Deep Residual Learning for Image Recognition

Abstract

越深的网络训练起来越困难。我们提出来一个残差学习网络框架来简化那些比之前使用的更深的网络的计算。我们明确地将层变为学习关于层输入的残差函数，而不是学习未参考的函数。我们提供了可理解的经验性的证据来证明这些残差网络更容易优化，可以从增加的深度中得到可观的精度收益。在 ImageNet 数据库我们使用了达到 152 层深度的网络来预测(是 VGG 的八倍)，但是仍然有较低的复杂度，这些残差网络在 ImageNet 测试集上达到了 3.57%的错误率。这个结果赢得了 ILSVRC2015 分类赛的第一名。我们也在 CIFAR-10 数据集上使用了 100 层和 1000 层进行了分析。

深度的代表性对于许多视觉识别任务来说是很重要的。仅仅因为我们极致的深度，在 COCO 物体探测中我们获得了接近 28%的提升。深度残差网络是我们向 ILSVRC 和 COCO2015 比赛提交的作品的基础。我们也赢得了 ImageNet 物体探测，ImageNet 物体定位，和 COCO 探测，COCO 分割任务的第一名。

Introduction

深度卷积神经网络导致了图片分类领域的一系列突破。深度网络自然整合了低中高等级的特征，分类器以端到端的多层结构。不同等级的特征可以通过堆叠层数来丰富。最近的证据揭示了神经网络的深度非常重要。在 ImageNet 数据库起领导型作用的网络都探索了非常深的模型，从十六层到三十层。许多其他的视觉识别任务也受益于非常深的网络。

在深度的意义的推动下，我们不禁提出这样的问题：是不是说堆叠更多的层数就可以得到更好的学习网络？对于这个问题的一个阻碍就是梯度消失问题。它从开始就阻碍了收敛。然而这个问题然而，这个问题通过标准初始化和中间标准化层在很大程度上已经解决，这使得数十层的网络能通过具有反向传播的随机梯度下降（SGD）开始收敛。

当更深的网络能够开始收敛，退化问题又被发现：随着网络深度的增加，精度开始饱和然后迅速下降。出乎意料的是，这种下降并不是因为过拟合，对一个深度合适的模型增加更多的深度导致了更高的训练误差，如同[11,42]报告的那样，我们的实验证明了这一现象，在表 1 中展示了一个典型例子。

训练精度的退化说明不是所有的系统都一样容易优化。让我们考虑一个浅层的结构，它的深层的结构在其上增加更多的层。存在通过构建得到更深层模型的解决方案：增加的层是横等映射，其它层是从已经学习好的浅层模型复制而来。但是实验说明给我们现在手边的解决方案不能够找到与构建的解决方案相对比较不错的方案(或者说不能够在合理的时间内做到)。

在这篇文章中，我们通过引入了一个深度残差网络框架来解决退化问题。我们明确地让这些层拟合残差映射，而不是希望每几个堆叠的层直接拟合期望的基础映射。形式上，将期

望的基础映射表示为 H(x)，我们将堆叠的非线性层拟合另一个映射 F(x):=H(x)−x。原始的映射重写为 F(x)+x。我们假设残差映射比原始的、未参考的映射更容易优化。在极端情况下，如果一个恒等映射是最优的，那么将残差置为零比通过一堆非线性层来拟合恒等映射更容易。

公式 F(x)+x 可以被认为是带有快捷连接的前向神经网络。快捷连接是跳过一层或多层的连接。在我们的实验中，快捷连接简单的执行恒等映射，它们的输出也只是加在了网络块的输出上。恒等快捷连接并没有增加额外的参数和计算复杂度。剩余的网络仍然可以通过反向传播的 SGD 算法进行端对端的训练，可以使用常见的库来实现不用修改求解器。

