图像识别的深度残差学习

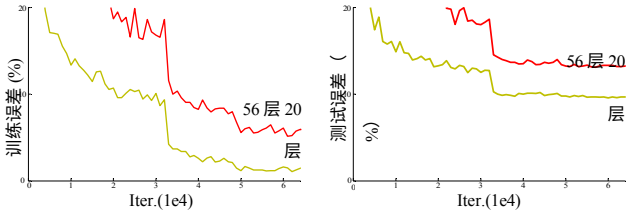
Kaiming He Xiangyu Zhang Shaoqing Ren JianSun
微软研究院
{kahe、v-xiangz、v-shren、jiansun}@microsoft.com

摘要

更深的神经网络更难训练。我们提出了一个残差学习框架，以简化比以前所用网络更深的网络的训练。我们明确地将各层重新表述为参照各层输入的学习残差函数，并将各层的输入和输出重新表述为残差函数。

而不是学习无参函数。我们提供了全面的经验证据，表明这些残差网络更容易优化，并能通过大幅增加深度获得准确性。在 ImageNet 数据集上，我们评估了深度高达 152 层的残差网络--比 VGG 网络[41]更深 8 倍，但复杂度仍然更低。这些残差网络的集合在 ImageNet 测试集上实现了 3.57% 的误差。这一结果赢得了 ILSVRC 2015 分类任务的第一名。我们还对 100 层和 1000 层的 CIFAR-10 进行了分析。

表征深度对于许多视觉识别任务来说都至关重要。正是由于我们的深度表征，我们在 COCO 物体检测数据集上获得了 28% 的相对提升。深度残差网络是我们在 ILSVRC 和 COCO 2015 竞赛中提交的成果的基础。[1] *我们还在 ImageNet 检测、ImageNet 本地化、COCO 检测和 COCO 分割任务中获得了第一名。*



1. 导言

深度卷积神经网络[22, 21]为图像分类带来了一系列突破[21, 50, 40]。深度网络以端到端的多层方式自然地集成了低/中/高层特征[50]和分类器，特征的 "层次 "可以通过堆叠层的数量（深度）来丰富。最近的证据[41, 44] 表明，网络深度至关重要，在具有挑战性的 ImageNet 数据集[36]上取得的领先成果[41, 44, 13, 16]都利用了 "非常深"[41]的模型，深度从 16 [41] 到 30 [16] 不等。许多其他非琐碎的视觉识别任务 [8, 12, 7, 32, 27] 也采用了 "深度 "模型。

[1] http://image-net.org/challenges/LSVRC/2015/ 和 http://mscoco.org/dataset/#detections-challenge2015。

图 1.20 层和 56 层 "普通 "网络在 CIFAR-10 上的训练误差（左）和测试误差（右）。较深的网络具有较高的训练误差，因此测试误差也较高。图 4 显示了 ImageNet 上的类似现象。

非常深奥的模型使我们受益匪浅。

在深度意义的驱动下，一个问题出现了：*学习更好的网络就像堆叠更多的层一样简单吗？* 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，归一化初始化[23, 9, 37, 13]和中间归一化层[16]在很大程度上解决了这一问题，使数十层的网络能够开始收敛，从而实现随机梯度下降（SGD）和反向传播[22]。

当更深的网络能够开始收敛时，一个*退化*问题就暴露出来了：随着网络深度的增加，准确度会达到饱和（这可能并不令人惊讶），然后迅速退化。令人意想不到的是，这种退化*并不是由过度拟合引起的，在*适当的深度模型中增加更多层会导致*更高的训练误差，*这一点在文献[11, 42]中已有报道，并在我们的实验中得到了充分验证。图 1 显示了一个典型的例子。

训练精度的降低表明，并非所有系统都同样容易优化。让我们考虑一个较浅的架构和一个增加了更多层次的较深架构。深层模型存在一个*构造*解：添加的层是*身份*映射，其他层是从学习到的浅层模型复制过来的。这种构造解的存在表明，深层模型产生的训练误差不应高于浅层模型。但实验表明，我们现有的求解器无法找到符合以下条件的解决方案

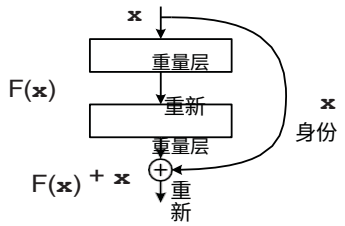


图 2.残余学习：构件

或比构建的解决方案更好（或无法在可行的时间内做到）。

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。我们不希望每个堆叠层直接拟合一个所需的底层映射，而是明确地让这些层 拟合一个残差映射。形式上，将所需的底层映射表示为 H(**x**)，我们让堆叠的非线性层拟合另一个映射 F(**x**) := H(**x**) -**x**。我们假设，优化残差映射比优化原始的、无参照的映射更容易。更极端的是，如果同一映射是最优的，那么将残差推至零将比通过堆叠非线性层来拟合同一映射更容易。

F(**x**) + **x** 的公式可以通过带有 "捷径连接 "的馈电神经网络来实现（图 2）。捷径连接 [2, 34, 49] 是指跳过一层或多层的连接。在我们的例子中，捷径连接只是执行*身份*映射，其输出被添加到堆叠层的输出中（图 2）。身份短切连接既不会增加额外的参数，也不会增加计算复杂度。整个网络仍然可以通过 SGD 和反向传播进行端到端训练，并且可以在不修改求解器的情况下使用普通库（*如*Caffe [19]）轻松实现。

我们在 ImageNet
[36] 来说明退化问题并评估我们的方法。我们的研究表明1) 我们的极深残差网络很容易优化，但当深度增加时，对应的 "普通 "网络（简单地堆叠层）会表现出更高的训练误差；2) 我们的深度残差网络很容易从深度的大幅增加中获得准确性的提升，其结果大大优于之前的网络。

类似的现象也出现在 CIFAR-10 数据集上[20]，这表明我们方法的优化难度和效果并不仅仅局限于特定的数据集。我们介绍了在该数据集上成功训练出的超过 100 层的模型，并探索了超过 1000 层的模型。

在 ImageNet 分类数据集 [36] 上，我们通过极深的残差网络获得了出色的结果。我们的 152 层残差网络是 ImageNet 上有史以来最深的网络，但其复杂度仍低于 VGG 网络[41]。 我们的网络组合在

并在 *ILSVRC 2015 分类竞赛中获得第一名*。这些极深的识别结果在其他识别任务中也具有出色的泛化性能，使我们进一步*赢得了 ILSVRC 和 COCO 2015 比赛的第一名：*在 ILSVRC 和 COCO 2015 比赛中，我们进一步赢得了 *ImageNet 检测*、*ImageNet 定位*、*COCO 检测和 COCO 分割的第一名*。这些强有力的证据表明，残差学习原理是通用的，我们期待它能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 [费雪向量[30]可以看作是 VLAD 的概率版本[18]。它们都是用于图像重估和分类的强大浅层表示法[4, 48]。对于矢量量化，对残差矢量进行编码 [17] 比对原始矢量进行编码更有效。

在低级视觉和计算机图形学中，为了求 解 偏微分方程（PDEs），广泛使用的多网格法[3]将系统重新表述为多个尺度的子问题，其中每个子问题负责较粗和较细尺度之间的残差解。Multigrid 的另一个替代方法是分层基础预处理 [45, 46]，它依赖于代表两个尺度之间残差向量的变量。研究表明 [3, 45, 46]，这些求解器的收敛速度比标准求解器快很多，因为标准求解器不知道解的残差性质。这些方法表明，良好的重新计算或预处理可以简化优化过程。

捷径连接。导致捷径连接的实践和理论[2, 34, 49]已经研究了很长时间。训练多层感知器（MLP）的早期做法是增加一个从网络输入连接到输出的线性层 [34, 49]。在文献[44, 24]中，一些中间层直接连接到辅助分类器，以解决梯度消失/爆炸的问题。文献[39, 38, 31, 47]提出了通过快捷连接实现层重定向、梯度和传播误差的方法。在文献[44]中，"起始 "层由一个捷径分支和几个更深的分支组成。

