VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan & Andrew Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford
{karen,az}@robots.ox.ac.uk

摘要

在这项工作中，我们研究了卷积网络深度对其在大规模图像识别设置中的准确性的影响。我们的主要贡献是使用非常小（3×3）的卷积滤波器的架构对深度增加的网络进行全面评估，这表明通过将深度推到 16-19 个权重层，可以实现对现有技术配置的显着改进。这些发现是我们提交 2014 年 ImageNet 挑战赛的基础，我们的团队分别在定位和分类任务中获得了第一和第二名。我们还展示了我们的表示在其他数据集上具有很好的泛化能力，并取得了最先进的结果。为了促进深度视觉表示在计算机视觉中的进一步研究，我们已经公开提供了我们两个表现最佳的 ConvNet 模型。

1 引言

卷积网络(ConvNets)最近在大规模图像和视频识别方面取得了巨大的成功，这得益于大规模公共图像库（如 ImageNet）和高性能计算系统（如 GPU 或大规模分布式集群）的出现。特别是，ImageNet 大规模视觉识别挑战(ILSVRC) 在深度视觉识别架构的进步中发挥了重要作用，它作为一个试验平台，测试了几代大规模图像分类系统，从高维浅层特征编码（ILSVRC-2011 的获胜者）到深度 ConvNets（ILSVRC-2012 的获胜者）。

随着 ConvNets 在计算机视觉领域变得越来越重要，人们已经进行了许多尝试来改进 Krizhevsky 等人的原始架构，以达到更高的准确性。例如，ILSVRC-2013 中表现最好的提交，利用了较小的感受窗口尺寸和较小的第一个卷积层步幅。另一方面的改进是对网络进行密集的训练和测试，覆盖整个图像和多个尺度。在本文中，我们解决了 ConvNet 架构设计的另一个重要方面-深度。为此，我们固定架构的其他参数，并通过添加更多的卷积层来逐步增加网络的深度，这是由于在所有层中使用非常小的(3×3)卷积滤波器而可行的。

因此，我们提出了更精确的 ConvNet 架构，它不仅在 ILSVRC 分类和定位任务上实现了最先进的准确性，而且还适用于其他图像识别数据集，即使用作相对简单的管道的一部分（例如，由线性 SVM 分类的深度特征，无需微调），它们也能获得出色的性能。我们发布了两款性能最佳的型号以促进进一步的研究。

本文的其余部分组织如下。第 2 节介绍了我们的 ConvNet 配置。然后在第 3 节中介绍了图像分类训练和评估的细节，并在第 4 节中对 ILSVRC 分类任务的配置进行了比较。第 5 节总结了本文。为了完整起见，附录 A 中描述和评估了我们的 ILSVRC-2014 目标定位系统，并在附录 B 中讨论了非常深层特征在其他数据集上的泛化性能。最后，附录 C 包含了主要的论文修订列表。

2 ConvNet 配置

为了衡量 ConvNet 深度在公平环境中所带来的改进，我们所有的 ConvNet 层配置都使用相同的规则，灵感来自 Ciresan 等。在本节中，我们首先描述我们的 ConvNet 配置的通用设计（第 2.1 节），然后详细说明评估中使用的具体配置（第 2.2 节）。最后，我们的设计选择将在 2.3 节进行讨论并与现有技术进行比较。

2.1 架构

在训练期间，我们的 ConvNet 的输入是固定大小的 224×224 RGB 图像。我们唯一的预处理是从每个像素中减去在训练集上计算的 RGB 均值。图像通过一堆卷积层，我们使用感受野很小的滤波器：3×3（这是捕获左/右，上/下，中心概念的最小尺寸）。在其中一种配

置中，我们还使用了卷积滤波器，可以看作输入通道的线性变换（后面是非线性）。卷积步长固定为 1 个像素；卷积层输入的空间填充使得在卷积后保持空间分辨率，即对于 3×3 的卷积层，填充（padding）为 1 个像素。空间池化由五个最大池化层进行，这些层在一些卷积层之后（不是所有的卷积层之后都是最大池化）。最大池化在 2×2 像素窗口上进行，步长为 2。

一系列卷积层（在不同架构中具有不同深度）之后是三个全连接（Fully-Connecte,FC）层：前两个每个都有 4096 个通道，第三个执行 1000 维 ILSVRC 分类，因此包含 1000 个通道（一个通道对应一个类别）。最后一层是 soft-max 层。所有网络中全连接层的配置是相同的。

所有隐藏层都配备了修正非线性（ReLU）。我们注意到，我们的网络（除了一个）都不包含局部响应归一化（Local Response Normalisation,LRN）：将在第 4 节看到，这种归一化并不能提高在 ILSVRC 数据集上的性能，但增加了内存消耗和计算时间。在应用的地方，LRN 层的参数是（Krizhevsky 等，2012）的参数。

2.2 配置

本文中评估的 ConvNet 配置在表 1 中列出，每列一个。接下来我们将按网络名称（A-E）来提及网络。所有配置都遵循 2.1 节提出的通用设计，并且仅是深度不同：从网络 A 中的 11 个加权层（8 个卷积层和 3 个全连接层）到网络 E 中的 19 个加权层（16 个卷积层和 3 个全连接层）。卷积层的宽度（通道数）相当小，从第一层中的 64 开始，然后在每个最大池化层之后增加 2 倍，直到达到 512。

