# 用于大规模图像识别的非常深入的卷积网络

## 摘要

本文研究了深度对卷积网络在大规模图像识别中准确率的影响。本文的主要贡献是，对使用很小(3×3)的卷积[滤波器](https://m.baidu.com/s?word=%E6%BB%A4%E6%B3%A2%E5%99%A8&" \t "/Users/hjl\\x/_blank)来增加深度的网络进行了一个全面的评估，表明了通过将深度提高到16 - 19个权重层，业界最好网络的性能也能够得到显著的提升。这些发现是我们参加ImageNet2014挑战赛的基础，我们也因此在定位和分类跟踪任务上分别获得了第一名和第二名的成绩。本文表明了我们的模型在其他数据集上同样表现得很好，并都达到了当前最佳的水平。我们已经公布了两种性能最好的卷积网络模型，希望能促进将深度视觉表达应用于计算机视觉的进一步研究。

## 1. 介绍

卷积网络(ConvNets)最近在大规模的图像和视频识别中获得了很大的成功，这可能得益于大型公共图像库，如 ImageNet，以及高性能计算系统，如GPU或大规模分布式集群。特别是在深度[视觉识别](https://m.baidu.com/s?word=%E8%A7%86%E8%A7%89%E8%AF%86%E5%88%AB&" \t "/Users/hjl\\x/_blank)结构发展中扮演重要角色的ImageNet大规模[视觉识别](https://m.baidu.com/s?word=%E8%A7%86%E8%A7%89%E8%AF%86%E5%88%AB&" \t "/Users/hjl\\x/_blank)大赛(ILSVRC)，从高维浅层特征编码 (ILSVRC-2011冠军)到深度卷积网络(ILSVRC-2012 冠军)，它为几代大规模图像分类系统提供了测试平台。

随着卷积网络在计算机视觉领域的应用越来越广泛，为了获得更高的准确率，越来越多的人尝试在[krizhevsky2012imagenet](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "/Users/hjl\\x/_blank)中的原始框架上进行改进。例如，ILSVRC-2013最好的参赛模型在第一个卷积层上使用了较小的接受域窗口以及较小的滑动步长。另一种改进方案是在整幅图像以及它的多个尺寸上，稠密的训练和测试网络。本文中，我们关注了卷积网络结构设计中的另一个重要因素——深度。为此，我们固定了网络框架的其他参数，然后通过增加更多的卷积层来增加网络的深度，这是可行的，因为我们在所有层都是用非常小(3×3)的卷积[滤波器](https://m.baidu.com/s?word=%E6%BB%A4%E6%B3%A2%E5%99%A8&" \t "/Users/hjl\\x/_blank)。

因此，我们提出了更加精确的卷积网络框架，它不仅在ILSVRC分类和定位任务上获得了最好的成绩，同时在其他图像识别数据集上也表现出了很优异的性能，即便只是用来作为一个相对简单框架的一部分(例如，无需微调的线性SVM深度特征分类)。我们公布了两个性能最好的[模型](http://www.robots.ox.ac.uk/~vgg/research/very_deep/" \t "/Users/hjl\\x/_blank)来促进进一步的研究。

本文其余部分组织如下。在Sect.2 中，描述了我们的卷积网络框架。Sect.3 介绍了图像分类任务的训练和评估的主要细节。Sect.4 在ILSVRC分类任务上对不同框架进行了对比。Sect.5 对本文进行了总结。为了文章的完整性，我们在Appendix A 中描述了我们的ILSVRC-2014目标定位系统，并在Appendix B 中讨论了非常深的特征对于其他数据集的泛化能力。最后，Appendix C 包含了本文的主要修订记录。

1. 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层配置都采用相同的原理设计，灵感来自Ciresan等人。 （2011）; Krizhevsky等人。 （2012年）。 在本节中，我们首先描述ConvNet配置的通用布局（第2.1节），然后详细介绍评估中使用的特定配置（第2.2节）。 然后讨论我们的设计选择，并与Sect中的现有技术进行比较。 2.3。

* 1. 建筑

在训练期间，我们的ConvNets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平均RGB值。 图像通过一堆卷积（转换）图层，我们使用具有非常小的感受野的滤镜：3×3（这是捕捉左/右，上/下概念的最小尺寸，

中央）。 在其中一种配置中，我们还使用了1×1卷积滤波器，可以看作是

输入通道的线性变换（后跟非线性）。 卷积的步伐是

固定为1个像素; 转换的空间填充。 层输入使得在卷积之后保留空间分辨率，即，对于3×3转换，填充是1个像素。 层。 空间池由五个最大池组执行，这些层跟随一些转换。 图层（并非所有转换图层都跟随最大池）。 最大池化在2×2像素窗口上执行，步幅为2。

一堆卷积层（在不同的架构中具有不同的深度）之后是三个完全连接（FC）层：前两个有4096个通道，第三个执行1000路ILSVRC分类，因此包含1000个通道（一个为每个班级）。 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。

所有隐藏层都配备有整流（ReLU（Krizhevsky等，2012））非线性。 我们注意到，我们的网络（除了一个）都没有包含本地响应规范化（LRN）规范化（Krizhevsky等，2012）：如将在Sect中所示。 如图4所示，这种归一化不会改善ILSVRC数据集的性能，但会导致内存消耗和计算时间增加。 在适用的情况下，LRN层的参数是（Krizhevsky等，2012）的参数。

* 1. 配置

本文评估的ConvNet配置在表1中列出，每列一个。 在下文中，我们将通过名称（A-E）来引用网络。 所有配置均遵循Sect中提出的通用设计。 2.1，并且仅在深度上有所不同：从网络A中的11个权重层（8个转换层和3个FC层）到网络E中的19个权重层（16个转换层和3个FC层）。 转的宽度。 层（通道数）相当小，从第一层中的64开始，然后在每个最大池层之后增加2倍，直到达到512。

在表2中，我们报告了每个配置的参数数量。 尽管深度很大，但我们网中的权重数量不大于更大转数的更浅网中的权重数。 层宽和感受野（144M权重（Sermanet等，2014））。

