非常深的卷积网络用于大规模图像识别

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

ABSTRACT

本文研究了卷积网络深度在大规模图像识别设置中对其精度的影响。本文的主要贡献是使用具有非常小(3 × 3)卷积滤波器的架构对深度不断增加的网络进行了彻底的评估，这表明，通过将深度推至16- 19权重层，可以实现对现有技术配置的显著改进。这些发现是我们ImageNetChallenge 2014年提交的基础，我们的团队在本地化和分类跟踪中获得了第一和第二名分别。所提出表示在其他数据集上的泛化性很好，取得了最先进的结果。公开了两个表现最好的卷积网络模型，以促进在计算机视觉中使用深度视觉表示的进一步研究。

1 简介

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(Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014; Simonyan & Zisserman, 2014) 由于大型公共图像存储库(如ImageNet (Deng et al., 2009))和高性能计算系统(如gpu或大规模分布式集群(Dean et al., 2012))，这已经成为可能。特别是，ImageNet大规模视觉识别挑战(ILSVRC) (Russakovsky et al., 2014)在深度视觉识别架构的进步中发挥了重要作用，作为几代大规模图像分类系统的测试平台，从高维浅层特征编码(Perronnin et al., 2010) (ILSVRC-2011的获胜者) 到deep ConvNets (Krizhevsky et al., 2012) (ILSVRC-2012的获胜者)。

随着卷积网络在计算机视觉领域越来越普及，人们进行了许多尝试，以改进Krizhevsky et al. (2012)的原始架构以实现更好的准确性。例如，提交给ILSVRC-2013 (Zeiler & Fergus, 2013; Sermanet et al., 2014)的表现最佳的提交利用了较小的接受窗口大小和较小的第一个卷积步幅图层。另一项改进涉及在整个图像和多个尺度上密集地训练和测试网络(Sermanet et al., 2014; Howard, 2014)。本文探讨了卷积网络架构设计的另一个重要方面——深度。为此，我们固定了架构的其他参数，并通过添加更多的卷积层来稳步增加网络的深度，这是可行的，因为在所有层中使用非常小的卷积滤波器(3 × 3)。

提出了更准确的卷积网络架构，不仅在ILSVRC分类和定位任务上达到了最先进的精度，而且也适用于其他图像识别数据集，即使在用作相对简单管道的一部分(e.g.深度特征由线性SVM分类，无需微调)。我们发布了两个表现最好的模型[1]，以促进进一步的研究。

论文的其余部分组织如下。在Sect. 2中，我们描述了我们的卷积网络配置。图像分类训练和评估的细节在Sect. 3中给出，在Sect. 4中对ILSVRC分类任务上的配置进行了比较。Sect. 5是论文的结尾。为了完整性，本文还描述和评估了Appendix A中的ILSVRC-2014对象定位系统，并讨论了非常深入的特征的泛化其他数据集在Appendix B。最后，Appendix C包含了主要论文的修订列表。

2 卷积网络配置

为了在公平的环境中测量增加的卷积网络深度带来的改善，我们所有的ConvNet层配置都使用相同的原则设计，灵感来自Ciresan et al. (2011); Krizhevsky et al. (2012)。在本节中，我们首先描述ConvNet配置的通用布局(Sect. 2.1)，然后详细描述评估中使用的特定配置(Sect. 2.2)。然后讨论我们的设计选择，并与现有技术进行比较Sect. 2.3。

*现隶属单位:谷歌DeepMind +目前从属关系:牛津大学和谷歌DeepMind
[1] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

2.1 架构

在训练过程中，我们卷积网络的输入是一个固定大小的224 × 224 RGB图像。我们所做的唯一预处理是从每个像素减去在训练集上计算的平均RGB值。图像通过卷积(conv.)层的堆栈传递，其中我们使用具有非常小的感受野的滤波器:3 × 3(这是捕捉概念的最小尺寸左/右，上/下，中)。在其中一种配置中，我们还使用1 × 1卷积滤波器，它可以被视为输入通道的线性变换(其次是非线性)。卷积步长固定为1像素;卷积层输入的空间填充是这样的，卷积后的空间分辨率得到保留，i.e. 3 × 3卷积层的内边距是1像素。空间池化由5个最大池化层进行，它们遵循一些卷积层(并非所有的卷积层都遵循最大池化)。最大池化在2 × 2像素窗口上执行，步幅2。

卷积层的堆叠(在不同的体系结构中具有不同的深度)之后是三个全连接(FC)层:第一个其中两个每个有4096个通道，第三个执行1000路ILSVRC分类，因此包含1000个通道(每个类一个)。最后一层是soft-max层。全连接层的配置在所有网络中都是相同的。

