# 深度残差学习在图像识别中的应用

深度神经网络更难训练。我们提出了一个残差学习框架，以简化网络的训练，这些网络比以前使用的网络要深得多。我们显式地将层重新表示为参考层输入的残差学习函数，而不是未引用的学习函数。我们提供了全面的经验证据，表明这些残差神经网络更容易优化，并且可以通过深度增加提升精度。在ImageNet数据集上，我们评估了深度高达152层的残差神经网络（它比VGG网络[41]深8倍），但复杂度仍然较低。这些残差网络的集成在ImageNet测试集上获得了3.57%的误差。这一结果在ILSVRC 2015分类任务中获得第一名。我们还对CIFAR-10进行了100层和1000层的分析。

深度对于许多视觉识别任务至关重要。仅由于我们极深的探索，我们在COCO目标检测数据集上获得了28%的相对改进。深层残差网络是我们提交给ILSVRC和COCO 2015竞赛的基础，我们还在ImageNet检测、ImageNet本地化、COCO检测和COCO细分任务上获得了第一名。

## 简介

深卷积神经网络[22，21]为图像分类[21，50，40]带来了一系列突破。深层网络以端到端的多层方式自然地集成了低/中/高级别特征[50]和分类器，并且特征的“级别”可以通过堆叠的层(深度)的数量来丰富。最近的证据[41，44]揭示了网络深度是至关重要的，在具有挑战性的ImageNet数据集[36]上的领先结果[41，44，13，16]都利用了深度从16[41]到30[16]的“非常深”[41]模型。许多其他非常重要的视觉识别任务[8，12，7，32，27]也从非常深入的模型中受益匪浅。

图1.具有20层和56层“普通”网络的CIFAR-10上的训练错误(左)和测试错误(右)。网络越深，训练误差越大，因此测试误差越大。ImageNet上的类似现象如图4所示。

在深度重要性的推动下，一个问题出现了：学习更好的网络是否像堆叠更多层一样容易？回答这个问题的一个障碍是臭名昭著的梯度消失/爆炸问题[1，9]，它从一开始就阻碍了收敛。然而，这个问题在很大程度上已经通过归一化初始化[23，9，37，13]和中间归一化层[16]来解决，这使得具有数十层的网络能够开始收敛于具有反向传播的随机梯度下降(SGD)[22]。

当更深的网络能够开始收敛时，降级问题就暴露了出来：随着网络深度的增加，精确度达到饱和(这可能并不令人惊讶)，然后迅速降级。出乎意料的是，这种退化并不是由过度拟合引起的，而且在适当深度的模型上增加更多的层会导致更高的训练误差，正如[11，42]中所报告的那样，并被我们的实验彻底验证。图1显示了一个典型的示例

(训练精度的)下降表明并非所有的系统都同样容易优化。让我们考虑一种较浅的层数，以及在其上增加更多层的更深层的模型。对于更深层的模型，有一种解决方案：添加的层是识别映射，其他层是从深度学习的浅层模型复制而来的。这种构造结果的存在表明，较深的模型应该不会比较浅的模型产生更高的训练误差。但是实验表明，我们现有的求解器不能找到比构造的解更好或更好的解(或者不能在可行的时间内找到)。

在本文中，我们通过引入深度残差学习框架来解决退化问题。我们不希望较少数量的层直接拟合所需的基本映射，而是显式地让这些层拟合残差映射。形式上，我们将基本映射表示为H(X)，使叠加的非线性层拟合的映射表示为F(X)：=H(X)−x，则基本映射为F(X)+x，我们假设残差比原始的无参考映射更容易优化。在极端情况下，如果映射是最优的，那么将残差推到零要比通过一堆非线性层来拟合单位映射要容易得多。

F(X)+x的公式可以通过具有“捷径连接”的前馈神经网络来实现(图2)。快捷连接[2、34、49]是跳过一个或多个层的连接。在我们的例子中，快捷连接只是执行识别映射，它们的输出被添加到堆叠层的输出(图2)。识别快捷连接既不会增加额外参数，也不会增加计算复杂性。整个网络仍然可以由SGD使用反向传播进行端到端的训练，并且可以使用通用库(例如，Caffe[19])轻松实现，而无需修改解算器