* 1. 讨论

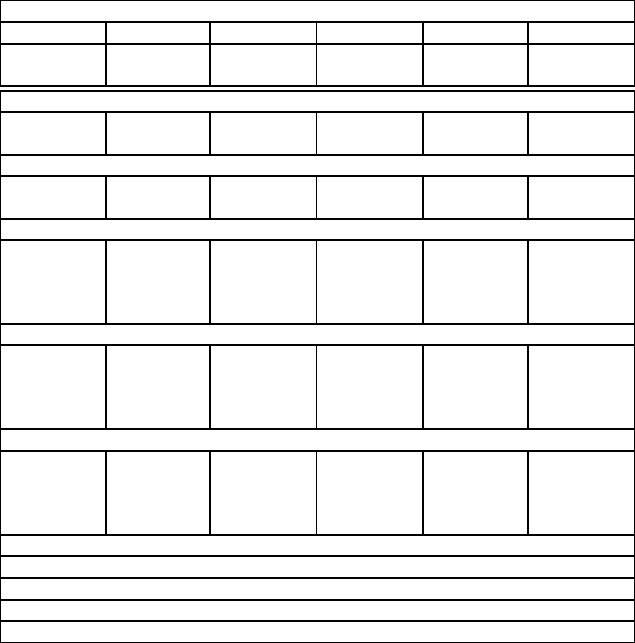
我们的ConvNet配置与ILSVRC-2012（Krizhevsky等，2012）和ILSVRC-2013竞赛（Zeiler＆Fergus，2013; Sermanet等，2014）的最佳表现中使用的配置截然不同。 而不是在第一个转义中使用相对较大的感受域。 层（例如11×11，步幅4（Krizhevsky等，2012），或7×7，步幅2（Zeiler＆Fergus，

2013; Sermanet et al。，2014）），我们在整个网络中使用非常小的3×3感受野，

它们与每个像素的输入进行卷积（使用步幅1）。 很容易看出一堆两个

1. ×3转。 层（其间没有空间池）具有5×5的有效感受野; 三

表1：ConvNet配置（显示在列中）。 随着更多层的添加（添加的层以粗体显示），配置的深度从左（A）增加到右（E）。 卷积层参数表示为“conv（接收场大小） - （信道数）”。 为简洁起见，未显示ReLU激活功能。



ConvNet配置

A A-LRN B C D E

11权重 11重量 13重量 16重量 16重量 19个重量层 层 层 层 层 图层输入（224×224 RGB图像）

conv3-64 conv3-64 conv3-64 conv3-64 conv3-64 conv3-64

**LRN 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 conv3-256 conv3-256 conv3-256 conv3-256

**conv1-256 conv3-256** conv3-256

**conv3-256**

maxpool

conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512

**conv1-512 conv3-512** conv3-512

**conv3-512**

maxpool

conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512 conv3-512

**conv1-512 conv3-512** conv3-512

**conv3-512**

maxpool FC-4096 FC-4096 FC-1000

软-MAX

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| A | B | C | D | E | F |
| 11 | 11 | 13 | 16 | 16 | 19 |
| input (224 × 224 RGB image) | | | | | |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

表2：参数数量（以百万计）。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 网络 | A，A-LRN | B | C | D | E |
| 参数数量 | 133 | 133 | 134 | 138 | 144 |

这些层具有7×7有效的感受野。 那么我们通过使用例如三个3×3转换的堆栈获得了什么。 层而不是单个7×7层？ 首先，我们合并了三个非线性整流层而不是一个非线性整流层，这使决策功能更具辨识力。

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，堆栈参数化为3(32C2= 27C 2

权重; 同时，单个7×7转。 层将需要72C2= 49C2参数，即

多81％。 这可以看作是对7×7转换进行正则化。 过滤器，迫使它们通过3×3滤波器进行分解（其间注入非线性）。

加入1×1转换。 层（配置C，表1）是一种增加决策功能的非线性而不影响转换器的接收场的方法。 层。 即使在我们的情况下，1×1卷积基本上是在相同维度的空间上的线性投影（输入和输出通道的数量相同），但是整流函数引入了额外的非线性。 应该注意1×1转。 最近在Lin等人的“Network in Network”架构中使用了这些层。 （2014）。

Ciresan等人先前已使用小尺寸卷积滤波器。 （2011年），但他们的网络明显不如我们的网络，他们没有评估大规模的ILSVRC数据集。 Goodfellow等人。 （2014）将深度ConvNets（11个重量层）应用于街道号码识别任务，并表明增加的深度导致更好的性能。 GoogLeNet（Szegedy等，2014），ILSVRC-2014分类任务的最佳表现，是独立于我们的工作而开发的，但类似之处在于它基于非常深的ConvNets

（22个重量层）和小卷积滤波器（除了3×3，它们也使用1×1和5×5

卷积）。 然而，它们的网络拓扑结构比我们的网络拓扑更复杂，并且空间分辨率更高。

在第一层中更加积极地减少特征图的lution以减少计算量。 正如将在Sect。中所示。 4.5，我们的模型优于Szegedy等人的模型。 （2014）在单网络分类准确性方面。

1. 分类框架

在上一节中，我们介绍了网络配置的详细信息。 在本节中，我们将描述ConvNet分类培训和评估的详细信息。

* 1. 训练

ConvNet培训程序通常遵循Krizhevsky等人的观点。 （2012）（除了从多尺度训练图像中对输入作物进行采样，如后面所述）。 即，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））优化多项逻辑回归目标来执行训练。 批量大小设置为256，动量为0.9。 训练通过体重衰减（L2罚分乘数设定为）进行正规化

5·10−4）和前两个完全连接层的压差正规化（压差比设置为0.5）。

学习率初始设置为10−2，然后当验证设置精度停止改善时降低10倍。 总的来说，学习率下降了3倍，并且学习

在370K迭代（74个时期）后停止。 我们推测，尽管与（Krizhevsky等人，2012）相比，参数数量更多，网络深度更大，但由于（a）由更大的深度和更小的转换强加的隐式正则化，网络需要更少的时期收敛. . 过滤器尺寸; （b）某些层的预初始化。

