VGGNet 两个特点：层数更深、卷积核更小

小卷积核带来参数量的减半，可以更稳定地增加层数同时又不必担心计算量的暴增。

池化核变小且为偶数

任务背景

**浅层学到的是纹理特征，而深层学到的是语义特征**”

对CNN提取到的特征做了聚合和一些变换

网络输入224\*224 输入前需要减去RGB的均值

2.1卷积

Kernel Size

Stride

两个3×3的卷积堆叠获得的感受野大小，相当一个5×5的卷积；而3个3×3卷积的堆叠获取到的感受野相当于一个7×7的卷积

卷积核大小对结果的影响

Very Deep Convolutional Networks For Large-scale Image Recognition.用于大规模图像识别的深度卷积网络

## 摘要

本文研究了在大规模图像识别环境中，卷积网络深度对其准确性的影响。本文的主要贡献在于，对使用很小(3×3)的卷积滤波器来增加深度的网络进行了一个全面的评估，结果表明，通过将深度提高到16-19个权重层，此前的技术性能能够实现显著地提升。这些发现是我们参加的2014年ImageNet挑战赛的基础，我们的团队也因此在定位和分类任务上分别获得了第一和第二名的成绩。本文表明，我们的模型表现在其他数据集上推广同样表现得很好，并都达到了当前最佳的水平。我们已经公开了两种性能最好的卷积网络模型，希望能够促进深度视觉模型在计算机视觉领域应用的研究。

## 1 简介

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，这的实现得益于大型公开图像库（如ImageNet）和高性能计算系统（如GPU或大规模分布式集群）。特别是在深度视觉识别结构发展中扮演重要角色的ImageNet大规模视觉识别大赛(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有效，为了获得更高的准确率，很多人多次尝试在Krizhevsky et al. (2012)的原始框架上进行改进，以获得更好的精度。例如，ILSVRC-2013最好的参赛模型在第一个卷积层上使用了更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，密集地训练和测试网络。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稳步增加网络的深度，通过在所有层都使用非常小(3×3)的卷积滤波器，这是可行的。

因此，我们意味深长地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来促进进一步的研究。

本文其余部分如下：Sect.2描述了我们的卷积网络的模型。Sect.3介绍了图像分类任务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，我们也在附录A中描述了我们的ILSVRC-2014目标定位系统，并在附录B中讨论了非常深的特征在其他数据集上泛化能力。最后，附录C包含了本文的主要修订记录。

## 2 卷积网络参数模型

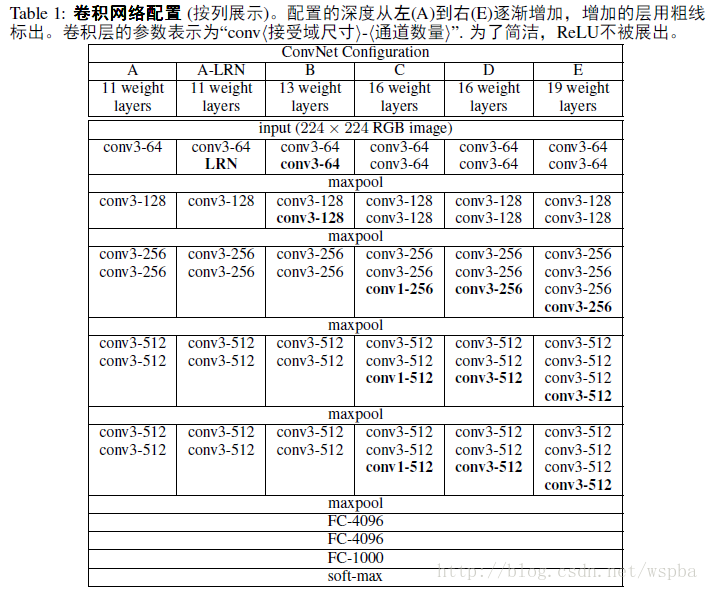
为了公平地评估增加的卷积网络深度带来的改进，我们所有的卷积层参数都以相同的原则设计的，这是受Ciresan et al.(2011)和Krizhevsky et al.(2012)的启发。本章中，Sect. 2.1首先介绍了我们的卷积网络的一般结构；然后在Sect. 2.2中描述了在评估中具体参数的细节；最后Sect. 2.3中讨论了我们的设计选择，并与先前最好的网络进行比较。

2.1 结构  
 在训练阶段，卷积网络的输入为尺寸固定为224×224的RGB图像。我们做的唯一的预处理是每个像素减去训练集中RGB的均值。图像通过一堆卷积层，卷积滤波器使用非常小的接受域：3×3(是能够获取左右、上下和中心的最小尺寸)。在我们的一种配置中，我们也使用1×1的卷积滤波器，这可以看作是对输入通道的一个线性变换(其后面接一个非线性变换)。卷积的步长固定为1个像素；对卷积层输入进行空间填充，以保持卷积后的空间分辨率，即3×3卷积层的空间填充为1像素。空间池化包含5个最大池化层，它们接在部分卷积层的后面(并不是所有的卷积层都接有最大池化层)。最大池化在2×2的滑动窗口上进行，滑动步长为2。