我们在ImageNet[36]上进行了全面的实验，以显示退化问题并评估我们的方法。结果表明：1)深度残差网络易于优化，但对应的“普通”网络(简单的堆叠层)在深度增加时表现出更高的训练误差；2)深残差网络可以很容易地从深度的大幅增加中获得精度提升，其结果明显好于以往的网络。

在CIFAR-10集[20]上也显示了类似的现象，这表明我们的方法的优化问题和效果不仅仅类似于特定的数据集。我们在这个数据集上提供了超过100层的成功训练的模型，并探索了超过1000层的模型。

在ImageNet分类器数据集[36]上，我们通过极深的残差网络获得了很好的结果。我们的152层残差网络是ImageNet上出现的最深的网络，但其复杂度仍低于VGG网络[41]。我们的集成在ImageNet测试集上有3.57%的TOP-5错误，并在ILSVRC 2015分类比赛中获得第一名。拥有较高深度的层数在其他识别任务上也有很好的泛化性能，并带领我们在ILSVRC和COCO 2015大赛中进一步获得了ImageNet检测、ImageNet本地化、CoCo检测和CoCo分割的第一名。这一强有力的证据表明，残差学习原理是通用的，我们期望它也适用于其他视觉和非视觉问题。

## 2相关工作

残差表示法。

在图像识别中，VLAD[18]是由相对于字典的残差向量编码的表示，并且Fisher向量[30]可以被表示为VLAD的概率版本[18]。它们都是用于图像检索和分类的强大的浅层表示[4，48]。对于矢量化，编码残差矢量[17]被证明比编码原始矢量更有效。

在低层视觉和计算机图形学中，为了求解偏微分方程(PDE)，广泛使用的多重网格法[3]将系统重新描述为多个规模上的子问题，其中每个子问题负责较粗和较细规模之间的残差解。多重网格的另一种选择是分层基础预处理[45，46]，它依赖于代表两个尺度之间的残差向量的变量。已经证明[3，45，46]这些求解器比不知道解的残差性质的标准求解器收敛得快得多。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。

导致捷径连接的实践和理论研究由来已久[2，34，49]。训练多层感知机(MLP)的早期做法是添加从网络输入到输出的线性层[34，49]。在[44，24]中，几个中间层直接连接到辅助分类器，用于处理消失/爆炸梯度。文献[39，38，31，47]提出了将中间层响应、梯度和传播误差实现快捷连接。在[44]中，“初始”层由一个快捷分支和几个深层分支组成。

在我们工作的同时，“高速网络”[42，43]提供了与门控功能[15]的快捷连接。这些门依赖于数据并且有参数，这与我们的识别快捷方式是无参数的形成对比。当门控捷径为“闭合”(接近于零)时，高速网络层代表非残差功能。相反，我们的公式总是学习残差函数；我们的识别快捷方式永远不会关闭，所有信息都会通过，还有其他剩余函数需要学习。此外，高速网络层并没有随着深度的极大增加(例如，超过100层)而表现出精确度的提高。

3深度残差学习

3.1.。残差学习

让我们考虑H(X)作为由少量堆叠层(不一定是整个网)拟合的底层映射，其中x表示这些层中的第一个层的输入。如果假设多个非线性层可以渐近逼近复杂函数，则等价于假设它们可以渐近逼近残差函数，即H(X)−x(假设输入和输出具有相同的维数)。因此，我们不是期望堆积层近似H(X)，而是显式地希望堆积层近似残差函数F(X)：=H(X)−x。因此，原始函数变成F(X)+x。虽然这两种形式应该能够渐进地近似期望的函数(如假设的)，但学习的难易程度可能不同。

这种重新表述的动机是关于退化问题的违反直觉的现象(图1，左)。正如我们在引言中讨论的那样，如果添加的层可以被构造为识别映射，则较深的模型的训练误差不应大于较浅的对应模型。退化问题表明，求解器在用多个非线性层近似单位映射时可能有困难。利用残差学习重构，如果单位映射是最优的，则解算器可以简单地将多个非线性层的权重推向零以接近单位映射。

