广义的剩余网络

Sergey Zagoruyko
sergey.zagoruyko@enpc.fr
尼古斯-科莫达基斯
nikos.komodakis@enpc.fr

巴黎东区大学，巴黎科技大学桥梁学院
法国，巴黎

摘要

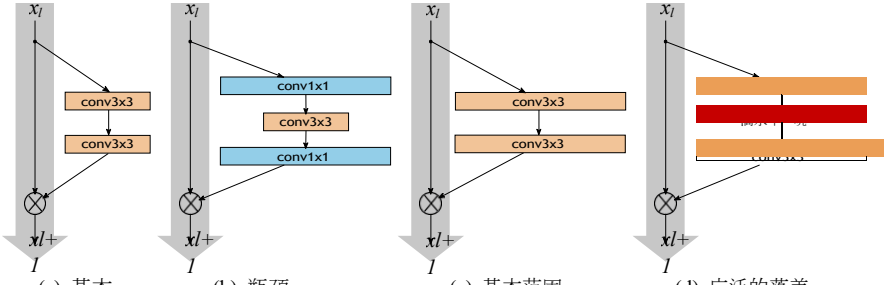
深度残差网络被证明能够扩展到数千层，并且仍然具有不断提高的性能。然而，每提高一个百分点的准确率，都要付出层数翻倍的代价，因此，训练非常深的残差网络有一个特征重用率下降的问题，这使得这些网络的训练速度非常慢。为了解决这些问题，本文对ResNet块的结构进行了详细的实验研究，在此基础上，我们提出了一种新的结构，即降低剩余网络的深度并增加宽度。我们把由此产生的网络结构称为宽残差网络（WRNs），并表明这些网络结构远远优于常用的薄和非常深的对应结构。例如，我们证明，即使是一个简单的16层深的宽残差网络，在精度和效率上也优于以前所有的深残差网络，包括千层深的网络，在CIFAR、SVHN、COCO上取得了新的最先进的结果，在ImageNet上也有明显的改进。我们的代码和模型可在https://github.com/szagoruyko/wide-residual-networks。

1 简介

在过去的几年里，卷积神经网络的层数逐渐增加，从AlexNet[16]、VGG[26]、Inception[30]到Residual[11]网络工程，对应于许多图像识别任务的改进。深度网络的优越性在近几年的一些工作中已经被发现[3, 22]。然而，训练深度神经网络有几个困难，包括爆炸/消失的梯度和退化。人们提出了各种技术来实现深层神经网络的训练，如精心设计的初始化策略[1，12]，更好的优化器[29]，跳过连接[19，23]，知识转移[4，24]和分层训练[25]。

最新的残差网络[11]取得了巨大的成功，赢得了ImageNet和COCO 2015比赛，并在一些基准中取得了最先进的成绩，包括ImageNet和CIFAR的物体分类，PASCAL VOC和MS COCO的物体检测和分割。与Inception架构相比，它们显示出更好的泛化能力，这意味着这些特征可以以更好的效率被用于迁移学习。另外，后续工作显示，残余链接加快了深度网络的收敛速度[31]。最近的后续工作探索了残差网络的激活顺序，提出了残差块中的身份映射[13]，改善了非常深的网络的训练。通过使用高速公路网

络，成功训练非常深的网络也被证明是可能的



(a) 基本的 (b) 瓶颈 (c) 基本范围 (d) 广泛的落差

图1：本文中使用的各种残差块。每个卷积之前都有批量归一化和ReLU（为清晰起见省略）。

[28]，这是一个在残差网络之前就已经提出的结构。残差网络和高速公路网络的本质区别在于，在后者中，残差链接是有门的，这些门的权重是学习的。

因此，到目前为止，对残差网络的研究主要集中在ResNet块内部激活的顺序和残差网络的深度。在这项工作中，我们试图进行一项超越上述几点的实验研究。通过这样做，我们的目标是探索一套更丰富的ResNet块的网络结构，并彻底检查除了激活顺序之外的其他几个不同方面如何影响性能。正如我们在下面解释的那样，这样的架构探索导致了新的有趣的发现，关于残余网络具有很大的实际意义。

剩余网络中的宽度与深度。在机器学习中，浅层网络与深层网络的问题已经讨论了很久[2,
18]，电路复杂度理论文献的指向性表明，浅层电路比深层电路需要的元件会成倍增加。残差网络的作者试图使它们尽可能薄，以利于增加它们的深度和拥有更少的参数，甚至引入了一个 "瓶颈 "块，使ResNet块变得更薄。

