ResNet-Deep Residual Learning for Image Recongnition(图像识别中的深度残差学习)

**摘要**

越深的网络通常越难以训练。我们提出了一种残差学习的框架来简化神经网络的训练，从而能够训练更深层的网络。我们重新配置了网络的结构，通过直接引用输入层来学习残差，而不是直接学习目标函数。我们有充分的证据能够证明，这些残差网络更容易收敛，而且随着网络深度的增加，能够获得精度的提升。在Image net的数据集上评估152层的残差网络——是VGGNet的深度的8倍，但是比VGGNet的复杂度还要低。通过继承这些残差网络，能够在ImageNet的测试集上实现3.57%的错误率。这个成绩在ILSVRC 2015图像分类任务中赢得了第一名。我们还在CIFAR-10数据集上实现了100到1000层的深度残差网路。

增加深度在许多图像识别任务中至关重要，仅仅通过增加网络的深度在COCO目标检测数据集上就能获得28%的性能提升。深度残差网路是我们在ILSVRC& COCO 2015竞赛中获胜的基础，同时，我们还赢得了ImageNet目标检测，COCO目标检测以及图像分割的第一名。

**1.绪论**

深度卷积神经网络为图像分类任务带来了一系列的突破。深层网络天然的能够整合低中高层次的特征，以端到端的多层网络进行分类，并且，通过卷积层的堆叠能够丰富层次特征，总揽全局。有证据表明，网络深度是是十分重要的，在ImageNet竞赛的领先的提交都采用了很深的模型，深度在16-30层。在许多其他的图像识别任务中，也都能能从深层模型中获益。

随着大家争相使用更深的模型，一个问题也随之而来，是否堆叠出更深的模型，就能活得更好的学习效果？要回答这个问题，我们首先要必须面对的是臭名昭著的梯度消失以及梯度爆炸问题，它们严重阻碍了网络的收敛。通过归一化（特征归一化以及层归一化），使用随机梯度下降算法能够在很大程度上解决这个问题。

随着更深层网络的出现，一个精度退化问题也随之而来：随着网络深度的逐步加深，精度上升逐渐饱和，随之开始快速下降。不幸的是，这种精度退化问题，并不是由过拟合引起的，并且添加更多的层到模型中反而会导致更高的训练误差，参考文献[11,42]提出了这个问题，我们在实验中证明了这个问题存在。如图1所示。

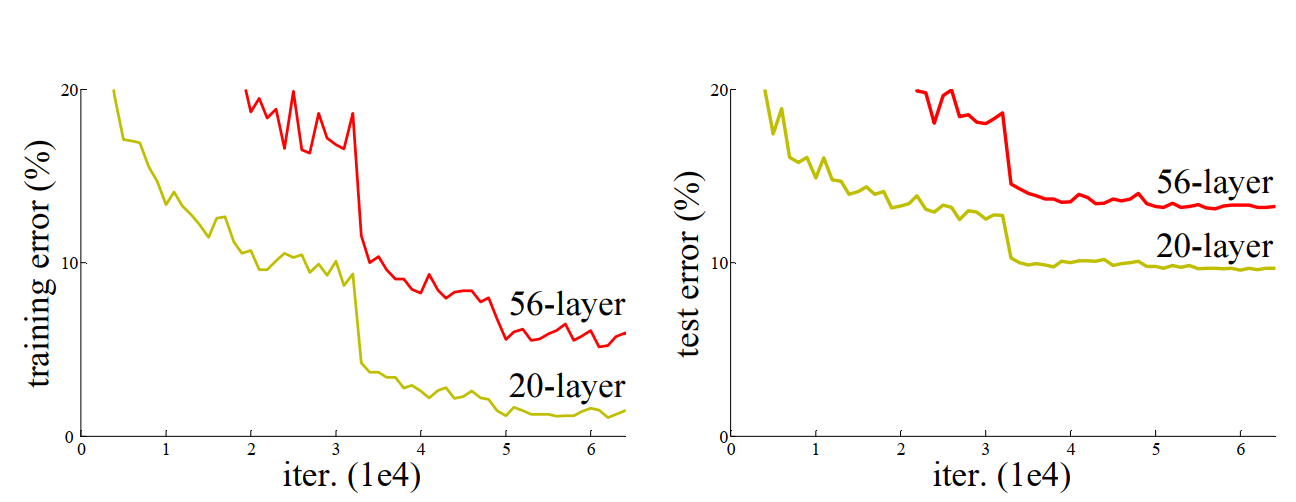


图1 训练误差（左）以及测试误差(右)