所有隐藏层都配备了整流(ReLU (Krizhevsky et al., 2012))非线性。我们注意到，我们的网络(除了一个)都不包含局部响应规范化(LRN)规范化(Krizhevsky et al., 2012):如Sect. 4所示，这样的规范化在ILSVRC数据集上并没有提高性能，但会导致内存消耗和计算时间增加。在适用的情况下，LRN层的参数为(Krizhevsky et al., 2012)。

2.2 配置

本文评估的ConvNet配置在Table 1中概述，每列一个。下面我们将以网队的名字来介绍他们(a—e)。所有配置遵循Sect. 2.1中介绍的通用设计，仅在深度上有所不同: 从网络A的11层权值层(8卷积，3 FC层)到网络E的19层权值层(16卷积，3 FC层)。卷积层的宽度(通道的数量)相当小，从第一层的64开始，然后在每个最大池化层后以2的倍数增加，直到到达512。

在Table 2中，我们报告了每个配置的参数数量。尽管深度很大，但我们网络中的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(144米重量(Sermanet et al., 2014))。

2.3 讨论

我们的卷积网络配置与ILSVRC-2012 (Krizhevsky et al., 2012)和ILSVRC-2013竞赛中表现最好的参赛作品中使用的卷积网络配置有很大不同(Zeiler & Fergus, 2013; Sermanet et al., 2014)。与其在第一个卷积层中使用相对较大的感受野(e.g. 11 × 11与stride 4在(Krizhevsky et al., 2012)中，或7 × 7与stride 2在(Zeiler & Fergus, 2013; Sermanet et al., 2014)中)，我们在整个网络中使用非常小的3 × 3感受野，在每个像素与输入进行卷积(步幅1)。很容易看到两个3 × 3 conv.层的堆栈(之间没有空间池化)具有5 × 5的有效感受野; 三个这样的层具有7 × 7有效感受野。那么，通过使用三层3 × 3 conv.层而不是单个7 × 7层，我们得到了什么? 首先，我们合并了三个非线性校正层而不是一个，这使决策函数更具判别力。其次，我们减少参数的数量:假设三层3 × 3卷积堆栈的输入和输出都有C通道，堆栈是参数化的通过$3\left(3^2C^2\right) = 27C^2$权重;同时，一个单一的7×7 conv.层将需要$7^2C^2 = 49C^2$参数，i.e. 81%更多。这可以被视为对7 × 7 conv.滤波器进行正则化，迫使它们通过3 × 3滤波器进行分解(在两者之间注入非线性)。

1 × 1 conv. layers(配置C, Table 1)的合并是一种在不影响的情况下增加决策函数的非线性的方法卷积层的感受野。即使在我们的例子中1 × 1卷积本质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)，由整流函数引入额外的非线性。值得注意的是，1 × 1 conv.层最近被用于Lin et al. (2014)的“网络中的网络”架构。

小型卷积滤波器以前已经被Ciresan et al. (2011)使用，但他们的网络明显不如我们的深度，他们也没有在大规模ILSVRC数据集上进行评估。Goodfellow et al. (2014)将深度卷积网络(11权重层)应用于街道编号识别任务，并表明深度的增加导致了更好的性能。GoogLeNet (Szegedy et al., 2014)是ILSVRC-2014分类任务中表现最好的条目，是独立于我们的工作开发的，但与之相似的是，它基于非常深的卷积网络(22个权重层)和小卷积滤波器(除了3 × 3，它们还使用1 × 1和5 × 5卷积)。然而，他们的网络拓扑比我们的更复杂，并且特征图的空间分辨率在第一层中被更积极地降低，以减少计算量。如Sect. 4.5所示，我们的模型在单网络分类精度方面优于Szegedy et al. (2014)。

Table 1: **ConvNet**配置(如列所示)。配置的深度从左(A)到右(E)增加，随着更多的层被添加(添加的层显示在粗体)。卷积层参数表示为‘conv ⟨感受野大小⟩ - ⟨通道数⟩ ‘。为简洁起见，这里没有给出ReLU激活函数。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight 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					

Table 2: 参数数量(以百万为单位)。

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

3 分类框架

在上一节中，我们介绍了网络配置的细节。在本节中，我们将描述分类卷积网络训练和评估的细节。

3.1 培训

ConvNet训练程序通常遵循Krizhevsky et al. (2012)(除了从多尺度训练图像中采样输入作物，稍后解释)。即，训练是通过使用动量小批量梯度下降(基于反向传播(LeCun et al., 1989))优化多项逻辑回归目标来进行的。批量大小设置为256，动量设置为0.9。训练通过权重衰减(L_2惩罚乘法器设置为$5 \cdot 10^{-4}$)和前两个全连接层的dropout正则化(dropout ratio设置为0.5)进行正则化。学习率最初设置为10^{-2}，然后下降当验证集精度停止提高时，损失了10倍。总的来说，学习率下降了3倍，在370 K次迭代后停止学习(74个时代)。我们推测，尽管与(Krizhevsky et al., 2012)相比，我们的网有更多的参数和更大的深度，但网需要的纪元更少由于(a)由更大的深度和更小的卷积滤波器尺寸施加的隐式正则化;(b)预先初始化某些层。