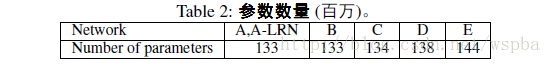
在一整堆的卷积层后面(不同结构有不同深度)接了3个全连接层：前两个全连接层各有4096个通道，第三层执行1000类的ILSVRC分类，因此包含1000个通道(每个通道代表一类)。最后一层是一个soft-max层。全连接层的参数配置在所有网络中都是相同的。

所有隐层都是用非线性修正(ReLU)。注意到我们的网络(除了一个)都不包含局部响应标准化(LRN)：在Sect.4 中会说明，这个标准化并不会提高网络在ILSVRC数据集上的性能，反而会增加内存消耗和计算时间。它适用的情况是Krizhevsky et al.(2012)中的LRN参数配置。

2.2 配置  
 本文所评估的卷积网络的配置参数如表1所示，每一列代表一种配置方式。接下来我们分别称它们为(A-E)。所有网络的配置遵循Sect 2.1中的通用设计，只有深度不同：从含有11个权重层的网络A(8个卷积层和3个全连接层)到含有19个权重层的网络E(16个卷积层和3个全连接层)。卷积层的宽度(通道的数量)非常小，从第一层的64开始，然后每经过一个最大池化层，数量增加一倍，直到数量达到512。



在表2中我们展示了每一种配置的参数。尽管网络很深，但是权重的数量并不比一个拥有大卷积层宽度和接受域尺寸的浅层网络多(例如Sermanet等人在2014的网络有144M个权重)。



2.3 讨论

我们的卷积网络配置与ILSVRC-2012和ILSVRC-2013竞赛中表现最好的模型配置完全不同。我们并没有在第一个卷积层使用相对大的接受域（例如[Krizhevsky等人在2012Imagenet](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "https://blog.csdn.net/wspba/article/details/_blank)中使用的接受域为11×11，滑动步长为4；[Sermanet等人在2013Overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)中使用的接受域为7×7，滑动步长为2），而是在整个网络使用了一个非常小的3×3的接受域，并且在输入的每一个像素上都进行了卷积(滑动步长为1)。很容易发现，两个3×3卷积层堆叠(中间没有空间池化)相当于5×5的有效接受域；三个这样的层相当于7×7 的有效接受域。   
 那么使用三个3×3卷积层的堆叠比一个单一的7×7卷积层的好处是什么呢？首先，我们使用了三个而不是一个单一的非线性修正层，这使得决策函数更具判别性。其次，我们减少参数的数目：假定输入和一个三层的3×3卷积叠加输出C通道，堆栈parametrised由3 32c2 = 27c2权重；同时，一个7×7转换层需要72c2 = 49c2参数，即81%。其次，我们减少了参数的数量：假设一个含有三个3×3卷积层堆叠的输入和输出都有C个通道，因此这个堆叠含有3(32C2)=27C2 个权重；而一个单一的7×7卷积层却需要72C2=49C2个参数，相对增加了81%。这相当于在7×7卷积滤波器上强加一个正则化，迫使它们通过3×3的滤波器来进行分解(中间有非线性的加入)。

1×1卷积层的使用(配置C，Table1)是在不影响卷积层接受域的情况下增加决策函数非线性的一种方法。尽管在这里，1×1的卷积层本质上相当于到相同维度空间的一个线性投影(输出和输出的通道数量相同)，但是修正函数引入了非线性。值得注意的是，1×1卷积层最近被使用在[lin2013network](https://arxiv.org/abs/1312.4400" \t "https://blog.csdn.net/wspba/article/details/_blank)的“Network in Network”结构中。

小尺寸的卷积滤波器之前在[ciresan2011flexible](https://www.researchgate.net/publication/220812758_Flexible_High_Performance_Convolutional_Neural_Networks_for_Image_Classification" \t "https://blog.csdn.net/wspba/article/details/_blank)中使用过，但是他们的网络远没有我们的深，而且他们也没有在大规模的ILSVRC数据集上进行评估。[goodfellow2013multi](https://arxiv.org/abs/1312.6082" \t "https://blog.csdn.net/wspba/article/details/_blank)在街道数字识别的任务中应用了深度卷积网络(11个权重层)，并证明了增加的深度使模型产生了更好的性能。[GoogLeNet](https://arxiv.org/abs/1409.4842" \t "https://blog.csdn.net/wspba/article/details/_blank)，ILSVRC-2014分类任务上最优秀的模型，虽然区别于我们的工作，但是相似的是，它也是基于非常深的卷积网络(22个权重层)以及很小的卷积滤波器(除了3×3的卷积层，它们同样还使用了1×1和mbox{5×5} 的卷积层)。他们的网络拓扑比我们的更加复杂，而且为了降低计算量，特征图的空间分辨率在第一层衰减的特别严重。Sect 4.5 将会展示，我们的模型在单一网络分类准确率上优于GoogLeNet的网络。

