。训练多层感知机的早起实践就是添加一个从网络输入到输出的线性层。在一些论文中（如GoogLeNet），一些中间层被直接连接到辅助分类器上去解决梯度消失或梯度爆炸问题。在一些论文中提出了通过快捷连接将层的输出、梯度和反向传播误差集中起来的方法。在GoogLeNet中，初始层由快捷连接和一些更深的分支组成。

伴随作者研究的是具有门控函数（比如循环神经网络中的门）的快捷连接的高速神经网络。这些门是数据驱动的且具有参数，这与作者的无参快捷连接形成了对比。在传统的门控函数中，当门控函数关闭时（接近0），网络中层代表非残差函数。相反，作者提出的ResNet总是学习非残差函数，它的快捷连接从未关闭，并且所有信息都可以通过，还需要额外的残差函数去学习。此外，高速神经网络也没有表现出当网络深度增加时有显著的精度提升。

## 深度残差网络

残差学习

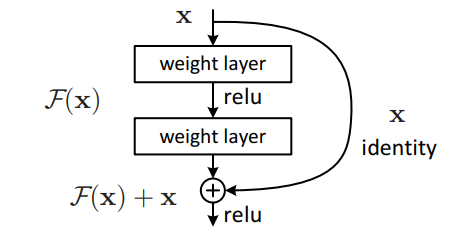
设H(x)需要拟合的从输入到输出的底层映射，它可能是通过某些层并非整个网络来拟合，x表示这些层中的第一层的输入。假设多个非线性函数可以渐进拟合复杂的函数，则可以假设它们也能渐进拟合残差函数：F(x) = H(x) – x（假设输入和输出的维度相同），因此，原始函数就变为F(x) + x。尽管两种形式都能拟合，但是学习的容易程度有所不同。

|  |
| --- |
| 万能近似定理：对于任意一个连续的函数，都可以用一个神经网络来近似表示。该定理表明神经网络具有较强的函数近似能力。 |

这种变化是为了解决退化问题。如果添加的层可以被构造为恒等映射，更深的模型的误差也不会比浅层网络的误差更大。退化问题表明，多层非线性网络对于恒等映射的拟合效果不尽人意。而在使用残差模块的情况下，如果恒等映射是最优的，网络可以将非线性层的权重变为0而仅仅使用恒等映射。

在现实情况下，恒等映射是最优的是不太可能的，所以残差模块需要进行修正。如果最优函数比零映射更接近恒等映射，则拥有恒等映射的网络更容易寻找到关于恒等映射的微小的扰动而不是学习为一个新函数。实验证明，总体上残差函数有更小的输出，表明了恒等映射发挥了合理的作用。

通过快捷连接的恒等映射

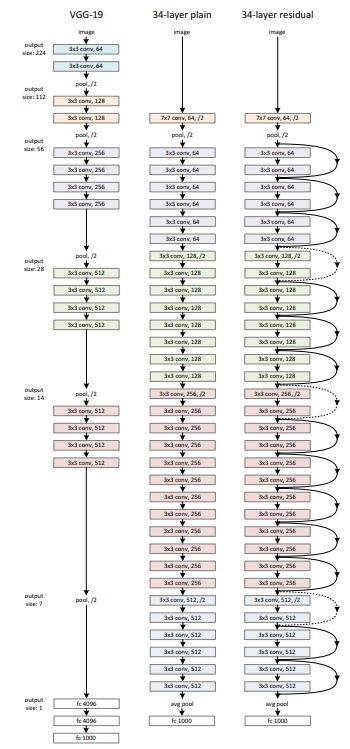


在本文中，网络的每层都是使用了残差模块，即y = F(x, {Wi}) + x，x是层的输入向量，y是层的输出向量，F(x, {Wi})代表需要学习的残差映射。如图中的例子，它有两层，F=W2σ(W1x)，σ表示ReLu激活函数，为了简化计算，偏置项被省略。F + x通过快捷连接来执行，并进行了逐元素相加，在相加之后进行了第二个非线性激活函数。