表 1：ConvNet 配置（以列显示）。随着更多的层被添加，配置的深度从左（A）增加到右（E）（添加的层以粗体显示）。卷积层参数表示为“conv<感受野大小>-<通道数>”。为了简洁起见，我们没有将 ReLU 的激活函数显示出来。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

在表 2 中，我们报告了每个配置的参数量。尽管深度很大，我们的网络中权重数量并不大于具有更大卷积层宽度和感受野的更浅网络中的权重数量（在（Sermanet 等人，2014）中为 144M 的权重）。

表 2：参数量（百万级别）

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

2.3 讨论

我们的 ConvNet 配置与 ILSVRC-2012（Alexnet）和 ILSVRC-2013（ZFNet）比赛表现最佳的参赛提交中使用的 ConvNet 配置有很大不同。不是在第一卷积层中使用相对较大的感受野（例如，在（Krizhevsky 等人，2012）中的 11×11，步长为 4，或在（Zeiler＆Fergus，2013；Sermanet 等，2014）中的 7×7，步长为 2），我们在整个网络使用非常小的 3×3 感受野，与输入的每个像素（步长为 1）进行卷积。显而易见的是两层级联 3×3 卷积拥有 5×5 的感受野；三层级联 3×3 拥有 7×7 的有效感受野。如果我们使用三层级联 3×3 卷积替换 7×7 卷积能够有什么好处？第一，用三层非线性整流层替换简单的一层，使得决策函数更具判别性。第二，降低了参数量：假设三层 3×3 卷积堆叠的输入和输出都有 C 个通道，则级联的参数量为 $3(3^2C^2) = 27C^2$，同时，单个 7×7 的卷积参数量为 $(7^2C^2) = 49C^2$，即参数多 81%。这可以看作是对 7×7 卷积滤波器进行正则化，迫使它们通过 3×3 滤波器（在它们之间注入非线性）进行分解。

结合 1×1 卷积层（配置 C，表 1）是在不影响卷积层感受野的情况下增加决策函数非线性能力的一种方式。即使在我们的案例下，1×1 卷积基本上是在相同维度空间上的线性投影（输入和输出通道的数量相同），由修正函数引入附加的非线性能力。值得注意的是，1×1 卷积层最近在 Lin 等人(2014)的“Network in Network(NIN)”架构中已经得到了使用。

Ciresan 等人（2011）以前使用小尺寸的卷积滤波器，但是他们的网络深度远远低于我们的网络，他们并没有在大规模的 ILSVRC 数据集上进行评估。Goodfellow 等人（2014）在街道号识别任务中采用深层 ConvNets（11 个权重层），显示出增加的深度导致了更好的性能。GooLeNet（Szegedy 等，2014），ILSVRC-2014 分类任务的表现最好的项目，是独立于我们工作之外的开发的，但是类似的是它是基于非常深的 ConvNets（22 个权重层）和小卷积滤波器（除了 3×3，它们也使用了 1×1 和 5×5 卷积）。然而，它们的网络拓扑结构比我们的更复杂，并且在第一层中特征图的空间分辨率被更积极地减少，以减少计算量。正如将在第 4.5 节显示的那样，我们的模型在单网络分类精度方面胜过 Szegedy 等人（2014）。

3 分类框架

在上一节中，我们介绍了我们的网络配置的细节。在本节中，我们将介绍分类 ConvNet 训练和评估的细节。

3.1 训练

ConvNet 训练过程通常遵循 Krizhevsky 等人（2012）（除了从多尺度训练图像中对输入裁剪图像进行采样外，如下文所述）。也就是说，通过使用具有动量的小批量梯度下降（基于反向传播（LeCun 等人，1989））优化多项式逻辑回归目标函数来进行训练。批量大小设为 256，动量为 0.9。训练通过权重衰减（L_2惩罚乘子设定为 $5 * 10^{-4}$）进行正则化，前两个全连接层执行 dropout 正则化（dropout 率设定为 0.5）。学习率初始设定为10^{-2}，然后当验证集准确率停止改善时，减少 10 倍。学习率总共降低 3 次，学习在 37 万次迭代后停止（74 个 epochs）。我们推测，尽管与 AlexNet 相比我们的网络参数更多，网络的深度更大，但网络需要更小的 epoch 就可以收敛，这是由于（a）由更大的深度和更小的卷积滤波器尺寸引起的隐式正则化，（b）某些层的预初始化。

网络权重的初始化是重要的，因为由于深度网络中梯度的不稳定，不好的初始化可能会阻碍学习。为了避免这个问题，我们开始训练配置 A（表 1），足够浅可以随机初始化进行训练。然后，当训练更深的架构时，我们用网络 A 的层初始化前四个卷积层和最后三个全连接层（中间层被随机初始化）。我们没有减少预初始化层的学习率，允许他们在学习过程中改变。对于随机初始化（如果应用），我们从均值为 0 和方差为10^{-2}的正态分布中采样权重。偏置初始化为零。值得注意的是，在提交论文之后，我们发现可以通过使用 Glorot＆Bengio（2010）的随机初始化程序来初始化权重而不进行预训练。