3分类网络框架

在上一章中，展示了我们网络配置的细节。在这一章，我们将对分类卷积网络的训练和评估的细节进行介绍。

3.1 训练   
 卷积网络的训练过程基本遵循[krizhevsky2012imagenet](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "https://blog.csdn.net/wspba/article/details/_blank)(除了从多尺度训练图像上采样裁切的输入图像，稍后会解释)。也就是说，通过使用含动量的小批量(mini-batch)梯度下降(基于反向传播)优化多元逻辑回归来对模型进行训练。小批量的尺寸为256，动量为0.9。通过权值衰减(L2惩罚系数设置为5)以及对前两个全连接层执行dropout(dropout比率设置为0.5)来对训练进行正则化。初始学习率设置为10−2，当验证集准确率稳定时将学习率除以10。学习率总共降低了3次，训练一共进行了370K次迭代(74个epoch)。我们猜测，尽管和[krizhevsky2012](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "https://blog.csdn.net/wspba/article/details/_blank)相比，我们的网络有更多的参数和更深的深度，但是网络只需要更少的epoch来收敛，这是因为：(a) 更深的深度和更小的卷积滤波器尺寸隐式的增强了正则化；(b) 某些层执行了预初始化。

网络权重的初始化是非常重要的，由于深度网络梯度的不稳定性，不合适的初始化将会阻碍网络的学习。为了避免这个问题，我们从训练配置A开始(Table1)，它足够浅以致于能够使用随机初始化进行训练。然后在训练更深的结构时，我们对前四个卷积层和最后三个全连接层使用网络A来初始化(中间层使用随机初始化)。我们没有降低预初始化层的学习率，允许它们在训练过程中进行改变。对于随机初始化，我们从0均值10−2方差的正态分布中对权重进行采样。偏置项初始化为0。值得注意的是，在文章提交后，我们发现可以使用[glorot2010understanding](https://www.researchgate.net/publication/215616968_Understanding_the_difficulty_of_training_deep_feedforward_neural_networks" \t "https://blog.csdn.net/wspba/article/details/_blank)中的随机初始化程序来对权重进初始化而不需要进行预训练。

随机的从经过尺寸缩放的训练图像上进行裁切，来获得卷积网络固定尺寸224×224的输入图像(每一张图像的每一个SGD迭代时裁切一次)。为了进一步对训练集进行数据增强，裁切图像进行随机的水平翻转和随机的RGB颜色转换。训练图像的尺寸缩放将在后面进行解释。

Training image size 用S代表经过各向同性缩放的(isotropically-rescaled)训练图像的最小边，卷积网络的输入图像就是从中裁切的(我们也将S成为训练尺寸)。若裁切图像的尺寸固定为224×224，原则上S可以取任意不小于224的值：如果S=224，裁切图像将会获取整幅图像的统计信息，可以完整的涵盖训练图像的最小边。如果S≫224，裁切图像将会对应图像的一小部分，包含一个小的对象或者对象的一个部分。

我们考虑使用两种方法来设定训练尺寸S。第一种方法是针对单尺度图像的训练，固定S(注意到，在采样的裁切图像内的内容仍然能表示多尺度图像的统计信息)。在我们的实验中，我们评估了两种固定尺寸训练获得的模型：S=256和S=384。我们先使用S=256来对一种卷积网络配置进行训练。为了加快S=384的网络的训练速度，使用S=256预训练的权重进行初始化，并且使用一个很小的初始学习率：10−3。

第二种设定S的方法是多尺度训练，每一幅图像单独的从[Smin,Smax]

中随机选取S来进行尺寸缩放(Smin=256，Smax=512)。由于图像中的对象可能是各种尺寸的，因此在训练中采用这种方法是有利的。这同样可以看作是一种尺寸抖动(scale jittering)的训练集数据增强，使得一个单一模型能够识别各种尺寸的对象。考虑到训练速度，我们使用固定S=384预训练模型相同的配置对一个单尺度模型的所有层进行微调，来训练多尺度模型。

3.2 TESTING   
 在测试阶段，对于一个训练好的卷积网络和一张输入图像，由以下方法进行分类。首先，图像的最小边被各向同性的缩放成预定义的尺寸，设为Q(我们也将此称为测试尺寸)。我们注意到Q并不一定要与训练尺寸S相同(Sect.4 将会展示，对于一个S使用不同的Q将有助于性能的提升)。然后，根据[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)中的方法将网络密集的应用在经过尺寸缩放的测试图像上。即先将全连接层转换成卷积层(第一个全连接层转换成7×7的卷积层，后两个全连接层转换成1×1的卷积层)，再将这样得到的全卷积网络运用在整幅图像上(未裁切的)。输出是一个分类得分图，通道的数量和类别的数量相同，空间分辨率依赖于输入图像尺寸。最终为了得到固定尺寸的分类得分向量，将分类得分图进行空间平均化(求和——池化)。我们同样使用水平翻转来对测试集进行数据增强；在原始图像和翻转图像上的soft-max分类概率的平均值作为这幅图像的最终得分。

由于测试阶段在整幅图像上使用了全卷积网络，因此不需要对图像进行多个裁切采样[krizhevsky2012imagenet](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "https://blog.csdn.net/wspba/article/details/_blank)，因为网络对每个裁切的重新计算会使效率降低。而与此同时，正如GoogLeNet 所做的，使用大量的裁切图像可以提高准确率，因为和全卷积网络相比，它能生成关于输入图像更好的采样。同样的，由于不同的卷积边界条件，多重裁切评估与密集评估是互补的：当在裁切图像上使用卷积网络时，卷积特征图使用0进行填充，而使用密集评估时，对于相同裁切图像的填充自然而然的来自图像邻近的部分(由于卷积和空间池化)，这大大增加了网络整体的接受域，因此获取了更多的上下文信息。虽然我们认为在实践中，多重裁切带来的计算时间并不能成为准确率上潜在收益的理由，但是作为参考，我们依然在评估时对每个尺寸使用了50张裁切图像(5×5个规则网格以及它们的翻转)，3种尺寸一共150张裁切图像，这和GoogLeNet 中使用4种尺寸一共144张裁切图像是可比的。