快捷连接没有引入额外的参数，也没有增加计算复杂度。这样可以与同等参数量、深度和计算量的普通网络作对比。

X和F的维度需要相等（即输入和输出的通道数），如果不同，可以在恒等映射上加一个线性投影来匹配输出的通道数，即y = F(x, {Wi}) + Wsx。虽然也可以在恒等映射上加一个权重矩阵，但是实验表明，恒等映射已经足够处理网络退化问题，所以仅在匹配输出大小时才在恒等映射上使用权重矩阵。

残差函数F的形式是灵活的。在本论文中使用了2~3层，其实更多层也是可能得，但是如果只有一层就变为了线性层，没有取得明显效果。为了简单起见，残差模块使用的全连接层，实际上也适用于卷积层。这时的逐元素相加就在两特征图上逐通道相加。

网络结构

当普通网络（和VGG类似的网络）和ResNet的层数都为34时，ResNet的参数数量仅为普通网络的18%。将上述的普通网络加入快捷连接，从而将网络转换为残差网络。当输入和输出的维度相同时直接添加恒等映射。输出的维度比输入的维度大时，作者考虑了两种方法：

* 快捷连接仍然为恒等映射，增加的维度用0填充。
* 对恒等连接做1×1卷积。

对于两种选择，当恒等连接通过两个不同维度的层时，使用的都是步长为2的池化层。

实施

ResNet采用和AlexNet、VGG相同的训练规则。图片被缩放到短边随机分布在[256, 480]的范围内。对一张图片或它的水平翻转后的图片进行随机224×224大小的裁剪，每个像素要减去像素平均值。使用了和AlexNet相同的标准颜色增强。在每个卷积层之后和激活函数之前采用了BN归一化。从头开始训练普通网络或残差网络。学习率从0.1开始，当错误率趋于0时除以10，并且模型被训练60万个迭代。权重衰减为0.0001，动量为0.9。没有使用丢弃层。

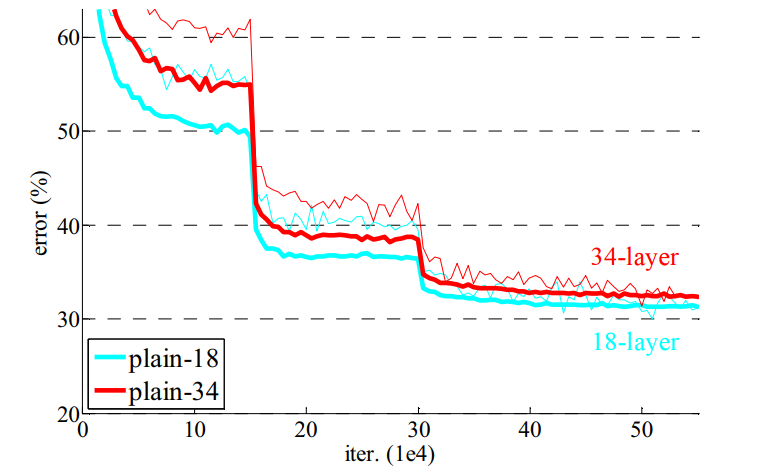
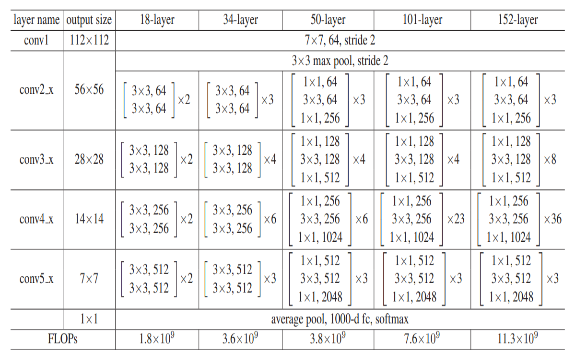
在测试中，为了对照研究作者采用了对每张图片进行10次裁剪的测试。为了得到更好的结果，作者对图片进行了不同程度的缩放，对每张图片的不同缩放取平均成绩。