然而，我们注意到，允许训练非常深的网络的残差块与身份映射同时也是残差网络的一个弱点。当梯度流经网络时，没有什么可以迫使它通过剩余块的权重，它可以在训练中避免学习任何东西，所以有可能要么只有少数块学习有用的表征，要么许多块分享很少的信息，对最终目标的贡献很小。这个问题在[28]中被表述为特征重用的递减。[14]的作者试图通过在训练中随机禁用残余块的想法来解决这个问题。这种方法可以看作是dropout[27]的一个特例，其中每个残余块都有一个身份标量权，在此基础上应用dropout。这种方法的有效性证明了上述假设。

在上述观察的激励下，我们的工作建立在[13]的基础上，试图回答深度残差网络应该多宽的问题，并解决训练的问题。在这种情况下，我们表明，与增加深度相比，拓宽ResNet块（如果做得好的话）为提高残差网络的性能提供了一种更有效的方法。特别是，我们提出了更宽的深度残差网络，比[13]有了明显的改善，层数减少了*50倍*，速度提高了2倍以上。我们把由此产生的网络结构称为*宽残差网络*。例如，我们的宽16层深度网络具有与1000层薄深度网络相同的精度和相当数量的参数，尽管其训练速度快了几倍。这种类型的实验

因此，似乎表明深层残差网络的主要力量是在残差块中，而深度的影响是补充性的。我们注意到，人们可以训练出甚至更好的具有两倍参数（甚至更多）的宽残差网络，这表明为了通过增加薄网络的深度来进一步提高性能，在这种情况下需要增加成千上万的层。

在ResNet块中使用dropout。Dropout首次被引入[27]，然后被许多成功的架构采用，如[16，26]等。它主要应用于有大量参数的顶层，以防止特征共适和过度拟合。随后，它主要被批量归一化[15]所取代，该技术通过归一化来减少神经网络激活的内部协变量偏移，使其具有特殊的分布。它也可以作为一个正则器，作者在实验中表明，具有批处理规范化的网络比具有剔除功能的网络取得了更好的准确性。在我们的案例中，由于残差块的扩大导致了参数数量的增加，我们研究了dropout对于规范化训练和防止过拟合的影响。以前，残差网络中的dropout在[13]中被研究，dropout被插入到区块的身份部分，作者显示了其负面效应。相反，我们在这里认为，dropout应该被插入卷积层之间。宽残差网络的实验结果表明，这导致了一致的收益，甚至产生了新的最先进的结果（*例如*，16层深的宽残差网络与dropout在SVHN上实现了1.64%的误差）。

综上所述，这项工作的贡献如下。

- 我们提出了一个详细的残留网络结构的实验研究，彻底检查了ResNet块结构的几个重要方面。
- 我们为ResNet块提出了一种新颖的*加宽*结构，使残余网络的性能得到显著提高。
- 我们提出了一种在深度残差网络中利用dropout的新方法，以便在训练过程中适当地规范它们并防止过度拟合。
- 最后，我们表明，我们提出的ResNet架构在几个数据集上取得了最先进的结果，极大地提高了剩余网络的准确性和速度。

2 广泛的剩余网络

具有身份映射的剩余区块可以用以下公式表示。

$$\mathbf{x}_{l+1} = \mathbf{x}_l + F(\mathbf{x}_l, W_l) \tag{1}$$

其中x_{l+1}和x_l是网络中*第l个*单元的输入和输出，F是一个残差函数，W_l是块的参数。残差网络由依次堆叠的残差块组成。

在[13]中，残留网络由两种类型的块组成。

- *基本*--连续两次3×3的卷积，在卷积之前进行批量归一化和ReLU：`conv3×3-conv3×3` 图1(a)
- *瓶颈*--用一个3×3卷积层包围着降维和扩展的1×1卷积层：`conv1×1-conv3×3-conv1×1`图1(b)

组名	输出尺寸	块类型 = $B(3, 3)$
conv1	32 × 32	[3×3, 16]
conv2	32×32	$\begin{bmatrix} 3\times3, & 16\times k \\ 3\times3, & 16\times k \end{bmatrix} \times N$
定罪3	16×16	$\begin{bmatrix} 3\times3, & 32\times k \\ 3\times3, & 32\times k \end{bmatrix} \times N$
conv4	8×8	$\begin{bmatrix} 3\times3, & 64\times k \\ 3\times3, & 64\times k \end{bmatrix} \times N$
游泳池	1 × 1	[8 ×8]

表1：
宽残差网络的结构。网络宽度由系数k决定。原始结构[13]相当于k=1。卷积组显示在括号中，其中N是组中的块数，由第一层进行的下采样
在conv3和conv4组。最后的分类层被省略了，以便清除。在所示的特定例子中，网络使用$B(3, 3)$类型的ResNet块。

与原始架构[11]相比，在[13]中，残差块中的批量归一化、激活和卷积的顺序从conv-BN-ReLU改为BN-ReLU-conv。由于后者被证明能更快地训练并取得更好的结果，我们不考虑原始版本。此外，所谓的 "瓶颈"块最初被用来使块的计算成本降低，以增加层的数量。由于我们想研究拓宽的效果，而 "瓶颈 "是用来使网络变薄的，所以我们也不考虑它，而是关注 "基本"的剩余结构。