网络权重的初始化很重要，因为由于深网中梯度的不稳定性，不良的初始化会使学习失速。 为了避免这个问题，我们开始训练配置A（表1），其足够浅以便随机初始化进行训练。 然后，当训练更深层的体系结构时，我们初始化前四个卷积层，最后三个完全连接的层用网A初始化（中间层随机初始化）。 我们没有降低预初始化层的学习率，允许它们在学习期间改变。

对于随机初始化（如果适用），我们从具有零均值和10−2方差的正态分布中对权重进行采样。 偏差初始化为零。 值得注意的是，在提交论文后，我们发现通过使用Glorot＆Bengio（2010）的随机初始化程序，可以在没有预训练的情况下初始化权重。

为了获得固定大小的224×224 ConvNet输入图像，从重新调整的训练图像中随机裁剪它们（每个SGD迭代每个图像一个裁剪）。 为了进一步增加训练集，作物经历了随机水平翻转和随机RGB色移（Krizhevsky等，2012）。

下面解释训练图像重新缩放。

**训**练图像大小。 设S是各向同性重新缩放的训练图像的最小边，从中裁剪ConvNet输入（我们也将S称为训练比例）。 虽然作物大小固定为224×224，但原则上S可以采用不小于224的任何值：对于S = 224，作物将捕获整个图像统计数据，完全跨越训练图像的最小边; 对于

S»224裁剪将对应于图像的一小部分，包含一个小物体或一个物体部分。

我们考虑设置训练量表S的两种方法。第一种是修正S，这对应于单一规模的训练（注意，采样作物中的图像内容仍然可以代表多尺度图像统计）。 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（已在现有技术中广泛使用（Krizhevsky等，2012; Zeiler＆Fergus，2013; Sermanet等，2014））和S在给定ConvNet配置的情况下，我们首先使用S = 256训练网络。为了加速S = 384网络的训练，使用权重对其进行初始化

预训练S = 256，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，其中每个训练图像通过从特定范围[S*分*，S*最大*]（我们使用S*分*= 256和S（S）随机采样S来单独重新调整。 TF92）= 512）。 由于图像中的物体可以具有不同的尺寸，因此在训练期间考虑这一点是有益的。 这也可以看作是通过尺度抖动的训练集增强，其中单个

训练模型以识别各种尺度的物体。 出于速度原因，我们通过使用相同的配置微调单级模型的所有层来训练多尺度模型，使用固定的S = 384进行预训练。

* 1. 测试

在测试时，给定经过训练的ConvNet和输入图像，按以下方式对其进行分类。 首先，它被各向同性地重新缩放到预定义的最小图像侧，表示为Q（我们也将其称为测试标度）。 我们注意到Q不一定等于训练量S（正如我们将在第4节中所示，每个S使用几个Q值可以提高性能）。 然后，网络

以类似于（Sermanet等，2014）的方式密集地施加在重新缩放的测试图像上。 即，首先将完全连接的层转换为卷积层（第一FC层到7×7转换层，最后两个FC层转换为1×1转换层）。 然后将得到的完全卷积网应用于整个（未剪切的）图像。 结果是一个类别得分图，其数量为

通道等于类的数量，以及可变的空间分辨率，取决于输入的图像大小。 最后，为了获得图像的类别得分的固定大小矢量，对类别得分图进行空间平均（求和）。 我们还通过水平翻转图像来增加测试集; 将原始图像和翻转图像的soft-max类后验平均化以获得图像的最终分数。

由于完全卷积网络应用于整个图像，因此无需在测试时对多个作物进行采样（Krizhevsky等，2012），因为它需要对每种作物进行网络重新计算，效率较低。 与此同时，使用大量作物，如Szegedy等人所做的那样。 （2014），可以提高准确度，因为与完全卷积网相比，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，多作物评估是对密集评估的补充：当将ConvNet应用于作物时，卷积的特征图用零填充，而在密集评估的情况下，同一作物的填充自然来临从图像的相邻部分（由于卷积和空间池），这大大增加了整个网络接收场，因此捕获了更多的上下文。 虽然我们认为实际上多种作物的计算时间增加了

不能证明准确性的潜在收益，参考我们还评估我们的网络，每个规模使用50个作物（5×5规则网格，2个翻转），总共150个作物超过3个等级，相当于144个作物超过4个等级由Szegedy等人使用。 （2014）。

* 1. 实施细节

我们的实现源自公开的C ++ Caffe工具箱（Jia，2013）（2013年12月扩展），但包含许多重要的修改，允许我们对安装在单个系统中的多个GPU进行培训和评估，以及作为训练和评估多尺度的全尺寸（未剪切）图像（如上所述）。 多GPU训练利用数据并行性，并通过将每批训练图像分成几个GPU批次来执行，并在每个GPU上并行处理。 在计算GPU批量梯度之后，对它们进行平均以获得完整批次的梯度。 梯度计算在GPU之间是同步的，因此结果与在单个GPU上进行训练时的结果完全相同。

虽然最近提出了加速ConvNet培训的更复杂的方法（Krizhevsky，2014），它采用了网络不同层的模型和数据并行性，但我们发现我们在概念上更简单的方案已经提供了3.75倍的加速。与使用单个GPU相比，现成的4-GPU系统。 在配备四个NVIDIA Titan Black GPU的系统上，根据架构的不同，培训一个网络需要2-3周。

1. 分类实验

**数**据集。 在本节中，我们将介绍所描述的ConvNet架构在ILSVRC-2012数据集（用于ILSVRC 2012-2014挑战）上实现的图像分类结果。 该数据集包括1000个类的图像，并分为三组：训练（1.3M图像），验证（50K图像）和测试（具有保持类标签的100K图像）。 使用两个度量评估分类性能：top-1和top-5错误。 前者是一个多类别的分类错误，即错误分类图像的比例; 后者是

在ILSVRC中使用的主要评估标准，并且计算为图像的比例，使得地面实况类别在前5个预测类别之外。

对于大多数实验，我们使用验证集作为测试集。 在测试装置上也进行了某些实验，并作为“VGG”团队参加ILSVRC-2014竞赛提交给官方ILSVRC服务器（Russakovsky等，2014）。