这种退化预示着，并不是所有的网络都易于优化。让我们考虑一个浅层结构以及在此基础之上构建的深层次结构。有一个解决方案是，将一些已学习的较浅的层进行简单的复制，额外增加恒等映射的层(Identity Mapping)。这个解决方案显示，一个更深的模型不会比一个浅层模型拥有更高的训练误差。但是实验显示，我们目前的解决方案，不能够找到比已有方案足够好的结果（或者在可接受的时间内，找到可接受的结果）。

在本文中，我们通过引入深度残差学习框架来解决这个问题。而不是寄希望于直接堆叠几个隐藏层来拟合目标，我们明确的让这些隐藏层去拟合一个残差。形式上，我们不是去拟合H(x)，而是让这些堆叠的非线性层去拟合残差F(x) := H(x) − x。一开始的顺序架构变成了F(x)+x。我们认为残差结构与比顺序结构，更容易优化。极端的情况，如果一个恒等映射是最优的，那么比起堆叠非线性层去拟合恒等映射，将残差F(x)优化到0要容易一些。

公式F(x)+x可以通过如图2中所示的，以捷径连结的方式直接连接到输出。捷径链接直接忽略了中间的一到几层。在我们的示例中，捷径链接是更容易作为恒等映射，其输出与非线性堆叠层的输出叠加在一起（图2）。恒等捷径连接既没有增加参数，也没有增加额外的计算复杂度。整个网络仍然能够通过随机梯度下降进行端到端的训练，现有的tensorflow、cafffe、pytorch都很容易实现，不必做更改。

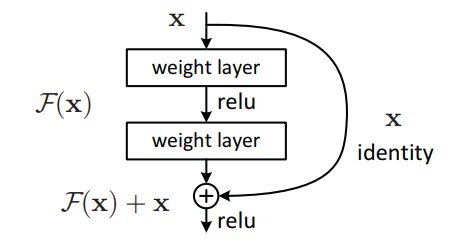


图2 残差学习中的一个残差块

我们在ImageNet的数据集上进行了全面的实验，来展示精度退化问题，并对我们提出的方法进行了评估。得出了如下结论：1)我们的深度残差网络十分易于优化，而“普通的”网络（简单的堆叠非线性层）在深度增加时会有很高的训练误差。2)我们的深度残差网络在增加深度时很容易得到精度的提升，能够得到比以前的网络更好的结果。

相同的实验结果也出现在了CIFAR-10数据集的测试，这表明传统方法的优化困境以及我们所提出方法的有效性并不是局限在特定的数据集。我们再次数据集上成功训练了超过100层的模型，并且探索了超过1000层的模型

在ImageNet的分类数据集上，我们通过深度残差网络取得了十分卓越的结果。在ImageNet数据集上我们提出采用了深达152层的残差网络，而且仍然比VGGNet拥有更低的计算复杂度。我们的模型在ImageNet的测试集上取得了top5 3.75%的错误率，赢得了ILSVRC2015分类竞赛的第一名。极深的残差网络表现形式同时拥有出色的泛化性能，让我们进一步赢得了，多个比赛的第一名。这充分的证明了残差学习是具有普适性的规则，我们希望其能够被应用到更多图像或者非图像的问题中去。

**2.相关工作**

**残差表示.** 在图像识别中，VALD（Vector of aggragate localy descriptor）[18]是一种通过对特征字典的残差向量进行编码的表示，费舍尔向量是VALD的概率版本。这两种算法都是在图像分类与检索中有效的浅层表示。对于矢量量化，编码残差向量比编码原始向量更有效。

在低层次的视觉和计算机图形学中，对于偏微分的计算，多重网格法呗广泛的应用，将偏微分的求解划分为了多个不能同尺度的子问题，这里的每一个子问题负责计算一个较大或者较小规模之间的残差。多重网格的一个替代方法是分层基预处理，依赖于两个尺度之间的向量残差表示。文献[3,45,46]表明这些使用了残差性质的解决方案的收敛速度要明显快于没有使用残差性质的传统方法。