为了获得固定尺寸的 224×224 的 ConvNet 输入图像，它们是从重新缩放的训练图像中随机裁剪得到的（每个 SGD 迭代每个图像一次裁剪）。为了进一步增加训练集的多样性，裁剪图像还经历了随机水平翻转和随机 RGB 颜色偏移。训练图像的重新缩放方法将在下面

解释。

训练图像的尺寸. 设S是经过等比例缩放的训练图像的最小边,从中裁剪出ConvNet输入(我们也将S称为训练尺度)。虽然裁剪尺寸固定为224×224，但原则上S可以取任何不小于224的值：对于$S = 224$，裁剪将捕捉整个图像的统计信息，完全覆盖训练图像的最小边；对于$S \gg 224$，裁剪将对应于图像的一小部分，包含一个小物体或物体的一部分。

我们考虑两种设置训练尺度S的方法。第一种是固定S,这对应于单尺度训练(请注意,采样裁剪内的图像内容仍然可以表示多尺度图像统计信息)。在我们的实验中，我们评估了在两个固定尺度下训练的模型：$S = 256$（在先前的研究中广泛使用）和$S = 384$。给定一个ConvNet配置，我们首先使用$S = 256$训练网络。为了加速$S = 384$网络的训练，它使用了使用$S = 256$预训练的权重进行初始化，并且我们使用了较小的初始学习率10^{-3}。

第二种设置S的方法是多尺度训练，其中每个训练图像通过从一定范围$[S_{min}, S_{max}]$中随机采样S进行个别缩放（我们使用$S_{min} = 256$和$S_{max} = 512$）。由于图像中的物体可能具有不同的尺寸，因此在训练过程中考虑这一点是有益的。这也可以看作是通过尺度抖动来增加训练集的数据增强,其中单个模型被训练以识别一系列不同尺度的物体。出于速度的考虑，我们通过使用相同配置的单尺度模型的所有层进行微调，以预训练的固定$S = 384$模型来训练多尺度模型。

3.2 测试

在测试阶段，给定一个经过训练的卷积神经网络和一个输入图像，它的分类过程如下。首先，将图像等比例缩放到预定义的最小图像边长，记为Q（我们也称之为测试尺度）。需要注意的是，Q不一定等于训练尺度S（正如我们将在第4节中展示的，对于每个S使用多个Q的值可以提高性能）。然后，网络密集地应用于缩放后的测试图像，类似于（Sermanet等，2014）。也就是说，首先将全连接层转换为卷积层（将第一个全连接层转换为7×7的卷积层，将最后两个全连接层转换为1×1的卷积层）。然后，将得到的全卷积网络应用于整个（未裁剪的）图像。结果是一个类别分数图，通道数等于类别数，空间分辨率可变，取决于输入图像的大小。最后，为了获得图像的固定大小的类别分数向量，对类别分数图进行空间平均（求和池化）。我们还通过水平翻转图像来增强测试集；将原始图像和翻转图像的soft-max类后验进行平均，以获得图像的最终分数。

由于全卷积网络应用于整个图像，因此在测试时不需要对多个裁剪进行采样（Krizhevsky等，2012），这样做效率较低，因为每次裁剪都需要重新计算网络。同时，使用大量的裁剪，如Szegedy等人（2014）所做的，可以提高准确性，因为它相对于全卷积网络对输入图像进行了更细致的采样。此外，多裁剪评估与密集评估相辅相成，因为它们具有不同的卷积边界条件：当将ConvNet应用于裁剪时，卷积特征图会用零进行填充，而在密集评估的情况下，相同裁剪的填充自然来自图像的相邻部分（由于卷积和空间池化），这大大增加了整个网络的感受野，从而捕获更多的上下文信息。虽然我们认为在实践中，多个裁剪的增加计算时间并不能证明准确性的潜在提升，但为了参考，我们还使用每个尺度50个裁剪(5×5的规则网格和2个翻转),总共在3个尺度上进行了150个裁剪的评估,与Szegedy等人（2014）使用的4个尺度上的144个裁剪相当。

3.3 实现细节

我们的实现来源于公开的C++ Caffe工具箱（2013年12月推出），但包含了一些重大的修改，使我们能够对安装在单个系统中的多个GPU进行训练和评估，也能训练和评估在多个尺度上（如上所述）的全尺寸（未裁剪）图像。多GPU训练利用数据并行性，通过将每批训练图像分成几个GPU批次，每个GPU并行处理。在计算GPU批次梯度之后，将其平均以获得完整批次的梯度。梯度计算在GPU之间是同步的，所以结果与在单个GPU上训练完全一样。

最近提出了更加复杂的加速ConvNet训练的方法（Krizhevsky，2014），它们对网络的不同层之间采用模型和数据并行,我们发现我们概念上更简单的方案与使用单个GPU相比，在现有的4-GPU系统上已经提供了3.75倍的加速。在配备四个NVIDIA Titan Black GPU的系统上，根据架构训练单个网络需要2-3周时间。

4 分类实验

数据集. 在本节中，我们介绍了描述的 ConvNet 架构（用于 ILSVRC 2012-2014 挑战）在 ILSVRC-2012 数据集上实现的图像分类结果。数据集包括 1000 个类别的图像，并分为三个子集：训练集（130 万张图像），验证集（5 万张图像）和测试集（留有类标签的 10 万张图像）。使用两个措施评估分类性能：top-1 和 top-5 错误率。前者是多类分类误差，即不正确分类图像的比例；后者是 ILSVRC 中使用的主要评估标准，并且计算为图像真实类别在前 5 个预测类别之外的图像比例。

