基于图片识别的深度残差学习

概述

更深的神经网络非常难以训练，我们发现了一个残差的学习框架让之前的更深的网络训练得更加容易。我们重新分开定制看层用残差函数和先前层的输入实现，来取代用无关联的函数学习。我们用了很好理解的经验验证发现，这些残差学习网络很容易优化并且可以在增加精确度的同时增加深度。在ImageNet上我们用了152层的残差网络，虽然比VGGnet层数多了八倍，但是却有更小的复杂性。这些残差层在ImageNet数据集上测试可以达到3.57%的错误率。这赢过了2015年ILSVRC的第一名。我们也发现了在CIFAR-10第100层和1000层上的分析方法。

深度的表现对于很多视觉识别任务是很重要的。仅仅是由于我们及其深的深度表现，我们达到了28%相关性提升在COCO目标检测数据集上。深度残差网络是ILSVRC和2015年COCO比赛上，让我们赢得了ImageNet检测，IMageNet本土化，COCO检测和COCO分割上的第一名。

介绍

深度卷积神经网络在图片分类取得了一系列重大的突破。深度网络通常整合了低中高的特征并且对不同类的风格进行端对端的分类，与此同时，特征的level可以用一堆层叠的层来填充。最近事实揭露了网络的深度是具有重要决定性的。在训练ImageNet数据集的学习结果都都被发现用了非常深的网络（16-30层）。很多没实现的视觉识别的任务也在很深的模型中得益。

鉴于深度的重要性，出现了一个问题：更好的学习网络和堆叠的层一样简单吗？回答这个问题的一个障碍就是一个关于梯度消失和爆炸，这个阻碍在最开始就融合了。然而，这个问题很大取决于初始正规化和中间的标准化层，这让保证了十层网络在反向传播中随机梯度下降开始融合。

当更深的网络能够开始融合，一个降解的问题暴露出来了：随着网络深度增加，精确度开始趋于饱和并且开始快速下降。不巧的是，这种下降不是因为过拟合，而是过多增加的网络层需要适应更深的模型导致更高的训练错误。如图一所示这证明了这个观点。

这种下降表明了不是所有的网络都是一样容易优化的。让我们考虑一下更浅的结构并且使他能够和增加更多层效果相当。这有一个办法来建立一个更深的模型：加入的层有了自己的编号谱，而其他层就从这个浅层的学习模型中复制过去。这个解决方法表明了更深的模型不应该由训练错误产生而是有与之相对于的更浅的模型来产生。但是实验表面我们目前手头上的求解方法没有办法在可行的时候这么做。

在这篇文章中，我们提到了用深度残差学习框架来解决问题。不同于希望每一叠加的层直接适应底层的图谱，我们将他们分开去适应残差图谱。算式上来说设当前希望得到的图谱是H(x)，我们让堆叠层非线性的层去适应其他的图谱F(x) = H(x) - x；最初的图谱在F(x)+x上重铸。我们猜测优化残差图谱比优化原先的没被用到的图谱要简单很多。更极端的，如果给图谱编号是最佳的，那么在堆叠的非线性层将残差的值为零比起适应一个映射好的图谱要更简单。

算式F(x)+x可以在有“捷径连接”的反馈神经网络上实现。捷径连接是跳过一个或者多个层。在我们这儿，捷径连接是简单的表现在映射图谱上，并且他们的输出也被加到了堆叠层的输出上了。映射捷径连接不仅仅没有加入额外的参数也么有加入更加复杂的计算。整个网络然而是用反向传播的SGD端到端实现，可以很容易的用通常的库不需要校准solver来实现。

我们目前综合了一个实验在ImageNet上表明了降解问题和我们的计算方法。我们表明了：1）我们极端深残差网络是很容易优化的，但是相对应的“平”网络（简单是堆叠层）展现了更高的训练错误率当深度增加。2）我们深度残差网络可以在深度增加时很容易使得精确度增加，产生的结果比之前的网络要好。