网络权重的初始化很重要，因为在深度网络中，由于梯度的不稳定性，糟糕的初始化可能会导致学习停滞。为了规避这个问题，我们开始训练配置A (Table 1)，它足够浅，可以用随机初始化进行训练。然后，当训练更深的时候架构方面，我们使用网络A的层初始化前4个卷积层和后3个全连接层(中间层为随机初始化)。我们没有降低预初始化层的学习率，允许它们在学习过程中改变。对于随机初始化(在适用的情况下)，我们从均值为零和方差

为10^{-2}的正态分布中采样权重。偏差初始化为零。值得注意的是，在论文提交后，我们发现可以使用Glorot & Bengio (2010)的随机初始化程序来初始化权重，而无需预训练。

为了获得固定大小的224×224 ConvNet输入图像，它们从重新缩放的训练图像中随机裁剪(每次SGD迭代每张图像一次裁剪)。为了进一步增强训练集，作物经历了随机的水平翻转和随机的RGB颜色偏移(Krizhevsky et al., 2012)。训练图像的缩放解释如下。

训练图像大小。 设S是各向同性缩放的训练图像的最小边，卷积网络输入从该图像中裁剪(我们也将S称为训练尺度)。当作物大小固定为224×224时，原则上S可以取不小于224的任何值:对于$S = 224$，作物将捕获全图像统计，完全跨越训练图像的最小边; 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。

我们考虑两种方法来设置训练规模S。第一个是修复S，它对应于单尺度训练(请注意，采样作物中的图像内容仍然可以表示多尺度图像统计)。在实验中，评估了在两个固定规模下训练的模型:$S = 256$(在现有技术中已广泛使用(Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014))和$S = 384$。给定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$S = 256$预训练的权重进行初始化，并且我们使用较小的初始学习率10^{-3}。

设置S的第二种方法是多尺度训练，每个训练图像都是从一定范围$[S_{min}, S_{max}]$中随机抽样S进行单独缩放(我们使用$S_{min} = 256$和$S_{max} = 512$)。由于图像中的对象可以是不同大小的，因此在训练过程中考虑这一点是有益的。这也可以被视为通过尺度抖动进行训练集增强，其中单个模型被训练以识别大范围内的物体。出于速度的原因，我们通过微调具有相同配置的单尺度模型的所有层来训练多尺度模型，用固定的$S = 384$进行预训练。

3.2 测试

在测试时，给定一个经过训练的卷积网络和一个输入图像，按以下方式分类。首先，它被各向同性地缩放到预定义的最小图像侧，记为Q(我们也称它为测试尺度)。我们注意到Q不一定等于培训规模S(我们将在Sect. 4中显示，为每个S使用多个Q值可以提高性能)。然后，以类似于(Sermanet et al., 2014)的方式将网络密集应用于重新缩放的测试图像。即，首先将全连接层转换为卷积层(第一个FC层转换为7×7卷积层，最后两个FC层转换为1×1卷积层)。然后将得到的全卷积网络应用于整个(未裁剪)图像。结果是一个类分数图，其中通道数量等于类的数量，并具有可变的空间分辨率输入图像的大小。最后，为了获得图像的类分数的固定大小的向量，对类分数图进行空间平均(和池化)。我们还通过水平翻转图像来扩充测试集;对原始图像和翻转图像的soft-max类后验进行平均得到图像的最终得分。

由于全卷积网络应用于整个图像，不需要在测试时对多个作物进行采样(Krizhevsky et al., 2012)，这是低效的，因为它需要对每个作物进行网络重新计算。与此同时，使用大量的农作物(如Szegedy et al. (2014)所做的)可以提高准确性，因为与全卷积网络相比，它可以对输入图像进行更精细的采样。此外，由于不同的卷积边界条件，多作物评估与密集评估是互补的:当将卷积网络应用于作物时，卷积特征图用零填充，而在密集评估的情况下，同一作物的填充自然来自图像的邻近部分(由于卷积和空间池化)，这大大增加了整体网络感受野，因此捕获了更多的上下文。虽然我们认为，在实践中，多种作物增加的计算时间并不能证明在准确性方面的潜在收益，但作为参考，我们还使用每尺度50作物(5×5规则网格与2翻转)来评估我们的网络，在3尺度上总共150作物，这与Szegedy et al. (2014)使用的144作物在4尺度上相当。