基本上有三种简单的方法来提高剩余块的表现力。

- 以增加每块的卷积层
- 通过增加更多的特征平面来扩大卷积层的范围
- 增加卷积层中的过滤器大小

由于在包括[26, 31]在内的一些工作中，小型过滤器被证明是非常有效的，我们不考虑使用大于3×3的过滤器。
让我们也引入两个因素，深化系数l和拓宽系数k，其中l是一个块中的卷积数，k乘以是卷积层中的特征数，因此基线 "基本"块对应于l=2，k=1。图1(a)和1(c)分别显示了 "基本 "和 "基本范围 "块的示意图。

我们的残差网络的一般结构如表1所示：它包括一个初始卷积层conv1，然后是3组（每组大小为N）残差块conv2、conv3和conv4，然后是平均池和最终分类层。在我们所有的实验中，conv1的大小是固定的，而引入的拓宽因子k则是对三组conv2-4中的残余块的宽度进行缩放（*例如*
原来的 "基本 "结构相当于k=1）。我们想研究的是代表
为此，我们进行了一些修改，并测试了残余区块的功能。
对 "基本 "架构的影响，将在以下几个小节中详细介绍。

2.1 残余块中的卷积类型

让$B(M)$表示残余块结构，其中M是一个带有块中卷积层内核大小的列表。例如，$B(3, 1)$表示具有3×3和1×1卷积层的剩余块（我们总是假设空间核为方形）。请注意，由于我们不考虑前面解释的 "瓶颈 "块，特征平面的数量总是
在整个区块中保持相同。我们想回答这样一个问题："基本
"残差结构的3×3卷积层中的每一层有多重要，它们是否可以
用计算成本较低的1×1层或甚至1×1层的组合来代替。
和3×3卷积层，*例如*B（1，3）或B（1，3）。这可以增加或减少区块的表示能力。因此，我们对以下组合进行了实验
(注意最后一个组合，即$B(3, 1, 1)$与有效的网中网相似。
[20]架构）。

1. $B(3, 3)$ - 原始 "基本 "区块
2. $B(3, 1, 3)$ - 有一个额外的1×1层
3. $B(1, 3, 1)$ - 所有卷的维度相同，"拉直 "瓶颈
4. $B(1, 3)$ - 该网络到处都有1×1-3×3的交替卷积。
5. $B(3, 1)$ - 与上一个区块的想法类似
6. $B(3, 1, 1)$ - 网中网风格区块

2.2 每个剩余块的卷积层数量

我们还对区块深化因子l*进行了*实验，看看它对性能有什么影响。比较必须在具有相同数量参数的网络之间进行，因此在这种情况下，我们需要建立具有不同l和d（其中d表示块的总数）的网络，同时确保网络复杂性保持大致不变。这意味着，例如，当l增加时，d应该减少。

2.3 残留块的宽度

除了上述修改外，我们还对块的拓宽系数k*进行了*实验。虽然参数的数量随着l（加深系数）和d（ResNet块的数量）线性增加，但参数的数量和计算复杂度是k的二次方。然而，拓宽层比有成千上万的小核更有计算效率，因为GPU在大张量的并行计算中效率更高，所以我们对最佳的d和k*的*比例感兴趣。

关于更宽的残差网络的一个论点是，在残差网络之前的几乎所有架构，包括最成功的Inception[30]和VGG[26]，与[13]相比都宽得多。例如，残差网络WRN-22-8和WRN-16-
10（关于这个符号的解释见下一段）在宽度、深度和参数数量上与VGG架构非常相似。

我们进一步将$k=1$的原始残差网络称为 "薄"，将$k=1$的网络称为$k>1$为 "宽"。在本文的其余部分，我们使用以下符号。WRN-n-k表示残差网络，其卷积层总数为n，拓宽系数为k（例如，具有40层且$k=2$倍于原始网络的网络将被表示为WRN-40-
2）。另外，在适用的情况下，我们会附加区块类型，*例如*WRN-40-2-B（3，3）。

块型	深度	# 参数	时间,s	CIFAR-10
$B(1, 3, 1)$	40	1.4M	85.8	6.06
$B(3, 1)$	40	1.2M	67.5	5.78
$B(1, 3)$	40	1.3M	72.2	6.42
$B(3, 1, 1)$	40	1.3M	82.2	5.86
$B(3, 3)$	28	1.5M	67.5	5.73
$B(3, 1, 3)$	22	1.1M	59.9	5.78