与我们的工作同时，"高速公路网络"[42, 43] 提出了具有门控功能的捷径连接[15]。这些门是与数据相关的，并且有参数，而我们的识别捷径则没有参数。当门控捷径 "关闭"（趋近于零）时，高速公路网络中的层代表*非残留功能*。相反，我们的公式总是能学习残差函数；我们的身份捷径永远不会关闭，所有信息都会通过，并学习额外的残差函数。此外，高

在深度极度增加的情况下（*如*超过 100 层），路由网络的准确性并没有提高。

---

3. 深度残差学习

3.1. 残余学习

让我们把 H(**x**) 看作由几个堆叠层（不一定是整个网络）拟合的底层映射，**x** 表示这些层中第一个层的输入。如果假设多个非线性层可以渐近地逼近复杂的函数[2]则等同于假设它们可以渐近地逼近残差函数，*即* H(**x**) - **x**（假设输入和输出的维数相同）。因此，我们并不期望堆叠层逼近 H(x)，而是明确地让这些层逼近残差函数 F(**x**) := H(**x**) - **x**。虽然两种形式都能渐近地逼近所需的函数（正如假设的那样），但学习的难易程度可能有所不同。

这种重新表述的动机是退化问题的反直觉现象（图 1 左）。正如我们在导言中所讨论的，如果所添加的层可以作为同一映射来构建，那么较深模型的训练误差应该不会大于其较浅层的对应部分。退化问题表明，求解器在通过多个非线性层逼近同一映射时可能会遇到困难。通过残差学习重构，如果同一映射是最优的，求解器可以简单地将多个非线性层的权重推向零，以接近同一映射。

在实际情况中，同一映射不太可能是最优的，但我们的重新表述可能有助于为问题提供先决条件。如果最优函数更接近同一映射而非零映射，那么求解器参照同一映射找到扰动应该比学习新函数更容易。我们通过实验（图 7）发现，学习到的残差函数一般反应较小，这表明同一映射提供了合理的前提条件。

3.2. 身份映射捷径

我们对每几个堆叠层采用残差学习。图 2 显示了一个构件。在本文中，我们考虑的构件形式定义如下

$$\mathbf{y} = \mathsf{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}。 \quad (1)$$

这里的 **x** 和 **y** 是外行人的输入和输出向量。函数 F(**x**, $\{W_i\}$) 表示要学习的残差映射。对于图 2 中有两层的例子，$\mathsf{F} = W_2\,\sigma(W_1\,x)$ 其中 σ 表示

[2] 不过，这一假设仍是一个悬而未决的问题。见 [28]。

ReLU [29] 和偏置被省略，以简化计算。$\mathcal{F} + \mathbf{x}$ 的运算是通过快捷连接和元素加法进行的。我们采用加法后的第二非线性（*即* $\sigma(\mathbf{y})$，见图 2）。

式（1）中的快捷连接既不引入外差参数，也不引入计算复杂性。这不仅在实践中很有吸引力，而且对我们比较普通网络和残差网络也很重要。我们可以将同时具有相同参数数、深度、宽度和计算成本（可忽略的元素相加除外）的普通/残差网络进行比较。在公式（1）中，$\mathbf{x}$ 和 $\mathcal{F}$ 的维度必须相等。

如果情况并非如此（*例如*，在更改输入/输出时通道），我们可以通过快捷连接进行线性投影 W_s，以匹配尺寸：

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s x。 \quad (2)$$

我们也可以在式(1)中使用方矩阵 W_s。但我们将通过实验证明，身份映射足以解决退化问题，而且经济实惠，因此 W_s 只在匹配维度时使用。

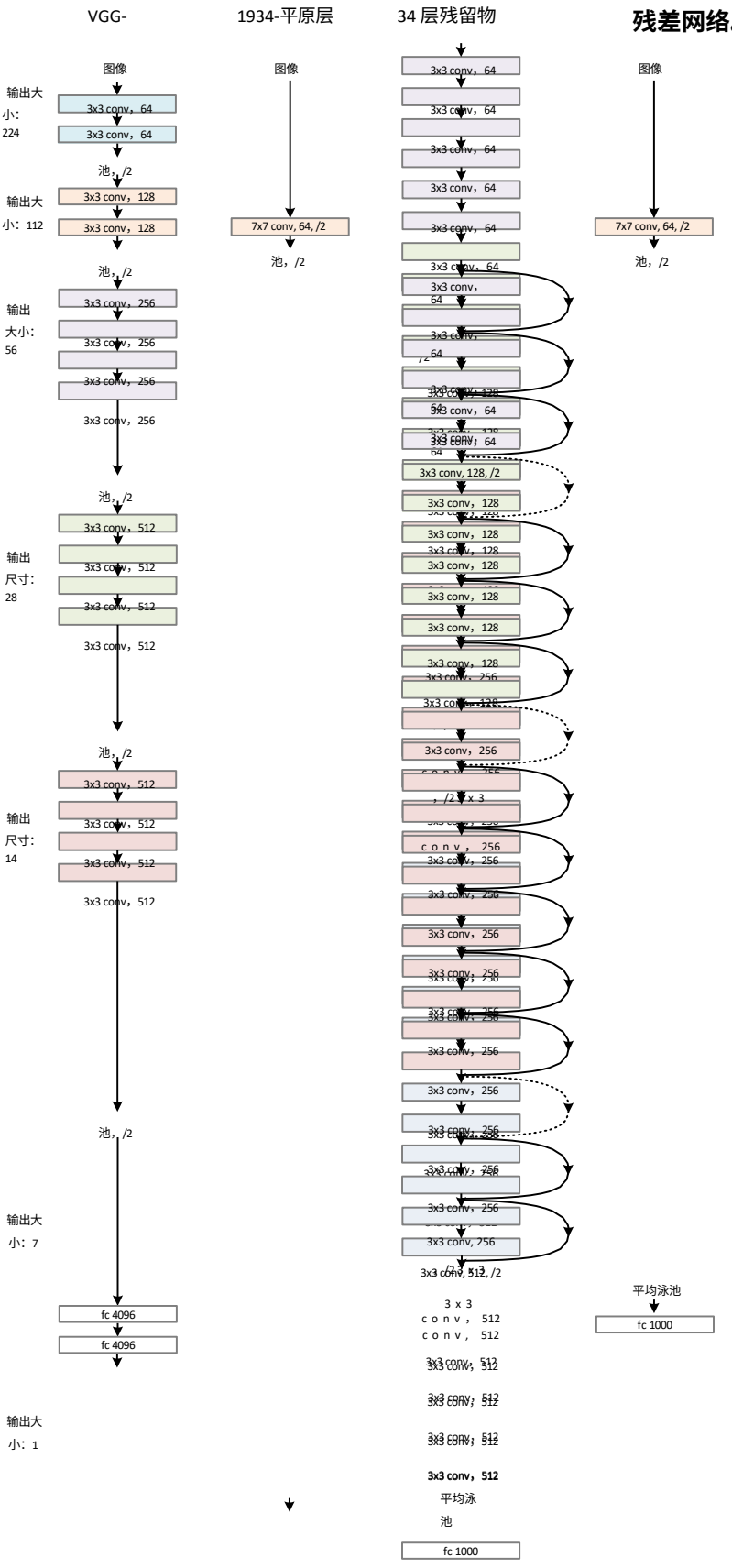
残差函数 $\mathcal{F}$ 的形式是灵活的。本文的实验涉及的函数 $\mathcal{F}$ 有两层或三层（图 5），也可能有更多层。但如果 $\mathcal{F}$ 只有单层，则公式（1）类似于线性层：$\mathbf{y} = W_1 \mathrm{x} + \mathbf{x}$，我们没有观察到其优势。我们还注意到，虽然为了简单起见，上述符号是关于全连接层的，但它们也适用于卷积层。函数 $\mathcal{F}(\mathbf{x}, \{W_i\})$ 可以代表多个卷积层。逐个通道对两个特征图进行元素相加。