3.3 实现细节

我们的实现来自于公开可用的c++ Caffe工具箱(Jia, 2013)(在2013年12月扩展)，但包含一些重大修改，允许我们在单个系统中安装的多个gpu上进行训练和评估，以及在多个尺度的全尺寸(未裁剪)图像上进行训练和评估(如上所述)。多GPU训练利用数据并行性，通过将每批训练图像分割为多个GPU批，在每个GPU上并行处理来实现。计算GPU批处理梯度后，对它们进行平均以获得整个批处理的梯度。梯度计算是跨GPU同步的，因此结果与在单个GPU上训练时完全相同。

虽然最近提出了加快卷积网络训练的更复杂的方法(Krizhevsky, 2014)，这些方法对不同层采用模型和数据并行对于网络，我们已经发现，与使用单个GPU相比，我们在概念上更简单的方案已经在现成的4-GPU系统上提供了3.75倍的加速。在配备了四个NVIDIA Titan Black gpu的系统上，训练单个网络需要2- 3周，具体取决于架构。

4 分类实验

数据集。 在本节中，我们介绍了所述卷积网络架构在ILSVRC-2012数据集(用于ILSVRC 2012- 2014挑战)上取得的图像分类结果。该数据集包括1000类的图像，并分为三个集合:训练(1.3 M个图像)，验证(50 K个图像)和测试(100 K个图像保留类标签)。通过top-1和top-5误差两个指标来评估分类性能。前者是多分类错误，i.e.的比例不正确分类图像;后者是ILSVRC中使用的主要评价标准，计算为基准真实类别在预测前5类之外的图像所占的比例。

对于大多数实验，我们使用验证集作为测试集。在测试集上进行了一定的实验，并提交给了ILSVRC官方服务器作为"VGG"团队参加ILSVRC-2014竞赛(Russakovsky et al., 2014)。

4.1 单尺度评价

首先，用Sect. 2.2中描述的层配置评估单个卷积网络模型的性能。测试图像大小设置如下:$Q = S$用于固定S, $Q = 0.5(S_{min} + S_{max})$用于抖动$S \in [S_{min}, S_{max}]$。的结果显示在Table 3。

首先，我们注意到使用局部响应归一化(A- lrn网络)并没有改善模型A 没有任何规范化层。因此，我们在更深的架构(B- E)中不采用规范化。

其次，我们观察到分类错误随着卷积网络深度的增加而减少:从A的11层到E的19层。值得注意的是，尽管深度相同，配置C(包含三个1×1 conv层)的性能比配置D，在整个网络中使用3×3 conv.层。这表明，虽然额外的非线性确实有帮助(C比B好)，但通过使用具有非平凡感受野的conv.滤波器(D比C好)来捕获空间上下文也很重要。当深度达到19层时，我们架构的错误率饱和，但更深的模型可能对更大的数据集有益。我们还将net B与具有5个5×5 conv.层的浅网进行了比较，后者是通过替换从B中派生出来的每一对3×3 conv.层与一个单独的5×5 conv.层(具有相同的感受野解释在Sect. 2.3)。浅网的top-1误差比B(在一个中心作物上)高7%，这证实了一个小的深网过滤器的性能优于具有较大过滤器的浅网络。

最后，训练时的尺度抖动($S \in [256; 512]$)比在最小边固定的图像($S = 256$或$S = 384$)上训练的结果要好得多，即使在测试时使用单一的量表。这证实了通过尺度抖动增强训练集确实有助于捕获多尺度图像统计。

Table 3: **ConvNet**在单一测试规模下的性能。

ConvNet config. (Table 1)	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	28.7	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**

4.2 多尺度评价

在评估了单一尺度的ConvNet模型后，我们现在评估测试时尺度抖动的影响。它包括在测试图像的几个缩放版本上运行模型(对应于Q的不同值)，然后对得到的类后验进行平均。考虑到训练和测试规模之间的较大差异会导致性能下降，对使用固定S训练的模型进行了评估三个测试图像大小，接近训练图像:$Q = \{S - 32, S, S + 32\}$。同时，训练时的尺度抖动允许网络在测试时应用到更大的尺度范围，从而训练出模型使用变量$S \in [S_{min}; S_{max}]$在更大的尺寸范围内进行评估$Q = \{S_{min}, 0.5(S_{min} + S_{max}), S_{max}\}$。

结果发表在Table 4上，表明测试时的缩放抖动可以带来更好的性能(与评估相比) 相同的模型在单一的比例，显示在Table 3)。和之前一样，最深的配置(D和E)表现最好，缩放抖动比

固定最小边的训练更好S。我们在验证集上的最佳单网络性能是24.8%/7.5% top-1/top-5错误(在Table 4中以粗体突出显示)。在测试集上，配置E达到7.3% top-5错误。

Table 4: **ConvNet**在多个测试尺度下的性能。

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	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 多作物评价

在Table 5中，我们将密集卷积网络评估与多种作物评估进行了比较(详细信息请参见Sect. 3.2)。通过对两种评估技术的软最大输出进行平均，评估了它们的互补性。可以看出，使用多种作物的表现略好于密集评估，这两种方法确实是互补的，因为它们的组合优于它们中的任何一种。如上所述，我们假设这是由于对卷积边界条件的不同处理。