## 实验结果

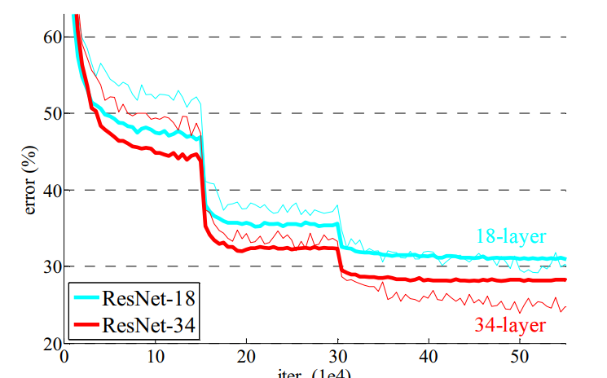
18层和34层的普通网络

更深的34层普通网络的错误率更高，原因是退化现象。

作者认为，这种现象的出现不太可能是由梯度消失引起的。这些普通网络通过BN归一化来进行训练，这就保证了前向传播具有非0的方差。作者还验证了反向传播的梯度也是正常的。所以问题不是出现在前向传播和反向传播消失引起的。事实上，34层的普通网络仍然能实现有竞争力的效果，这表示网络在某种程度上是工作的。作者推断深层的网络可能拥有指数级低的收敛速度，这会影响训练误差的降低，这个问题还有待研究。



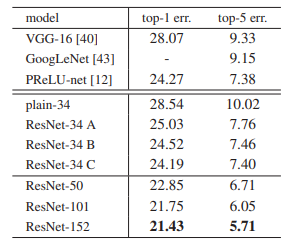
残差网络

接下来作者评估了18层和34层残差网络，基础结构与上面普通网络相同，只是对3×3过滤器上加了一个快捷连接。在第一次比较时，对所有快捷连接使用恒等连接，并使用0填充用于增加维度，所以相较于普通网络没有额外的参数。从结果中可以看出，34层的ResNet的效果比18层的ResNet效果更好，这表明退化问题得到了解决。

18层普通网络比18层ResNet收敛得更快。当网络的层数不太深时，虽然普通网络也能找到比较好的结果，但是ResNet在早期阶段提供更快收敛。

恒等快捷连接和投影快捷连接

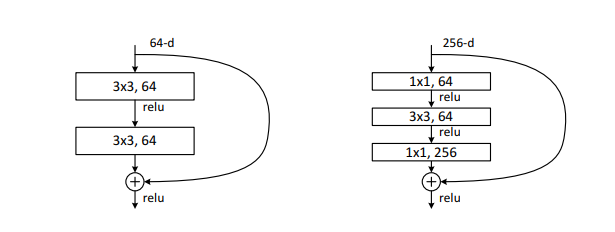
作者研究了投影快捷连接：

* A：0填充用于增加维度，所有快捷连接都是无参的，
* B：投影快捷连接用于增加维度，其他快捷连接都是恒等连接
* C：所有快捷连接都是投影快捷连接

上表展示了三种选项比普通的网络效果都好。B比A的效果稍微好，作者推断这是因为A中的0填充因为没有进行残差学习而丢失了一部分信息；C比B好，因为通过13个投影连接层引入了额外的参数。但是它们的差异很小，表明在快捷连接上做处理对于解决退化问题并不重要。所以作者在本论文的剩余部分没有使用C来减少空间时间复杂度和模型大小。

|  |
| --- |
| 投影快捷连接：在快捷连接上添加了参数层。 |

更深的bottleneck结构

由于担心花费过多的训练时间，作者对更深的网络做了修改。对每个残差函数F，作者使用了3层网络而不是2层网络。三层网络使用了分别使用了1×1、3×3、1×1卷积网络，其中1×1卷积用于降维和升维，这样使3×3卷积层有更小的输入和输出维度。两种设计有相近的时间复杂度。无参的恒等连接对于bottleneck体系结构尤其重要，如果使用投影，则时间复杂度和模型大小都要加倍，因为快捷连接连接到两个高维的网络层。