**捷径连接.** 捷径连接[2，34，49]的实践和理论已经研究了很长时间。一个早期的实践是将一个线性层从多层感知机网络的输入直接加到输出。在文献[44,24]中，一些中间层被直接连接到后端的分类器，以此来解决梯度消失与爆炸的问题。

与我们的工作同时，“高速网络”提出了带门功能的捷径连接。这些门能够通过参数进行控制打开或者关闭。而我们的捷径连接是一直打开的。

**3.深度残差学习**

**3.1残差学习**

H(x)表示一个几层的神经网络的目标函数，其中，x表示第一层的输入。如果多层非线性层的假设能够逐渐逼近复杂函数，那么等价于它能逼近残差函数H(x)-x（假设输入和输出是相同的维度）。因此，比起通过堆叠许多非线性层去逼近H(x)，我们让这些网络层去逼近残差函数F(x):= H(x)-x。原始的目标函数变成了H(x) = F(x) + x。尽管这个两个形式都能逐渐的逼近目标函数，但是学习的难易程度是不同的。

之所以重新设计网络结构，是因为图1中违反直觉的反常现象。正如我们在绪论中提到的，如果增加一些恒等映射层，那么一个更深层次的网络结构至少不会比其浅层的结构具有更高的训练误差。退化问题表明通过增加额外的非线性层难以逼近恒等映射。通过残差学习的重构，如果恒等映射是最优的，则可以简单地将多个非线性层的权重调整到零，那么目标函数逼近的就是恒等映射。

在真实的案例中，恒等映射不可能是最优的，但是，我们的残差形式有助于解决这个问题。如果优化函数能够比零映射更逼近恒等变换，那么比起学习一个新的函数，通过参考恒等映射应该更容易发现微笑的差异。我们通过实验展示（图7），所学习的残差函数通常具有较小的响应，这表明身份映射提供了合理的预处理。

**3.2 通过捷径实现恒等映射**

我们对每几层网络就引入一个残差学习，一个残差块如图2中所示。正式地，在本文中，我们一如下的方式定义一个残差块：

y=F(x, {Wi})+x （1）

这里的x和y指的是输入和输出这些层的向量。F(x, {Wi})指的是需要学习的残差映射。在图2的示例中，有两层来学习残差，F =W2σ (W1X),这里的σ指的是ReLU激活函数，为了简化符号省略了偏差。F+X通过捷径连接来实现，逐元素相加。再相加之后，通过第二个非线性激活函数。

等式1中捷径连接没有引入额外的参数和计算复杂度。这不仅在实践中具有吸引力，而且在我们比较普通网络和残差网络时也很重要。我们可以公平地比较同时具有相同数量的参数，深度，隐藏层单元个数和计算成本（除了可以忽略的逐元素加法）的普通/残差网络。

等式1中X和F的维度必须相等。如果不相等（如，改变输入输出的通道数），我们可以对X进行一个线性投影Ws :

y=F(x, {Wi})+WsX （2）

我们也可以在等式1中使用平方矩阵Ws。我们将通过实验证明，恒等映射对于解决退化问题是有效而且经济的。Ws仅仅适用于维度匹配。

残差函数F的形式是灵活的，在本文的实验中，采用的残差的函数F有两层或者三层（图5），采用更多的层也是可能的。但是如果在等式1中的残差函数F只有一层：y=W1X+X，我们在实验中没有看到作用。

我们也注意到，尽管上面的符号是关于全连接层的，但是他们同样适用于卷积层。函数F(x; fWig)可以表示多个卷积层。 在两个特征图上逐个通道地执行逐元素加法。

**3.3 网络架构**

我们测试过多个普通网络与残差网络，观察到了一致的现象。为了提供讨论实例，我们描述了ImageNet的两个模型，如下所示。

**普通网络**. 我们简单的基线模型（图3，中间）主要受到VGG网络原理的启发[41]（图3，左）。卷积层大多采用3×3的卷积核，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，这些卷积层具有相同数量的卷积核； （ii）如果特征图的大小减半，则过卷积核的数量将增加一倍，以保持每层的时间复杂度。我们直接对卷积层进行步长为2的池化操作。网络以全局平均池化层和使用softmax激活的1000维全连接层结束。 在图3（中）所示，带权重层的总数为34。

值得注意的是，我们的模型比VGG网络[41]具有更少的卷积核和更低的复杂度（图3，左）。 我们的34层基准模型具有36亿个FLOP（乘法加法），仅占VGG-19（196亿个FLOP）的18％。