* 1. 单一评估

我们首先使用Sect中描述的层配置，以单一尺度评估各个ConvNet模型的性能。 2.2。 测试图像尺寸设定如下：对于固定的S，Q = S，并且对于抖动的S∈[S*分*，S*最大*]，Q = 0.5（S*分*+ S*最大*）。 结果如表3所示。

首先，我们注意到使用本地响应规范化（A-LRN网络）并没有改进模型A而没有任何规范化层。 因此，我们不在更深层的架构（B-E）中采用规范化。

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，但配置C（包含三个1×1转换层），性能比配置D差，后者使用3×3转换。 整个网络中的层。 这表明虽然额外的非线性确实有帮助（C是

优于B），使用conv捕获空间上下文也很重要。 具有非平凡感受野的过滤器（D优于C）。 当深度达到19层时，我们的架构的错误率会饱和，但更深的模型可能对更大的数据集有益。 我们还将净B与浅网进行了比较，其中有5个5×5转。 层，通过替换从B派生

每对3×3转。 具有单个5×5转换的层。 图层（具有相同的感受域）

在Sect。中解释。 2.3）。 测量浅网的前1个误差比其高7％

B（在中心作物上），证实带有小过滤器的深网优于具有较大过滤器的浅网。

最后，在训练时间（S∈[256; 512]）的比例抖动导致比具有固定最小边（S = 256或S = 384）的图像训练显着更好的结果，即使单个比例是

在测试时使用。 这证实了通过尺度抖动的训练集增强确实有助于捕获多尺度图像统计。

# 表3：单一测试规模的ConvNet性能。

|  |  |  |  |
| --- | --- | --- | --- |
| ConvNet配置。 （表格1） | 最小的图像侧列车（S） 测试（Q） | top-1 val。 错误（％） | 前5名。 错误（％） |
| A | 256 256 | 29.6 | 10.4 |
| A-LRN | 256 256 | 29.7 | 10.5 |
| B | 256 256 | 28.7 | 9.9 |
|  | 256 256 | 28.1 | 9.4 |
| C | 384 384 | 28.1 | 9.3 |
|  | [256;512] 384 | 27.3 | 8.8 |
|  | 256 256 | 27.0 | 8.8 |
| D | 384 384 | 26.8 | 8.7 |
|  | [256;512] 384 | 25.6 | 8.1 |
|  | 256 256 | 27.3 | 9.0 |
| E | 384 384 | 26.9 | 8.7 |
|  | [256;512] 384 | **25.5** | **8.0** |

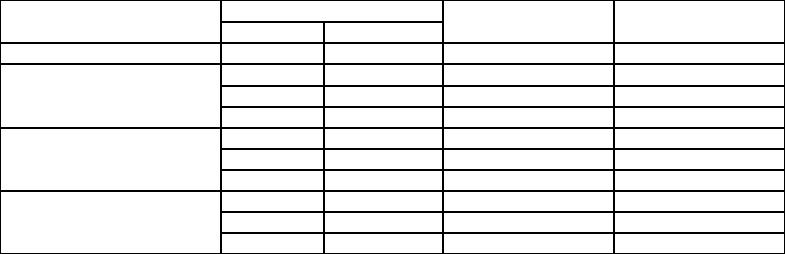
* 1. 多尺度评估

在单一规模评估ConvNet模型后，我们现在评估测试时刻度抖动的影响。 它包括在测试图像的几个重新缩放版本上运行模型（对应于不同的Q值），然后对得到的类后验进行平均。 考虑到训练和测试量表之间的巨大差异导致性能下降，使用固定S训练的模型在三个测试图像大小上进行评估，接近训练一：Q =

{S-32，S，S + 32}。 同时，在训练时刻度抖动允许网络在测试时应用于更广泛的尺度，因此模型训练变量S∈[S*分*; 在更大范围的尺寸Q = {S*分*，0.5（S*分*+ S*最大*），S*最大*}上评估S*最大*]。

表4中显示的结果表明，在测试时刻度的抖动导致更好的性能（与在单一规模上评估相同模型相比，如表3所示）。 与以前一样，最深的配置（D和E）表现最佳，并且比例抖动优于使用固定最小侧S的训练。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（表4中以粗体突出显示）。 在测试集上，配置E达到7.3％的前5个错误。

# 表4：多个测试规模的ConvNet性能。



ConvNet配置。 （表格1）

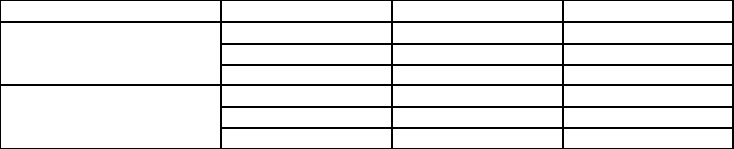
最小的图像面

top-1 val。 错误（％） 前5名。 错误（％）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 火车（S） | 测试（Q） |  | |
| B | 256 | 224,256,288 | 28.2 | 9.6 |
|  | 256 | 224,256,288 | 27.7 | 9.2 |
| C | 384 | 352,384,416 | 27.8 | 9.2 |
|  | [256; 512] | 256,384,512 | 26.3 | 8.2 |
|  | 256 | 224,256,288 | 26.6 | 8.6 |
| D | 384 | 352,384,416 | 26.5 | 8.6 |
|  | [256; 512] | 256,384,512 | **24.8** | **7.5** |
|  | 256 | 224,256,288 | 26.9 | 8.7 |
| E | 384 | 352,384,416 | 26.7 | 8.6 |
|  | [256; 512] | 256,384,512 | **24.8** | **7.5** |

* 1. 多目标评估

在表5中，我们将密集的ConvNet评估与多作物评估进行比较（详见第3.2节）。 我们还通过平均其软最大输出来评估两种评估技术的互补性。 可以看出，使用多种作物的表现略好于密集评价，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，我们假设这是由于对卷积边界条件的不同处理。