在实际情况中，身份映射不太可能是最优的，但我们的重新表述可能有助于预先确定问题的条件。如果最优函数更接近于恒等式映射而不是零映射，则求解器应该更容易参考恒等式映射找到扰动，而不是将该函数作为新函数学习。我们通过实验(图7)表明，学习的残差函数通常有小的响应，这表明身份映射提供了合理的预处理。

3.2.使用快捷方式进行识别映射

我们对每几个叠加层采用残差学习。构建块如图2所示。形式上，我们认为构建块定义为：



这里x和y是所考虑的层的输入和输出向量。函数F(x，{Wi})表示要学习的残差映射。对于具有两层的图2中的示例，F=W2σ(W1x)，其中σ表示RELU[29]，并且为了简化表示法而省略了偏置。运算F+x通过快捷连接和逐个元素相加来执行。我们采用加法后的二阶非线性(即σ(Y)，见图2)。

公式(1)中的快捷连接既不引入额外的参数，也不增加计算复杂度。这不仅在实践中很有吸引力，而且在比较平面网络和剩余网络时也很重要。我们可以公平地比较同时具有相同数量的参数、深度、宽度和计算成本的平坦/残差网络(除了可以忽略的元素相加)。

在公式(1)中，x和F的尺寸必须相等。如果不是这种情况(例如，当改变输入/输出通道时)，我们可以通过快捷连接执行线性投影Ws以匹配尺寸。



我们也可以使用公式(1)中的方阵Ws。但是我们将通过实验证明，同一性映射对于解决退化问题是足够的，并且是经济的，因此只有在匹配维度时才使用Ws。

剩余函数F的形式是灵活的。本文中的实验涉及一个函数F，它有两层或三层(图5)，而更多层是可能的。但如果F只有一层，则公式(1)类似于线性层：y=W1x+x，我们没有观察到它的优点。

我们还注意到，尽管为了简单起见，上述符号是关于全连通层的，但它们也适用于卷积层。函数F(x，{Wi})可以表示多个卷积层。逐个通道地对两个特征映射执行逐个元素的相加。

3.3.网络架构

我们测试了各种平面/残留网，观察到了一致的现象。为了提供实例供讨论，我们描述了ImageNet的两个模型，如下所示。

纯网络。我们的平原基线(图3，中间)主要受到VGG网[41](图3，左)的哲学启发。卷积层大多具有3×3个滤波器，并遵循两个简单的设计规则：(I)对于相同的输出特征图大小，各层具有相同数目的滤波器；(Ii)如果特征图大小减半，则滤波器数目增加一倍，以保持每层的时间复杂度。我们通过跨度为2的卷积层直接执行下采样。网络以全局平均池层和使用Softmax的1000路全连接层结束。图3(中间)的加权层总数为34层。

值得注意的是，我们的模型比VGG网[41](图3，左)具有更少的过滤器和更低的复杂度。我们的34层基线有36亿个Flops(乘加)，仅为VGG-19(196亿个Flops)的18%。

图3.ImageNet的示例网络体系结构。左：参考VGG-19型[41](196亿FLOP)。中间：具有34个参数层(36亿FLOP)的平面网络。右图：具有34个参数层(36亿FLOP)的剩余网络。虚线快捷键增加了尺寸。表1显示了更多详细信息和其他变体。

剩余网络。在上述普通网络的基础上，我们插入捷径连接(图3，右侧)，将网络转换为其对应的剩余版本。当输入和输出的尺寸相同时，可以直接使用身份快捷方式(公式(1))(图3中的实线快捷方式)。当维度增加时(图3中的虚线快捷方式)，我们考虑两个选项：(A)快捷方式仍然执行身份映射，并为增加维度填充额外的零个条目。此选项不引入额外参数；(B)公式(2)中的投影快捷方式用于匹配尺寸(通过1×1卷积完成)。对于这两种选项，当快捷键跨越两种大小的要素地图时，它们的执行步长为2