3.3. 网络架构

我们对各种普通/残差网络进行了测试，发现了一致的现象。为了提供讨论的实例，我们描述了 ImageNet 的两个模型如下。

普通网络。我们的普通基线（图 3，中）主要受 VGG 网络[41]（图 3，左）的理念启发。卷积层大多采用 3×3 过滤器，并遵循两条简单的设计规则：(i) 在输出特征图大小相同的情况下，各层的过滤器数量相同；(ii) 如果特征图大小减半，则过滤器数量加倍，以保持每层的时间复杂性。我们直接通过步长为 2 的卷积层进行降采样。网络的末端是全局平均池化层和 1000 路带 softmax 的全连接层。图 3（中）中加权层的总数为 34。

值得注意的是，与 VGG 网络 [41] 相比，我们的模型具有*更少的*滤波器和更*低的*复杂度（图 3 左）。我们的 34 层基线具有 36 亿 FLOPs（乘法加法），仅为 VGG-19 的 18%（196 亿 FLOPs）。



残差网络。基于上述平原网络，我们

方案不引入额外参数；（B）使用公式（2）中的投影快捷方式来匹配维数（通过 1×1 卷积完成）。对于这两个选项，当快捷方式穿过两个尺寸的特征图时，它们的步长都是 2。

3.4. 实施情况

我们采用 [21, 41] 中的做法来实现 ImageNet。调整图像大小时，在[256, 480]范围内对图像短边进行采样，以增强比例[41]。从图像或其水平翻转图像中随机取样 224×224 裁剪，并减去每个像素的平均值[21]。使用 [21] 中的标准颜色增强。在每次卷积后和激活前，我们都会采用批量归一化（BN）[16]。我们按照文献[13]中的方法初始化权重，并从头开始训练所有普通/残差网络。我们使用迷你批量大小为 256 的 SGD。学习率从 0.1 开始，当误差趋于平稳时再除以 10，模型训练的迭代次数最多为 60×10^4。我们使用的权重衰减为 0.0001，动量为 0.9。按照文献[16]的做法，我们不使用 dropout [14]。

在测试中，为了进行比较研究，我们采用了标准的 10 裁剪测试 [21]。为了达到最佳效果，我们采用了文献[41, 13]中的全卷积形式，并对多个尺度的得分进行了平均（图像大小经过调整，短边在 {224, 256, 384, 480, 640}范围内）。

4. 实验

4.1. 图像网络分类

我们在 ImageNet 2012 分类数据集 [36] 上对我们的方法进行了评估，该数据集包含 1000 个类别。模型在 128 万张训练图像上进行训练，并在 5 万张验证图像上进行评估。我们还在测试服务器报告的 10 万张测试图像上获得了最终结果。我们同时评估了前 1 名和前 5 名的错误率。

插入捷径连接（图 3 右），将网络转换为对应的残差版本。当输入和输出的维数相同时，可以直接使用标识快捷方式（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），我们考虑两种方案：（A）快捷方式仍然执行同一映射，但在维度增加时填充额外的零条目。这种

普　　网　　。　　们

通　　络　　我　　首

图 3ImageNet 的网络架构示例。**左图**：参考 VGG-19 模型[41]（196 亿 FLOPs）。**左中**：具有 34 个参数层的普通网络（36 亿 FLOPs）。**右图**：34 个参数层的残差网络（36 亿 FLOPs）。虚线快捷方式增加了维度。**表 1** 显示了更多细节和其他变体。

先评估了 18 层和 34 层 普通网络。图 3（中）为 34 层普通网络。图

18 层平网的形式与之类似。去尾结构见表 1。

表 2 中的结果显示，较深的 34 层普通网络的验证误差高于较浅的 18 层普通网络。为了揭示原因，我们在图 4（左）中比较了它们在训练过程中的训练/验证误差。我们观察到了降级问题--即

层名	输出尺寸	18 层	34 层	50 层	101 层	152 层
定罪1	112×112	7×7，64，第 2 步				
conv2_x	56×56	3×3 最大泳池，跨步 2				
		3×3, 64 3×3, 64 ×2	3×3, 64 3×3, 64 ×3	1×1, 64 3×3, 64 1×1, 256 ×3	1×1, 64 3×3, 64 1×1, 256 ×3	1×1, 64 3×3, 64 1×1, 256 ×3
conv3_x	28×28	3×3, 128 3×3, 128 ×2	3×3, 128 3×3, 128 ×4	1×1, 128 3×3, 128 1×1, 512 ×4	1×1, 128 3×3, 128 1×1, 512 ×4	1×1, 128 3×3, 128 1×1, 512 ×8
conv4_x	14×14	3×3, 256 3×3, 256 ×2	3×3, 256 3×3, 256 ×6	1×1, 256 3×3, 256 1×1, 1024 ×6	1×1, 256 3×3, 256 1×1, 1024 ×23	1×1, 256 3×3, 256 1×1, 1024 ×36
conv5_x	7×7	3×3, 512 3×3, 512 ×2	3×3, 512 3×3, 512 ×3	1×1, 512 3×3, 512 1×1, 2048 ×3	1×1, 512 3×3, 512 1×1, 2048 ×3	1×1, 512 3×3, 512 1×1, 2048 ×3
	1×1	平均池、1000 分滤波、软上限				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

表 1.ImageNet 的架构。括号中显示的是构建模块（另见图 5），以及堆叠的模块数量。向下采样由 conv3 1、conv4 1 和 conv5 1 执行，步长为 2。

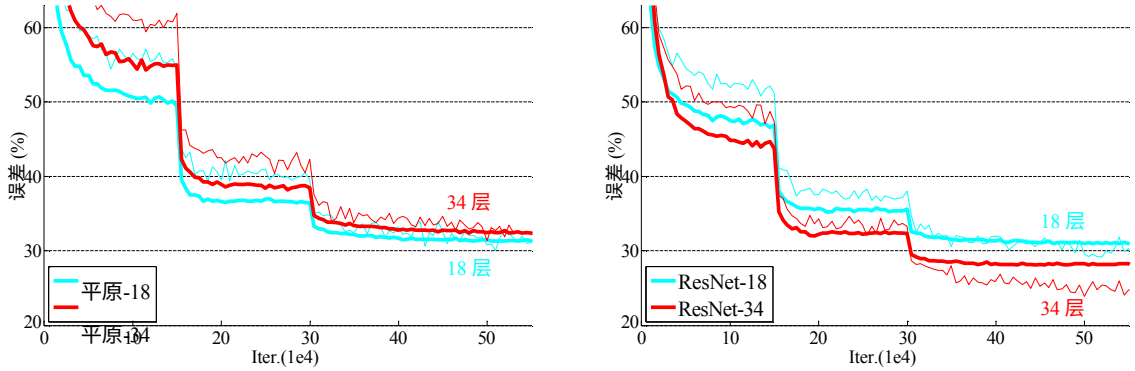


图 4.在 **ImageNet** 上进行的训练。细曲线表示训练误差，粗曲线表示中心作物的验证误差。左图：18 层和 34 层的普通网络。右图18 层和 34 层的残差网络。在这幅图中，与普通网络相比，残差网络没有额外的参数。

	平实	ResNet
18 层	27.94	27.88
34 层	28.54	**25.03**

表 2.ImageNet 验证的前 1 名误差（%，10 次裁剪测试）。与普通网络相比，ResNets 没有额外的参数。图 4 显示了训练过程。

尽管 18 层普通网络的解空间是 34 层网络的子空间，但在整个训练过程中，34 层普通网络的*训练*误差更大。

我们认为，这种优化困难*不太可能*是梯度消失造成的。这些平原网络是用 BN [16] 训练的，它能确保前向传播信号的方差不为零。我们还验证了后向传播梯度在 BN 中表现出健康的规范。因此，前向信号和后向信号都不会消失。事实上，34 层平网仍能达到相当的精度（表 3），这表明求解器在一定程度上是有效的。我们推测，深层平原网的收敛率可能呈指数级降低，从而影响了解算器的精度。