表5：ConvNet评估技术比较。 在所有实验中，训练量表S从[256; 并且考虑了三个测试标度Q：{256,384,512}。

|  |  |  |  |
| --- | --- | --- | --- |
| ConvNet配置。 （表格1） | 评估方法 | top-1 val。 错误（％） | 前5名。 错误（％） |
|  | 稠密 | 24.8 | 7.5 |
| D | 多作物 | 24.6 | 7.5 |
|  | 多作物和密集 | **24.4** | **7.2** |
|  | 稠密 | 24.8 | 7.5 |
| E | 多作物 | 24.6 | 7.4 |
|  | 多作物和密集 | **24.4** | **7.1** |

* 1. ConvNet融合

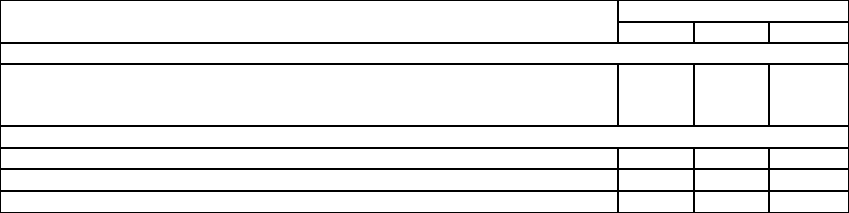
到目前为止，我们评估了各个ConvNet模型的性能。 在这部分实验中，我们将几个模型的输出结合起来，通过平均它们的soft-max类后验。 由于模型的互补性，这改善了性能，并且在2012年的最高ILSVRC提交（Krizhevsky等，2012）和2013（Zeiler＆Fergus，2013; Sermanet等，2014）中使用。

结果显示在表6中。到ILSVRC提交时，我们只训练了单尺度网络，以及多尺度模型D（通过仅微调全连接层而不是所有层）。 由此产生的7个网络集合具有7.3％的ILSVRC测试错误。 提交后，我们考虑了仅有两个表现最佳的多尺度模型（配置D和E）的集合，使用密集评估将测试误差降低到7.0％，使用组合密集和多作物评估将测试误差降低到6.8％。 作为参考，我们性能最佳的单一模型可实现7.1％的误差（模型E，表5）。

* 1. 与现有技术的比较

最后，我们将结果与表7中的现有技术进行了比较。在ILSVRC-2014挑战的分类任务中（Russakovsky等，2014），我们的“VGG”团队获得了第二名

表**6：多个ConvNet融合结果。**



结合ConvNet模型

错误

2 5 5测试val值

提交人的ILSVRC（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）

后提交

24.7

7.5

7.3

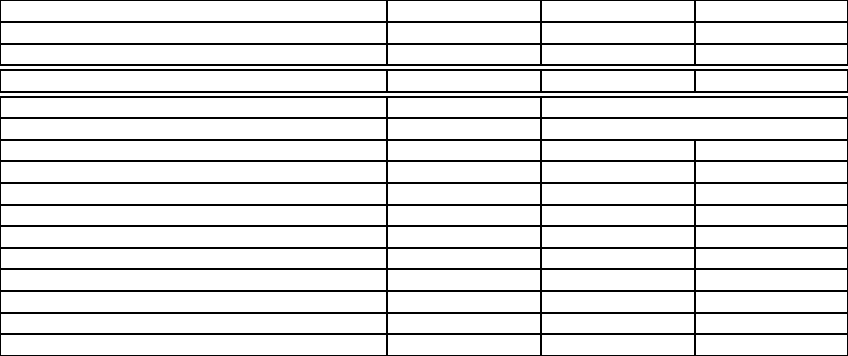
|  |  |  |  |
| --- | --- | --- | --- |
| （D / [256; 512] / 256,384,512），（E / [256; 512] / 256,384,512），密集eval。 | 24.0 | 7.1 | 7.0 |
| （D / [256; 512] / 256,384,512），（E / [256; 512] / 256,384,512），多作物 | 23.9 | 7.2 | - |
| （D / [256; 512] / 256,384,512），（E / [256; 512] / 256,384,512），多作物和密集的eval。 | **23.7** | **6.8** | **6.8** |

使用7个模型的集合，测试误差为7.3％。 提交后，我们将错误率降低到

6.8％使用2个模型的集合。

从表7中可以看出，我们非常深刻的ConvNets明显优于上一代模型，后者在ILSVRC-2012和ILSVRC-2013竞赛中取得了最佳成绩。 我们的结果在分类任务获胜者（GoogLeNet误差为6.7％）方面也具有竞争力，并且大大优于ILSVRC-2013获胜提交的Clarifai，其中外部训练数据达到11.2％，没有它的11.7％。 这是非常值得注意的，考虑到仅通过组合两个模型就可以实现我们的最佳结果 - 显着低于大多数ILSVRC提交的模型。 就单网性能而言，我们的架构实现了最佳结果（7.0％测试错误），优于单个GoogLeNet 0.9％。 值得注意的是，我们并没有偏离LeCun等人的经典ConvNet架构。 （1989），但通过大幅增加深度来改善它。

表7：与ILSVRC分类中的现有技术的比较。 我们的方法表示为“VGG”。 仅报告没有外部训练数据的结果。



方法

top-1 val。 错误（％）top-5 val。 错误（％）top-5测试错误（％）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VGG（2个网，多种作物和密集的eval。） | **23.7** | **6.8** |  | **6.8** |
| VGG（1净，多作物和密集评估。） | 24.4 | 7.1 |  | 7.0 |
| VGG（ILSVRC提交，7网，密集评估。） | 24.7 | 7.5 |  | 7.3 |
| GoogLeNet（Szegedy等，2014）（1网） | - |  | 7.9 |  |
| GoogLeNet（Szegedy等，2014）（7网）  MSRA（He et al。，2014）（11网） | -  - | - | **6.7** | 8.1 |
| MSRA（He et al。，2014）（1 net） | 27.9 | 9.1 |  | 9.1 |
| Clarifai（Russakovsky等，2014）（多个网） | - | - |  | 11.7 |
| Clarifai（Russakovsky等，2014）（1网） | - | - |  | 12.5 |
| Zeiler＆Fergus（Zeiler＆Fergus，2013）（6网） | 36.0 | 14.7 |  | 14.8 |
| Zeiler＆Fergus（Zeiler＆Fergus，2013）（1网） | 37.5 | 16.0 |  | 16.1 |
| OverFeat（Sermanet等，2014）（7网） | 34.0 | 13.2 |  | 13.6 |
| OverFeat（Sermanet et al。，2014）（1 net） | 35.7 | 14.2 |  | - |
| Krizhevsky等人。 （Krizhevsky等，2012）（5网） | 38.1 | 16.4 |  | 16.4 |
| Krizhevsky等人。 （Krizhevsky等，2012）（1网） | 40.7 | 18.2 |  | - |