我们在 ImageNet 上进行综合性的实验展示了退化问题和评估了我们的方法。我们发现：1）我们极深的残差网络易于优化，但当深度增加时，对应的“简单”网络（简单堆叠层）表现出更高的训练误差；2）我们的深度残差网络可以从大大增加的深度中轻松获得准确性收益，生成的结果实质上比以前的网络更好。

相同的结果同样展示在 CIFAT-10 数据集上，说明了我们的方法的操作难度和效果不仅仅针对于特定的数据集。在这个数据集上我们成功的训练了超过一百层的模型，探索超过一千层的模型。

在 ImageNet 分类数据集上，通过极其深度的残差网络我们获得了优秀的结果，我们的 152 层的残差网络是 ImageNet 有史以来提交的最深的网络，同时复杂度低于 VGG 网络。我们的模型集合在 ImageNet 测试集上有 3.57%的 top-5 错误率，赢得了 ILSVRC2015 分类赛的第一名。极深的网络代表他在其他识别任务中也有着优异的泛化性能，使我们赢得了更多的第一名：包括 ILSVRC & COCO 2015 竞赛中的 ImageNet 检测，ImageNet 定位，COCO 检测和 COCO 分割。强有力的证据说明残差学习网络是通用的。我们希望它能够被应用到其他的视觉或者非视觉任务中去。

2.相关工作

残差表示。在图片识别中 VLAD 是一种通过关于字典的残差向量进行编码的表示形式。费舍尔向量可以被看做是 VLAD 的概率版本。对于图片分类和识别他们都具有强大的浅层代表。对于矢量量化，参数差向量编码比原始向量编码更加有效率。

在低等级的视觉和计算机图形学中，为了解决部分差异等式，广泛应用的多网格方法将系统重构为多个尺度上的子问题。每个子问题负责较粗尺度和较细尺度的残差。Multigrid 的替代方法是层次化基础预处理，它依赖于表示两个尺度之间残差向量的变量。已经被证明这些求解器比不知道解的残差性质的标准求解器收敛得更快。这些方法表明，良好的重构或预处理可以简化优化。

快捷连接。导致快捷连接的实践和理论已经研究了很久。训练多层感知机（MLP）的早期实践是添加一个线性层来连接网络的输入和输出，在[44,24]，一些中间层直接和辅助分类器相连，为了解决梯度消失问题。这篇论文的[39,38,31,47]提出了一些通过快捷连接实现中间层响应，梯度，前项误差的方法。在[44]一个“inception”层由一个快捷分支和一些更深的分支组成。

和我们同时进行的工作，“highway networks” [41, 42]提出了门功能[15]的快捷连接。这些门是数据相关且有参数的，与我们不具有参数的恒等快捷连接相反。当门控快捷连接“关闭”（接近零）时，高速网络中的层表示非残差函数。相反，我们的公式总是学习残差函数；我们的恒等快捷连接永远不会关闭，所有的信息总是通过，还有额外的残差函数要学习。此外，高速网络还没有证实极度增加的深度（例如，超过 100 个层）带来的准确性收益。

3.深度残差学习

3.1 残差学习

我们考虑 H(x)作为几个网络(不必要是整个网络)要拟合的映射，x 定义为第一层的输入。如果假设多个非线性层可以近似拟合一个复杂的函数。这相当于说他们可以近似拟合残差函数，例如：H(x)-x(假设输入和输出是相同的维度)。所以与其令这些网络近似 H(x),我们让这些网络去拟合一个残差函数 F(x)=H(x)-x。原始的函数因此变为 F(x)+x，尽管这两种形式都能够近似我们希望的函数，但是学习的难易程度可能不同。

关于退化问题的反直觉现象激发了这次重构。和我们在介绍中讨论的一样，如果额外增加的层可以被认为是恒等映射，更深的模型应该比浅层的模型有着更低的训练错误率。退化问题说明了求解器可能在用多个非线性层拟合恒等映射上有困难。使用残差学习，如果恒等映射是最佳的，求解器将只需把多个非线性层的权重变为零去拟合恒等映射。