3.3 IMPLEMENTATION DETAILS   
本文的实验基于公开的Caffe(C++)工具包(2013年12月版本)，但进行了一些重要修改，让我们能够在一个系统的多块GPU上进行训练和评估，以及对对（如上所述）多个尺度的全尺寸图像（未裁切）进行训练和评估   
多GPU训练利用数据并行，通过将每批图像分给多批GPU来进行并行运算。当每批的梯度计算好之后，再求平均值作为总的梯度。在多个GPU上的梯度计算是同步的，因此和使用单一GPU进行训练的结果一致。

虽然[krizhevsky2014one](https://arxiv.org/abs/1404.5997" \t "https://blog.csdn.net/wspba/article/details/_blank)提出了更加复杂的方法来加快卷积网络的训练，它在网络中不同层上利用模型和数据并行。但是我们发现单是提出的相对简单的方案在现成的4块GPU系统上已经比单GPU提升了3.75倍的计算速度。在配备4块NVIDIA Titan Black GPU的系统上，训练一个单一网络需要2——3周的时间

4 CLASSIFICATION EXPERIMENTS

Dataset 本章我们展示了本文所描述的卷积网络在ILSVRC-2012数据集(被用在ILSVRC2012——2014挑战赛上)上的图像分类结果。数据集包含1000类的图像，并分为训练集(1.3

M张图像)、验证集(50K张图像)和测试集(100

K张抽出标签的图像)。我们使用两种方法来对分类性能进行评估：top-1和top-5错误率。前者为一个多类分类错误率，即错误分类图像的比例；后者是在ILSVRC上的主要评估标准，即真实类别不在top-5预测类别之中的图像的比例。

对于大部分实验，我们使用验证集作为测试集。某些实验也在测试集上进行，并提交给官方ILSVRC服务器作为“VGG”团队参加ILSVRC-2014竞赛。

4.1 SINGLE SCALE EVALUATION   
我们开始评估在单一尺度上使用Sect 2.2中层配置的独立卷积网络模型的性能。测试图像尺寸的设置按照：对于固定的S

，设置Q=S；对于抖动的S∈[Smin,Smax]，设置Q=0.5(Smin+Smax)

。结果如Table3 展示。

首先，我们注意到使用局部响应标准化(A-LRN网络)并没有提高不使用任何标准化层的模型A的性能。因此我们不在更深的结构(B–E)中使用标准化操作。

其次，我们观测到分类的错误率随着卷积网络深度的增加而降低了：从含有11层的A到含有19层的E。值得注意的是，虽然配置C和配置D具有相同的深度，含有3个1×1

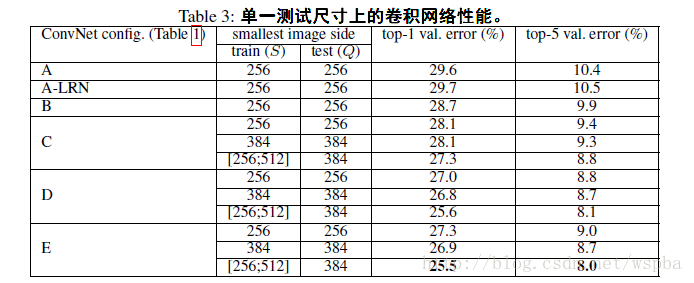
卷积层的配置C比整个网络都使用3×3卷积层的配置D的性能要差。结果表明了增加的非线性层对网络性能的提升是有帮助的(C好于B)，并且使用具有非平凡接受域的卷积滤波器对于获取空间上下文信息是很重要的(D好于C)。当我们的模型深度达到19层时，错误率达到饱和，但是对于更大的数据集，更深的模型可能会更有效果。我们也对网络B和一个使用5个5×5卷积层的浅层网络进行比较(通过将B中的3×3卷积层对替换成单独的5×5卷积层，正如Sect 2.3，它们具有相同的接受域)。在top-1错误率上，浅层网络比网络B要高出7%

(在中心裁切图像上)，证明了使用小滤波器的深层网络比使用大滤波器的浅层网络性能更好。

最后，即使在测试阶段使用单一尺度进行测试，但是在训练阶段使用尺寸抖动(S∈[256;512]

)比图像最小边使用固定尺寸(S=256 or S=384

) 的结果要好得多。这表明了在训练集上通过尺寸抖动的数据增强对于获取多尺度图像统计信息确实是有帮助的。



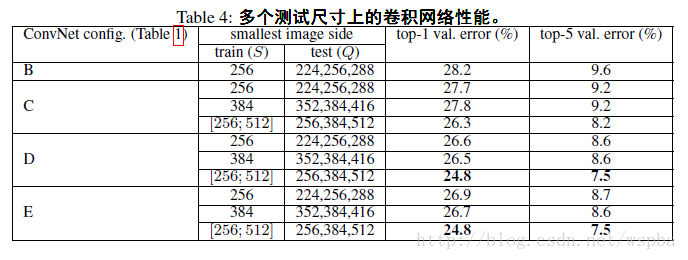
4.2 MULTI-SCALE EVALUATION   
评估了卷积网络模型在单一尺度上的性能之后，我们现在来评估在测试阶段使用尺寸抖动的效果。先在几个经过尺寸缩放的测试图像上运行模型(对应于不同的Q