表2：测试误差（%，5次运行的中位数），在CIFAR-10上的残差网络，$k = 2$和不同的块类型。时间栏衡量一个训练纪元。

l	CIFAR-10
1	6.69
2	5.43
3	5.65
4	5.93

表3：对CIFAR-的测试误差（%，5次运行的平均值）。WRN-40-2的10个 (2.2M) 与各种L。

2.4 剩余区块的辍学情况

由于拓宽增加了参数的数量，我们想研究正则化的方法。残差网络已经有了批量规范化，提供了一个规范化的效果，但是它需要大量的数据增强，我们想避免这种情况，而且这并不总是可能的。如图1(d)所示，我们在每个残差块之间和ReLU之后添加一个dropout层，以扰乱下一个残差块的批量归一化，并防止其过拟合。在非常深的残差网络中，这应该有助于处理递减的特征重用问题，强制在不同的残差块中学习。

3 实验结果

在实验中我们选择了著名的CIFAR-10、CIFAR-100、SVHN和ImageNet图像分类数据集。CIFAR-10和CIFAR-100数据集[17]由32×32的彩色图像组成，分别来自10个和100个类别，分为50,000张训练图像和10,000张测试图像。对于数据我们进行水平翻转，并从图像中随机裁剪，每边填充4个像素，用原始图像的反射来填补缺失的像素。我们没有像[9]中提出的那样使用大量的数据增强。SVHN是谷歌街景门牌号图像的数据集，包含大约600,000个数字图像，来自一个明显困难的现实世界问题。在SVHN的实验中，我们没有做任何图像预处理，只是将图像除以255，以提供[0,1]范围内的图像作为输入。除了ImageNet之外，我们所有的实验都是基于[13]的架构，带有预激活的剩余块，我们把它作为基线。对于ImageNet，我们发现在少于100层的网络中使用预激活并没有任何明显的区别，因此我们决定在这种情况下使用原始的ResNet架构。除非另有提及，对于CIFAR，我们遵循[8]的图像预处理，并进行ZCA美白。然而，在一些CIFAR实验中，我们使用了简单的均值/std归一化，这样我们可以直接与[13]和其他使用这种预处理的ResNet相关工作进行比较。

在下文中，我们描述了我们对不同ResNet块结构的研究结果，同时也分析了我们提出的宽残差网络的性能。我们注意到，在所有与 "块中的卷积类型 "和 "每块的卷积数量

"相关的实验中，我们使用了k=2，并且与[13]相比，降低了深度，以加快训练速度。

一个区块中的旋转类型

我们首先报告了使用不同块类型B的训练网络的结果（报告的结果是在CIFAR-10上）。我们对B（1，3，1）、B（3，1）、B（1，3）和B（3，1，1）区块使用WRN-40-2，因为这些区块只有一个3×3卷积。为了保持参数的数量可以比较的是，我们训练了其他层数较少的网络。WRN-28-B(3，3)和WRN-22-2 B(3，1，3).我们在表2中提供了包括5次运行的测试准确率中位数和每个训练纪元的时间在内的结果。B(3，3)块以很小的幅度成为最好的，B(3，1)和B(3，1，3)在准确率上非常接近B(3，3)，但参数和层数都较少。B(3，1，3)以很小的幅度比其他的快。

基于上述情况，具有相当数量参数的区块被证明给出了或多或少的相同结果。由于这一事实，我们在下文中只关注具有3×3卷积的WRN，以便与其他方法保持一致。

每个区块的卷积数

我们接下来进行与改变深化系数l（代表每块卷积层的数量）有关的实验。我们在表3中显示了指示性的结果，在这种情况下，我们用3×3卷积的WRN-40-2来训练几个网络不同的深化系数$l \in [1，2，3，4]$，相同的参数数（2.2×10^6）和相同数量的卷积层。

可以注意到，B（3，3）结果是最好的，而B（3，3，3）和B（3，3，3，3）的性能最差。我们推测，这可能是由于在最后两个阶段，由于剩余连接数的减少，优化的难度增加了。

案例。此外，B(3)被证明是相当糟糕的。结论是B(3，3)在每块的卷积数方面是最佳的。由于这个原因，在剩下的实验中，我们只考虑具有B(3，3)型块的宽残差网络。

残留块的宽度

当我们试图增加拓宽参数k时，我们必须减少总层数。为了找到一个最佳比例，我们对k从2到12，深度从16到40进行了实验。结果显示在表4中。可以看出，所有40层、22层和16层的网络都能看到当宽度增加1到12倍时，有一致的收益。另一方面，当保持相同的固定加宽系数k=8或k=10，并将深度从16到28变化时，有一个一致的改进，然而当我们进一步将深度增加到40时，准确性就会下降（*例如*，WRN-40-8的准确性会输给WRN-22-8）。

我们在表5中显示了更多的结果，在那里我们比较了薄的和宽的残余网络。可以看出，宽的WRN-40-4与瘦的ResNet-1001相比更有优势，因为它在CIFAR-10和CIFAR-100上都取得了更好的准确性。然而，有趣的是，这些网络的参数数量相当，分别为8.9×10^6和10.2×10^6，这表明在这个水平上，深度与宽度相比没有增加正则化效应。如同我们在基准测试中进一步显示的那样标志着，WRN-40-4的训练速度快了8倍，所以很明显，原始薄残差网络的深度和宽度比例远非最佳。