相似的现象也在CIFAR10上表现出来。这表明了优化的困难和我们方法的有效性不会依赖于特定的数据集。我们现在很成功的训练出了模型在100层，探索模型在超过1000层。

在ImageNet数据集上，我们用极端深度残差网络获得了很好的结果。我们的152层残差网络是在ImageNet上训练的迄今为止最深的网络，也比VGG要有更低的复杂性。我们最终的ImageNet数据集top-5错误率在3.57%，赢得了ILSVRC2015分类比赛的第一名。极端深度残差网络也有很好的表现在很多识别任务上，这使我们在ImageNet检测，IMageNet本土化，COCO检测和COCO分割上的比赛中得到第一。这更强烈的表明了残差学习原则是通用的。我们希望它也能运用到更多其他的视觉和非视觉问题上。

相关的工作

残差表现：在图片识别中，VLAD是尊重字典用残差向量进行编码的表现，Fisher Vector是在可能的情况下VLAD的算式。他们两个都在图片训练和分类上有很强大的浅的表现。在向量上，残差向量编码比原有向量编码要更有效。

在低级视觉和计算机图形学上，为了解决部分不同等式（PDEs）。最常用的方法多重网格方法对系统并发为在多重领域上进行重构，每个并发问题都对在粗略的有限的领域之间 残差解决有影响。一种对多重网格的改变是将之前网络的依赖于夹在两个领域之间的残差向量的变量的偏差进行分级，这样的solver比起标准的没有意识到残差自然的解决方法solver要收敛的更快，这些方法表明了一个好的重算式或先前的环境可以在优化时变得简洁。

捷径连接：捷径连接的实践和理论被研究很久了。早年训练多层感知机的实验是通过加入一个线性层连接网络的输入输出。在。。论文中提到一些中间层内直接连接了辅助分类器为了解决消失和爆炸梯度。这些论文提到了在中间层响应，求梯度和传播错误的方法，用捷径连接实现的。在这个论文中，一个“起始”层是由一个捷径分之和一些更深的分支组成。

结合我们的工作，“高速网络”利用门函数提供了捷径连接。这些门是依赖于数据并且含有参数的。相对的我们的映射捷径确实不需要参数的。当捷径门关闭了（通道值为0）在高速网络上的层就代表了无残差函数。相反的，我们的算式总是学习残差函数，我们的映射捷径是永远不会关闭的，并且在学习了额外的残差函数，所有的信息都能通过。此外，高速网络在增加了深度后（100层以上）也还没被证明精确度增加的情况。

3 深度残差学习

3.1 残差学习

让我们假设H(x)是一个底层的图谱利用前一层提供的输入x，适应了一些堆叠层（并不需要是整个网络）。如果假设多重非线性层可以近似渐进复杂的函数，那么它就等同于推测它们可以近似渐进与残差函数。H(x)-x（假设输入和输出是同一个维度）所以比起希望对叠层近似H(x)，我们更希望让这些层近似于一个残差函数F(x)=H(x)-x。这个原始的函数因此变成了F(x)+x。即使两种形式都可以近似渐进与我们所希望的函数（如假设一般）。那么学习的难易就可能不同了。

这个重构算式是被一个关于降解问题的有悖于常理的现象所启发的。就像我们在介绍里面讨论到的，如果增加层数可以用映射图谱建造起来，那么一个没有训练错误的更深的模型就不如它相对应的更浅的模型。这个降解问题表明了solver可能在利用多重非线性层估计映射图片时会有困难。随着残差学习算式的重构，如果映射图谱可以达到最佳，那么solver在驱使多重非线性层从零达到映射图谱的权重就会简单多了。

在实际情况下，映射图谱达到最佳是不可能的。但是我们的重构算式可能能够帮忙解决这个问题，如果最佳函数是从零开始接近映射图谱，那么对于solver来说，找到参考的一个映射图谱的扰动方法会更加容易，比起学习一个新的函数。图7我们用实验表面了。通常情况下，学习残差函数有很小的响应，这表明了映射图谱提供了合理的预处理。