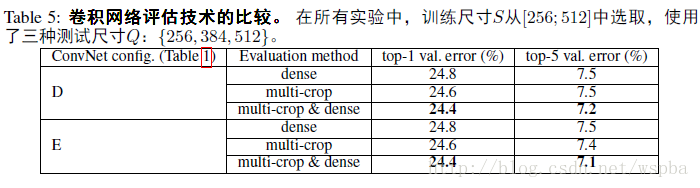
值)，然后计算每个类概率的平均值。考虑到当训练尺寸和测试尺寸存在巨大差别时将会导致模型性能下降，因此在模型训练时，我们固定S的值，使用接近于S的三个测试尺寸Q={S−32,S,S+32}来进行评估。同时在训练阶段使用尺寸抖动使得模型在测试阶段能够应用在尺寸范围更大的图像上，因此模型的训练尺寸使用S∈[Smin;Smax]，然后在更大范围的测试尺寸Q={Smin,0.5(Smin+Smax),Smax}

上进行评估。

Table4 中的结果表明，使用尺寸抖动的模型在测试阶段获得了更好的性能(与使用单一尺度的相同模型Table3 相比)。与之前的一样，最深的配置(D和E)表现最好，同时训练阶段使用尺寸抖动比最短边S

使用固定尺寸的效果更好。我们最好的单一模型在验证集上的top-1/top-5错误率为24.8%/7.5%(Table4 中加粗部分)。在测试集上，配置E达到了7.3%

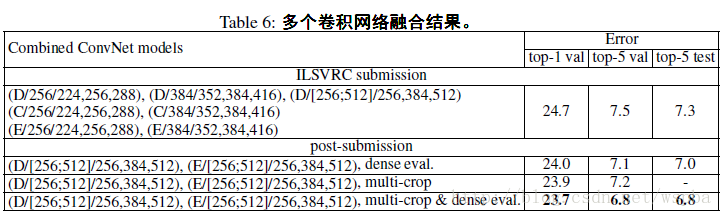
的top-5错误率。   


4.3 MULTI-CROP EVALUATION   
Table5 中我们对密集卷积网络评估和多重裁切评估进行了比较(见Sect 3.2)。我们同样还评估了两种技术通过计算两者soft-max输出平均值的互补结果。可以看出，使用多重裁切比密集评估的效果略好，并且两种方法是完全互补的，因为两者组合的效果比每一种都要好。根据以上结果，我们假设这是由对于卷积边界条件的不同处理方法造成的。   


4.4 CONVNET FUSION   
至此，我们评估了独立卷积网络模型的性能。在这部分实验中，我们通过计算多个模型soft-max分类概率的平均值来对它们的输出进行组合。由于模型的互补性，性能有所提高，这也用在ILSVRC的最佳结果中。

结果如Table6 所示。在提交ILSVRC参赛模型时，我们只训练了单尺度网络和一个多尺度模型D(只对全连接层进行调优而不是所有层)。7个模型组合的结果在ILSVRC错误率上为7.3%

。在模型提交结束后，我们研究了只使用两个性能最好的多尺度模型的组合(配置D和E)，当使用密集评估时错误率降到了7.0%，而使用密集和多重裁切组合时，错误率仅为6.8%。作为参考，我们性能最好的单一模型错误率为7.1%

(模型E，Table5)。   


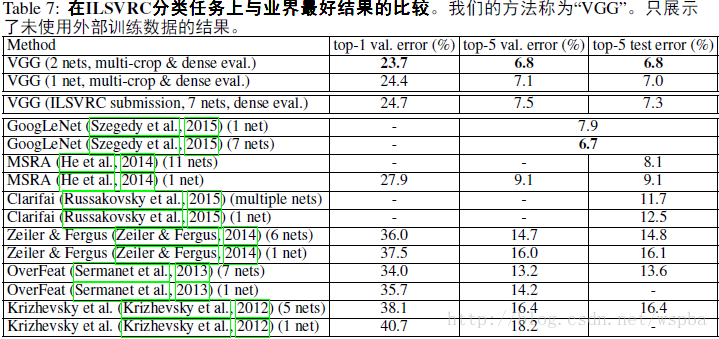
4.5 COMPARISON WITH THE STATE OF THE ART   
最后，对我们的结果与Table7 中业界最好的结果进行了比较。在ILSVRC-2014竞赛的分类任务上，我们的“VGG”团队使用了7个模型的组合以7.3%

的错误率获得了第二名。参赛结束后，我们使用了2个模型的组合将错误率降到了6.8%

。

从Table7 可以看出，我们非常深的卷积网络比在ILSVRC-2012和ILSVRC-2013中成绩最好的模型效果明显要好。我们的结果与分类任务的冠军旗鼓相当(GoogLeNet为6.7%