Table 5: 卷积网络评估技术比较。 在所有实验中，训练量表S从$[256; 512]$中采样，并考虑三个测试量表Q: $\{256, 384, 512\}$。

ConvNet config. (Table 1)	Evaluation method	top-1 val. error (%)	top-5 val. error (%)
D	dense	24.8	7.5
	multi-crop	24.6	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，我们评估了单个卷积网络模型的性能。在这部分实验中，我们通过对多个模型的soft-max类后验进行平均来组合它们的输出。由于模型的互补性，这提高了性能，并在2012年的顶级ILSVRC提交中使用(Krizhevsky et al., 2012)和2013年(Zeiler & Fergus, 2013; Sermanet et al., 2014)。

结果显示在Table 6。在ILSVRC提交时，我们只训练了单尺度网络和多尺度模型D(通过只微调全连接层而不是所有层)。得到的7个网络集成的ILSVRC测试误差为7.3%。在提交之后，我们只考虑了两个表现最好的多尺度模型(配置D和E)的集成，这将测试误差降低到7.0%采用密集评价，6.8%采用密集和多作物联合评价。作为参考，我们表现最好的单个模型实现了7.1%误差(模型E, Table 5)。

Table 6: 多个**ConvNet**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**

4.5 与最新技术的比较

最后，我们将我们的结果与Table 7中的最新技术进行了比较。在ILSVRC-2014挑战赛(Russakovsky et al., 2014)的分类任务中，我们的"VGG"团队使用集成学习获得了第二名，测试错误率为7.3% 7种型号。在提交后，我们使用两个模型的集成将错误率降低到6.8%。

从Table 7可以看出，我们的very deep ConvNets的性能明显优于上一代模型，取得了最好的结果参加ILSVRC-2012和ILSVRC-2013竞赛。与分类任务获胜者(GoogLeNet, 6.7%错误)相比，该结果也具有竞争力优于ILSVRC-2013获奖作品Clarifai, Clarifai在外部训练数据下实现了11.2%，在没有外部训练数据的情况下实现了11.7%。这是值得注意的，考虑到我们的最佳结果是通过组合两个模型实现的——比大多数ILSVRC提交中使用的要少得多。就单网络性能而言，我们的架构实现了最佳结果(7.0%测试错误)，性能超过了单个GoogLeNet 0.9%。值得注意的是，我们没有离开LeCun et al. (1989)的经典卷积网络架构，而是通过大幅增加深度来改进它。

Table 7: 与现有的**ILSVRC**分类方法进行了比较。我们的方法记为"VGG "。只报告在没有外部训练数据的情况下获得的结果。

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.)	24.4	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 (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	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 (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 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，我们评估了用于大规模图像分类的非常深的卷积网络(多达19个权重层)。实验结果表明，表示深度有利于分类精度使用传统的卷积网络架构(LeCun et al., 1989; Krizhevsky et al., 2012)可以实现在ImageNet challenge数据集上的最先进性能深度大幅增加。在附录中，所提出模型在广泛的任务和数据集上都有很好的泛化能力，匹配或超过了围绕较少深度图像表示建立的更复杂的识别管道。结果再次证实了深度在视觉表示中的重要性。

致谢

这项工作得到了ERC的资助。228180. 我们感谢NVIDIA公司捐赠用于本研究的gpu的支持。

REFERENCES

Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. *CoRR*, abs/1412.0623, 2014.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC.*, 2014.

Cimpoi, M., Maji, S., and Vedaldi, A. Deep convolutional filter banks for texture recognition and segmentation. *CoRR*, abs/1411.6836, 2014.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In *IJCAI*, pp. 1237–1242, 2011.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., and Ng, A. Y. Large scale distributed deep networks. In *NIPS*, pp. 1232–1240, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. The Pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop of Generative Model Based Vision*, 2004.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524v5, 2014. Published in Proc. CVPR, 2014.

Gkioxari, G., Girshick, R., and Malik, J. Actions and attributes from wholes and parts. *CoRR*, abs/1412.2604, 2014.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proc. AISTATS*, volume 9, pp. 249–256, 2010.

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In *Proc. ICLR*, 2014.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

He, K., Zhang, X., Ren, S., and Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729v2, 2014.

Hoai, M. Regularized max pooling for image categorization. In *Proc. BMVC.*, 2014.

Howard, A. G. Some improvements on deep convolutional neural network based image classification. In *Proc. ICLR*, 2014.

Jia, Y. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Lin, M., Chen, Q., and Yan, S. Network in network. In *Proc. ICLR*, 2014.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. CVPR*, 2014.

Perronnin, F., Sánchez, J., and Mensink, T. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *CoRR*, abs/1403.6382, 2014.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc. ICLR*, 2014.

Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014. Published in Proc. NIPS, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S. CNN: Single-label to multi-label. *CoRR*, abs/1406.5726, 2014.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. Published in Proc. ECCV, 2014.

A 本地化

在本文的主体部分，我们考虑了ILSVRC挑战的分类任务，并对不同深度的卷积网络架构进行了彻底的评估。在本节中，我们转向挑战的本地化任务，我们在2014年以25.3%错误赢得了该挑战。它可以被视为目标检测的一个特殊情况，其中应该为前5个类别中的每个类别预测单个目标边界框，而不管该类的实际目标数量。为此，我们采用了Sermanet et al. (2014)的方法，ILSVRC-2013本地化挑战的获胜者，并进行了一些修改。我们的方法在Sect. A.1上进行了描述，并在Sect. A.2上进行了评估。

A.1 定位卷积网络

为了进行对象定位，我们使用一个非常深的卷积网络，其中最后一个全连接层预测边界框位置而不是类分数。边界框由存储其中心坐标、宽度和高度的4-D向量表示。可以选择边界框预测是否在所有类别中共享(单类回归，SCR (Sermanet et al., 2014))或特定于类(逐类回归，PCR)。在前一种情况下，最后一层是4-D，而在后一种情况下是4000-D(因为数据集中有1000个类)。除了最后一个边界框预测层，我们使用卷积网络结构D (Table 1)，它包含16个权重层，在分类任务中被发现表现最好(Sect. 4)。

培训。 本地化卷积网络的训练类似于分类卷积网络(Sect. 3.1)。主要区别在于，我们将逻辑回归目标替换为欧氏损失，该损失惩罚了预测边界框参数与真实值的偏差。我们训练了两个本地化模型，每个模型都是单一规模的:$S = 256$和$S = 384$(由于时间限制，我们没有在ILSVRC-2014提交中使用训练规模抖动)。训练用相应的分类模型初始化(在相同的规模上训练)，初始学习率设置为10^{-3}。我们探索了微调所有层和只微调前两个全连接层，如(Sermanet et al., 2014)所做的。最后一个全连接层为随机初始化并从头开始训练。

测试。 我们考虑两种测试方案。第一种用于比较验证集上的不同网络修改，只考虑地面真值类的边界框预测(以排除分类错误)。边界框是通过仅将网络应用于图像的中心裁剪来获得的。

第二个，完全成熟的测试过程是基于对整个图像的本地化卷积网络的密集应用，类似于分类任务(Sect. 3.2)。不同的是，最后一个全连接层的输出是一组预测的边界框，而不是类分数图。为了得出最终的预测，我们使用Sermanet et al. (2014)的贪婪合并过程，它首先合并空间上接近的预测(通过对它们的坐标进行平均)，然后对它们进行评级基于从分类卷积网络获得的类分数。当使用几个本地化卷积网络时，我们首先获取它们的边界框预测集的并集，然后在该并集上运行合并过程。我们没有使用Sermanet et al. (2014)的多重池化偏移技术，该技术增加了边界框预测的空间分辨率并可以进一步提高结果。

A.2 定位实验

在本节中，我们首先确定最佳的本地化设置(使用第一个测试协议)，然后在一个完全成熟的场景中评估它(第二种协议)。定位误差根据ILSVRC准则测量(Russakovsky et al., 2014), i.e.如果边界框预测与真实值的交并比，则认为是正确的边界框在0.5上方。

设置比较。 从Table 8可以看出，每类回归(PCR)优于类无关的单类回归(SCR)，这与Sermanet et al. (2014)的发现不同，PCR的表现优于SCR。我们还注意到，微调本地化任务的所有层比只微调全连接层的效果明显更好(如(Sermanet et al., 2014)所示)。在这些实验中，最小图像侧设置为$S = 384$;$S = 256$的结果显示了相同的行为，但为了简洁起见没有显示。

全面评估。 在确定了最佳的定位设置(PCR，所有层的微调)后，我们现在将其应用于全面的场景中，在这里，使用我们性能最好的分类系统(Sect. 4.5)预测top-5类标签，并使用合并多个密集计算的边界框预测Sermanet et al. (2014)的方法。从Table 9可以看出，与使用中心裁剪相比，将本地化卷积网络应用于整个图像大大提高了结果(Table 8)，尽管使用了预测的前5个类别标签，而不是基本事实。类似于分类任务(Sect. 4)，在多个规模上进行测试，并结合多个网络的预测进一步提高了性能。

Table 8: 不同修改的本地化错误 简化的测试协议:从单个中心图像裁剪中预测边界框，并使用ground truth类。所有ConvNet层(除了最后一个)都具有配置D (Table 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
	PCR	34.3
all	PCR	**33.1**

Table 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	26.7
384	352,384	27.5	-
fusion: 256/256 and 384/352,384		**26.9**	**25.3**