对于大多数实验，我们使用验证集作为测试集。在测试集上也进行了一些实验，并将其作为 ILSVRC-2014 竞赛“VGG”小组的输入提交到了官方的 ILSVRC 服务器。

4.1 单尺度评估

我们首先评估单个 ConvNet 模型在单尺度上的性能，其层结构配置如 2.2 节中描述。测试图像大小设置如下:对于固定 S 的 $Q = S$，对于当 S 在$[S_{min}, S_{max}]$范围内变动时，$Q = 0.5(S_{min} + S_{max})$。结果如表 3 所示。

表 3：在单一测试尺度的 ConvNet 性能

ConvNet config. (Table1)	smallest image side		top-1 val. error(%)	top-5 val. error(%)
	train(S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	287	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

首先，我们注意到使用局部响应归一化（A-LRN 网络）并没有改善没有任何归一化层的 A 模型。因此，在更深的架构（B-E）中我们没有采用归一化。

其次，我们观察到分类误差随着 ConvNet 深度的增加而减小：从 A 中的 11 层到 E 中的 19 层。值得注意的是，尽管深度相同，配置 C（包含三个 1×1 卷积层）比在整个网络层中使用 3×3 卷积的配置 D 更差。这表明，虽然额外的非线性确实有帮助（C 优于 B），但也可以通过使用具有非平凡感受野（D 比 C 好）的卷积滤波器来捕获空间上下文。当深度达到 19 层时，我们架构的错误率达到饱和，但更深的模型可能有益于较大的数据集。我们还将网络 B 与具有 5×5 卷积层的浅层网络进行了比较，浅层网络可以通过用单个 5×5 卷积层替换 B 中每对 3×3 卷积层得到（其具有相同的感受野如第 2.3 节所述）。测量的浅层网络 top-1 错误率比网络 B 的 top-1 错误率（在中心裁剪图像上）高 7%，这证实了具有小滤波器的深层网络优于具有较大滤波器的浅层网络。

最后，尽管在测试时只使用单一尺度，但训练时的尺度抖动（S 在[256, 512]范围内）比固定最小边长（S=256 或 S=384）的图像训练导致的结果显著更好。这证实了通过尺度抖动进行训练集增强确实有助于捕捉多尺度图像统计信息。

4.2 多尺度评估

在单尺度上评估 ConvNet 模型后，我们现在评估测试时尺度抖动的影响。它包括在一个测试图像的几个重新缩放的版本上运行一个模型（对应于不同的 Q 值），然后对所得到的类别后验进行平均。考虑到训练和测试尺度之间的巨大差异会导致性能下降，用固定 S 训练的模型在三个测试图像尺度上进行了评估，接近于训练一次：$Q = \{S - 32, S, S + 32\}$。同时，训练时的尺度抖动允许网络在测试时应用于更广的尺度范围，所以用变量 $S \in [S_{min}; S_{max}]$训练的模型在更大的尺寸范围 $Q = \{S_{min}, 0.5(S_{min} + S_{max}), S_{max}\}$上进行评估。

表 4 中给出的结果表明，测试时的尺度抖动导致了更好的性能（与在单一尺度上相同模型的评估相比，如表 3 所示）。如前所述，最深的配置(D 和)执行最佳，并且尺度抖动优于使用固定最小边 S 的训练。我们在验证集上的最佳单网络性能为 24.8%/7.5% top-1/top-5 的错误率(在表 4 中用粗体突出显示)。在测试集上，配置 E 实现了 7.3% top-5 的错误率。

表 4：在多个测试尺度上的 ConvNet 性能

ConvNet config. (Table1)	smallest image side		top-1 val. error(%)	top-5 val. error(%)
	train(S)	test (Q)		
B	256	224,256,288	282	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256;512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256;512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256;512]	256,384,512	**24.8**	**7.5**

4.3 多裁剪图像评估

在表 5 中，我们将稠密 ConvNet 评估与多裁剪图像评估进行比较（细节参见第 3.2 节）。我们还通过平均其 soft-max 输出来评估两种评估技术的互补性。可以看出，使用多裁剪图像表现比密集评估略好，而且这两种方法确实是互补的，因为它们的组合优于其中的每一种。如上所述，我们假设这是由于卷积边界条件的不同处理。

表 5：ConvNet 评估技术比较。在所有的实验中训练尺度 S 从[256；512]采样，我们考虑了三个测试量表 Q：{256, 384, 512}。

ConvNet config. (Table1)	Evaluation method	top-1 val. error(%)	top-5 val. error(%)
D	dense	24.8	7.5
	multi-crop	246	7.5
	multi-crop & dense	**24.4**	**7.2**
E	dense	24.8	7.5
	multi-crop	24.6	7.4
	multi-crop & dense	**24.4**	**7.1**

4.4 卷积网络融合

到目前为止，我们评估了 ConvNet 模型的性能。在这部分实验中，我们通过对 soft-max 类别后验进行平均，结合了几种模型的输出。由于模型的互补性，这提高了性能，并且在了 2012 年（Krizhevsky 等，2012）和 2013 年（Zeiler＆Fergus，2013；Sermanet 等，2014）ILSVRC 的顶级提交中使用。