的错误率)，并且明显比ILSVRC-2013的冠军Clarifai的表现好得多，它使用外部训练数据时的错误率为11.2%，而不使用外部数据时为11.7%。更加重要的是，我们最佳的结果是通过对两个模型的组合——这明显比大多数ILSVRC参赛模型要少。在单一网络性能上，我们的模型取得了最好的结果(7.0%的测试错误率)，比单一的GoogLeNet低0.9%

。值得注意的是，我们并没有摒弃经典的卷积网络框架，并通过显著增加深度对它的性能进行了提升。   


5 CONCLUSION

本文中，我们评估了非常深的卷积网络(达到了19个权重层)在大规模图像分类上的性能。结果表明，表达的深度有利于分类准确率的提升，在传统的卷积网络框架中使用更深的层能够在ImageNet数据集上取得优异的结果。

在附录中也展示了我们的模型在广泛的任务和数据集上良好的泛化能力，性能达到甚至超过了围绕较浅深度的图像表达建立的更复杂的识别流程。我们的实验结果再次确认了视觉表达中深度的重要性。

APPENDIX A LOCALISATION

本文我们主要针对ILSVRC竞赛的分类任务，对不同深度的卷积网络框架进行了全面的评估。本章我们主要介绍定位任务，在这项上我们以25.3%

的错误率获得了冠军。这可以看做是目标检测的一种特殊情况，对于top-5类中的每一类，需要预测一个单一的目标边界框，而不需考虑每一类目标的实际数量。为此，我们在ILSVRC-2013定位任务冠军模型[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)的基础上，进行了一些修正。Sect A.1描述了我们的方法，Sect A.2对我们的方法进行了评估。

A.1 LOCALISATION CONVNET   
在目标定位中，我们使用了非常深的卷积网络，最后一个全连接层用来预测边界框的位置而不是分类得分。边界框由一个4-D向量表示，包括了它的中心坐标、宽度和高度。关于边界框的预测有两种方案：是在所有类之间共享的(单类回归，SCR)还是针对特定类的(每类回归，PCR)。对于前一种，最后一层的输出维度是4-D；而对于后一种，最后一层的输出维度是4000-D(因为数据集中共有1000中类别)。除了最后一个边界框预测层，前面我们使用配置D的卷积网络，它包含16个权重层，在分类任务中性能是最好的(Sect.4)。

Training 定位卷积网络的训练和分类卷积网络的训练类似(Sect 3.1)。主要的区别是我们将逻辑回归替换成了欧式损失(Euclidean loss)，用来惩罚预测的边界框参数与真值之间的偏差。我们训练了两个定位模型，分别使用单一尺寸：S=256

和 S=384 (由于时间限制，在ILSVRC-2014上我们并没有在训练时使用尺寸抖动)。训练使用对应的分类模型进行初始化(使用相同尺寸训练)，初始学习率为10−3

。我们研究了两种调优方式：在所有层进行调优和只在前两个全连接层进行调优。最后一个全连接层使用随机初始化并从头开始训练。

Testing 我们考虑两种测试方案。第一种用来比较使用不同网络修正(即调优方式不同，所有层调优/全连接层调优)在验证集上的区别，只考虑对真实类别的边界框预测(排除了分类错误).只在图像的中心裁切上使用网络来获得边界框。

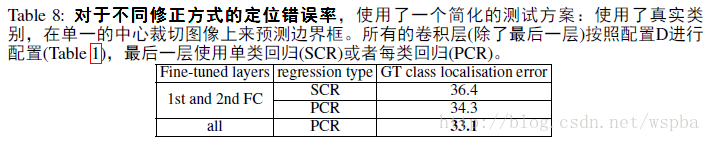
第二种，全面的、基于定位卷积网络在整幅图像上密集应用的测试程序，这类似于分类任务(Sect 3.2)。不同的是，最后一个全连接层的输出是由一组边界框的预测取代了分类得分图。为了获得最终预测结果，我们使用了[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)中的贪婪合并程序，先合并空间上接近的预测(通过计算坐标的平均值)，然后使用由分类卷积网络获得的分类得分对他们进行评级。当使用了一些定位卷积网络时，我们首先得到它们对边界框预测的合集，然后在这个合集上运行合并程序。我们并没有使用[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)中用来增加边界框预测空间分辨率的多重池化偏移技术来进一步提升结果。   
(译者注：即第一种是在真实类上预测边界框，为了确定模型配置；第二种在确定模型配置后，使用预测类，在预测类上预测边界框。)

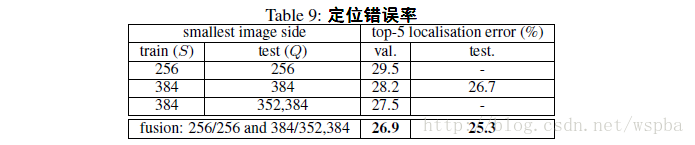
A.2 LOCALISATION EXPERIMENTS   
本章，我们先来确定性能最好的定位网络设定(使用第一种测试方案)，再使用一个全面的方案(第二种方案)来对它进行评估。定位错误率根据ILSVRC标准进行衡量，即，如果预测的边界框与真实边界框的重叠率大于0.5