与最新技术的比较。 我们将我们最好的本地化结果与最先进的Table 10进行比较。凭借25.3%的测试误差，我们的"VGG"团队赢得了ILSVRC-2014 (Russakovsky et al., 2014)的本地化挑战。值得注意的是，我们的结果大大优于ILSVRC-2013冠军Overfeat (Sermanet et al., 2014)，尽管我们使用了更少的规模，并且没有使用分辨率增强技术。我们设想，如果将这种技术纳入我们的方法中，可以实现更好的本地化性能。这表明我们非常深的卷积网络带来的性能进步——我们用更简单的本地化方法获得了更好的结果，但更强大的表示。

Table 10: 与现有的**ILSVRC**本地化技术进行了比较。我们的方法记为"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

B 非常深层特征的泛化

在前面的小节中，我们讨论了在ILSVRC数据集上训练和评估非常深的卷积网络。在本节中，我们评估了在ILSVRC上预训练的卷积网络，将其作为其他较小数据集的特征提取器，在这些数据集上，由于过度拟合，从头开始训练大型模型是不可行的。最近，人们对这样的用例产生了很大的兴趣(Zeiler & Fergus, 2013; Donahue et al., 2013; Razavian et al., 2014; Chatfield et al., 2014)，因为事实证明，在ILSVRC上学习的深度图像表示，可以很好地泛化到其他数据集，在这些数据集上，它们的表现大大超过了手工制作的表示。研究了所提出模型是否比最先进方法中使用的更浅层模型有更好的性能。在此评估中，我们考虑了两个在ILSVRC (Sect. 4)上具有最佳分类性能的模型——配置'Net-D '和'Net-E '(我们公开提供)。

为了利用在ILSVRC上预训练的卷积网络在其他数据集上进行图像分类，删除了最后一个全连接层(执行1000路ILSVRC分类)，并使用倒数第二层的4096-D激活作为图像特征，这些特征在多个位置和尺度上聚合。得到的图像描述符是L_2归一化的，并与线性SVM分类器结合，在目标数据集上进行训练。为简单起见，预训练的卷积网络权重保持固定(不进行微调)。

特征聚合的方式与我们的ILSVRC评估过程类似(Sect. 3.2)。即，首先重新缩放图像，使其最小边等于Q，然后在图像平面上密集应用网络(当所有权重层都被视为卷积层时，这是可能的)。然后，我们对生成的特征图进行全局平均池化，生成一个4096维的图像描述符。然后将该描述子与水平翻转图像的描述子进行平均。如Sect. 4.2所示，在多个尺度上进行评估是有益的，因此我们在多个尺度上提取特征Q。由此产生的多尺度特征可以跨尺度堆叠或合并。堆叠允许后续的分类器学习如何在一定范围内优化组合图像统计信息;然而，这是以增加描述子维度为代价的。我们将在下面的实验中回到对这种设计选择的讨论。还评估了使用两个网络计算的特征后期融合，这是通过堆叠各自的图像描述符来执行的。

基于**VOC-2007**和**VOC-2012**的图像分类。 首先对PASCAL VOC-2007和VOC-2012基准(Everingham et al., 2015)的图像分类任务进行了评估。这些数据集分别包含10K和22.5K图

Table 11: 在**VOC-2007**、**VOC-2012**、**Caltech-101**和**Caltech-256**数据集上与现有图像分类方法进行对比。我们的模型被表示为"VGG "。标记为*的结果是使用ConvNets在扩展的ILSVRC数据集(2000类)上进行预训练得到的。

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	-	$\mathbf{93.4 \pm 0.5}$	-
Wei et al. (Wei et al., 2014)	81.5 (85.2*)	81.7 (**90.3***)	-	-
VGG Net-D (16 layers)	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	$\mathbf{86.2 \pm 0.3}$

像，每个图像都标注了一个或多个标签，对应20个对象类别。VOC组织者提供预定义的训练、验证和测试数据(VOC-2012的测试数据不公开;相反，它提供了一个官方的评估服务器)。识别性能使用跨类别的平均精度均值(mAP)来衡量。

值得注意的是，通过检查VOC-2007和VOC-2012验证集上的性能，我们发现在多个尺度上计算的聚合图像描述符，通过平均的方式与通过堆叠的聚合方式的性能类似。我们假设这是由于在VOC数据集中，对象出现在各种尺度上，因此没有分类器可以利用的特定尺度的语义。由于平均具有不扩大描述符维度的优点，我们能够在广泛的尺度上聚合图像描述符:$Q \in \{256, 384, 512, 640, 768\}$。值得注意的是，与较小范围的$\{256, 384, 512\}$相比，改进相当有限(0.3%)。