在真实的条件下，恒等映射是最佳的不太可能，但是我们的重构可能有助于问题的预处理。如果最佳的函数接近一个恒等映射而不是一个零映射，求解器更容易找到关于恒等映射的抖动，而不是将其作为一个新的函数来学习。通过实验我们展示了学习的残差函数通常有更小的响应，表明恒等映射提供了合理的预处理。

3.2.通过捷径实现恒等映射

我们每隔几个层使用残差学习。构建如图 2 所示的块。在本文我们定义一个块如：

$$y = F(x,\{Wi\})+x \qquad (1)$$

在这里 x 和 y 分别是我们所考虑的这些层的输入和输出向量。函数$F(x,\{Wi\})$代表了要学习的残差网络。例如在表二中有两层。$F = W_2\sigma(W_1x)$，σ代表 RELU，为了简化公式移去了偏置。F+x 操作通过快捷连接和各个元素相加来执行。在相加之后我们采纳了第二种非线性（即σ(y)，看图 2）。

等式一中的快捷连接没有引入额外的参数和计算复杂度。这不仅仅在实践中非常有吸引力而且对于我们的空白组和残差网络的对照也非常重要。我们可以公平地比较具有相同数量的参数，深度，宽度和计算消耗的残差网络和空白网络。(除了不可忽略的元素加法)

在等式(1)中 x 和 F 的维度必须相同。如果不同的话(例如输出/输出的通道改变)，我们可以通过快捷连接使用一个线性投影 W_S 来匹配维度：

$$y = F(x, \{Wi\}) + W_S x \qquad (2)$$

我们也可以在等式一中使用一个方阵 W_s。但是我们将会通过实验说明恒等映射解决退化问题是高效经济的，因此 W_s 只是用来匹配维度的。

残差函数 F 的形式是灵活的。本文的实验中残差函数为有两层或三层的网络。更多的层也是可以的。但是如果 F 只有单独的一层，等式一就如同一个线性层，$y = W_1 x + x$,这种形式尚且没有看到优势。

我们也注意到为了简便尽管上面的符号是关于全连接层的，他们同样也适用于卷积层。公式$F(x, \{Wi\})$可以被多层卷积网络代替。元素加法在两个特征图上逐通道进行。

3.3. 网络结构

我们测试了多种空白/残差网络，观察到了不变的现象。为了证明讨论的例子，下面我们描述 ImageNet 的两个模型。

Plain Network 我们的空白网络的基准主要是受到了 VGG 网络的启发。卷积层大部分有 3*3 的卷积核和如下两个简单的设计规则：(i)为了能够有相同的输出特征图尺寸，所有的层有着相同数量的滤波器。(ii)如果特征图尺寸变为原来的一半，那么滤波器的数量就加倍，这样可以保证每一层的时间复杂度。我们通过步长为 2 的卷积层直接执行下采样。网络以全局平均池化层和具有 softmax 的 1000 维全连接层结束。图 3（中间）的加权层总数为 34。

值得注意的是我们的模型比 VGG 网络有着更少的滤波器和更低的复杂度，我们的 34 层基准有 3.6 亿浮点运算。这仅仅只有 VGG19 的 18%。

残差网络。基于上述的空白网络，我们加入快捷连接(图三右)将网络转换为相应的残差网络。当输入和输出维度相同时我们可以直接使用恒等连接(等式 1)，当维度增加时，我们考虑两种选择：(A)快捷连接仍然作为恒等映射，使用额外的零来填充增加的维度，这个选择没有引入额外的参数。(B)使用等式二中的连接来匹配维度。对于这两个选项，当快捷连接跨越两种尺寸的特征图时，它们执行时步长为 2。