另外，宽的WRN-28-10在CIFAR-10上比薄的ResNet-1001好0.92%（在训练过程中的小批量大小相同），在CIFAR-100上好3.46%，层数少36倍（见表5）。我们注意到，ResNet-1001的4.64%的结果是在批量大小为64的情况下获得的，而我们在所有的实验中都使用了批量大小为128的结果（即所有其他结果

深度	k	# 参数	CIFAR-10	CIFAR-100
40	1	0.6M	6.85	30.89
40	2	2.2M	5.33	26.04
40	4	8.9M	4.97	22.89
40	8	35.7M	4.66	-
28	10	36.5M	**4.17**	20.50
28	12	52.5M	4.33	**20.43**
22	8	17.2M	4.38	21.22
22	10	26.8M	4.44	20.75
16	8	11.0M	4.81	22.07
16	10	17.1M	4.56	21.59

表4：在CIFAR-10和CIFAR-100（ZCA预处理）上各种广义网络的测试误差（%）。

表5中报告的是批次大小为128的情况）。这些网络的训练曲线见图2。

尽管之前有观点认为深度会产生正则化效应，而宽度会导致网络过度拟合，但我们还是成功地训练出了比ResNet-1001多几倍参数的网络。例如，宽幅WRN-28-10（表5）和宽幅WRN-40-10（表9）的参数分别是ResNet-1001的3.6倍和5倍，并且都以很大的优势超过了它。

	深度-k	# 参数	CIFAR-10	CIFAR-100
NIN[20]			8.81	35.67
DSN [19]			8.22	34.57
FitNet[24]			8.39	35.04
公路[28]			7.72	32.39
ELU [5]			6.55	24.28
原始ResNet[11]	110	1.7M	6.43	25.16
	1202	10.2M	7.93	27.82
深度[14]	110	1.7M	5.23	24.58
	1202	10.2M	4.91	-
pre-act-ResNet[13]	110	1.7M	6.37	-
	164	1.7M	5.46	24.33
	1001	10.2M	4.92(4.64)	22.71
WRN (我们的)	40-4	8.9M	4.53	21.18
	16-8	11.0M	4.27	20.43
	28-10	36.5M	**4.00**	**19.25**

表5：不同方法在CIFAR-10和CIFAR-100上的测试误差，有适度的数据增强（翻转/翻译）和平均/STD归一化。对于这些结果，我们不使用辍学。在第二列中，k是一个拓宽因子。[13]的结果显示，小批处理量为128（和我们的一样），括号内为64。我们的结果是通过计算5次运行的中位数得到的。

总的来说，我们观察到CIFAR均值/STD预处理允许训练更广泛和更深入的网络，具有更好的准确性，在CIFAR-100上使用WRN-40-10与56×10^6参数达到18.3%（表9），比ResNet-1001总共提高4.4%。

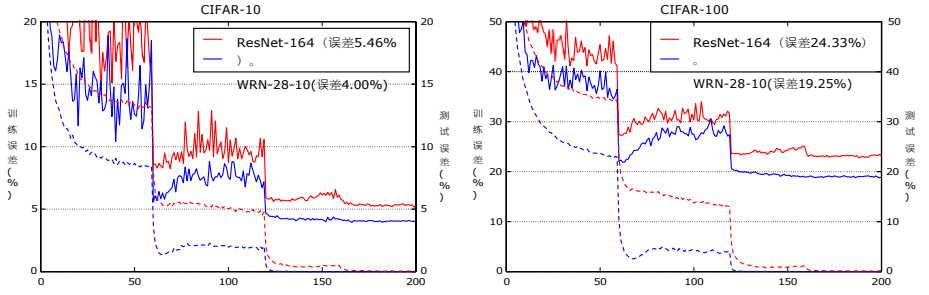


图2： CIFAR-10和CIFAR-100上的薄型和宽型残差网络的训练曲线。实线表示测试误差（y轴在右边），虚线表示训练损失（y轴在左边）。

并在这个数据集上建立了一个新的最先进的结果。

总结一下。

- 在不同深度的残余网络中，拓宽持续改善性能。
- 增加深度和宽度都有帮助，直到参数的数量变得太高，需要更强的正则化。
- 残差网的深度很高，似乎没有正则化效应，因为参数数量与薄网络相同的宽网络可以学习。
  相同或更好的表征。此外，宽网络可以成功地学习比瘦网络多2倍或更多的参数，这就需要将瘦网络的深度增加一倍，使它们的训练成本高得难以想象。