，则被认为是正确的。

Settings comparison Table8 中可以看出，每类回归(PCR)由于类无关的单类回归(SCR)，而这和[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)中的发现正好相反。我们也注意到，对于定位任务，在所有层进行调优的结果明显比只在全连接层调优的结果好得多(与[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)相同)。在这些实验中，图像的最小边设置为S=384

；S=256

的结果类似，为了简洁不进行展示。   


Fully-fledged evaluation   
在得到了最佳的定位网络设定(PCR，在所有层调优)后，我们使用一个完整的方案来进行评估，使用我们性能最好的分类系统(Sect 4.5)来对top-5类别进行预测，然后使用[sermanet2013overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)的方法来对多个密集计算预测的边界框进行合并。从Table9 中可以看出，在整幅图像上使用定位卷积网络，比在中心裁切图像上(Table8 )的结果明显要好, 尽管使用的是top-5的预测类而不是真实类。与分类任务(Sect.4)相同，在多个尺度上进行测试然后组合多个模型的预测结果，能够对性能产生进一步的提升。   


Comparison with the state of the art Table10 将我们的定位模型结果与业界最好的结果进行比较。我们的“VGG”团队以25.3%

的测试错误率获得了ILSVRC-2014定位比赛的冠军。值得注意的是，我们的结果比ILSVRC-2013的冠军[Overfeat](https://arxiv.org/abs/1312.6229" \t "https://blog.csdn.net/wspba/article/details/_blank)要好得多，尽管我们使用了较少的尺度，也没有使用他们的分辨率增强技术。我们认为如果这一技术能容纳到我们的方法中，可以获得更好的定位效果。这表明了我们非常深的卷积网络能给性能带来很大的提升——虽然我们只使用了一个很简单的定位方法，但是由我们更强力的表达，我们得到了更好的结果。   


APPENDIX B GENERALISATION OF VERY DEEP FEATURES

在前几章，我们讨论了非常深的卷积网络在ILSVRC数据集上的训练和评估。在本章，使用我们在ILSVRC上预训练的卷积网络作为其他更小数据集的特征提取器，因为小数据集上从头开始训练大的模型会由于过拟合而变得很困难。最近，有很多很有趣的案例发现，从ILSVRC上学到的深度图像表达，在其他数据集上的泛化能力很好，它们比手工表达的效果要好很多。根据这一系列工作，我们也开始研究我们的模型是否能比业界最好方法所使用的的浅层模型效果更好。在这项评估中，我们考虑两个在ILSVRC上分类性能最好的模型(Sect.4)——配置“Net-D”和“Net-E”(我们都进行了公开)。

为了在其它数据集上使用在ILSVRC上预训练的模型来进行图像分类，我们去除了最后一个全连接层(用来进行1000类的ILSVRC分类)，使用倒数第二层的4096-D激活作为图像特征，并在多个位置和尺度之间汇总。由此产生的图像描述子使用L2

正则化并与一个在目标数据集上训练的线性SVM分类器组合。简单来说，预训练的卷积网络的权重保持固定(不进行调优)。

特征的汇总方式与我们的ILSVRC评估程序(Sect 3.2)相似。也就是说，先将图像尺寸缩放到使它的最短边等于Q

，然后在图像平面上密集的应用网络(当所有权重层被视为卷积层时，这是有可能的)。然后在产生的特征图上使用全局平均池化，便产生了一个4096-D的图像描述子。再将这个描述子与在对应的水平翻转图像上产生的描述子取平均。如Sect 4.1所展示的，多个尺度上的评估是有利的，因此我们也在多个尺度Q

上提取特征。由此产生的多尺度特征既可以尺度间堆叠也可以池化。堆叠也使得之后的分类器能够学习如何最优的将多个尺度的图像统计特征进行组合；但是这也会由增加的描述子维度而增加计算成本。我们在下面的实验中回到对这个设计选择的讨论。我们也会评估由两个模型计算特征的后期融合，这是通过它们各自的图像描述子的堆叠完成的。

Image Classification on VOC-2007 and VOC-2012 我们开始评估PASCAL VOC-2007和VOC-2012基准数据集的图像分类任务。这两个数据集分别包含10K和22.5K张图像，每一张图像标注有一个或多个标签，对应有20个目标种类。VOC组织者将数据集分为训练集、验证集和测试集(VOC-2012测试集并没有公开，而提供了一个官方的评估服务器)。识别性能使用多类上的平均精度(mAP)来评估。值得注意的是，通过验证在\mbox{VOC-2007}和VOC-2012验证集上的性能，我们发现通过计算多个尺度间的平均来融合描述子与通过堆叠来融合的效果相似。我们认为这是因为VOC数据集中，目标通常呈现在多个尺度上，因此并没有分类器可以利用的特定的尺度语义。由于计算平均并不会增加描述子的维度，我们能够融合大范围尺度的图像描述子：Q∈{256,384,512,640,768}

。值得注意的是，它比使用小范围尺度{256,384,512}的提升是非常小的(0.3%

)。

在测试集上的性能以及与其他方法的比较见Table11。我们的网络“Net-D”和“Net-E”在VOC数据集上展现了相同的性能，它们的组合对结果略有提高。我们在ILSVRC数据集上的预训练模型，展现了优异的图像表达能力，比之前最好的结果[chatfield2014return](https://arxiv.org/abs/1405.3531" \t "https://blog.csdn.net/wspba/article/details/_blank)高出了6%