减少训练误差[3].未来将对造成这种优化困难的原因进行研究。

残差网络接下来，我们对 18 层和 34 层残差网络（*ResNets*）进行评估。基线架构与上述普通网络相同，只是在每对 3×3 过滤器中添加了一个捷径连接，如图 3（右）所示。在第一项比较（表 2 和图 4 右）中，我们对所有快捷方式都使用了身份映射，并对递增维度使用了零填充（选项 A）。因此，与普通滤波器相比，它们*没有额外的参数*。

从表 2 和图 4 中，我们可以得出三大结论。首先，残差学习的情况正好相反--34 层 ResNet 优于 18 层 ResNet（2.8%）。更重要的是，34 层 ResNet 的训练误差大大降低，并可推广到验证数据。这表明，在这种情况下，退化问题得到了很好的解决，我们设法通过增加深度来提高准确性。

其次，与普通的同类产品相比，34 层的

[3] 我们尝试了更多的训练迭代次数（3 *倍*），但仍然发现了退化问题，这表明仅仅使用更多的迭代次数是无法解决这一问题的。

模型	TOP-1 Er.	前 5 名错误。
VGG-16 [41]	28.07	9.33
GoogLeNet [44]	-	9.15
PReLU-net [13]	24.27	7.38
平原-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

表 3.ImageNet 验证的错误率（%，**10 个作物**测试）。VGG-16 基于我们的测试。ResNet-50/101/152 属于选项 B，该选项只使用递增维度的投影。

方法	TOP-1 Er.	前 5 名错误。
VGG [41] （ISVRC'14）	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

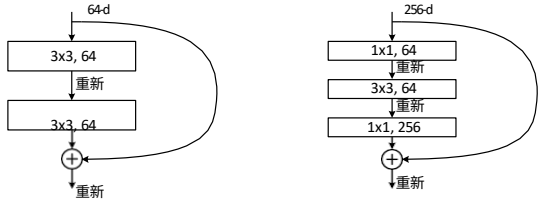
表 4.ImageNet 验证集上**单一模型**结果的错误率（%）（测试集上报告的† 除外）。

方法	前 5 名错误。**测试**
VGG [41] （ISVRC'14）	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**

表 5.**集合的**错误率（%）。前 5 名的误差是在 ImageNet 测试集上的误差，由测试服务器报告。

由于成功减少了训练误差，ResNet 将 top-1 误差降低了 3.5%（表 2）（图 4 右侧 *与*左侧 *对比*）。这一对比验证了残差学习在极深系统中的有效性。

最后，我们还注意到，18 层普通网/残差网的精确



度相当（表 2），但 18 层残差网的收敛速度更快（图 4 右侧与左侧 *对比*）。当网络 "不太深 "时（此处为 18 层），当前的 SGD 求解器仍能为普通网络找到良好的解决方案。在这种情况下，ResNet 通过在早期阶段提供更快的收敛速度来简化优化。

身份 *与*投射捷径。我们已经证明

图 5.ImageNet 的深度残差函数 *F*。左图：图 3 中 ResNet- 的构建模块（在 56×56 个特征图上）。
34.右图：ResNet-50/101/152 的 "瓶颈 "构件。

无参数、身份捷径有助于训练。接下来我们研究投影捷径（公式(2)）。在表 3 中，我们比较了三种方案：(A) 零填充快捷方式用于递增维度，所有快捷方式均不含参数（与表 2 和右图 4 相同）；(B) 投影快捷方式用于递增维度，其他快捷方式均为身份快捷方式；(C) 所有快捷方式均为投影快捷方式。表 3 显示，这三种方案都比普通方案好很多。我们认为这是因为 A 中的零填充维度确实没有剩余学习。C 略好于 B，我们将其归因于许多（13 个）投影捷径引入的额外参数。但是，A/B/C 之间的微小差异表明，投影捷径对解决退化问题并不重要。因此我们在本文其他部分不使用选项 C，以减少内存/时间复杂性和模型大小。身份捷径对于不增加模型的复杂性尤为重要。下面将介绍瓶颈架构。

深度瓶颈架构。接下来，我们将介绍 ImageNet 的深度网络。考虑到我们所能承受的训练时间，我们将构建模块修改为*瓶颈*设计[4].对于每个残差函数 F，我们使用 3 层堆栈，而不是 2 层（图 5）。这三层分别是 1×1、3×3 和 1×1 卷积层，其中 1×1 层负责减少然后增加（恢复）维度，剩下的 3×3 层则是瓶颈层，其输入/输出维度较小。图 5 显示了一个例子，两种设计的时间复杂度相似。

无参数标识捷径对于瓶颈结构尤为重要。如果将图 5（右图）中的身份捷径替换为投影，可以看出时间复杂度和模型大小都增加了一倍，因为捷径与两个高维端相连。因此，身份捷径为瓶颈设计带来了更高效的模型。

50 层 ResNet：我们将

[4] 较深的非瓶颈 ResNets（*如*图 5 左侧）也能通过增加深度获得精度（如 CIFAR-10 所示），但不如瓶颈 ResNets 经济。因此，使用瓶颈设计主要是出于实际考虑。我们还注意到，瓶颈设计也存在普通网的 退化问题。

在 34 层网络中加入这 3 层瓶颈区块，形成 50 层的 ResNet（表 1）。我们使用选项 B 来增加维数。该模型有 38 亿 FLOPs。

101 层和 152 层 ResNet：我们通过使用更多的 3 层块构建了 101 层和 152 层 ResNet（表 1）。值得注意的是，虽然深度显著增加，但 152 层 ResNet（113 亿 FLOPs）的*复杂度仍低于* VGG-16/19 网（153/196 亿 FLOPs）。

50/101/152 层 ResNets 比 34 层 ResNets 的精确度高出很多（表 3 和表 4）。我们没有观察到退化问题，因此，深度的大幅增加带来了显著的精度提升。深度对所有评估指标都有好处（表 3 和表 4）。

与最先进方法的比较。在表 4 中，我们与之前的最佳单一模型结果进行了比较。我们的基线 34 层 ResNet 达到了非常高的准确率。我们的 152 层 ResNet 的单一模型前五名验证误差为 4.49%。这一单一模型结果优于之前所有的集合结果（表 5）。我们将六个不同深度的模型组合成一个集合（提交时只有两个 152 层的模型）。这使得测试集的前 5 名错误率为 **3.57%**（表 5）。*该作品获得了 2015 年 ILSVRC 第一名。*

4.2. CIFAR-10 和分析

我们在 CIFAR-10 数据集[20]上进行了更多研究，该数据集由 10 个类别的 5 万张训练图像和 1 万张测试图像组成。我们介绍了在训练集上训练并在测试集上评估的实验。我们的重点是研究超深度网络的行为，而不是推崇最先进的结果，因此我们有意使用了如下简单的架构。

普通/残差架构如图 3（中/右）所示。网络输入为 32×32 的图像，并减去每像素的平均值。第一层为 3×3 对话。然后，我们在大小分别为{32、16、8}的特征图上堆叠 *6n* 层，层数为 3×3，每种大小的特征图堆叠 2n 层。滤波器的数量分别为{16、32、64}。网络以全局平均池化、10 路全连接层和 softmax 结束。总共有 6n+2 个堆叠加权层。下表概括了该架构：

输出地图大小	32×32	16×16	8×8
# 层	*1+2n*	*2n*	*2n*
# 过滤器	16	32	64

当使用捷径连接时，它们连接到 3×3 层（共 3n 个捷径）。在这个数据集上，我们在所有情况下都使用了身份快捷方式（*即*选项 A）、

方法			误差 (%)
最大输出 [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
	# 层	# 参数	
FitNet [35]	19	2.5M	8.39
公路 [42, 43]	19	2.3M	7.54 (7.72±0.16)
公路 [42, 43]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	**6.43** (6.61±0.16)
ResNet	1202	19.4M	7.93