3.2 捷径下的映射图谱

我们将残差学习用到了很少的堆叠层上。图2显示了一个构建的块。总的来说，在这个论文中我们考虑了一个构件块被定义为：y = F(x,{Wi}) + x，在这里的x和y是所考虑层的输入和输出。函数F(x,{Wi})代表了需要学习的残差图谱。在图2所示的例子中有两层，F = W2×sigma×(W1\*x)，这里的sigma就是省略后的ReLU和偏差。操作F+x体现了捷径连接和逐元素添加。我们用了第二次非线性性在这个添加后。

介绍中的捷径连接不仅没有额外的参数而且也没有增加计算的复杂性。这不仅仅在实践中很好也对我们比较普通网络和残差网络非常重要。我们可公平的比较普通和残差网络同时在有相同多的参数，深度，宽度和计算损耗（除开了微不足道的逐元素添加）。

x和Fd 维度应该是一样的，如果不是这样（例如改变输入和输出的通道），我们可以通过一个由捷径连接来做的线性项Ws来匹配维度。Y = F(x,{Wi}) + Ws\*x。我们也可以用方阵Ws。但是我们想要在实验中呈现的是映射图谱对于解决降解问题是足够的并且也是经济的，而且Ws只是仅仅在匹配维度上使用。

残差函数的形式F是灵活的。在这论文上的实验加入的函数F有两或三层，当然，更多层也是可以的。但是如果F仅仅只有一层，式子(1)就会和线性层y=W1×x+x相似，那么我们就观察不到它的优点了。

我们也注意到，即使上面对于全连接层比较简单，他们仍旧可以用在卷积层。函数F(x, {Wi})可以代表多重卷积层。逐参数增加也是在两个特征图谱上实现，从通道到通道。

3.3 网络结构

我们也测试了不同的普通和残差网络，观察到不同的现象。为了证明这里讨论的问题，我们在下面描述了两种模型对于ImageNet的结构。

普通网络。我们的普通网络主要是用了VGG网络来实现。卷积层几乎都是3\*3的过滤器并且遵循两种设计原则：(i)如果层有相同输出特征图谱大小，那么这些层就有相同数目的过滤器；(ii)如果特征图谱大小减半，过滤器的数量通过先前每一层的复杂程度来考虑。我们也在卷积层用步长为2的过滤器直接进行降采样。网络是以一个全局平均池化层和一个1000维含有softmax的全连接层结束的。整个权值层总共有34层如图3所示。

需要注意的是，我们的模型有更少的过滤器和更小的复杂性比起VGG网络。我们主要的34层只要36亿FLOPs是VGG的18%.

残差网络。在普通网络的基础上，我们加入了捷径连接让整个网络变成相应的残差网络。当输入和输出在相同维度上的时候，映射捷径可以直接用。当维度增加的时候（图中虚线），我们可以用两种方法：(A)在映射图谱上用额外的零实体填充增加的维度后实现捷径连接。这个方法不会产生额外的参数。(B)在这个文章闪有说到匹配维度(用1\*1的卷积)。这两个方法，当在两个特征图谱上实现捷径连接时，他们的步长都为2.

3.4 实现

我们在ImageNet上实现方法是根据论文。。将图片沿着短边在增强范围内随机采样重构。224\*224大小的图片是从图片上或是其水平翻转并且减掉图片像素的均值后随机采样得到的。图片增强是用了标准的颜色。我们根据论文在每一层卷积的激励前先做了批量正常化处理。我们根据论文初始化了权值，训练了所有抓取的普通/残差网络。我们在256大小的最小批量上进行随机梯度下降。。Learning rate宣威0.1并且当错误率趋于平缓时learning rate会除上10，并且模型训练了60\*10\*e(4)次迭代。我们用了0.0001的权值衰减和momentum为0.9。我们没有在实践中用到dropout。

在测试的时候，比较了标准的10-crop测试。最好的激活是我们用了全连接形式根据论文。并且求得了多重范围的平均结果。(图片在{}上沿着短边重构)。

4 实验

4.1 ImageNet 分类

我们将我们的方法用到了包含了1000类的ImageNet2012分类数据集上。模型用了128万训练图片，并且正在5万张图片上做验证。我们也在10万张图片上做了测试得到了结果，计算了top-1和top-5.