结果如表 6 所示。在 ILSVRC 提交的时候，我们只训练了单规模网络，以及一个多尺度模型 D（仅在全连接层进行微调而不是所有层）。由此产生的 7 个网络组合具有 7.3%的 ILSVRC 测试误差。在提交之后，我们考虑了只有两个表现最好的多尺度模型（配置 D 和 E）的组合，它使用密集评估将测试误差降低到 7.0%，使用密集评估和多裁剪图像评估将测试误差降低到 6.8%。作为参考，我们表现最佳的单模型达到 7.1%的误差（模型 E，表 5）。

4.5 与最新技术比较

最后，我们在表 7 中与最新技术(SOTA)比较我们的结果。在 ILSVRC-2014 挑战的分类任务中，我们的“VGG”团队获得了第二名，使用 7 个模型的组合取得了 7.3%测试误差。提交后，我们使用 2 个模型的组合将错误率降低到 6.8%。

表 6：多个卷积网络融合结果

Combined ConvNet models	Error

	top-1 val	top-5 val	top-5 test
ILSVRC submission			
(D/256/224,256,288)，(D/384/352,384,416)，(D/[256;512]/256,384,512) (C/256/224,256,288)，(C/384/352,384,416) (E/256/224,256.288)，(E/384/352,384.416)	24.7	7.5	7.3
post-submission			
(D/[256:512]/256,384,512)，(E/[256;512]/256,384,512)，dense eval	24.0	7.1	7.0
(D/[256;512]/256,384,512)，(E/[256;512]/256,384,512)，multi-crop	23.9	7.2	-
(D/[256;512]/256,384,512)，(E/[256;512]/256,384,512)，multi-crop & dense eval	**23.7**	**6.8**	**6.8**

从表 7 可以看出，我们非常深的 ConvNets 显著优于前一代模型，在 ILSVRC-2012 和 ILSVRC-2013 竞赛中取得了最好的结果。我们的结果对于分类任务获胜者（GoogLeNet 具有 6.7%的错误率）也具有竞争力，并且大大优于 ILSVRC-2013 获胜者 Clarifai 的提交，其使用外部训练数据取得了 11.2%的错误率，没有外部数据则为 11.7%。这是非常显著的，考虑到我们最好的结果是仅通过组合两个模型实现的——明显少于大多数 ILSVRC 提交。在单网络性能方面，我们的架构取得了最好节果（7.0%测试误差），超过单个 GoogLeNet 0.9%。值得注意的是，我们并没有偏离 LeCun 等人经典的 ConvNet 架构，但通过大幅增加深度改善了它。

表 7：在 ILSVRC 分类中与最新技术比较。我们的方法表示为“VGG”。报告的结果没有使用外部数据。

Method	top-1 val. error(%)	top-5 val. error(%)	top-5 test error(%)
VGG(2 nets,multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG(1 net, multi-crop & dense eval.)	244	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014)(1 net)	-	7.9	
GoogLeNet(Szegedy et al.,2014)(7 nets)	-	**6.7**	
MSRA(Heetal.,2014)(11nets)	-	-	8.1
MSRA(Heetal.,2014)(1net)	27.9	9.1	9.1
Clarifai(Russakovsky et al.,2014)(multiple nets)	-	-	11.7
Clarifai(Russakovsky et al.,2014)(1 net)	-	-	12.5
Zeiler & Fergus(Zeiler & Fergus,2013)(6 nets)	36.0	14.7	14.8
Zeiler & Fergus(Zeiler & Fergus,2013)(1 net)	37.5	16.0	16.1
OverFeat(Sermanetetal.,2014)(7 nets)	34.0	13.2	13.6
OverFeat(Sermanetetal.,2014)(1 net)	35.7	14.2	-
Krizhevsky et al.(Krizhevsky et al.,2012)(5 nets)	38.1	16.4	16.4
Krizhevsky et al.(Krizhevsky et al.,2012)(1 net)	40.7	18.2	-

5 结论

在这项工作中，我们评估了非常深的卷积网络（最多 19 个权重层）用于大规模图像分类。结果表明，表示深度对分类精度有益，并且可以使用传统的 ConvNet 架构（LeCun 等，1989；Krizhevsky 等，2012）在 ImageNet 挑战数据集上实现最先进的性能。在附录中，我们还展示了我们的模型在各种任务和数据集上具有良好的泛化能力，并且在构建基于较浅图像表示的更复杂的识别流程时，与之匹配或超越其性能。我们的结果再次证实了在视觉表示中深度的重要性。

致谢

这项工作得到了 ERC 拨款 VisRec 编号 228180 的支持。我们感谢 NVIDIA 公司对用于本研究的 GPU 的支持。

附录

A 定位

本文主要讨论了 ILSVRC 挑战赛中分类任务的问题，并对不同深度的 ConvNet 架构进行了全面评估。在本节中，我们转向挑战赛的定位任务，我们在 2014 年以 25.3%的错误率获胜。这可以看作是目标检测的特殊情况，其中应预测每个前五类别的单个对象边界框，而不考虑该类别的实际对象数量。为此，我们采用了 Sermanet 等人（2014 年）在 ILSVRC-2013 定位挑战赛中的方法，并进行了一些修改。我们的方法在 A.1 节中描述，并在 A.2 节中进行了评估。