表 6.**CIFAR-10** 测试集的分类误差。所有方法都使用了数据增强。对于 ResNet-110，我们运行了 5 次，并按照 [43] 中的方法显示了 "最佳值（*平均值*±统计值）"。

因此，我们的残差模型在深度、宽度和参数数量上与普通模型完全相同。

我们使用了 0.0001 的权重衰减和 0.9 的动量，并采用了 [13] 和 BN [16] 中的权重初始化，但没有丢弃。这些模型在两台 GPU 上以 128 个小批量规模进行训练。我们从 0.1 的学习率开始，在 32k 和 48k 次迭代时将学习率除以 10，并在 64k 次迭代时终止训练，这是由 45k/5k train/val 分割决定的。我们采用 [24] 中的简单数据增强方法进行训练：每边填充 4 个像素，从填充图像或其水平翻转图像中随机抽取 32×32 的裁剪。测试时，我们只评估原始 32×32 图像的单一视图。

我们比较了 $n = \{3、5、7、9\}$，从而得出 20、32、44 和 56 层网络。图 6（左）显示了普通网络的行为。深度平原网络会受到深度增加的影响，当深度增加时会表现出更高的训练误差。这种现象与 ImageNet（图 4 左）和 MNIST（见 [42]）上的情况类似，表明这种优化困难是一个基本问题。

图 6（中）显示了 ResNets 的行为。与 ImageNet 案例（图 4 右）类似，我们的 ResNets 也能克服优化困难，并在深度增加时显示出准确率的提高。

我们进一步探讨了 $n = 18$ 的情况，即 110 层的 ResNet。在这种情况下，我们发现 0.1 的初始学习率稍大，无法开始收敛。[5].因此，我们使用 0.01 为训练预热，直到训练误差低于 80%（约 400 次迭代），然后调回 0.1 继续训练。其余的学习计划如前所述。这个 110 层的网络收敛效果很好（图 6，中间）。与其他深层和薄层网络相比，它的参数*更少*。

[5] 初始学习率为 0.1，经过几个历元后开始收敛（*误差小于 90%*），但仍能达到相似的准确率。

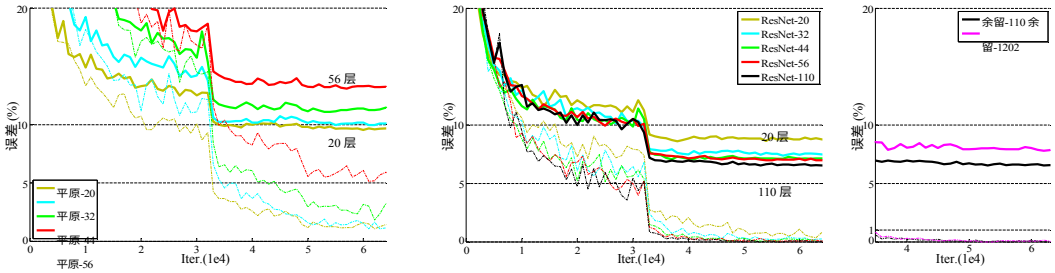


图 6.在 **CIFAR-10** 上进行的训练。虚线表示训练误差，粗线表示测试误差。**左图**：普通网络。plain-110 的误差高于 60%，故未显示。**中间**：ResNets：ResNets.右图具有 110 层和 1202 层的 ResNets。

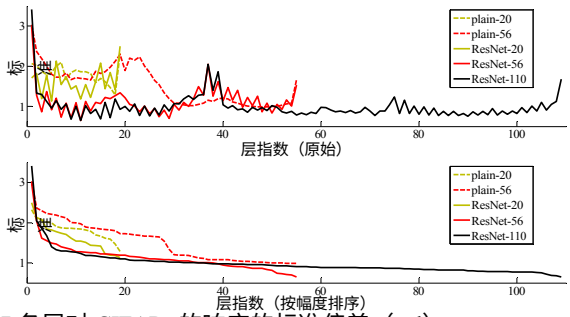


图 7.各层对 CIFAR- 的响应的标准偏差（std）10.响应是每个 3×3 层在 BN 之后和非线性之前的输出。**上图**：各层按原始顺序排列。**下图**：响应按降序排列。

表 6），但也跻身最先进结果之列（6.43%，表 6）。

层响应分析。图 7 显示了各层响应的标准偏差 (std)。这些响应是每个 3×3 层在 BN 之后和其他非线性（ReLU/添加）之前的输出。对于 ResNets，这种分析方法揭示了残差函数的响应强度。图 7 显示，ResNets 的响应强度普遍小于普通网络。这些结果支持了我们的基本动机（第 3.1 节），即残差函数一般可能比非残差函数更接近零。我们还注意到，图 7 中 ResNet-20、56 和 110 的比较结果表明，较深的 ResNet 的响应幅度较小。当层数较多时，单个 ResNet 层对信号的改变往往较小。

探索 1000 多个层。我们将探索一个超过 1000 层的深度模型。我们设置 $n = 200$，从而得到一个 1202 层的网络，训练方法如上所述。我们的方法*没有*显示出*任何优化困难*，这个 10^3 -层网络能够达到*训练误差* <0.1%（图 6 右）。 其测试误差仍然相当不错

训练数据	07+12	07++12
测试数据	挥发性有机化合物 07 测试	挥发性有机化合物 12 测试
VGG-16	73.2	70.4
ResNet-101	**76.4**	**73.8**

(7.93%，表 6）。

但是，这种深度模型还存在一些问题。这个 1202 层网络的测试结果比我们的 110 层网络差，尽管两者都

表 7.使用**基线**快速 R-CNN 在 PASCAL VOC 2007/2012 测试集上的物体检测 mAP（%）。更好的结果请参见表 10 和表 11。

公制	mAP@.5	mAP@[.5, .95]
VGG-16	41.5	21.2
ResNet-101	**48.4**	**27.2**

表 8.使用**基线**快速 R-CNN 在 COCO 验证集上的物体检测 mAP（%）。更佳结果另见表 9。

有相似的训练误差。我们认为这是由于过度拟合造成的。对于这个小数据集来说，1202 层网络可能过于庞大（19.4M）。为了在该数据集上获得最佳结果（[10, 25, 24, 35]），我们采用了诸如 maxout [10] 或 dropout [14] 等强正则化方法。在本文中，我们没有使用 maxout/dropout，只是简单地通过深层和薄层架构设计施加正则化，而没有分散对优化难点的关注。但结合更强的正则化可能会对结果产生影响，我们将在未来对此进行研究。

4.3. PASCAL 和 MS COCO 上的目标检测

我们的方法在其他识别任务中具有良好的泛化性能。表 7 和表 8 显示了 2007 年和 2012 年 PASCAL VOC 的物体识别基线结果。

[5] 和 COCO [26]。我们采用 *Faster R-CNN* [32] 作为去检测方法。在此，我们关注的是 ResNet-101 取代 VGG-16 [41] 所带来的改进。使用这两种模型的检测实现（见附录）是相同的，因此收益只能归因于更好的网络。最值得注意的是，在具有挑战性的 COCO 数据集上，我们发现 COCO 的标准指标（mAP@[.5.X]）提高了 6.0%、

.95]），相对提高了 28%。这一进步完全归功于学习到的表征。

基于深度残差网络，我们在 ILSVRC 和 COCO 2015 比赛中获得了多个赛道的第一名：ImageNet检测、ImageNet定位、COCO检测和COCO分割。详情见附录。

参考资料

[1] Y.Bengio, P. Simard, and P. Frasconi.用梯度下降学习长期依赖是困难的。*IEEE Transactions on Neural Networks*, 5(2):157-166, 1994.