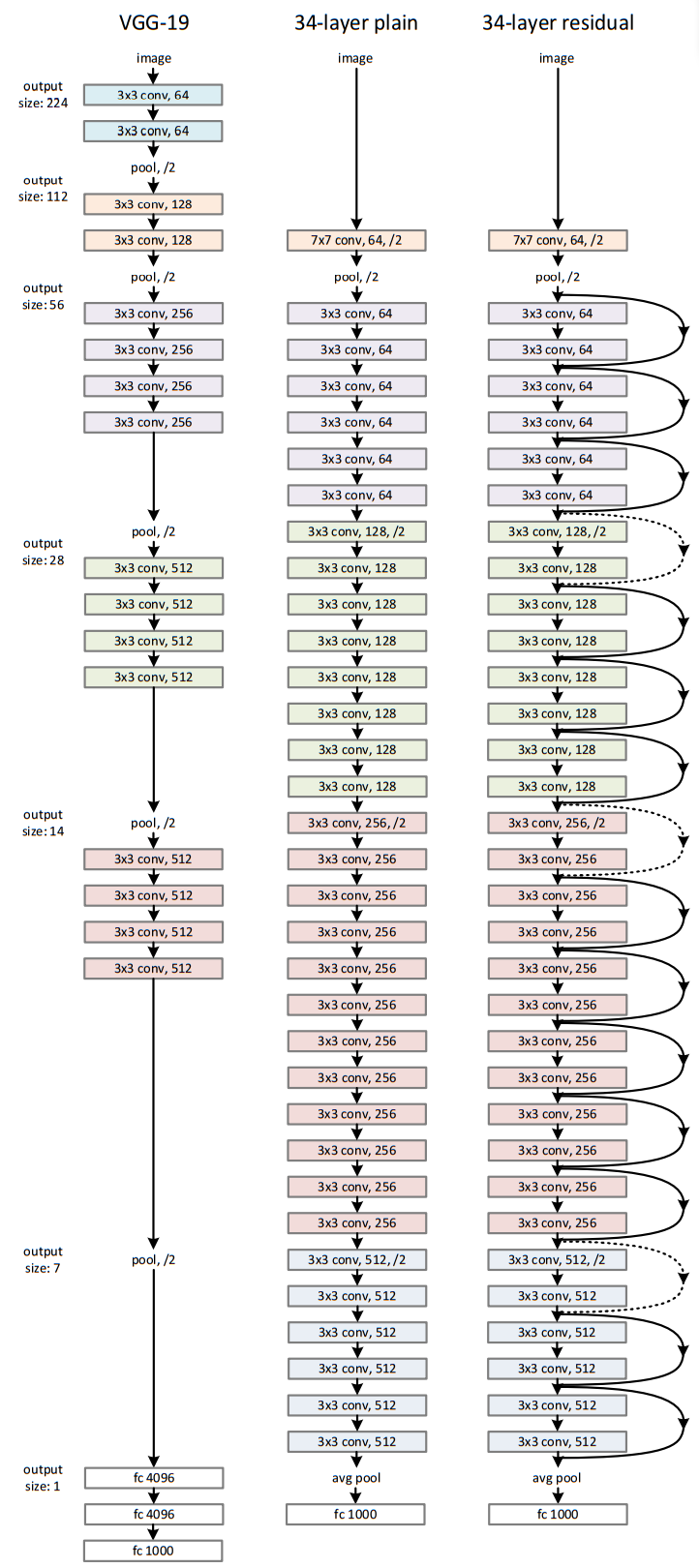


图3. 在ImageNet数据集上使用的示例网络架构。 左：作为参考的VGG-19模型[41]（196亿个FLOP）。 中：包含34个参数层（36亿个FLOP）的普通网络。 右：一个具有34个参数层的残差网络（36亿个FLOP）。 虚线快捷方式会增加尺寸。 表1显示了更多详细信息和其他变体。

**残差网络.** 基于上图的普通网络，我们插入了捷径连接，将网络结构变成了残差版本。当输入输出维度相等时，恒等捷径可以直接使用（图3中的光滑曲线）。如果维度增加（图3中的虚线），我们采用两种措施：（A）仍然采用恒等映射，通过补0来增加维度，不会引入额外的参数。（B）采用等式2中的投影操作。对于这两种操作，当捷径连接跨过两种尺寸的特征贴图时，步幅为2。

**3.4 实现**

我们对ImageNet的实现遵循文献[21，41]中的做法。重新调整图片的大小，将图片按照较短的边随机放缩到256x256或者480x480。从图像或其水平翻转中随机采样224×224的像素，并减去每像素均值[21]。同时使用文献[21]中的色彩增强。我们在每一个卷积层之后，激活层之前采用批归一化。我们以文献[13]中的方式初始化权重参数，并从头开始训练所有普通/残差网络。我们使用SGD，每一个批量大小为256。学习率从0.1开始，当误差达到稳定水平时除以10，然后对模型进行多达60×104的迭代训练。我们采用权重衰减0.0001以及动量为0.9。我们不使用随即丢弃。