A.1 定位 ConvNet

为了进行对象定位，我们使用了一个非常深的 ConvNet，其中最后一个全连接层预测边界框位置，而不是类别得分。边界框由一个 4 维向量表示，存储其中心坐标、宽度和高度。有两种选择，边界框预测是否在所有类别之间共享（单类回归，SCR (Sermanet 等人，2014 年)）或者是类别特定的（每类回归，per-class regression,PCR）。在前一种情况下，最后一层是 4 维的，而在后一种情况下，它是 4000 维的（因为数据集中有 1000 个类别）。除了最后的边界框预测层外，我们使用了 ConvNet 架构 D（表 1），它包含 16 个权重层，在分类任务中表现最好（第 4 节）。

训练. 定位 ConvNet 的训练与分类 ConvNet 的训练类似（第 3.1 节）。主要区别在于，我们用欧几里德损失替换了逻辑回归目标，该损失惩罚预测的边界框参数与真实值之间的偏差。我们训练了两个定位模型，每个模型只在一个尺度上进行训练：$S = 256$ 和 $S = 384$（由于时间限制，我们没有在 ILSVRC-2014 提交中使用训练尺度抖动）。训练是以相应的分类模型（在相同尺度上训练）为初始状态进行的，初始学习率设置为10^{-3}。我们尝试了全层微调和仅微调前两个全连接层的方式，与（Sermanet 等人，2014 年）中所做的方式相同。最后一个全连接层是随机初始化并从头开始训练的。

测试. 我们考虑了两个测试方案。第一种方法用于比较验证集上的不同的网络修改，并只考虑地面真实类的边界框预测（以排除分类误差）。边界框仅通过将网络应用于图像的中央作物而获得的。

第二个完整的测试过程基于对整个图像进行定位卷积神经网络的密集应用，类似于分类任务(3.2 节)。不同之处在于，最后一个全连接层的输出是一组边界框预测，而不是类别得分图。为了得出最终的预测，我们使用了 Sermanet 等人（2014）的贪婪合并程序，该程序首先合并空间上接近的预测（通过平均其坐标），然后根据从分类 ConvNet 中获得的类分数对它们进行评级。当使用多个定位卷积神经网络时，首先将它们的边界框预测集合取并集，然后在并集上运行合并过程。我们没有使用 Sermanet et al.(2014)的多池化偏移技术，该技术增加了边界框预测的空间分辨率，可以进一步改善结果。

A.2 定位实验

在本节中，我们首先确定表现最佳的定位设置（使用第一个测试方案），然后在一个完整的场景中对其进行评估（第二个方案）。定位误差是根据 ILSVRC 准则（Russakovsky 等人，2014 年）进行测量的，即如果边界框预测与真实边界框的交并比(intersection over union,IoU)大于 0.5，则被认为是正确的。

设置比较. 从表 8 中可以看出，每类回归（PCR）优于类别不可知的单类回归（class-agnostic single-class regression,SCR），这与 Sermanet 等人（2014 年）的研究结果不同，他们发现 PCR 被 SCR 超越。我们还注意到，对于定位任务，微调所有层的结果明显比仅微调全连接层（如 Sermanet 等人，2014 年所做）要好。在这些实验中，最小图像边长设置为 $S = 384$；当 $S = 256$ 时，结果表现相同，为了简洁起见未写出。

表 8：使用简化的测试方案进行不同修改的定位误差：从单个中心图像裁剪预测边界框，并使用真实类别。所有 ConvNet 层（最后一层除外）具有配置 D（表 1），而最后一层执行单类回归（SCR）或每类回归（PCR）

Fine-tuned layers	regression type	GT class localisation error
1st and 2nd FC	SCR 36.4	SCR 36.4
	PCR 34.3	PCR 34.3
all	PCR 33.1	all PCR 33.1

完整的评估. 在确定了最佳的定位设置（PCR，所有层的微调）之后，我们现在将其应用于完整的场景，其中使用我们最佳的分类系统(第 4.5 节)预测前 5 个类标签，并使用 Sermanet et al. (2014)的方法合并多个密集计算的边界框预测。如 Table 9 所示，将定位 ConvNet 应用于整个图像相对于使用中心裁剪（表 8）显著改善了结果，尽管使用的是前 5 个预测的类标签而不是真实标签。类似于分类任务（第 4 节），在多个尺度上进行测试并结合多个网络的预测进一步提高了性能。

表 9:定位误差

smallest image side		top-5 localisation error (%)	
train(S)	test(Q)	val.	test.
256	256	29.5	-
384	384	28.2	267
384	352,384	27.5	-
fusion: 256/256 and 384/352,384		**26.9**	**25.3**

与现有技术的比较. 我们将我们最好的定位结果与现有技术在 Table 10 中进行比较。经过测试， 我们的 “VGG” 团队在 ILSVRC-2014 的定位挑战中以 25.3% 的测试误差获胜。值得注意的是，我们的结果比 ILSVRC-2013 的获胜者 Overfeat 的结果要好得多(Sermanet et al., 2014)，尽管我们使用的比例较小，并且没有采用他们的分辨率增强技术。我们预计，如果将这种技术纳入我们的方法中，可以实现更好的定位性能。这表明了我们的非常深入的 ConvNets 带来的性能改进-我们通过更简单的定位方法得到了更好的结果，但具有更强大的表示能力。