[2] C.M. Bishop.*模式识别神经网络》*，牛津大学出版社，1995 年。牛津 大学出版社，1995 年。

[3] W.W. L. Briggs, S. F. McCormick, et al. *A Multigrid Tutorial*.Siam, 2000.

[4] K.Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman.细节决定成败：最新特征编码方法评估。*BMVC*, 2011.

[5] M.Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman.Pascal Visual Object Classes (VOC) Challenge.*IJCV*， 第 303-338 页，2010 年。

[6] S.Gidaris 和 N. Komodakis.通过多区域和 语义分割感知 cnn 模型进行物体检测。*ICCV*, 2015.

[7] R.Girshick.快速 R-CNN.In *ICCV*, 2015.

[8] R.Girshick, J. Donahue, T. Darrell, and J. Malik.用于精确物体检测和语义分割的 丰富特征层次。 *CVPR*, 2014.

[9] X.Glorot 和 Y. Bengio.了解 深度前馈神经网络的训练难度。*AISTATS*, 2010.

[10] I.J. Goodfellow、D. Warde-Farley、M. Mirza、A. Courville 和 Y.Bengio.Maxout networks. *ArXiv:1302.4389*, 2013.

[11] K.He and J. Sun.时间受限的卷积神经网络 cost.In *CVPR*, 2015.

[12] K.He, X. Zhang, S. Ren, and J. Sun.用于视觉识别的深度 卷积网络中的空间金字塔池。In *ECCV*, 2014.

[13] K.He, X. Zhang, S. Ren, and J. Sun.深入研究整流器：超越人类水平的图像网分类性能。 *ICCV*，2015。

[14] G. E. Hinton、N. Srivastava、A. Krizhevsky、I. Sutskever 和 R.R. Salakhutdinov.通过防止特征检测器的共 适应改进神经网络。*arXiv:1207.0580*，2012。

[15] S.Hochreiter 和 J. Schmidhuber.Long short-term memory.*神经 计算*，9（8）：1735-1780，1997.

[16] S.Ioffe 和 C. Szegedy.批量归一化：通过减少内部协变量偏移加速深度 网络训练。In *ICML*, 2015.

[17] H.Jegou, M. Douze, and C. Schmid.最近 邻居搜索的乘积量化。*tpami*, 33, 2011.

[18] H.Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C.Schmid.将局部图像描述符聚合成紧凑代码 *TPAMI*, 2012.

[19] Y.Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick、S.Guadarrama 和 T. Darrell。Caffe：用于 快速特征嵌入的卷积架构。*arXiv:1408.5093*，2014。

[20] A.Krizhevsky.从微小图像中学习多层特征 ages.*技术报告*，2009 年。

[21] A.Krizhevsky, I. Sutskever, and G. Hinton.使用深度卷积神经网络的图像网分类 。In *NIPS*, 2012.

[22] Y.LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard、W.Hubbard, and L. D. Jackel.反向传播应用于手写邮政编码识别- . *神经计算*，1989 年。

[23] Y.LeCun, L. Bottou, G. B. Orr, and K.-R.Mu¨ller.Efficient backprop. *神经网络：贸易技巧》*，第 9-50 页。Springer, 1998.

[24] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu.Lee、S. Xie、P. Gallagher、Z. Zhang 和 Z. Tu.Deeply- supervised nets. *arXiv:1409.5185*, 2014.

[25] M.Lin, Q. Chen, and S. Yan.网络中的网络。*ArXiv:1312.4400*, 2013.

[26] T.-Y. LinLin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan、P.Dolla´r, and C. L. Zitnick.微软 COCO： 上下文中的通用对象。In *ECCV*.2014.

[27] J.Long, E. Shelhamer, and T. Darrell.用于语义分割的全卷积网络 。In *CVPR*, 2015.

[28] G. Montu´far、R. Pascanu、K. Cho 和 Y. Bengio。关于深度神经网络 线性区域的数量。In *NIPS*, 2014.

[29] V.Nair 和 G. E. Hinton.整流线性单元改进受限 波尔兹曼机。*ICML*, 2010.

[30] F.Perronnin 和 C. Dance.用于 图像分类的视觉词汇表费雪核。2007年，*CVPR*。

[31] T.Raiko, H. Valpola, and Y. LeCun.通过 感知器中的线性变换让深度学习更简单。*AISTATS*, 2012。

[32] S.Ren, K. He, R. Girshick, and J. Sun.更快的 R-CNN：利用区域建议网络实现实时物体检测。In *NIPS*, 2015.

[33] S.Ren, K. He, R. Girshick, X. Zhang, and J. Sun.卷积特征图上的物体检测 网络。*arXiv:1504.06066*, 2015.

[34] B.D. Ripley.*Pattern Recognition and neural networks*. 剑 桥 University Press, 1996.

[35] A.Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y.Bengio.网：薄深度网的提示。In *ICLR*, 2015.

[36] O.O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、
Z.Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv:1409.0575*, 2014.

[37] A.M. Saxe, J. L. McClelland, and S. Ganguli.深度线性神经网络学习非线性动力学的精确解。 *arXiv:1312.6120*, 2013.

[38] N.N. Schraudolph.通过因子中心化加速梯度下降 分解。技术报告，1998 年。

[39] N.N. Schraudolph.神经网络梯度因子居中In *Neural Networks：贸易技巧》*，第 207-226 页。Springer, 1998.

[40] P.Sermanet、D. Eigen、X. Zhang、M. Mathieu、R. Fergus 和 Y. Le- Cun。Overfeat：使用卷积网络的集成识别、定位和检测 。*国际卷积网络会议*，2014 年。

[41] K.Simonyan 和 A. Zisserman.用于大规模图像识别的深度卷积网络 。In *ICLR*, 2015.

[42] R.K. Srivastava, K. Greff, and J. Schmidhuber.高速公路网络。*arXiv:1505.00387*, 2015.

[43] R.K. Srivastava, K. Greff, and J. Schmidhuber.训练非常深的 网络。*1507.06228*, 2015.

[44] C.Szegedy 、 W. Liu 、 Y. Jia 、 P. Sermanet 、 S. Reed 、 D. Anguelov、D. Er-han、V. Vanhoucke 和 A. Rabinovich。更深入的卷积- tions。In *CVPR*, 2015.

[45] R.Szeliski.使用层次基函数的快速曲面插值 *tions.TPAMI*, 1990.

[46] R.Szeliski.局部适应的分层基础预处理在 *siggraph*，2006 年。

[47] T.Vatanen, T. Raiko, H. Valpola, and Y. LeCun.将随机梯度推向二阶方法--非线性变换的后向传播学习。在*神经信息 处理*中，2013 年。

[48] A.Vedaldi 和 B. Fulkerson.VLFeat： VLFeat: An open and portable library of computer vision algorithms, 2008.

[49] W.W. Venables and B. Ripley.用 S-plus 进行现代应用统计》。1999.

[50] M.D. Zeiler 和 R. Fergus.可视化和理解卷 积 神经网络。*ECCV*, 2014.