3.4 实现

我们对于 ImageNet 的实现遵循[21,41]的实践。图片使用较短边缩放随机在[256 480]进行取样来尺度扩充。一个 224*224 的裁剪从一个图片或者其竖直翻转上随机取样。每个像素都减去均值。[21]中标准的颜色扩充也被用到。在每个卷积和激活之前我们都是用来批处理化。我们像[13]中一样初始化权重，从零开始训练所有的空白/残差网络。使用最小批尺寸为 256 的随机梯度下降法，学习率从 0.1 开始，当错误率平稳时学习率除以 10。并且模型训练高达 60×104 次迭代。我们使用的权重衰减为 0.0001，动量为 0.9。根据[16]的实践，我们不使用 dropup[13]。

在测试中，为了对照学习我们是用来标准的 10-crop 测试。为了最好的结果，我们是用来如同[41,13]的全连接形式。将不同尺度的得分平均（图像归一化，短边位于{224, 256, 384, 480, 640}中）。

4.实验

4.1 ImageNet 分类

我们在由一千种图像组成的 ImageNet-2012 分类数据集上评估我们的方法。模型在一百二十八万图像上训练，在五万张图片组成的验证集上评估。我们也使用一万张测试集的图片来作为最终结果，评估了 top-1 和 top-5 错误率。

空白网络，我们首先评估了十八层和三十四层的空白网络。三十四层的网络在图三中间。十八层网络有着相同的形式。详细的结构在表一中。

表二中的结果说明更深的三十四层空白网络比十八层网络有着更高的验证集错误率。为了解释原因，在图四左边我们比较了他们训练过程中的训练集和验证集的错误率，我们观察到了退化问题。三十四层的空白网络在整个训练过程中有着更高的训练错误率，虽然十八层网络的解空间是三十四层网络的子空间。

我们认为这种优化难度不是由梯度消失导致的。空白网络使用反向传播训练，这保证了前项传播信号有非零方差。我们也证明了反向传播的梯度符合 BN 标准。所以没有前向和反向的信号消失，事实上三十四层的空白网络仍然能够达到有竞争力的精度，说明求解器在某种程度仍在工作。我们推测深度空白网络可能有较低的指数收敛率，这影响力训练错误的减少，这种优化困难的原因还需要研究。

残差网络。接下来我们评估了十八层和三十四层残差网络，他们的基本结构与上面的空白网络一致，处理在没对三乘三的滤波器之间添加了快捷连接，在第一个比较中我们对所有的快捷连接使用了恒等映射和对所有的增加的维度使用了零填充，所以和空白网络比较并没有额外的参数增加。

从表二和图四中我们得到了三个主要观察结果，第一，残差学习的结果有所改变，三十四层的残差网络比十八层的网络结果要好(2.8%)，更重要的是，三十四层的残差网络展示出明显较低的训练误差和验证集误差，这说明退化问题被很好的解决了，我们可以收获由深度增加而带来的精度增益。

第二，与空白网络对比，三十四层的残差网络的 top-1 错误率减少了 3.5%，这得益于成功减少的训练集错误率，这个对比说明了残差网络在极致深度网络的效率。

最后我么有额注意到了十八层空白/残差网络有着相同的精度，但是十八层残差网络收敛更快。当网络不是很深的时候，SGD 求解器仍然能够为空白网络寻找好的解。在这个情况下，残差网络通过在早期提供更快的收敛来使优化变得容易。

恒等和投影快捷连接。我们已经证明参数没有增加，恒等连接对训练有帮助。接下来我们讨论投影快捷连接(等式二)，在表三种我们比较了三个选择，(A)增加的维度使用零填充快捷连接，所有的连接都是没有参数增加的(B)增加的维度使用投影快捷连接，其他的连接为恒等连接(C)所有的连接都是投影连接。