剩余区块的辍学情况

我们在所有的数据集上训练网络，并在卷积之间的残余块中插入了辍学。我们使用交叉验证法来确定放弃的概率值，在CIFAR上为0.3，在SVHN上为0.4。此外，与没有滤波的基线网络相比，我们不必增加训练历时的数量。

辍学使WRN-28-10在CIFAR-10和CIFAR-100上的测试误差分别减少了0.11%和0.4%（超过5次运行的中位数和平均/std预处理），并且对其他ResNets也有改进（表6）。据我们所知，这是第一个在CIFAR-100上接近20%误差的结果，甚至超过了有大量数据增强的方法。在CIFAR-10上，WRN-16-4的准确性只有轻微的下降，我们推测这是由于参数的数量相对较少。

我们注意到，在第一次学习率下降后的残差网络训练中，损失和验证误差突然开始上升，并在高值上震荡，直到下一次学习率下降。我们发现这是由权重衰减引起的，然而使其降低会导致准确率的大幅下降。有趣的是，在大多数情况下，dropout部分地消除了这种影响，见图2，3。

辍学的影响在SVHN上变得更加明显。这可能是由于我们没有做任何数据增量和批量归一化的过度，所以辍学增加了

深度	*k*	辍学者	CIFAR-10	CIFAR-100	SVHN
16	4		5.02	24.03	1.85
16	4		5.24	23.91	1.64
28	10		4.00	19.25	-
28	10		**3.89**	**18.85**	-
52	1		6.43	29.89	2.08
52	1		6.28	29.78	1.70

表6： 剩余区块中的辍学效应。(平均值/std预处理，CIFAR数字基于5次运行的中位数)

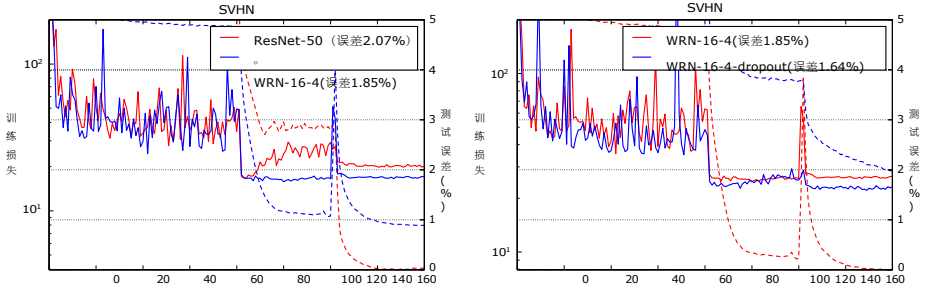


图3：
SVHN的训练曲线。左边是薄和宽的网络，右边是辍学的影响。实线表示测试误差（y轴在右边），虚线表示训练损失（y轴在左边）。

正则化效应。这方面的证据可以在图3的训练曲线上找到，在图3中，没有辍学的损失下降到非常低的数值。结果显示在表6中。我们观察到在薄型和宽型网络上使用dropout的明显改进。薄的50层深度网络甚至超过了具有随机深度的薄的152层深度网络[14]。我们还在SVHN上用dropout训练了WRN-16-8（表9），它在SVHN上取得了1.54%的成绩--这是我们所知的最好的公开结果。如果没有放弃，它的成绩为1.81%。

总的来说，尽管有与批量规范化相结合的争论，Dropout显示出它是稀疏和宽广网络规范化的一种有效技术。它可以用来进一步改善拓宽的结果，同时也是对它的补充。

图像网和COCO实验

对于ImageNet，我们首先用非瓶颈的ResNet-18和ResNet-34进行实验，尝试将它们的宽度从1.0逐渐增加到3.0。结果显示在表7中。增加宽度可以逐渐提高这两个网络的准确性，具有相当数量参数的网络可以取得类似的结果，尽管它们的深度不同。尽管这些网络有大量的参数，但它们比瓶颈网络更胜一筹，这可能是由于瓶颈结构更适合于Ima-Rock。

在这个过程中，我们发现了一些问题，比如说："为什么我们不把ResNet分类的任务交给一个更复杂的网络呢？为了测试这一点，我们采用了ResNet-50，并试图通过增加内部3×3层的宽度使其变宽。在拓宽系数为2.0的情况下，所得到的WRN-50-2-瓶颈超过了ResNet-50。

152的层数减少3倍，而且速度明显加快。WRN-50-2-瓶颈仅是

比起表现最好的预激活ResNet-200，虽然参数略多，但速度几乎是2倍（表8）。总的来说，我们发现，与CIFAR不同。

ImageNet网络在相同的深度下需要更多的宽度来达到相同的精度。然而，很明显，由于计算上的原因，没有必要拥有超过50层的残留网络。

我们没有尝试训练更大的瓶颈网络，因为这需要8个GPU的机器。

宽度		1.0	1.5	2.0	3.0
WRN-18	top1,top5 #参数	30.4, 10.93 11.7M	27.06, 9.0 25.9M	25.58, 8.06 45.6M	24.06, 7.33 101.8M
WRN-34	TOP1,TOP5 #参数	26.77, 8.67 21.8M	24.5, 7.58 48.6M	23.39, 7.00 86.0M	