3．4 实现

我们对ImageNet的实现遵循[21，41]中的实践。在[256,480]中对图像的短边进行随机采样以进行比例放大[41]，从而调整图像的大小。从图像或其水平翻转中随机采样224×224个裁剪，减去每个像素的平均值[21]。使用了[21]中的标准颜色增强。我们采用批次归一化(BN)[16]，即在每次卷积之后和激活之前，紧跟在[16]之后。我们如[13]中那样初始化权重，并从头开始训练所有平面/残差网络。我们使用SGD，小批量为256。学习速率从0.1开始，在误差平台期除以10，模型的训练次数可达60×104次。我们使用0.0001的权重衰减和0.9的动量。我们不使用辍学[14]，遵循[16]中的做法。

在测试中，为了进行比较研究，我们采用标准的10作物测试[21]。为了获得最好的结果，我们采用了像[41，13]中那样的完全卷积形式，并在多个尺度上平均得分(图像被调整大小，使得较短的一侧在{224,256,384,480,640}中)。

4 实验

4．1 图片网络分类器

我们在包含1000个类的ImageNet 2012分类数据集[36]上对我们的方法进行了评估。该模型在128万幅训练图像上进行训练，并在5万幅验证图像上进行评估。我们还获得了测试服务器报告的10万张测试图像的最终结果。我们评估前1名和前5名的错误率。

纯网络公司。我们首先评估了18层和34层的平面网。34层的平面网如图3(中间)所示。18层的平面网也是类似的形式。有关详细架构，请参见表1。

表2中的结果表明，较深的34层平面网比较浅的18层平面网具有更高的验证误差。为了揭示原因，在图4(左)中，我们比较了他们在训练过程中的训练/验证错误。我们观察到了退化问题--尽管18层平面网络的解空间是34层平面网络的子空间，但34层平面网络在整个训练过程中都有较高的训练误差。

表1.ImageNet的体系结构。方括号中显示了积木(另见图5)，其中堆叠了积木的数量。下采样由卷积3\_1、卷积4\_1和卷积5\_1执行，步长为2。

图4.关于ImageNet的培训。细曲线表示训练误差，粗曲线表示中心作物的验证误差。左：18层和34层的平面网络。右图：18层和34层的ResNet。在这个图中，与普通网络相比，残差网络没有额外的参数。

我们认为，这种优化困难不太可能是由梯度消失引起的。这些平面网络用BN[16]进行训练，确保前向传播的信号具有非零方差。我们还验证了向后传播的梯度对BN表现出健康范数。因此，向前或向后的信号都不会消失。事实上，34层的平面网仍然能够达到有竞争力的精度(表3)，这表明求解器在某种程度上是有效的。我们推测，深平面网络可能具有指数级的低收敛速度，这影响了训练误差的降低。这种优化困难的原因将在未来进行研究。

剩余网络公司。接下来，我们评估了18层和34层残差网络(ResNets)。基线架构与上面的平面网相同，只是每对3×3过滤器都增加了一个快捷连接，如图3(右)所示。在第一个比较中(右侧的表2和图4)，我们对所有快捷方式使用身份映射，对增加维度使用零填充(选项A)。因此，与普通的对应物相比，它们没有额外的参数。

我们从表2和图4中有三个主要观察到的结果。第一，情况与剩余学习相反-34层的ResNet比18层的ResNet好(2.8%)。更重要的是，34层ResNet表现出相当低的训练误差，并且可推广到验证数据。这表明退化问题在这种情况下得到了很好的解决，我们设法通过增加深度来获得精度提升。

其次，与普通的对应网络相比，34层ResNet将TOP-1错误减少了3.5%(表2)，这是由于成功减少了训练错误(图4右对左)。这种比较验证了残差学习在极深系统上的有效性。

最后，我们还注意到，18层的平面/剩余网相对准确(表2)，但18层的ResNet收敛更快(图4右对左)。当网格“不是太深”(这里是18层)时，当前的SGD解算器仍然能够找到平坦网格的良好解。在这种情况下，ResNet通过在早期阶段提供更快的收敛来简化优化。