A. 物体检测基线

在本节中，我们将介绍基于基线 Faster R-CNN [32] 系统的检测方法。模型由 ImageNet 分类模型初始化，然后根据物体检测数据进行微调。在 ILSVRC 和 COCO 2015 检测竞赛期间，我们使用 ResNet-50/101 进行了实验。

与 [32] 中使用的 VGG-16 不同，我们的 ResNet 没有隐藏 fc 层。我们采用 "基于 Conv 特征图的网络"（NoC）[33] 的理念来解决这一问题。我们使用图像上跨度不大于 16 像素的层计算全图像共享 conv 特征图（*即* conv1、conv2 x、conv3 x 和 conv4 x，ResNet-101 中共有 91 个 conv 层；表 1）。我们将这些层视为类似于 VGG-16 中的 13 个 conv 层，这样，ResNet 和 VGG-16 的 conv 特征图的总步长（16 像素）就相同了。区域提议网络（RPN，可生成 300 个提议）共享这些层

[32] 和快速 R-CNN 检测网络 [7]。在 conv5 1 之前执行 RoI 池[7]，在此 RoI 池特征上，每个区域采用 conv5 x 及以上的所有层，扮演 VGG-16 的 fc 层的角色。最后的分类层由两个同胞层（分类和盒式回归 [7]）取代。

对于 BN 层的使用，在预训练之后，我们在 ImageNet 训练集上计算每一层的 BN 统计量（均值和方差）。然后，在对物体检测进行微调时固定 BN 层。这样，BN 层就变成了具有恒定偏移和尺度的线性激活，BN 统计数据不会因微调而更新。我们固定 BN 层主要是为了减少 Faster R-CNN 训练中的内存消耗。

PASCAL VOC

根据文献[7, 32]，对于 PASCAL VOC 2007 *测试*集，我们使用 VOC 2007 中的 5k *trainval* 图像和 VOC 2012 中的 16k *train- val* 图像进行训练（"07+12"）。对于 PASCAL VOC 2012 *测试*集，我们使用 VOC 2007 中的 10k *trainval+test* 图像和 VOC 2012 中的 16k *trainval* 图像进行训练（"07++12"）。训练 Faster R-CNN 的超参数与 [32] 相同。表 7 显示了结果。ResNet-101 的 mAP 比 VGG-16 提高了 *3%*。这一提高完全是由于 ResNet 学习到了更好的特征。

MS COCO

MS COCO 数据集 [26] 包含 80 个对象类别。我们对 PASCAL VOC 指标（mAP @ IoU = 0.5）和标准 COCO 指标（mAP @ IoU = 0.5）进行了评估。

.5:.05:.95).我们使用训练集上的 8 万张图像进行训练，使用评估集上的 4 万张图像进行评估。我们的 COCO 检测系统与 PASCAL VOC 类似。我们使用 8GPU 实现对 COCO 模型进行训练，因此 RPN 步骤的迷你批次大小为

8 幅图像（*即*每个 GPU 1 幅图像），快速 R-CNN 步骤的迷你批次大小为 16 幅图像。RPN 步骤和快速 R-CNN 步骤均以 0.001 的学习率进行 240k 次迭代训练，然后以 0.0001 的学习率进行 80k 次迭代训练。

表 8 显示了 MS COCO 验证集的结果。ResNet-101 与 VGG-16 相比，mAP@[.5, .95] 提高了 6%，相对提高了 28%，这完全归功于更好的网络所学习到的特征。值得注意的是，mAP@[.5, .95] 的绝对增幅（6.0%）几乎与 mAP@.5 的增幅（6.9%）相当。这表明，深度网络可以提高识别和定位能力。

B. 物体检测改进

为完整起见，我们报告了为比赛所做的改进。这些改进基于深度特征，因此应从残差学习中获益。

MS COCO

*方框细化*我们的方框细化部分沿用了 [6] 中的迭代定位。在 Faster R-CNN 中，最终输出是一个回归盒，它不同于其提议盒。因此，在推理时，我们从回归框中汇集一个新特征，得到一个新的分类分数和一个新的回归框。我们将这 300 个新的预测结果与最初的 300 个预测结果结合起来。使用 0.3 的 IoU 阈值[8]，对预测框的联合集进行非最大抑制 (NMS)，然后进行框投票[6]。方框再精细化可将 mAP 提高约 2 个点（表 9）。

*全局语境*我们在快速 R-CNN 步骤中结合了全局上下文。给定全图 conv 特征图后，我们通过全局空间金字塔汇集法 [12]（使用 "单层 "金字塔）汇集一个特征，该方法可作为 "RoI "汇集法实施，使用整个图像的边界框作为 RoI。这一汇集的特征被输入到后 RoI 层，以获得全局上下文特征。该全局特征与原始的每个区域特征相结合，然后进入同级分类和盒式回归层。这种新结构是端到端训练。全局上下文将 mAP@.5 提升了约 1 个点（表 9）。

*多尺度测试。*在上文中，所有结果都是通过单尺度训练/测试获得的，如文献[32]，其中图像的较短边为 $s = 600$ 像素。多尺度训练/测试在文献[12, 7]中是通过从特征金字塔中选择一个尺度来实现的，在文献[33]中是通过使用最大输出层来实现的。在我们目前的实施中，我们按照 [33] 的方法进行了多尺度*测试*；由于时间有限，我们没有进行多尺度训练。此外，我们只对快速 R-CNN 步骤（尚未对 RPN 步骤）进行了多尺度测试。利用训练好的模型，我们在图像金字塔上计算 conv 特征图，其中图像的短边为 $s \in \{200, 400, 600, 800, 1000\}$。

训练数据	COCO 火车		COCO trainval	
测试数据	COCO 值		COCO test-dev	
mAP	@.5	@[.5, .95]	@.5	@[.5, .95]
基线 更快的 R-CNN (VGG-16)	41.5	21.2		
基线 更快的 R-CNN (ResNet-101)	48.4	27.2		
+ 框的细化	49.9	29.9		
+背景	51.1	30.0	53.3	32.2
+多尺度测试	53.8	32.5	**55.7**	**34.9**
建筑群			**59.0**	**37.4**

表 9.使用 Faster R-CNN 和 ResNet-101 对 MS COCO 进行物体检测的改进。

系统	网	数据	mAP	自行车 鸟 船 瓶子 公共汽车 carcat chair cow table dog horse mbike person plant sheep sofa train tv
底线	VGG-16	07+12	73.2	76.5 79.0 70.9 65.5 52.1 83.1 84.7 86.4 52.0 81.9 65.7 8 4 . 8 8 4 . 6 77.5 76.7 38.8 73.6 7 3 . 9 83.0 72.6
底线	ResNet-101	07+12	76.4	79.8 80.7 76.2 68.3 55.9 8 5 . 1 85.3 **89.8** 56.7 87.8 69.4 88.3 88.9 80.9 78.4 41.7 7 8 . 6 79.8 85.3 72.0
基线+++	ResNet-101	COCO+07+12	**85.6**	**90.0 89.6 87.8 80.8 76.1 8 9 . 9 89.9 8 9 . 6 75.5 90.0 8 0 . 7 89.6 9 0 . 3 8 9 . 1 88.7 65.4 88.1 85.6 8 9 . 0 86.8**

表 10.2007 年 PASCAL VOC 测试集的检测结果。基线为 Faster R-CNN 系统。基线+++"系统包括表 9 中的方框细化、上下文和多尺度测试。

系统	网	数据	mAP	自行车 鸟 船 瓶子 公共汽车 carcat chair cow table dog horse mbike person plant sheep sofa train tv
底线	VGG-16	07++12	70.4	84.9 79.8 74.3 53.9 49.8 77.5 75.9 88.5 45.6 77.1 55.3 86.9 81.7 80.9 79.6 40.1 72.6 60.9 8 1 . 2 61.5
底线	ResNet-101	07++12	73.8	86.5 81.6 77.2 58.0 51.0 78.6 76.6 93.2 48.6 80.4 59.0 92.1 85.3 84.8 80.7 48.1 77.3 66.5 84.7 65.6
基线+++	ResNet-101	COCO+07++12	**83.8**	**92.1 88.4 84.8 75.9 71.4 86.3 87.8 94.2 66.8 89.4 69.2 93.9 91.9 90.9 8 9 . 6 67.9 88.2 76.8 9 0 . 3 80.0**

表 11.PASCAL VOC 2012 测试集（`http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=4`）的检测结果。基线为 Faster R-CNN 系统。基线+++"系统包括表 9 中的方框细化、上下文和多尺度测试。

我们按照 [33] 的方法从金字塔中选择两个相邻的尺度。在这两个尺度的特征图上进行 RoI 池和后续层[33]，并按照[33]中的 maxout 方法进行合并。多尺度测试使 mAP 提高了 2 个点以上（表 9）。