在测试中，为了进行比较研究，我们采用了标准的10种裁剪测试方法[21]。 为了获得最佳结果，我们采用文献[41，13]中的全卷积形式，并在多个尺度上平均分数（图像经过调整大小，使得较短的边在{224; 256; 384; 480; 640}中）。

**4.实验**

**4.1ImageNet分类**

我们在包含1000个类的ImageNet2012数据集上评估我们的方法。模型在128w张图片的训练集上训练，使用5w张图片的验证集进行验证评估。在10w张图片的测试集上进行了测试，获得了最终结果，评估了top-1和top-5的错误率。

**普通网络.** 我们首先评估了18层和34层的普通网络。34层的普通网络如图3（中）所示。18层网络有类似的结构，表1中有详细的架构。

表2中结果表明，比起更浅的18层网络，34层的普通网络具有更更高的验证误差。为了找到原因，在图4（左）中我们比较了我们比较了他们在训练过程中的训练以及验证误差。我们观察到了退化问题——在整个训练过程中，34层的普通网络具有更高的训练误差，尽管18层普通网络的解空间是34层普通网络解空间的子空间。

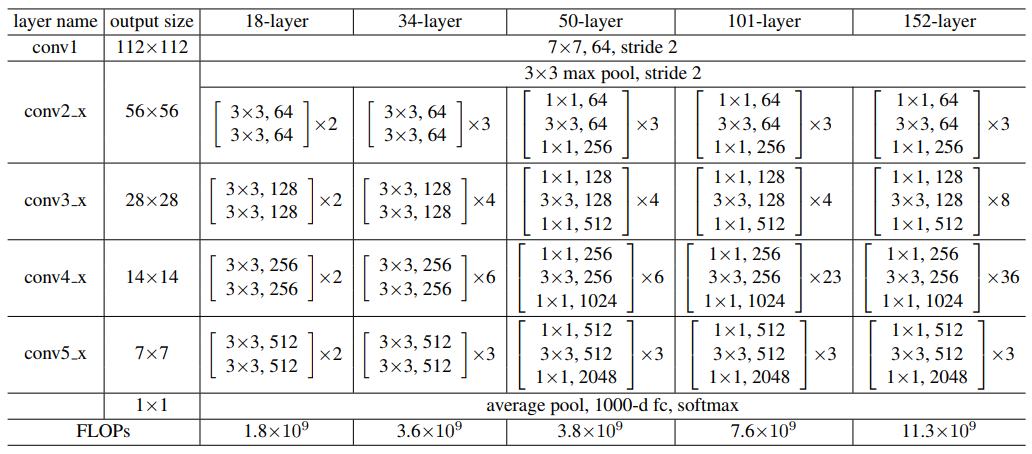


表1. ImageNet数据集上采用的残差网络结构，中括号内展示的是残差块（也在图5），将一定数量的残差块堆叠在一起，对卷积层3\_1,4\_1,5\_1尽心步长为2的下采样。