1. 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。 已经证明，表示深度有利于分类准确性，并且使用传统的ConvNet架构可以实现ImageNet挑战数据集上的最先进性能（LeCun等，1989; Krizhevsky等， 2012）深度大幅增加。 在附录中，我们还展示了我们的模型很好地概括了广泛的任务和数据集，匹配或优于围绕不太深的图像表示构建的更复杂的识别管道。 我们的结果再次证实了深度在视觉表现中的重要性。

致谢

这项工作得到了ERC拨款VisRec no的支持。 228180.我们非常感谢NVIDIA公司的支持，捐赠了用于本研究的GPU。

参考

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 net- works. In *NIPS*, pp. 1106–1114, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropa- gation 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., Sa ́nchez, 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.

9

Published as a conference paper at ICLR 2015

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.

1. 本土化

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ConvNet架构进行了全面评估。 在本节中，我们转向挑战的本地化任务，我们在2014年赢得了25.3％的错误。 它可以看作是对象检测的一种特殊情况，其中应该为每个前5个类预测单个对象边界框，而不管该类的实际对象数。 为此，我们采用了Sermanet等人的方法。 （2014年），ILSVRC-2013本地化挑战的获奖者，进行了一些修改。 我们的方法在Sect。 A.1并在Sect中评估。 A2。

* 1. 本地化CONVNET

为了执行对象本地化，我们使用非常深的ConvNet，其中最后一个完全连接的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。 可以选择是否在所有类中共享边界框预测（单类回归，SCR（Sermanet等，2014））或类特定（每类回归，PCR）。 在前一种情况下，最后一层是4-D，而后者是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ConvNet架构D（表1），它包含16个权重层，并且被发现在分类任务中表现最佳（第4节）。

**训**练。 本地化ConvNets的培训类似于ConvNets分类（第3.1节）。 主要区别在于我们用欧几里德损失代替逻辑回归目标，这会损害预测的边界框参数与地面实况的偏差。 我们训练了两个本地化模型，每个模型在一个尺度上：S = 256和S = 384（由于时间限制，我们没有使用ILSVRC-2014提交的训练尺度抖动）。 培训是

用相应的分类模型（在相同尺度上训练）初始化，并将初始学习率设置为10−3。 我们探索了微调所有层并仅微调前两个完全连接的层，如（Sermanet等，2014）所做。 最后一个完全连接的层随机初始化并从头开始训练。

**测**试。 我们考虑两种测试协议。 第一个用于比较验证集上的不同网络修改，并仅考虑基础真值类的边界框预测（以分解分类错误）。 通过仅将网络应用于图像的中心裁剪来获得边界框。

第二个完全成熟的测试程序基于将定位ConvNet密集应用于整个图像，类似于分类任务（第3.2节）。 不同之处在于，最后一个完全连接的层的输出是一组边界框预测，而不是类得分图。 为了得出最终预测，我们利用了Sermanet等人的贪婪合并程序。 （2014），首先合并空间接近的预测（通过平均他们的坐标），然后根据从分类ConvNet获得的类别分数对它们进行评级。 当使用几个本地化ConvNets时，我们首先结合它们的边界框预测集合，然后在union上运行合并过程。 我们没有使用多重池

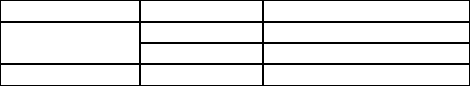
Sermanet的偏移技术等。 （2014），它增加了边界框预测的空间分辨率，并可以进一步改善结果。

* 1. 本地化实验

在本节中，我们首先确定性能最佳的本地化设置（使用第一个测试协议），然后在完全成熟的场景（第二个协议）中对其进行评估。 根据ILSVRC标准（Russakovsky等，2014）测量定位误差，即如果边界框预测与地面实界边界框的交联超过0.5，则认为边界框预测是正确的。

**设**置比较。 从表8可以看出，每类回归（PCR）优于类别不可知的单类回归（SCR），这与Sermanet等人的研究结果不同。 （2014），其中PCR优于SCR。 我们还注意到，对于本地化任务微调所有层导致明显比仅微调完全连接层的结果更好（如（Sermanet等，2014）中所做的那样）。 在这些实验中，最小图像侧设定为S = 384; S = 256的结果表现出相同的行为，为简洁起见未示出。

表8：使用简化测试协议进行的不同修改的本地化错误：从单个中心图像裁剪预测边界框，并使用地面实况类。 所有ConvNet层（最后一层除外）都具有配置D（表1），而最后一层执行单类回归（SCR）或每类回归（PCR）。



微调层回归类型GT类定位错误

第1和第2 FC PC1PCR

所有 PCR

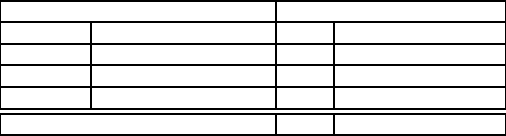
36.4

34.3

**33.1**

**完**全成熟的评估。 确定了最佳的本地化设置（PCR，微调所有层）后，我们现在将其应用于完全成熟的场景，其中使用我们表现最佳的分类系统（第4.5节）预测前5类标签，使用Sermanet等人的方法合并多个密集计算的边界框预测。 （2014）。 从表9可以看出，与使用中心作物相比，将定位ConvNet应用于整个图像显着改善了结果（表8），尽管使用前5个预测类标签而不是基础事实。 与分类任务（第4节）类似，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：本地化错误**