50层的ResNet

将34层的ResNet的2层结构换为3层结构，从而得到一个50层网络，使用方案B来增加维度，模型共有38亿参数。

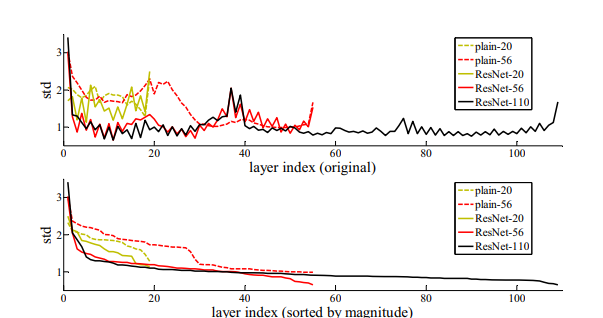
101层和152层ResNet

使用更多的3层结构来构建网络。虽然深度显著增加，但是152层的ResNet的复杂度仍然比VGG-16/19低。

在CIFAR-10上的测试和分析

网络的输入是32×32，每个像素都减去所有像素的平均值。第一层是3×3卷积层，然后在大小为32、16、8的特征图上各自使用了6n个3×3卷积层，每个特征图的大小为2n。通过步长为2来进行下采样，网络的结尾使用了全局平均池化，10路全连接层和oftmax。整个网络共有6n+2个有权重的层。每两层之间使用一个3×3快捷连接，总共3n个快捷连接，使用A方案来解决恒等连接输入和输出维度不同的情况。

层响应的分析

层响应是在BN标准化之后、其他非线性激活函数之前的每个3×3层的输出。对于ResNet，这个分析表明了残差函数的响应强度。这些结果表明残差函数比非残差函数更容易拟合0。作者还注意到更深的ResNet有更小的响应幅度。当层数较多时，单个ResNet层倾向于修改更小的信号。

对超过1000层网络的探索

对于1202层的网络并没有出现优化困难的问题（即网络退化），而且训练误差小于0.1%。测试集上的效果也不错，但是效果不如110层的ResNet，尽管它们的训练误差相似。作者认为出现这样的原因是过拟合，1202层的网络对于小的数据集参数量有点太大了。所以像maxout或dropout这样非常强的正则化应该被采用。对于超深网络的正则化会在后续进行研究。

# 问题回答

Q1 论文试图解决什么问题？

当网络的深度加深时出现的网络退化问题。

Q2 这是否是一个新的问题？

在VGG的论文中也出现过网络层数增加时，网络的效果没有提升，当时作者认为数据集的规模太小，不适合更深模型的使用。

Q3 这篇文章要验证一个什么科学假设？

使用残差模块可以解决网络退化问题。

Q4 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

对于网络深度加深时，会出现梯度爆炸和梯度消失问题，可以通过各种初始化和标准化技巧来处理数据，如BN可以使网络达到二十多层；也会出现网络退化问题，可以通过残差结构来解决，本文的作者何凯明等就是值得关注的研究员。

Q5 论文中提到的解决方案之关键是什么？

残差模块，通过增加一个快捷连接（使用恒等映射即可）来实现。

Q6 论文中的实验是如何设计的？

通过对比没有使用的残差模块的PlainNet和使用了残差模块的ResNet，它们的网络参数相同（恒等映射没有增加参数），但是在34层网络上ResNet没有出现网络退化，并且取得了更好的成绩。然后作者探索了更深的网络，为了减少更深网络中的参数数量，使用了bottleneck，并分析了它们的网络效果。

Q7 用于定量评估的数据集是什么？代码有没有开源？

ImageNet、CIFAR-10等数据集；代码开源。

Q8 论文中的实验及结果有没有很好地支持需要验证的科学假设？

通过使用了恒等映射的ResNet和没有使用的PlainNet对比中可以看出，使用了残差模块的ResNet可以更快地收敛，而且也没有出现网络退化问题；当网络大于50层时，使用bottleneck的ResNet的不仅没有出现网络退化问题，而且参数量也没有超过VGG，网络的效果大幅提升。

Q9 这篇论文到底有什么贡献？

这篇论文提出的残差模块使得堆叠多层网络成为可能，以往的VGG等卷积网络虽然深度比以往深，但是也局限于十几层，而使用了残差结构的ResNet可以使网络堆叠到几百层甚至上千层。

Q10 下一步呢？有什么工作可以继续深入？

对于超深网络的正则化；之后提出的ResNeXt结合了ResNet的shortcut模块与Inception的split-transform-merge思想，在保持参数量基本相等的情况下，提升了性能，同时又省去了inception网络里手工设计分支的步骤。