表7：ILSVRC-2012对各种拓宽因素的非瓶颈ResNets的验证误差（单一作物）。尽管网络的层数少了2倍，但具有相当数量的参数的网络达到了类似的精度。

模型	top-1 err, %	前5名错误，%。	#params	时间/批次 16
共和国网-50	24.01	7.02	25.6M	49
里斯网-101	22.44	6.21	44.5M	82
ResNet-152	22.16	6.16	60.2M	115
WRN-50-2-瓶颈	21.9	6.03	68.9M	93
前ResNet-200	21.66	5.79	64.7M	154

表8：ILSVRC-2012的瓶颈ResNets的验证误差（单一作物）。更快的WRN-50-2瓶颈超过了层数少3倍的ResNet-152，并接近于ResNet-200之前。

我们还用WRN-34-2参加了COCO 2016物体检测挑战赛，使用了MultiPathNet[32]和LocNet[7]的组合。尽管只有34层，这个模型达到了最先进的单模型性能，甚至超过了基于ResNet-152和Inception-v4的模型。

最后，我们在表9中总结了我们在各种常用数据集上的最佳WRN结果。

数据集	模型	辍学者	试验性香水。
CIFAR-10	WRN-40-10		3.8%
CIFAR-100	WRN-40-10		18.3%
SVHN	WRN-16-8		1.54%
图像网（单幅作物）	WRN-50-2-瓶颈		21.9%排名第一，5.79%排名第五
COCO测试-std	WRN-34-2		35.2 mAP

表9：各种数据集的最佳WRN性能，单次运行结果。COCO模型基于WRN-34-2（更宽的基本块），使用基于VGG-16的AttractioNet建议，并有一个LocNet风格的定位部分。据我们所知，这些是CIFAR-10、CIFAR-100、SVHN和COCO（使用非集合模型）的最佳公布结果。

计算效率

具有小内核的薄而深的残差网络由于其顺序结构而违背了GPU组合的本质。增加宽

度有助于有效平衡

正如我们的基准所显示的那样，宽幅网络比细幅网络的效率高很多倍。我们使用cudnn v5和Titan X来测量几个网络的前向+后向更新时间，最小批量为32，结果见图4。我们表明，我们最好的CIFAR宽WRN-28-10比薄ResNet-1001快1.6倍。此外，宽WRN-40-4，具有与ResNet-1001大致相同的精度，但速度是8倍。

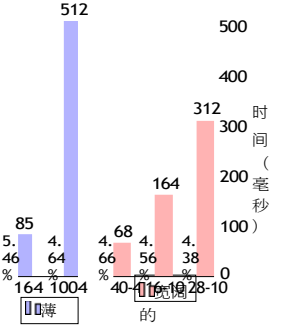


图4：
广义和狭义网络中每批32人的前向+后向更新时间（X轴表示网络深度和拓宽系数）。条形图旁边的数字表示CIFAR-10的测试误差，上面是时间（ms）。测试时间是这些基准的一个比例的分数。例如，请注意，宽的WRN-40-4比瘦的ResNet-1001快8倍，同时具有大约相同的精度。

实施细节

在我们所有的实验中，我们使用SGD与Nesterov动量和交叉熵损失。初始学习率被设置为0.1，权重衰减为0.0005，阻尼为0，动量为0.9，最小批次大小为128。在CIFAR上，学习率在60、120和160 epochs时下降了0.2，我们总共训练了200个epochs。在SVHN上，初始学习率被设置为0.01，我们在80和120个历时中降低0.1，总共训练了160个历时。我们的实现是基于Torch[6]的。我们使用[21]来减少我们所有网络的内存足迹。在ImageNet实验中，我们使用了`fb.resnet.torch`实现[10]。我们的代码和模型可在https://github.com/szagoruyko/wide-residual-networks。

4 结论

我们提出了一项关于残差网络的宽度以及残差结构的使用的研究。基于这项研究，我们提出了一个宽残差网络架构，在几个常用的基准数据集（包括CIFAR-10、CIFAR-100、SVHN和COCO）上提供了最先进的结果，并在ImageNet上有了显著的改进。我们证明了只有16层的宽幅网络在CIFAR上的表现明显优于1000层的深度网络，以及50层的网络在ImageNet上的表现优于152层的网络，从而证明了残差网络的主要力量在于残差块，而不是像之前所说的极端深度。另外，宽大的残差网络的训练速度

要快几倍。我们认为，这些耐人寻味的发现将有助于深度神经网络研究的进一步进展。

5 鸣谢

我们感谢初创公司VisionLabs和Eugenio Culurciello让我们使用他们的集群，没有他们，ImageNet实验就不可能进行。我们还感谢Adam Lerer和Sam Gross的有益讨论。工作由欧盟项目FP7-ICT- 611145 ROBOSPECT支持。

参考文献

[1] Yoshua Bengio和Xavier Glorot.了解训练深度前馈神经网络的难度。在*AISTATS 2010会议上*，第9卷，第249-256页，2010年5月。

[2] Yoshua Bengio和Yann LeCun.向人工智能扩展学习算法。In Léon Bottou, Olivier Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Ma- chines*.MIT Press, 2007.