B 对非常深的特征的概括

在前面的章节中，我们讨论了在 ILSVRC 数据集上训练和评估非常深的 ConvNet。在本节中，我们评估了在 ILSVRC 上预训练的 ConvNet 在其他较小的数据集上作为特征提取器的效果，因为在这些数据集上从头训练大型模型会因过拟合而不可行。最近，对于这样的用例引起了很大的兴趣，因为事实证明，在 ILSVRC 上学习的深度图像表示在其他数据集上具有很好的泛化性能，而且在性能上远远超过手工制作的表示方法。在这个工作的基础上，我们研究了我们的模型是否比在最新方法中使用的更浅的模型具有更好的性能。在这个评估中，我们考虑了在 ILSVRC 上具有最佳分类性能的两个模型（见第 4 节）－配置“Net-D”和“Net-E”（我们已经公开发布）。

表 10：与 ILSVRC 本地化的技术水平进行比较。我们的方法被记号为“VGG”。

Method	top-5 val. error(%)	top-5 test error(%)
VGG	**26.9**	**25.3**
GoogLeNet (Szegedy et al., 2014)	-	26.7
OverFeat (Sermanet et al., 2014)	30.0	29.9
Krizhevsky et al. (Krizhevsky et al., 2012)	-	34.2

为了在其他数据集上利用在 ILSVRC 上预训练的 ConvNet 进行图像分类，我们移除了最后的全连接层（该层执行 1000 类 ILSVRC 分类），并使用倒数第二层的 4096 维激活作为图像特征，这些特征在多个位置和尺度上聚合。得到的图像描述符进行 L2 归一化，并与在目标数据集上训练的线性 SVM 分类器相结合。为简单起见，我们保持预训练 ConvNet 权重不变（不进行微调）。

特征的聚合方式与我们在 ILSVRC 上的评估过程（见第 3.2 节）类似。即首先将图像缩放，使其最小的一边等于 Q，然后密集地应用网络于图像平面上（当所有权重层被视为卷积时，这是可能的）。然后，我们对得到的特征图执行全局平均池化，产生一个 4096 维的图像描述符。该描述符然后与水平翻转图像的描述符进行平均。正如在第 4.2 节中所示，多尺度的评估是有益的，所以我们在多个尺度 Q 上提取特征。由此产生的多尺度特征可以跨尺度进行堆叠或合并。堆叠允许后续的分类器学习如何最佳地组合一系列尺度上的图像统计特征，但这会增加描述符的维度。我们将在下面的实验中回到关于这个设计选择的讨论中。我们还评估了特征的晚期融合，使用两个网络进行计算，这是通过叠加它们各自的图像描述符

来执行的。

表 11：与 VOC-2007、VOC-2012、Caltech-101 和 Caltech-256 图像分类技术的比较。我们的模型被记为“VGG”。标记为*的结果是使用在扩展的 ILSVRC 数据集（2000 个类）上预训练的 ConvNets 获得的。

Method	VOC-2007 (mean AP)	VOC-2012 (mean AP)	Caltech-101 (mean class recall)	Caltech-256 (mean class recall)
Zeiler & Fergus(Zeiler&Fergus,2013)	-	79.0	86.5 ± 0.5	74.2 ± 0.3
Chatfield et al. (Chatfield et al., 2014)	82.4	83.2	88.4 ± 0.6	77.6 ± 0.1
He et al. (He et al., 2014)	82.4	-	**93.4 ± 0.5**	-
Wei et al. (Wei et al., 2014)	81.5(85.2*)	81.7(**90.3***)	-	-
VGGNet-D(16layers)	89.3	89.0	91.8 ± 1.0	85.0 ± 0.2
VGG Net-E(19 layers)	89.3	89.0	92.3 ± 0.5	85.1 ± 0.3
VGG Net-D & Net-E	**89.7**	**89.3**	92.7 ± 0.5	**86.2 ± 0.3**

在 VOC-2007 和 VOC-2012 上的图像分类. 我们首先对 PASCAL VOC-2007 和 VOC-2012 基准测试的图像分类任务进行评估。这些数据集分别包含 10K 和 22.5K 张图像，每张图像都带有一个或多个标签，以及对应 20 个物体类别。VOC 组织者提供了预定义的训练、验证和测试数据集（VOC-2012 的测试数据并不公开可用，而是提供了官方评估服务器）。采用平均精确度（mean average precision,mAP）来衡量识别性能。

值得注意的是，通过研究 VOC-2007 和 VOC-2012 的验证集上的性能，我们发现将在多个尺度上计算的图像描述符通过平均聚合与通过堆叠聚合的性能相似。我们假设这是由于在 VOC 数据集中，对象以多种尺度出现，因此没有特定的尺度特定语义可以被分类器利用。由于平均的好处是不会增加描述符的维度，我们能够在广泛的尺度范围内聚合图像描述符：$Q \in \{256,384,512,640,768\}$。然而，值得注意的是，在较小的范围$\{256, 384, 512\}$上的改进相当小（0.3%）。