身份识别与投影快捷方式。我们已经展示了参数自由、身份快捷方式对培训有帮助。接下来，我们研究投影快捷方式(公式(2))。在表3中，我们比较了三个选项：(A)零填充快捷键用于增加尺寸，并且所有快捷键都是无参数的(右如表2和图4)；(B)投影快捷键用于增加尺寸，其他快捷键是相同的；以及(C)所有快捷键都是投影。

表3显示，这三个选项都比普通的对应选项要好得多。B比A略好。我们认为这是因为A中的零填充维度确实没有剩余的学习。C略好于B，我们将其归因于许多(13个)投影快捷方式引入的额外参数。但A/B/C之间的微小差异表明，投影快捷方式对于解决退化问题并不是必不可少的。因此，我们在本文的其余部分不使用选项C，以降低内存/时间复杂度和模型大小。为了不增加下面介绍的瓶颈体系结构的复杂性，身份快捷方式尤其重要。

更深层次的瓶颈架构。接下来，我们将描述我们对ImageNet的更深层次的网络。出于对我们能负担得起的培训时间的考虑，我们将构建块修改为瓶颈设计4。对于每个剩余函数F，我们使用3层而不是2层的堆栈(图5)。这三层是1×1、3×3和1×1卷积，其中1×1层负责降低维度，然后增加(恢复)维度，使得3×3层成为输入/输出维度较小的瓶颈。图5显示了一个示例，其中两种设计具有相似的时间复杂度

无参数标识快捷方式对于瓶颈体系结构尤为重要。如果将图5(右)中的身份快捷方式替换为投影，则可以显示时间复杂度和模型大小增加了一倍，因为快捷方式连接到两个高维末端。因此，身份快捷方式为瓶颈设计提供了更高效的模型。

50层ResNet：我们将34层网络中的每个2层块替换为此3层瓶颈块，从而形成50层ResNet(表1)。我们使用选项B来增加维度。这个模型有38亿次失败。

101层和152层的ResNet：我们使用更多的3层块来构建101层和152层的ResNet(表1)。值得注意的是，尽管深度显著增加，152层ResNet(113亿Flops)的复杂度仍然低于VGG-16/19网络(15.3/196亿Flops)。

50/101/152层的ResNet比34层的ResNet精确得多(表3和表4)。我们没有观察到退化问题，因此从显著增加的深度中获得了显著的精度提升。深度对所有评估指标都有好处(表3和表4)。

与最先进的方法进行比较。在表4中，我们与之前最好的单一模型结果进行了比较。我们的基线34层ResNet已经达到了非常有竞争力的精确度。我们的152层ResNet的单模TOP-5验证误差为4.49%。这个单一模型的结果比之前所有的集成结果都要好(表5)。我们将六个不同深度的模型组合成一个整体(在提交时只有两个152层的模型)。这会导致测试集上3.57%的TOP-5错误(表5)。这项参赛作品获得了2015年ILSVRC的第一名。

4．2

我们在CIFAR-10数据集[20]上进行了更多的研究，该数据集包括10个类别的50k训练图像和10k测试图像。我们给出了在训练集上训练的实验，并在测试集上进行了评估。我们的重点是极深网络的行为，而不是推动最先进的结果，因此我们有意使用简单的架构，如下所示。

普通/剩余架构遵循图3(中/右)的形式。网络输入是32×32的图像，减去每个像素的平均值。第一层为3×3卷积。然后，我们在大小分别为{32，16，8}的特征地图上使用具有3×3卷积的6n层的堆栈，每个特征地图大小对应2n层。过滤器的数量分别为{16，32，64}个。子采样通过步长为2的卷积执行。网络以全局平均池、10路完全连接层和Softmax结束。共有6n+2个叠加权重层。下表汇总了体系结构：

当使用快捷连接时，它们连接到3×3层对(共3n个快捷连接)。在此数据集上，我们在所有情况下都使用标识快捷方式(即选项A)，因此我们的残差模型与普通模型具有完全相同的深度、宽度和参数数量。