表三说明了这三个选择都比他们的空白网络要好的多，B 比 A 稍微好一点，我们认为这是由于 A 中的零填充维度没有残差学习，C 比 B 要好的多，我们将其归因于一些投影连接引入了额外的参数。但是 A/B/C 这些小的不同说明投影连接对于解决退化问题并不重要。所以为了减少时间内存复杂度和模型尺寸在这篇论文的其他部分我们并没有使用 C 选项。恒等快捷连接对于不增加下面介绍的瓶颈结构的复杂性尤为重要。

更深的瓶颈结构。接下来我们介绍对于 ImageNet 更深的网络。因为考虑到在训练中我

们可以承受的时间，我们将构建块修改为瓶颈连接。对于每个残差函数 F，我们使用一个三成的网络而不是两层，三层分别是 1*1,3*3 和 3*3 的卷积，1*1 的网络主要用来减少然后增加维度，使得 3*3 层成为具有较小输入/输出维度的瓶颈，图五中展示了一个例子，这两种设计有着相同的时间复杂度。

无参数增加的恒等连接对于瓶颈结构尤其重要。如果图五中的恒等连接被投影连接替代，时间复杂度和模型吃错都会加倍，因为快捷连接是连接两个高维的端，因此恒等连接为瓶颈设计提供了更加高效的模型。

五十层残差网络：我们用 3 层瓶颈块替换 34 层网络中的每一个 2 层块，得到了一个 50 层 ResNet（表 1）。我们使用选项 B 来增加维度。该模型有 38 亿 FLOP。

101 层和 152 层 ResNet: 我们通过使用更多的 3 层瓶颈块来构建 101 层和 152 层 ResNets（表 1）。值得注意的是，尽管深度显著增加，但 152 层 ResNet（113 亿 FLOP）仍然比 VGG-16/19 网络（153/196 亿 FLOP）具有更低的复杂度。

50/101/152 层 ResNet 比 34 层 ResNet 的准确性要高得多（表 3 和 4）。我们没有观察到退化问题，因此可以从显著增加的深度中获得显著的准确性收益。所有评估指标都能证明深度的收益（表 3 和表 4）。

和之前最好的方法进行比较。在表四中我们和之前最好的单个网络结构进行比较。我们的基本三十四层残差网络已经达到了非常有竞争力的精确度。我们的一百五十二层残差网络单个模型的 top-5 验证集错误率为 4.49%，这个单个网络的结果比之前所欲融合的结果。我们融合了不同深度的六个模型来产生一个集合(在提交时只有两个 152 层的网络)，这在测试集上得到了 3.5%的 top-5 错误率（表 5）。这次提交在 2015 年 ILSVRC 中荣获了第一名。

4.2.CIFAR-10 分析

我们在 CIFAR-10 数据集上展开了更多的研究，它由五万张训练集图片和一万张测试集图片组成，分为十类。我们在训练集进行实验，用测试集进行评估。我们的重点放在了极致深度的网络的表现，而不是为了推动最先进的结果，因此我们趋向于使用下面所示的简单的结构。

空白/残差网络和图三的形式一致(中间/右侧)，网络输入为 32*32 的图像。每个像素都减去均值。第一层是 3*3 的卷积层，然后我们在大小为{32,16,8}的特征图上分别使用使用一个 6n 层的 3*3 的卷积层，对每个特征图尺寸使用 2n 个卷集成。滤波器的数量分别为{16,32,64}。下采样为步进为二的卷积，网络以一个全局平均池化，一个十维全连接层，和 softmax 层结束，总共有 6n+2 个带权重的层，下面的表总结了这个结构：

当使用快捷连接时，他们连接到成对的 3*3 卷积层上(共 3n 条快捷连接)。在这个数据集上所有情况我们都使用恒等映射连接。所以我们的残差网络和空白网络有着相同的深度，宽度和参数数量。