我们在18层和34层用了普通网络,34层是图三中间的模型。18层和他一样的形式。从表1可以看到结构的细节。

表2的结果表示，更深的34层有更高的验证错误比起更浅的18层。为了找到这个原因，我们比较了训练和验证的错误在训练过程中。我们观察到了降解问题。34层网络有跟高的训练错误在整个训练中，即使18层网络只是34层网络的子空间。

我们发现了这个优化困难可能不是梯度消失造成的。这个网络是用BN训练的，这保证了前向传播型号没有零变量。我们也确定了向后传播梯度在BN上呈现是正常的。所以不是前向也不是后向信号使得梯度消失。事实上，34层网络仍然能够达到对抗精确，在一定程度上表面了solver是有用的。我们推测深度网络可能有选择成倍低的收敛率，这影响了训练错误的减少。优化困难还需今后继续考究。

残差网络。接下来我们计算了18层和34层的残差网络。主要的结构和上面的网络是一样的除了在每一层3\*3过滤器上增加的捷径连接。第一次对比（右边的表2和图4）。我们用了映射图片在所有捷径上填0为了增加维度（方法A）。所有我们比起上一个网络没有多余的参数。

我们有三种主要的观察形式在图4和表2.首先，在残差网络上，情况正好相反34层的Resnet比18层resnet要更好。更重要的是，34层ResNet呈现了更小的训练错误而降解问题在这个设置中很好的解决，我们达到了在使增加深度让精确度增加。

第二，比起对应的普通网络，34层ResNet降低了top-1的错误率达3.5%，成功的降低了训练错误。这比较与之前的残差学习要更有效一些。最后我们也发现了在对比18层普通和残差网络训练精确度时18层ResNet收敛的更快。当网络不是过于深的时候，目前的SGDsolver就可以在普通网络上有一个好结果了。在这个例子中，ResNet要更好且更快的收敛在更早的平台上。

特性vs投影捷径。我们也发现在无需过多参数的情况下，映射捷径在训练中是有帮助的。接下来我们发现了投影捷径。在表3中，我们比价了三种方法：(A)增添0的捷径是在增加维度时使用并且所有的捷径不含任何参数。(B)投影捷径也在增加维度上使用并且所有的捷径都是有特性的。(C)所有捷径都是投影。

表3显示了这三个选项都比起相应的普通的更好些，B比起A要稍微好些。我们发现，这是由于填充0来增加维度，A确实没有任何的残差学习。C比起B来说要好很多，我们增加了额外的函数在很多投影捷径上。但是A/B/C上有个小小的区别表面了投影捷径对于解决降解问题并不是必须的。所有我们不需要在剩下的论文里讨论C方法，来介绍内存和时间的复杂性和模型大小。映射捷径是非常重要的在增加瓶颈结构的复杂度在下面所介绍。

更深的瓶颈结构。接下来我们要描述的是我们在ImageNet上更深的网络结构。由于对训练时间消耗的关注，我们修正了搭建的模型为瓶颈设计。每一个残差函数F，我们都用三层来代替两层，三层分别为1\*1,3\*3和1\*1卷积，这里的1\*1是为了降低并且增加维度，作为3\*3层的瓶颈呈现更小的输入和输出。图5举出了这个例子，这两种设计都有相同的时间复杂性。

不需要参数的特性捷径对于瓶颈结构非常重要。如果特性捷径在图5中被投影取代，这表示了一个就是时间复杂性和模型大小都需要考虑，当捷径连接了两个高维度端。所有特性捷径对于瓶颈模型设计是非常有效的。