我们报告了在测试集上的性能，并与其他方法进行了比较（详见表 Table 11）。我们的网络“Net-D”和“Net-E”在 VOC 数据集上的表现相同，它们的结合略微提高了结果。我们的方法在以 ILSVRC 数据集为基础的图像表示中树立了新的业界标准，超过了以前最佳结果(Chatfield et al., 2014)超过 6%。值得注意的是，Wei et al. (2014)的方法在 VOC-2012 上获得了更好的 mAP，但它是在一个扩展的 2000 类 ILSVRC 数据集上进行预训练的，该数据集包括与 VOC 数据集语义相近的额外 1000 个类别。它还从与目标检测辅助分类流水线的融合中获益。

在 Caltech-101 和 Caltech-256 上的图像分类. 在这一部分，我们在 Caltech-101 和 Caltech-256 图像分类基准测试中评估了非常深的特征。Caltech-101 包含 9000 张图像，分为 102 类（101 个物体类别和 1 个背景类），而 Caltech-256 则更大，有 31000 张图像和 257 类。这些数据集上的标准评估协议是生成几个随机分割成训练和测试数据，并报告分割的平均识别性能，该性能通过平均类别召回率来衡量（这纠正了每类测试图像数量不同的问题）。在 Caltech-101 上，我们遵循了 Chatfield et al. (2014); Zeiler & Fergus (2013); He et al. (2014)的方法，生成了 3 个随机分割成训练和测试数据的数据集，使每个分割包含每类 30 个训练图像，并且每类最多 50 个测试图像。在 Caltech-256 上，我们也生成了 3 个数据集，每个数据集包含每类 60 个训练图像（其余用于测试）。在每个分割中，训练图像的 20%被用作超参数选择的验证集。

我们发现与 VOC 不同的是，在 Caltech 数据集上，通过在多个尺度上计算并堆叠描述符要比平均或最大池化更好。这可以通过 Caltech 图像中的对象通常占据整个图像来解释，因此多尺度图像特征在语义上是不同的（捕捉整个对象 vs.对象的部分），堆叠允许分类器利用这样的尺度特定表示。我们使用了三个尺度 $Q \in \{256,384,512\}$。

我们将我们的模型与彼此和现有技术进行了比较（详见表 Table 11）。如表所示，更深的 19 层 Net-E 的性能优于 16 层 Net-D，它们的组合进一步提高了性能。在 Caltech-101 上，我们的表示与 He et al. (2014)的方法竞争，然而在 VOC-2007 上，它的性能明显较差。在 Caltech-256 数据集上，我们的特征性能远远超过了最先进的方法（Chatfield 等人，2014 年），差距达到了很大的幅度（8.6%）。

在 VOC-2012 上的动作分类任务. 我们还在 PASCAL VOC-2012 的动作分类任务上评估了我们表现最佳的图像表示方法（Net-D 和 Net-E 特征的叠加），该任务的目标是根据一个正

在执行动作的人的边界框，在单张图像上预测其动作类别。该数据集包含 4600 个经过标记的训练图像，分为 11 个类别。与 VOC-2012 对象分类任务类似，性能是通过 mAP 来衡量。我们考虑了两种训练设置： (i) 在整个图像上计算 ConvNet 特征，忽略提供的边界框； (ii) 计算整个图像和所提供的边界框上的特征，并将其进行叠加，得到最终的表示结果。并将结果与表 12 中的其他方法进行了比较。

即使不使用所提供的边界框，我们的表示也达到了 VOC 动作分类任务的最新水平，并且在同时使用图像和边界框时，结果会进一步改进。与其他方法不同，我们没有合并任何特定于任务的启发式方法，而是依赖于非常深的卷积特征的表示能力。

其他识别任务. 自从我们的模型公开发布以来，它们已经被研究界广泛使用于各种图像识别任务，并始终在效果上优于更浅的表征。例如，通过将 Krizhevsky et al. (2012)的 ConvNet 替换为我们的 16 层模型，Girshick et al. (2014) 实现了物体检测结果的最新状态。在语义分割 (Long et al., 2014)、图像标题生成 (Kiros et al., 2014; Karpathy & Fei-Fei, 2014)、纹理和材料识别 (Cimpoi et al., 2014; Bell et al., 2014)方面，相对于 Krizhevsky et al. (2012) 的更浅的架构也观察到了类似的收益。

表 12：与 VOC- 2012 单图像动作分类的技术比较。我们的模型被记为“VGG”。标记为*的结果是通过在扩展的 ILSVRC 数据集（1512 个类）上预训练的 ConvNets 获得的。

Method	VOC-2012 (mean AP)
(Oquab et al., 2014)	70.2*
(Gkioxari et al., 2014)	73.6
(Hoai, 2014)	76.3
VGG Net-D & Net-E, image-only	**79.2**
VGG Net-D & Net-E, image and bounding box	**84.0**

C 论文修订

在这里，我们列出了主要的论文修订列表，以便读者方便地了解重大变化。

v1 初始版本。介绍了在 ILSVRC 提交之前进行的实验。

v2 添加了使用尺度抖动进行训练集扩充的提交后 ILSVRC 实验，这提高了性能。

v3 添加了对 PASCAL VOC 和 Caltech 图像分类数据集的泛化实验（附录 B）。这些实验使用的模型是公开可用的。

v4 将论文转换为 ICLR-2015 提交格式。还添加了多裁剪分类的实验。

v6 定稿版 ICLR-2015 会议论文。添加了与浅层网络 net B 的比较以及 PASCAL VOC 动作分类基准的结果。