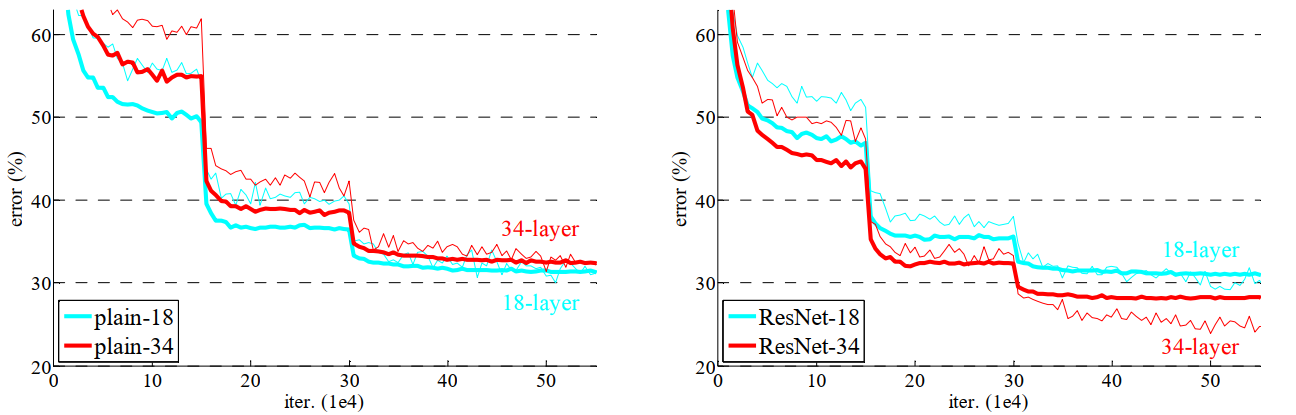


图4. ImageNet上的训练结果，细曲线表示训练误差，粗曲线表示验证误差

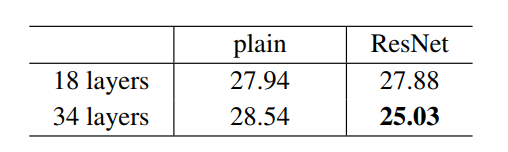


表2. top-1错误率

我们认为这种优化困境，不太可能是由梯度消失引起的，普通网络在训练时，添加了批归一化操作（卷积层与激活层之间），批归一化能够保证前向传播信号有非零方差。（均值为0，方差为1）。我们也验证了在有批归一化的情况之下，梯度的反向传播具有正确健康的形式。所以前向或者反向信号均没有消失。事实上，34层普通网络仍然可以达到竞争性的精度（表3），这表明网络在一定程度上可以工作。我们推测深层普通网络网络的收敛速度可能存在指数级的衰减，这会降低训练误差的收敛速度。这种优化困境的原因将会在将来进行研究。

**残差网络.** 接下来我们评估了18层以及34层的残差网络（ResNets），基线模型的网络结构与上面的普通网络相同，捷径连接被加到了每一个3X3的卷积核上，如图3（右）所示。在第一个比较中（表2和图4右），我们将恒等映射用于所有捷径链接，增加尺寸使用补0的方式（选项A）。 因此，与普通网络相比，它们没有额外的参数。

我们从表2和图4中得到了三个主要观察结果。首先，这种退化问题通过残差学习得到了扭转——34层ResNet优于18层ResNet（错误率降低了2.8％）。更重要的是，34层ResNet表现出低得多的训练误差，并且可以泛化到验证集。 这表明在这种情况下可以很好地解决退化问题，并且我们可以通过增加的深度来获得准确性的提高。

第二，预期对应的普通网络结构相比，34层的ResNet的top-1错误率降低了3.5%（表2），成功的降低了训练误差（图4中 右vs左）。这个 对比表明了残差学习在极深网络中的有效性。

最后，我们注意到18层的普通网络以及残差网络具有相似的训练误差，但是ResNet能够更快的收敛性（图4中 右vs左）。当网络“不是太深”（此处为18层）时，当前的SGD算法仍然能够为普通网络模型找到良好的解。 在这种情况下，ResNet通过在更快的收敛来简化优化。

**恒等映射与投影.** 我们曾经展示了不需额外参数的恒等捷径来帮助训练。接下来，我们研究投影捷径（等式2）.在表3中，我们比较了三种形式：（A）补零捷径用于增加维度，并且所有捷径都是无参数的（与表2和右图4相同）；（B）对应维度增加的用projections shortcuts，没有增加的用identity shortcuts；（C）所有的捷径都使用投影。

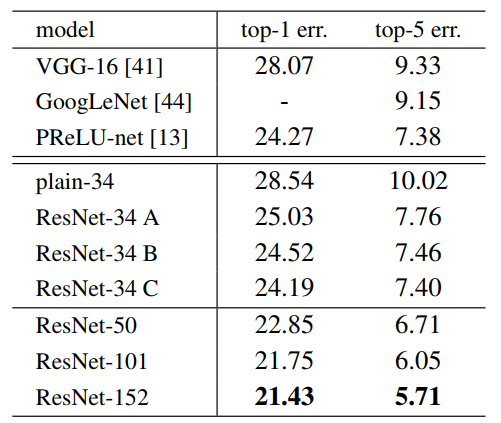


表3. 错误率，VGG-16是基线模型，ResNet-50/101/152使用使用选项B进行维度扩展

表3中展示了，三种操作均优于对应的普通网络。B略好于A。我们认为，补零的方式没有起到残差学习的作用。C略好于B，我们将此归因于许多（十三）投影捷径引入的额外参数。但是，A / B / C之间的细微差异表明，投影捷径对于解决退化问题并不是必不可少的。 因此，在本文的其余部分中，我们不会使用操作C来减少内存/时间的复杂性和模型大小。恒等捷径方式对于不增加下面介绍的Bottleneck架构的复杂性尤其重要。

**更深的Bottleneck架构.** 接下来，我们介绍更深的网络结构，考虑到我们能够等待的训练时间，我们更改残差结构为瓶颈式设计。对于每一个残差函数F，我们使用3层来代替前面的两层（图5所示）。三层分别是，1X1，3X3，1X1的卷积层，最一个个1X1的卷积层负责还原数据的维度。3X3的卷积层的数据维度小于输入输出的数据维度，类似于一个瓶颈。图5中展示了一个示例，具有相同的时间复杂度。