最小的图像面

前5个本地化错误（％）

融合：256/256和384 / 352,384 **26.9**

**25.3**

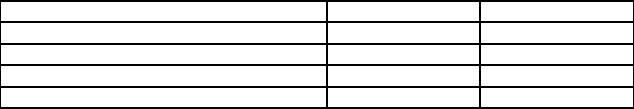
|  |  |  |  |
| --- | --- | --- | --- |
| 火车（S） | 测试（Q） | VAL。 | 测试。 |
| 256 | 256 | 29.5 | - |
| 384 | 384 | 28.2 | 26.7 |
| 384 | 352,384 | 27.5 | - |

**与**现有技术的比较。 我们将最佳定位结果与表10中的最新技术水平进行比较。由于25.3％的测试误差，我们的“VGG”团队赢得了ILSVRC-2014的本地化挑战（Russakovsky等，2014）。 值得注意的是，我们的结果比ILSVRC-2013获胜者Overfeat（Sermanet等，2014）的结果要好得多，尽管我们使用较少的量表并且没有采用其分辨率增强技术。 我们设想如果将这种技术结合到我们的方法中，可以实现更好的定位性能。 这表明我们非常深刻的ConvNets带来的性能提升 - 我们通过更简单的本地化方法获得了更好的结果，但是更强大的表示。

1. 非常深的特征的概括

在前面的部分中，我们讨论了ILSVRC数据集上非常深的ConvNets的培训和评估。 在本节中，我们评估我们在ILSVRC上预先培训的ConvNets作为功能

表**10：与ILSVRC定位中的现有技术的比较。 我们的方法表示为“VGG”。**



方法

前5名。 错误（％）top-5测试错误（％）

|  |  |  |
| --- | --- | --- |
| VGG | **26.9** | **25.3** |
| GoogLeNet（Szegedy等，2014） | - | 26.7 |
| OverFeat（Sermanet等，2014） | 30.0 | 29.9 |
| Krizhevsky等人。 （Krizhevsky等，2012） | - | 34.2 |

其他较小的数据集上的提取器，由于过度拟合，从头开始训练大型模型是不可行的。 最近，人们对这种用例感兴趣（Zeiler＆Fergus，2013; Donahue等，2013; Razavian等，2014; Chatfield等，2014），因为它证明了这一用例在ILSVRC上学习的表示，很好地概括为其他数据集，它们在很大程度上优于手工制作的表示。 在这一系列工作之后，我们将研究我们的模型是否比最先进方法中使用的更浅层模型产生更好的性能。 在本次评估中，我们考虑了两种在ILSVRC上具有最佳分类性能的模型（第4节） - 配置“Net-D”和“Net-E”（我们公开发布）。

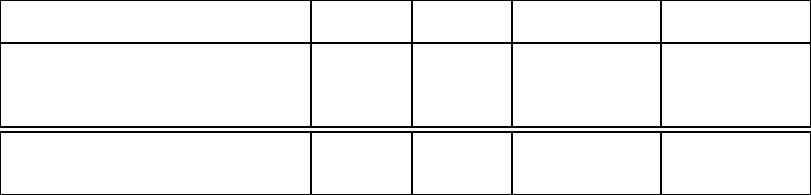
为了利用在ILSVRC上预训练的ConvNets，对其他数据集进行图像分类，我们删除最后一个完全连接的层（执行1000路ILSVRC分类），并使用倒数第二层的4096-D激活作为图像特征，它们在多个位置和比例上汇总。 得到的图像描述符是L2 - 标准化并与线性组合

SVM分类器，在目标数据集上训练。 为简单起见，保留了预先训练的ConvNet权重

固定（不进行微调）。

特征的聚合以与ILSVRC评估程序类似的方式进行（第3.2节）。 即，首先对图像进行重新缩放，使其最小边等于Q，然后在图像平面上密集地施加网络（当所有权重层都被视为卷积时，这是可能的）。 然后，我们在生成的特征映射上执行全局平均池，从而生成4096-D图像描述符。 然后使用水平翻转图像的描述符对描述符进行平均。 正如在Sect。中所示。 4.2，对多尺度的评估是有益的，因此我们在几个尺度上提取特征Q.得到的多尺度特征可以跨尺度堆叠或汇集。 堆叠允许后续分类器学习如何在一系列尺度上最佳地组合图像统计; 然而，这是以增加的描述符维度为代价的。 我们将在下面的实验中回到对此设计选择的讨论。 我们还评估了使用两个网络计算的特征的后期融合，其通过堆叠它们各自的图像描述符来执行。

表11：与VOC-2007，VOC-2012，Caltech-101和Caltech-256的图像分类的现有技术的比较。 我们的模型表示为“VGG”。 标有\*的结果是使用在扩展的ILSVRC数据集（2000类）上预训练的ConvNets实现的。



方法

VOC-2007 VOC-2012 加州理工学院-101

加州理工学院-256

（平均AP）（平均AP）（平均类别回忆）（平均类别回忆）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Zeiler＆Fergus（Zeiler＆Fergus，2013） | - | 79.0 | 86.5 ± 0.5 | 74.2 ± 0.3 |
| Chatfield等。 （Chatfield等，2014） | 82.4 | 83.2 | 88.4 ± 0.6 | 77.6 ± 0.1 |
| 他等人。 （He et al。，2014） | 82.4 | - | **93.4** ± **0.5** | - |
| 魏等人。 （Wei et al。，2014） | 81.5 (85.2∗) | 81.7 (**90.3**∗) | - | - |
| VGG Net-D（16层） | 89.3 | 89.0 | 91.8 ± 1.0 | 85.0 ± 0.2 |
| VGG Net-E（19层） | 89.3 | 89.0 | 92.3 ± 0.5 | 85.1 ± 0.3 |
| VGG Net-D和Net-E | **89.7** | **89.3** | 92.7 ± 0.5 | **86.2** ± **0.3** |