我们使用0.0001的权重衰减和0.9的动量，并且采用[13]和BN[16]中的权重初始化，但是没有丢失。这些型号在两个GPU上以128的小批量进行训练。我们从0.1的学习率开始，在32k和48k迭代时除以10，在64k迭代时结束训练，这是根据45k/5k列车/Val拆分确定的。我们遵循[24]中的简单数据增强进行训练：每边填充4个像素，并从填充的图像或其水平翻转中随机采样32×32个裁剪。为了进行测试，我们只评估了原始32×32图像的单一视图。

我们比较n={3，5，7，9}，得出20、32、44和56层网络。图6(左)显示了普通网络的行为。深度平坦的网络会受到深度增加的影响，并且在更深的时候表现出更高的训练误差。这种现象与ImageNet(图4，左)和MNIST(见[42])上的情况相似，表明这样的优化困难是一个基本问题。

图6(中间)显示了ResNet的行为。同样类似于ImageNet的情况(图4，右)，我们的ResNet设法克服了优化困难，并证明了随着深度的增加，精确度会有所提高。

我们进一步探索了导致110层ResNet的n=18。在这种情况下，我们发现初始学习率0.1略大，无法开始收敛5。因此，我们使用0.01来预热训练，直到训练误差低于80%(大约400次迭代)，然后返回到0.1并继续训练。学习日程的其余部分和之前一样。这个110层的网络汇聚得很好(图6，中间)。与Fitnet[35]和Highway[42](表6)等其他深薄网络相比，它的参数更少，但仍是最先进的结果(6.43%，表6)。

层响应分析。图7显示了层响应的标准差(STD)。响应是在BN之后和其他非线性(REU/加法)之前的每个3×3层的输出。对于ResNet，这种分析揭示了剩余函数的响应强度。图7显示ResNet的响应通常比其普通对应的响应要小。这些结果支持我们的基本动机(3.1节)，即剩余函数通常比非剩余函数更接近于零。我们还注意到，较深的ResNet有较小的响应幅度，图7中ResNet-20、56和110之间的比较证明了这一点。当有更多的层时，单个ResNet层对信号的修改往往较少。

探索1000多层。我们探索了一个超过1000层的深度极大的模型。我们设置n=200，这导致1202层网络，该网络如上所述被训练。我们的方法没有优化困难，并且这种103层网络能够实现<0.1%的训练误差(图6，右)。它的测试误差仍然很好(7.93%，表6)。

但在如此激进的深度模型上，仍然存在一些悬而未决的问题。这个1202层网络的测试结果比我们的110层网络的测试结果要差，虽然两者的训练误差差不多。我们认为这是因为过于贴合。对于这个小的数据集，1202层网络可能不必要地大(19.4m)。应用诸如maxout[10]或dropout[14]之类的强正则化以在该数据集上获得最佳结果([10，25，24，35])。在本文中，我们没有使用最大输出/丢弃，而是通过设计简单地通过深薄体系结构施加正则化，而没有分散对优化困难的关注。但结合更强的正则化可能会提高结果，这一点我们将在未来进行研究。

4．3基于PASCAL和MS COCO的目标检测

该方法对其他识别任务具有较好的泛化性能。表7和表8显示了Pascal VOC 2007和2012[5]以及COCO[26]上的目标检测基线结果。我们采用更快的R-CNN[32]作为检测方法。这里我们感兴趣的是用ResNet-101取代VGG-16[41]的改进。使用这两种模型的检测实施(请参阅附录)是相同的，因此收益只能归功于更好的网络。最值得注意的是，在具有挑战性的COCO数据集上，我们获得了COCO的标准指标(MAP@[.5，.95])增加了6.0%，这是28%的相对改进。这一收获完全归功于所学到的表述。

基于深度残差网络，我们在ILSVRC和COCO 2015大赛中获得了多个赛道的第一名：ImageNet检测、ImageNet本地化、COCO检测和COCO分割。详情载于附录。

Cnn卷积神经网络

Dnn深度神经网络

梯度消失和梯度爆炸原因及其解决方案

层数越多，越容易发生梯度爆炸和梯度消失

如果每个隐藏层都使用sigmod函数，且层数过多，会发生梯度消失的问题。

解决方案：一般是使用激活函数（如 ReLU）等方式。