50层ResNet。我们将34层网络的每两个层都替代成了三层瓶颈结构，结果在表1的50层ResNet。我们用B方法来增加维度。这个模型有38亿FLOPs。

101层和152层ResNet：我们用更多三层结构搭建了101层和152层ResNet。异常的是，即使深度很好的增加了，但是152层ResNet还是比VGG-16/19要更小的复杂性。ResNet的50/101/152层比起34层由于考虑的多所有要更精确。我们不需要观察降解问题并且可以得到适当增加层数所带来的精确度的增加。深度带来的好处都是在评测时可以看到的。

和最先进的方法对比。在表4巴拉巴拉说明优势

4.2 CIFAR-10和分析

我们也在拥有了5万训练图片和1万测试图片具有10类的CIFAR10上进行研究。我们训练也测试了这些数据集。我们主要专注于极端深度网络的结果，但是没有拿出最先进的结果，所有我们用了下面的简单是结构。

图三，网络的输入是32\*32的图片且每张图片每个像素都减去了图片像素的均值。第一层是3\*3的卷积层，然后我们用了一些6n层的3\*3的卷积层分别在大小为{32,16,8}的特征映射上做，而每个特征映射的大小有2n层。过滤器的大小分别是{16,32,64}.在卷积后面有步长为2的二次采样。网络是以全局平均池化和10维的全连接层和softmax结束的。整个网络一共有6n+2个权值层。下面的表是对结构的总结。当用到了捷径连接时，我们连接了一些3\*3的层总共就是3n次捷径。在整个数据集上，我们用了特征捷径在所有情况下（也就是A选项）。所以我们的残差模型就有准确的深度宽度和参数个数相较于plain。

我们用了权值衰减0.0001和momentum为0.9并且初始化权值做了BN但是没有做dropout。这些模型用了128大小每次在两个GPU上训练。我们最初的learning rate是0.1，在第32000次和48000次迭代时除以10.在最终到了64000次迭代，。。。我们准信了一个简单的数据增强在训练时：在每个边增加4像素，在32\*32的原图和水平翻转后样本上随机采样。测试时，我们仅仅用了原始的32\*32图片测试。

我们比较了n={3,5,7,9}用了20,32,44和56层网络。图6就是网络的表现。深度平原网在增加深度的时候会呈现更高的训练错误。这个现象在ImageNet和MNIST上也可以看到了优化困难是一个最基本的问题。

图6中间显示了ResNet的表现。同时和IMageNet相似的是，ResNet在深度增加时，还需克服优化困难和样本精确度增加。

我们更远的探索了n=18,110层ResNet。在这，我们发现初始化learning rate为0.1对于开始收敛的时候就稍微过大了。所有我们用0.01当训练错误在80%差不多400次迭代的时候就补偿为0.01。剩下的学习计划在之前完成了。这个110层网络收敛的很好。它有更少的岑姝比起其他深度网络如FitNet和Highway。目前是最先进的结果。

层响应分析。图7显示了层响应的标准差。响应是在BN之后非线性激励之前的3\*3网络输出。对于ResNet，这次分析揭露了残差函数响应的优点。图7表面了ResNet通常比起相应的平原网要有更小的响应。这些结构支持了我们的基础动机，残差函数可能生成的比起非残差函数要更接近。我们也注意到更深的ResNet有更小部分是响应比起20层56层层和110层。当有更多层是，一个单独的ResNEt层就对信号修正更少。

在超过1000层的ResNet。1202层，大概说的是效果也不错。训练错误差不多，但是在测试的时候没有110层的好。是应为过拟合了。对于小数据集上1202层太大了。所以为了更好，在上面加了强正规化如maxout和dropout。这个论文里面没有maxout和dropout，但是加了简单是脉冲正规化在设计通过深度和单模型，所有没有把重心放在这。加入正规化可能有效，所以今后可以多研究。