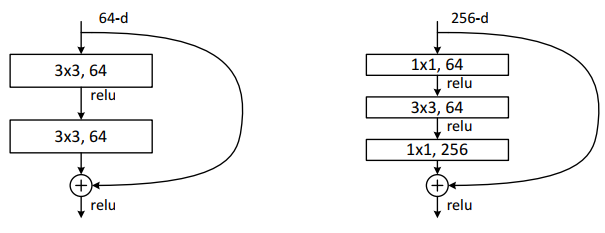


图5更深的残差函数

无参数恒等捷径对于bottleneck结构特别重要。 如果将图5（右）中的恒等捷径替换为投影捷径，则时间复杂度和模型大小增加了一倍，因为捷径连接到两个高维端。 因此，恒等捷径可以为Bottleneck结构提供更为高效的设计。

**50层ResNet.** 我们将34层的ResNet中的2层结构替换为3层结构，构造一个50层的ResNet（表1）。使用操作B来扩展维度，这个模型有38亿FLOPs（floating-point operations per second）。

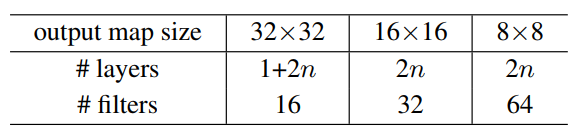
**101层和152层ResNet.** 我们使用更多的3层残差结构构建101层以及152层的ResNet（表1）。值得注意的是，尽管深度显着增加，但152层ResNet（113亿个FLOPs）的计算复杂度仍低于VGG-16/19 net（153/196亿FLOPs）。

50/101/152层ResNet比34层ResNet的准确度更高（表3和表4）。我们没有观察到退化问题，因此可以通过增加深度来获得了显着的精度提升。 所有评估指标都证明了增加ResNet深度的好处（表3和表4）。

**与最新方法的比较.** 在表4中，我们与以前的最佳单模型结果进行了比较。 我们作为基准的34层ResNet获得了非常具有竞争力的准确性。 我们的152层ResNet的单模型top-5验证错误为4.49％。 这个单一模型的结果优于所有先前的整体结果（表5）。 我们将六个不同深度的模型组合在一起以形成一个集成模型（提交时只有两个152层模型）。 这在测试集上达到了3.5%的top-5错误（表5）。 此项获得了ILSVRC 2015的第一名。

**4.2. CIFAR-10数据集以及分析.** 我们对CIFAR-10数据集进行了更多研究，该数据集包含10个类别，5万个训练图像和1万个测试图像。我们在训练集上训练，并在测试集上进行评估。我们关注的是极深网络的表现，而不是关注最先进的结果，因此我们特意使用了如下的简单架构。

普通/残差架构遵循图3中（中/右）的形式。网络的输入是32x32大小的图片，每个像素减去像素均值（去均值）。第一层是3x3的卷积层。然后，堆叠6n个网络层，在大小{32，16，8}尺寸的特征图上分别使用3x3的卷积，每个特征图尺寸有2n层。卷积核的数量分别是{16，32，64}。卷积之后的下采样步长为2。网络的最后是一个全局平均的池化层，一个10维输出的全连接层，以及softmax激活。共有堆叠了6n+2个带权网络层。下表总结了网络架构：



使用捷径连接时，它们连接到成对的3×3层（总共3n个捷径链接）。在本数据集中，我们使用的是恒等捷径，因此，残差模型拥有与普通网络严格相等的深度，宽度以及参数数量。

我们使用0.0001的权重衰减系数，动量0.9，使用文献[13]中的权重初始化方式，以及批归一化，不使用随机丢弃。这些模型以每批128个样本在两个GPU上训练。初始学习率为0.1，在32k以及48k迭代时除以10，在第64k迭代时结束训练，以45k/5K的比例划分训练集以及验证集。我们按照[24]中的方式进行简单数据增强：在图片每侧填充4个像素，并从填充的图像或其水平翻转中随机采样32×32大小的裁剪。 在测试中，我们仅评估原始32×32图像的单个视图。

我们比较n={3，5，7，9}时，构建的20，32，44以及56层的网络。图6（左）展示了普通网络的表现。普通网络随着深度的增加，会有更高的训练误差。这个现象ImageNet和MNIST数据集上是相似的，说明这是一个广泛存在的基本问题。