*使用验证数据。*接下来，我们使用 80k+40k trainval 集进行训练，并使用 20k test-dev 集进行评估。测试-开发集没有公开的地面实况，结果由评估服务器报告。在这种设置下，结果是 mAP@.5 为 55.7%，mAP@[.5, .95] 为 34.9%（表 9）。这就是我们的单一模型结果。

*集合。*在 Faster R-CNN 中，系统设计用于学习区域建议和对象分类器，因此可以使用集合来增强这两项任务。我们使用集合来提出区域建议，而建议的联合集则由每个区域分类器的集合来完成。表 9 显示了我们基于 3 个网络的组合得出的结果。在 test-dev 集上，mAP 为 59.0%，mAP 为 37.4%。*这一结果赢得了 COCO 2015 检测任务的第一名。*

PASCAL VOC

根据上述模型，我们重新研究了 PASCAL VOC 数据集。利用 COCO 数据集上的单一模型（表 9 中的 55.7% mAP@.5），我们在 PAS CAL VOC 数据集上对该模型进行了微调。此外，我们还采用了方框细化、连贯文本和多尺度测试等改进方法。这样做

	val2	测试
GoogLeNet [44] (ILSVRC'14)	-	43.9
我们的单一模型（ILSVRC'15）	60.5	58.8
我们的合奏（ILSVRC'15）	**63.6**	**62.1**

表 12.我们在 ImageNet 检测数据集上的结果（mAP，%）。我们的检测系统是 Faster R-CNN [32]，使用 ResNet-101 进行了表 9 所列的改进。

在 PASCAL VOC 2007（表 10）和 PASCAL VOC 2012（表 11）中，我们分别实现了 85.6% 和 83.8% 的 mAP[6].PASCAL VOC 2012 的结果比之前最先进的结果高出 10 个百分点[6]。

图像网络检测

ImageNet 检测（DET）任务涉及 200 个对象类别。准确度由 mAP@.5 评估。我们针对 ImageNet DET 的对象检测算法与表 9 中针对 MS COCO 的算法相同。网络在 1000 个类别的 ImageNet 分类集上进行了预训练，并在 DET 数据上进行了微调。我们按照 [8] 将验证集分为两部分（val1/val2）。我们使用 DET 训练集和 val1 集对检测模型进行微调。val2 集用于验证。我们不使用其他 ILSVRC 2015 数据。我们使用 ResNet-101 的单一模型具有

[6] http://host.robots.ox.ac.uk:8080/anonymous/3OJ4OJ.html，提交日期：2015-11-26。

LOC 方法	LOC 网络	测试	GT CLS 上的 LOC 错误	分类网络	预测 CLS 的前 5 个 LOC 误差
VGG's [41]	VGG-16	1 种	33.1 [41]		
RPN	ResNet-101	1 种	13.3		
RPN	ResNet-101	浓密	11.7		
RPN	ResNet-101	浓密		ResNet-101	14.4
RPN+RCNN	ResNet-101	浓密		ResNet-101	**10.6**
RPN+RCNN	建筑群	浓密		建筑群	**8.9**

表 13.ImageNet 验证的定位误差（%）。在 "GT 类的定位误差"（[41]）一栏中，使用的是地面实况类。在 "测试 "一栏中，"1-crop "表示在 *224×224* 像素的中心裁剪上进行测试，"dense "表示密集（全卷积）和多尺度测试。

在 DET 测试集上，我们的 3 个模型集合的 mAP 为 58.8%，mAP 为 62.1%（表 12）。*这一结果赢得了 ILSVRC 2015 ImageNet 检测任务的第一名*，超过第二名 **8.5 分**（绝对值）。

C. 图像网络定位

ImageNet 定位（LOC）任务 [36] 要求对物体进行分类和定位。按照文献 [40, 41]，我们假定首先采用图像级分类器预测图像的类别标签，而定位算法只考虑根据预测的类别预测边界框。我们采用 "每类再回归"（PCR）策略[40, 41]，为每一类学习一个边界框回归器。我们对网络进行 Im- ageNet 分类预训练，然后对其进行定位微调。我们在提供的 1000 个类别的 ImageNet 训练集上训练网络。

我们的定位算法基于 [32] 的 RPN 框架，并做了一些修改。与
[而我们用于定位的 RPN 则是*按*类别设计的。与文献[32]一样，该 RPN 以两个 1×1 卷积层结束，分别用于二元分类（*cls*）和盒式回归（*reg*）。与文*献[*32]不同的是，卷积层和回归层都是*按类别划分的。*具体来说，*cls* 层有 1000-d 的输出，每个维度都是*二元逻辑回归，*用于预测是否属于某个对象类别；*reg* 层有 1000×4-d 的输出，由 1000 个类别的盒式回归器组成。与文献[32]一样，我们的边界框回归是参照每个位置的多个平移不变的 "锚 "框。

方法	五大本地化错误	
	缬氨酸	测试
战胜[40]（ILSVRC'13）	30.0	29.9
GoogLeNet [44] (ILSVRC'14)	-	26.7
VGG [41]（ISVRC'14）	26.9	25.3
我们的（ILSVRC'15）	**8.9**	**9.0**

与 ImageNet 分类训练（第 3.4 节）一样，我们随机抽取 224×224 农作物作为数据扩增样本。我们使用 256 幅图像的迷你批量进行微调。为避免负样本占据主导地位，我们对每张图像随机抽取 8 个锚点，其中正锚点和负锚点的抽样比例为 1:1[32]。测试时，网络将对图像进行全卷积处理。

表 13 比较了定位结果。根据文献 [41]，我们首先使用地面实况类别作为分类预测，进行 "oracle "测试。VGG 的论文[41]重新进行了 "oracle "测试。

表 14.在 ImageNet 数据集上与最先进方法的定位误差（%）比较。

在使用地面实况类时，中心点误差为 33.1%（表 13）。在相同设置下，我们使用 ResNet-101 网的 RPN 方法将中心点误差大幅降低至 13.3%。这一对比证明了我们框架的卓越性能。在密集（完全卷积）和多尺度测试中，我们的 ResNet-101 使用地面实况类的误差为 11.7%。使用 ResNet-101 预测类别（前五名分类误差为 4.6%，表 4），前五名定位误差为 14.4%。

上述结果仅基于 Faster R-CNN [32] 中的*提议网络*（RPN）。我们可以使用 Faster R-CNN 中的*检测网络*（Fast R-CNN [7]）来改善结果。但我们注意到，在该数据集上，一幅图像通常只包含一个主要对象，而提议区域彼此高度重叠，因此具有非常相似的 RoI 池特征。因此，快速 R-CNN [7] 以图像为中心的训练会产生变化较小的样本，这可能不是随机训练所希望的。受此启发，在目前的实验中，我们使用以 RoI 为中心的原始 R-CNN [8] 代替快速 R-CNN。

我们的 R-CNN 实现过程如下。我们在训练图像上应用按类训练的 RPN 来预测基本真实类的边界框。这些预测的边界框起到了与类别相关的提案的作用。对于每张训练图像，提取得分最高的 200 个提案作为训练样本，以训练 R-CNN 分类器。图像区域从提案中裁剪出来，扭曲为 224×224 像素，然后输入 R-CNN 分类网络 [8]。该网络的输出由 *cls* 和 *reg* 的两个同级 fc 层组成，也是按类别形式。这个 R-CNN 网络以 RoI 为中心，在训练集上使用 256 个小批量进行微调。测试时，RPN 为每个预测类别生成得分最高的 200 个建议，R-CNN 网络用于更新这些建议的得分和方框位置。

这种方法将前五名的定位误差降低到了 10.6%（表 13）。这是我们在验证集上的单一模型结果。使用网络集合进行分类和定位，我们在测试集上的前 5 名定位误差为 9.0%。这一结果明显优于 ILSVRC 14 的结果（表 14），误差相对减少了 64%。*这一结果赢得了 2015 年 ILSVRC ImageNet 定位任务的第一名。*