我们使用一个 0.0001 的权重衰减和 0.9 的动量，使用[13]中的权重初始化和 BN 但是没有使用 dropout。这些模型使用 128 的小批量尺寸来训练在两块 GPU 上。学习率从 0.1 开始，在 32K 和 48K 次迭代时除以十，在 64K 次迭代时停止训练。这是由 45K/5K 的训练集和验证集分割决定的。我们使用了[24]中的简单的数据扩充来进行训练。每一边使用了四个像素的填充，并从填充图像或其水平翻转图像中随机采样 32×32 的裁剪图像。对于测试，我们只评估原始 32×32 图像的单一视图。

我们比较 n={3,5,7,9}，分别得到 20,32,44 和 56 层的网络，图 6(左)展示了空白网络，深度空白网络因为增加的深度呈现出更高的训练集误差，这个现象和在 ImageNet 和 MINST 上的相似。说明这样的优化困难是一个基本的问题。

图六(中间)展示了残差网络的表现。也如同 ImageNet 的表现一样，我们的残差网络设

法克服了优化困难，随着深度的增加展示出来精度的增益。

我们更进一步使 n=18，得到了一个 110 层的残差网络，在这个情况下我们发现初始的 0.1 的学习率有对于开始收敛来说一点大，所以我们使用了 0.01 的学习率开始训练直到训练错误率低于 80%(400 此迭代后)，然后仍然使用 0.1 的学习率进行训练。学习的其他计划和前面的一直。一百一十层网络收敛的很好。它与其它的深且窄的网络例如 FitNet[34]和 Highway41 相比有更少的参数，但结果仍在目前最好的结果之间（6.43%，表 6）

层响应分析。图 7 显示了层响应的标准偏差（std）。这些响应每个 3×3 层的输出，在 BN 之后和其他非线性（ReLU/加法）之前。对于 ResNets，该分析揭示了残差函数的响应强度。图 7 显示 ResNet 的响应比其对应的简单网络的响应更小。这些结果支持了我们的基本动机（第 3.1 节），残差函数通常具有比非残差函数更接近零。我们还注意到，更深的 ResNet 具有较小的响应幅度，如图 7 中 ResNet-20，56 和 110 之间的比较所证明的。当层数更多时，单层 ResNet 趋向于更少地修改信号。

探索超过 1000 层。我们探索超过 1000 层的过深的模型。我们设置 n=200，得到了 1202 层的网络，其训练如上所述。我们的方法显示没有优化困难，这个 103 层网络能够实现训练误差<0.1％（图 6，右图）。其测试误差仍然很好（7.93％，表 6）。

但是，这种极深的模型仍然存在着开放的问题。这个 1202 层网络的测试结果比我们的 110 层网络的测试结果更差，虽然两者都具有类似的训练误差。我们认为这是因为过拟合。对于这种小型数据集，1202 层网络可能是不必要的大（19.4M）。在这个数据集应用强大的正则化，如 maxout[9]或者 dropout[13]来获得最佳结果（[9,25,24,34]）。在本文中，我们不使用 maxout/dropout，只是简单地通过设计深且窄的架构简单地进行正则化，而不会分散集中在优化难点上的注意力。但结合更强的正规化可能会改善结果，我们将来会研究。

4.3. 在 PASCAL 和 MS COCO 上的目标检测

我们的方法在其他的识别任务上也有很好的表现。表 7 和表 8 显示了 PASCAL VOC 2007 和 2012[5]以及 COCO[26]的目标检测基准结果。我们采用更快的 R-CNN[32]作为检测方法。在这里，我们感兴趣的是用 ResNet-101 替换 VGG-16[40]。使用这两种模式的检测实现（见附录）是一样的，所以收益只能归因于更好的网络。最显著的是，在有挑战性的 COCO 数据集中，COCO 的标准度量指标（mAP@[.5，.95]）增长了 6.0％，相对改善了 28％。这种收益完全是由于学习表示。

基于深度残差网络，我们在 ILSVRC & COCO 2015 竞赛的几个任务中获得了第一名，分别是：ImageNet 检测，ImageNet 定位，COCO 检测，COCO 分割。更多细节请看附录。