[3] Monica Bianchini和Franco Scarselli.论浅层和深层神经网络分类器的复杂性。In *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*, 2014.

[4] T.Chen, I. Goodfellow, and J. Shlens.Net2net:通过知识转移加速学习。在*国际学习代表会议上*，2016年。

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter.通过指数线性单元（elus）进行快速准确的深度网络学习。*CoRR*, abs/1511.07289, 2015.

[6] R.Collobert, K. Kavukcuoglu, and C. Farabet.Torch7：一个类似matlab的机器学习环境。在*BigLearn，NIPS研讨会*，2011。

[7] Spyros Gidaris 和 Nikos Komodakis.Locnet:提高物体检测的定位精度。In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[8] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio.Maxout网络。In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning（ICML'13）*, pages 1319-1327, 2013.

[9] Benjamin Graham.分数最大集合。*arXiv:1412.6071*, 2014。

[10] 山姆-格罗斯和迈克尔-威尔伯。训练和调查残差网，2016年。URL https://github.com/facebook/fb.resnet.torch。

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.用于图像识别的深度残差学习.*CoRR*, abs/1512.03385, 2015.

[12] 何开明，张翔宇，任少卿，和孙健。深入研究整流器。超越人类水平的图像网分类性能。*CoRR*, abs/1502.01852, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.深度残差网络中的身份映射。*CoRR*, abs/1603.05027, 2016.

[14] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger.具有随机深度的深度网络。*CoRR*, abs/1603.09382, 2016.

[15] Sergey Ioffe and Christian Szegedy.批量归一化。通过减少内部协变量偏移来加速深度网络训练。在David Blei和Francis Bach编辑的《*第32届国际机器学习会议（ICML-15）论文集*》中，第448-456页。JMLR研讨会和会议记录，2015。

[16] A.Krizhevsky, I. Sutskever, and G. Hinton.用深度对话神经网络进行图像网分类。In *NIPS*, 2012.

[17] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.Cifar-10（加拿大高级研究学院）。2012.URL http://www.cs.toronto.edu/~kriz/ cifar.html。

[18] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio.对具有许多变化因素的问题的深度架构的经验评估。In Zoubin Ghahramani, editor, *Proceedings of the 24th International Con- ference on Machine Learning（ICML'07）*, pages 473-480.ACM, 2007.

[19] C.-Y.Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu.Deeply-Supervised Nets.2014.

[20] Min Lin, Qiang Chen, and Shuicheng Yan.网络中的网络.*CoRR*, abs/1312.4400, 2013.

[21] Francisco Massa.Optnet - 减少火炬神经网络的内存使用，2016年。URL https://github.com/fmassa/optimize-net。

[22] Guido F. Montúfar, Razvan Pascanu, KyungHyun Cho, and Yoshua Bengio.On the number of linear regions of deep neural networks.In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada,* pages 2924-2932, 2014.

[23] Tapani Raiko, Harri Valpola, and Yann Lecun.深度学习通过感知器的线性转换变得更容易。In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 924-932, 2012.

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio.FitNets：薄型深度网的提示。技术报告Arxiv报告1412.6550，arXiv，2014。

[25] J.Schmidhuber.使用历史压缩原则学习复杂的、扩展的序列.*Neural Computation*, 4(2):234-242, 1992.

[26] K.Simonyan and A. Zisserman.用于大规模图像识别的极深卷积网络。In *ICLR*,

2015.

[27] N.Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov.辍学。防止神经网络过拟合的一个简单方法。*JMLR*, 2014.

[28] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber.公路网络。*CoRR*, abs/1505.00387, 2015.

[29] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton.论深度学习中初始化和动力的重要性。在Sanjoy Dasgupta和David Mcallester编辑的*第30届国际机器学习会议（ICML-13）论文集中*，第28卷，第1139-1147页。JMLR研讨会和会议记录，2013年5月。

[30] C.Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich.用卷积法做得更深入。In *CVPR*, 2015.

[31] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke.Inception-v4，inception-resnet和剩余连接对学习的影响。abs/1602.07261，2016。

[32] S.Zagoruyko, A. Lerer, T. -Y.Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár.一个用于物体检测的多路径网络。In *BMVC*, 2016.