。值得注意的是[wei2014cnn](https://arxiv.org/abs/1406.5726" \t "https://blog.csdn.net/wspba/article/details/_blank)的方法，使用了2000类的ILSVRC拓展数据集进行预训练，增加了额外的1000个类别，这与VOC数据集很相近，它在VOC-2012上的mAP比我们的结果高出了1%

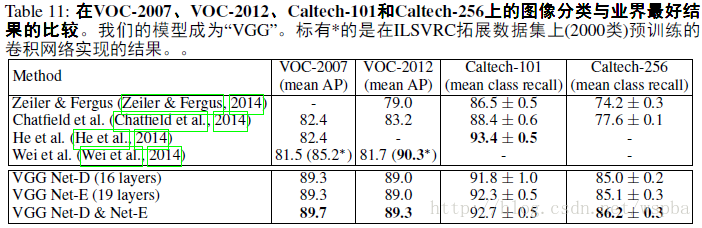
。它也得益于一个辅助目标检测的分类框架流程的融合。

Image Classification on Caltech-101 and Caltech-256 本章，我们在Caltech-101和Caltech-256的图像分类基准上对非常深的特征进行评估。Caltech-101包含9k张图像，一共有102类标签(101个目标种类和一个背景类别)，Caltech-256有31K张图像和257种类别。   
一个标准的评估方案就是将这些数据集随机分成多份训练和测试数据，然后报告多份之间的平均识别性能，识别性能有平均类召回率来衡量(是对每类测试图像数量不相同的补偿)。在Caltech-101上我们随机生成了3份训练和测试数据，每一份中每类含有30张训练图像，以及多达50张的测试图像。在Caltech-256上我们也随机生成3份数据，每一份每类含有60张训练图像(剩下的都用来测试)。在每一份中，20%的训练图像作为验证集，用来进行超参数选择。

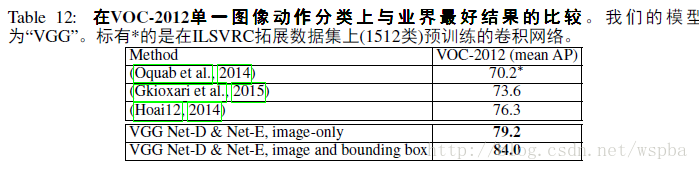
我们发现与VOC不同的是，在Caltech数据集上，将多个尺度上计算的描述子进行堆叠的效果比求平均或者最大池化的效果好。这可能是因为，在Caltech图像中，目标通常占据整幅图像，因此多个尺度上的图像特征是语义不同的(整个目标 vs 目标的一部分)，而堆叠使得分类器能够利用这样特定尺度的表达。我们使用三种尺度：Q∈{256,384,512}

。

Table11 中对我们的模型与业界最好的结果进行了比较。可以看出来，更深的19层的Net-E比16层的Net-D效果更好，它们俩的组合能够进一步的提升性能。在Caltech-101上，我们的表达与[he2014spatial](https://link.springer.com/chapter/10.1007/978-3-319-10578-9_23" \t "https://blog.csdn.net/wspba/article/details/_blank)中的方法性能相近，但是它在VOC-2007上的表现明显比我们的模型差得多。在Caltech-256上，我们的特征比业界最好的结果[chatfield2014return](https://arxiv.org/abs/1405.3531" \t "https://blog.csdn.net/wspba/article/details/_blank)有了很大的提升(8.6%

)。   


Action Classification on VOC-2012 我们同样在PASCAL VOC-2012动作分类任务上评估了我们性能最好的图像表达(Net-D和Net-E特征的堆叠)，包含了在给定人物动作的边界框时对动作类别的预测。数据集包含4.6K张训练图像，共有11类标签。与VOC-2012目标分类任务相似，性能用mAP 来评估。我们考虑以下两种训练设定：   
(i) 在整幅图像上计算卷积网络特征，并忽略提供的边界框；   
(ii) 在整幅图像上和在提供的边界框上计算特征，然后将它们堆叠来获得最终的表达。   
在Table12 中对我们的方法和其他方法的结果进行了比较。

我们的表达在不使用提供的边界框时，在VOC动作分类任务上达到了业界最好的水平，而同时使用图像和边界框时的结果得到了进一步的提升。不同于其他方法，我们并没有使用任何任务特定的探索法，只是依赖于非常深的卷积特征的表达能力。   


Other Recognition Tasks 自从我们的模型公开以来，它们被研究机构广泛的使用在图像识别任务中，并一致的优于浅层表达的性能。例如，[girshick2014rich](https://www.computer.org/csdl/proceedings/cvpr/2014/5118/00/5118a580-abs.html" \t "https://blog.csdn.net/wspba/article/details/_blank)通过将[krizhevsky2012imagenet](http://datascienceassn.org/sites/default/files/ImageNet Classification with Deep Convolutional Neural Networks.pdf" \t "https://blog.csdn.net/wspba/article/details/_blank)中的卷积网络替换成我们的16层模型，达到了目标检测业界最好的结果。在语义分割、图像字幕生成}以及纹理和材料识别都有着类似的结果