在Table 11上报告了测试集的性能，并与其他方法进行了比较。所提出的网络'Net-D '和'Net-E '在VOC数据集上表现出相同的性能，它们的组合略微提高了结果。所提出方法在ILSVRC数据集上进行了预训练，在图像表示方面达到了新的技术水平，比Chatfield et al. (2014)之前的最佳结果高出6%以上。需要注意的是，Wei et al. (2014)的方法在扩展的2000类ILSVRC数据集上进行了预训练，该数据集包括额外的1000个类别，在语义上接近VOC数据集中的类别，在VOC-2012上实现了1% better mAP。它还受益于与目标检测辅助分类管道的融合。

基于**Caltech-101**和**Caltech-256**的图像分类。 在本节中，我们在Caltech-101(Fei-Fei et al., 2004)和Caltech-256(Griffin et al., 2007)图像分类基准上评估非常深入的特征。Caltech-101包含9K张图像，分为102个类(101个目标类和一个背景类)，而Caltech-256更大，有31K张图像和257个类。对于这些数据集，一个标准的评估方案是随机地将数据划分为训练数据和测试数据，并报告这些划分的平均识别性能，这是通过平均类召回率(它补偿了每个类不同数量的测试图像)来衡量的。在Chatfield et al. (2014); Zeiler & Fergus (2013); He et al. (2014)之后，在Caltech-101上，我们生成了训练数据和测试数据的3个随机分割，这样每个分割每个类包含30个训练图像，每个类最多包含50个测试图像。在Caltech-256上，我们还生成了3个切分，每个类包含60个训练图像(其余用于测试)。在每次划分中，20%的训练图像被用作超参数选择的验证集。

我们发现，与VOC不同，在Caltech数据集上，在多个尺度上计算的描述符堆叠表现优于平均或最大池化。这可以解释为，在加州理工学院图像中，对象通常占据整个图像，因此多尺度图像特征在语义上是不同的(捕获整个对象*vs*.对象部分)，堆叠允许分类器利用这种特定尺度的表示。我们使用了三种量表$Q \in \{256, 384, 512\}$。

我们的模型相互比较，并在Table 11上进行了技术现状的比较。可以看出，更深的19层Net-E比16层Net-D性能更好，两者的结合进一步提高了性能。在Caltech-101上，所提出的表示方法与He et al. (2014)方法具有竞争力，然而，该方法在VOC-2007上的表现明显低于所提出的网络。在Caltech-256上，我们的特征比最先进的特征(Chatfield et al., 2014)有很大的优势(8.6%)。

基于**VOC-2012**的行为分类。 我们还评估了在PASCAL VOC-2012动作分类任务(Everingham et al., 2015)上表现最佳的图像表示(Net-D和Net-E特征的叠加)，该任务包括在给定执行动作的人的边界框的情况下，从单个图像预测动作类。该数据集包含4.6K训练图像，被标记为11个类别。与VOC-2012目标分类任务类似，使用mAP来测量性能。我们考虑了两种训练设置: (i)计算整个图像上的卷积网络特征，忽略提供的边界框; (ii)计算整个图像上的特征和提供的边界框上的特征，并堆叠它们以获得最终表示。结果与Table 12上的其他方法进行了比较。

Table 12: 在**VOC-2012**的单图像行为分类中与现有方法进行对比。我们的模型被表示为"VGG "。标记为*的结果是使用ConvNets在扩展的ILSVRC数据集(1512类)上进行预训练得到的。

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**

即使不使用提供的边界框，所提出的表示也在VOC动作分类任务上达到了最先进的水平，并且在同时使用图像和边界框时，结果得到了进一步的改善。与其他方法不同，我们没有纳入任何特定任务的启发式方法，而是依赖于非常深的卷积特征的表示能力。

其他识别任务。 自该模型公开发布以来，它们已被研究界积极用于广泛的图像识别任务，始终优于更浅层的表示。例如，Girshick et al. (2014)通过用我们的16层模型替换Krizhevsky et al. (2012)的卷积网络来实现目标检测结果的状态。在语义分割(Long et al., 2014)、图像标题生成(Kiros et al., 2014; Karpathy & Fei-Fei, 2014)、纹理和材质识别(Cimpoi et al., 2014; Bell et al., 2014)等方面也观察到了与Krizhevsky et al. (2012)更浅层架构相比的类似增益。

C 论文修改

在这里，我们列出了主要的论文修订，概述了重大变化，以方便读者。

v1初始版本。介绍了在ILSVRC提交之前进行的实验。

v2增加了提交后的ILSVRC实验，使用尺度抖动增强训练集，提高了性能。

v3在PASCAL VOC和Caltech图像分类数据集上增加了泛化实验(Appendix B)。用于这些实验的模型是公开可用的。

论文已转换为ICLR-2015提交格式。还增加了对多种作物进行分类的实验。

v6相机准备ICLR-2015会议论文。添加了网络B与浅层网络的比较以及PASCAL VOC动作分类基准的结果。