**V**OC-2007和VOC-2012的图像分类。 我们首先评估PASCAL VOC-2007和VOC-2012基准的图像分类任务（Everingham等，2015）。 这些数据集分别包含10K和22.5K图像，每个图像用一个或多个标签注释，对应于20个对象类别。 VOC组织者提供预先划分的培训，验证和测试数据（VOC-2012的测试数据不公开;相反，提供官方评估服务器）。 使用不同类别的平均精度（mAP）来测量识别性能。

值得注意的是，通过检查VOC-2007和VOC-2012验证集的性能，我们发现聚合图像描述符，通过平均计算多个尺度，执行sim-

通过堆叠来进行聚合。 我们假设这是因为在VOC数据集中，对象出现在各种尺度上，因此没有特定的特定于尺度的语义，分类器可以利用它们。 由于平均具有不夸大描述符维度的好处，因此我们能够在各种范围内聚合图像描述符：

Q∈{256,384,512,640,768}。 值得注意的是，{256,384,512}的较小范围的改善相当微不足道（0.3％）。

报告测试集性能并与表11中的其他方法进行比较。我们的网络“Net-D”和“Net-E”在VOC数据集上表现出相同的性能，它们的组合略微改善了结果。 我们的方法在ILSVRC数据集上预先训练的图像表示中设定了新的技术水平，优于Chatfield等人之前的最佳结果。 （2014年）超过6％。 应该注意的是Wei等人的方法。 （2014），在VOC-2012上实现了1％更好的mAP，在2000级ILSVRC扩展数据集上预先培训，其中包括额外的1000个类别，在语义上接近于VOC数据集中的类别。 它还受益于与物体检测辅助分类管道的融合。

**C**altech-101和Caltech-256的图像分类。 在本节中，我们评估了Caltech-101（Fei-Fei等，2004）和Caltech-256（Griffin等，2007）图像分类基准的非常深入的特征。 Caltech-101包含标记为102个类别（101个对象类别和背景类别）的9K图像，而Caltech-256更大，具有31K图像和257个类别。 这些数据集的标准评估协议是生成几个随机分成训练和测试数据，并报告分裂的平均识别性能，这是通过平均类别回忆（每类补偿不同数量的测试图像）来衡量的。 关注Chatfield等人。 （2014）; Zeiler＆Fergus（2013）; 他等人。 （2014年），在Caltech-101上，我们生成了3个随机分组到训练和测试数据中，因此每个分组包含每个类30个训练图像，每个类最多50个测试图像。 在Caltech-256上，我们还生成了3个分割，每个分割包含60个训练图像（其余用于测试）。 在每次拆分中，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，在加州理工学院的数据集中，在多个尺度上计算的描述符堆叠比平均或最大池化更好。 这可以通过以下事实来解释：在Caltech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（捕获整个对象与对象部分），并且堆叠允许分类器利用这种特定于比例的表示。 我们使用三个尺度Q∈{256,384,512}。

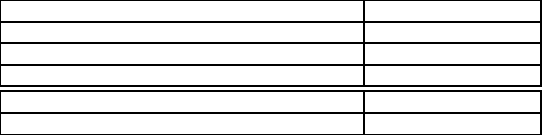
我们的模型相互比较，表11中的技术水平。可以看出，更深的19层Net-E比16层Net-D表现更好，它们的组合进一步提高了性能。 在Caltech-101上，我们的陈述与He等人的方法相竞争。 （2014年），然而，其表现明显差于我们在VOC-2007上的网络。 在Caltech-256上，我们的功能大大优于现有技术（Chatfield等，2014），大幅提升（8.6％）。

**V**OC-2012的行动分类。 我们还评估了PASCAL VOC-2012行动分类任务（Everingham等，2015）中表现最佳的图像表示（Net-D和Net-E特征的堆叠），其中包括从单个预测动作类图像，给出执行动作的人的边界框。 该数据集包含4.6K训练图像，标记为11个类。 与VOC-2012对象分类任务类似，使用mAP测量性能。 我们考虑了两种训练设置：（i）在整个图像上计算ConvNet特征并忽略提供的边界框; （ii）计算整个图像和所提供的边界框上的特征，并堆叠它们以获得最终表示。 将结果与表12中的其他方法进行比较。

即使不使用提供的边界框，我们的表示也实现了VOC动作分类任务的现状，并且当使用图像和边界框时，结果得到进一步改善。 与其他方法不同，我们没有包含任何特定于任务的启发式方法，而是依赖于非常深度卷积特征的表示能力。

**其**他识别任务。 自公开发布我们的模型以来，它们一直被研究界广泛用于各种图像识别任务，始终优于更浅层的表示。 例如，Girshick等。 （2014）通过替换Krizhevsky等人的ConvNet来实现目标检测结果的状态。 （2012）与我们的16层模型。 与Krizhevsky等人的更浅层架构相似。 （2012）已经

表12：与VOC-2012的单图像动作分类中的现有技术的比较。 我们的模型表示为“VGG”。 标有\*的结果是使用在扩展的ILSVRC数据集（1512类）上预训练的ConvNets实现的。



方法 VOC-2012（平均AP）

（Oquab等，2014） 70.2∗

（Gkioxari等，2014） 73.6

（Hoai，2014） 76.3

VGG Net-D和Net-E，仅限图像 **79.2**

VGG Net-D和Net-E，图像和边界框 **84.0**

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）。

C 论文修订

在这里，我们列出了主要的纸张修订清单，概述了为方便读者而进行的实质性修改。

**v**1初始版本。 介绍在ILSVRC提交之前进行的实验。

**v**2使用规模抖动添加提交后的ILSVRC实验和训练集扩充，从而提高性能。

**v**3在PASCAL VOC和Caltech图像分类数据集上添加泛化实验（附录B）。 用于这些实验的模型是公开的。

**v**4该论文转换为ICLR-2015提交格式。 还增加了多种作物的实验用于分类。

**v**6相机就绪ICLR-2015会议论文。 添加净B与浅网的比较以及PASCAL VOC行动分类基准的结果。