图6（中）展示了ResNet的行为。现象也类似于ImageNet（图4，右）。我们的ResNet克服了优化困境，并且随着深度的增加精度也在上升。

我们进一步探索了n=18时的110层的ResNet。在这个模型中，我们发现采用0.1的初始学习率太大以至于无法开始收敛（振荡）。因此我们使用0.01来预热训练，直到训练误差低于80％（约400次迭代），然后回到0.1并继续训练。其余的学习过程与之前一样。 这个110层的网络可以很好地收敛（图6，中间）。 与其他深层和窄网络（例如FitNet [35]和Highway [42]）相比，它具有更少的参数（表6），目前仍属于最好的结果（6.43％，表6）。

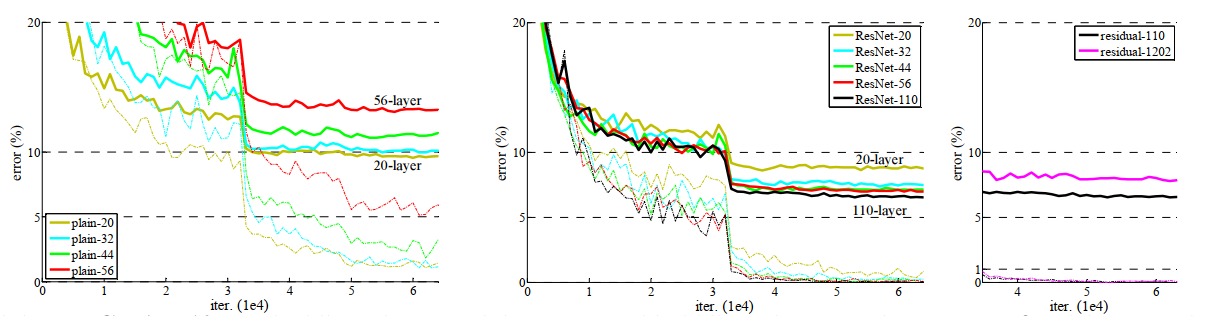


图6.有关CIFAR-10数据集的训练。 虚线表示训练误差，而粗线表示测试误差。 左：普通网络。 Plain-110的错误高于60％，没有显示在图中。 中：ResNets。 右：具有110和1202层的ResNet。

**每层响应分析**. 图7展示了每层响应的标准偏差。响应是每个3x3层的输出，批归一化之后，非线性激活函数之前。对于ResNet，此分析揭示了残差函数的响应强度。图7显示ResNet的响应通常比对应的普通网络响应小。 这些结果支持了我们的基本动机（第3.1节），即与非残差函数相比，残差函数通常可能更接近于零。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的比较所证明的。当有更多层时，ResNets的单个层往往会修改信号较少。

**探索超过1000层的网络.** 我们探索了一个超过1000层的深度模型。 我们将n设置为200，从而得出1202层网络，该网络如上所述进行了训练。 我们的方法没有优化困难，该103层网络能够实现训练误差<0.1％（图6，右）。 其测试误差仍然相当不错

（7.93％，表6）。但是，在如此激进的深度模型上仍然存在未解决的问题。 尽管这两个1202层网络的训练误差相似，但其测试结果却比我们的110层网络的测试结果差。我们认为这是过拟合的原因。对于这个小的数据集，1202层网络可能会太大了（19.4M）。 应用强正则化（例如maxout [10]或dropout [14]）以在此数据集上获得最佳结果（[10、25、24、35]）。在本文中，我们不使用maxout / dropout，而只是通过通过设计深度和精简架构进行正则化，注意力主要集中在对优化困难上。 但是，结合更强的正则化可能会改善结果，我们将在以后进行研究。

**4.3 PASCAL和MS COCO上的目标检测.** 我们的方法在其他识别任务上具有良好的泛化性能。 表7和8显示了PASCAL VOC 2007和2012[5]和COCO[26]上的对象检测基线

结果。我们采用Faster R-CNN [32]作为检测方法。在这里，我们对用ResNet-101替换VGG-16 [41]的改进感兴趣。使用这两种模型的检测实现方式（请参见附录）是相同的，因此只能将收益归因于更好的网络。 最引人注目的是，在具有挑战性的COCO数据集上，我们的COCO标准指标（mAP@ [.5，.95]）增加了6.0％，相对改进了28％。 该收益完全归因于所学的表示。

基于深层残差网络，我们在ILSVRC和COCO2015竞赛的多个赛道上均获得了第一名：ImageNet检测，ImageNet定位，COCO检测和COCO分割。详细信息在附录中。