



可视化和理解卷积网络

Matthew D. Zeiler和Rob Fergus

美国纽约大学计算机科学系
， 美国

`{zeiler,fergus}@cs.nyu.edu`

摘要。大型卷积网络模型最近在ImageNet基准测试中表现出令人印象深刻的分类性能，Krizhevsky等人[18]。然而，人们对它们为什么表现得如此出色，以及如何改进它们，还没有明确的认识。在本文中，我们探讨了这两个问题。我们引入了一种新的可视化技术，使人们能够深入了解中间特征层的功能和分类器的运作。在诊断的作用下，这些可视化技术使，我们找到了在ImageNet分类基准上优于Krizhevsky等人的模型架构。我们还进行了一项消融研究，以发现不同模型层的性能贡献。我们表明，我们的ImageNet模型在其他数据集上有很好的通用性：当softmax分类器被重新训练时，它在Caltech-101和Caltech-256数据集上令人信服地击败了当前最先进的结果。


1 简介

自从LeCun等人[20]在20世纪90年代初引入卷积网络以来，卷积网络在诸如手写数字分类和人脸检测等任务中表现出了卓越的性能。在过去的18个月里，有七篇论文表明，它们在更具挑战性的视觉分类任务上也能提供出色的表现。Ciresan等人[4]在NORB和CIFAR-10数据集上展示了最先进的性能。最值得注意的是，Krizhevsky等人[18]在ImageNet 2012分类基准上展示了创纪录的性能，他们的convnet模型实现了16.4%的错误率，而第二名的结果是26.1%。继这项工作之后，Girshick等人[10]在PASCAL VOC数据集上显示了领先的检测性能。有七种因素导致了性能的大幅提高：(i) 有了更大的训练集，有了数以百万计的标记实例；(ii) 强大的GPU实现，使训练非常大的模型成为可能；(iii) 更好的模型正则化策略，如Dropout[14]。

尽管取得了这一令人鼓舞的进展，但人们对这些复杂模型的内部运作和行为仍然缺乏深入了解，也不知道它们是如何取得如此好的性能的。从科学的角度

来看，这是很不令人满意的。如果不清楚它们是如何和为什么工作的，那么更好的模型的开发就会沦为试错。在本文中，我们介绍了一个可视化的

D.Fleet等人(Eds.)。ECCV 2014, Part I, LNCS 8689, pp.818-833, 2014.

 Springer International Publishing Switzerland 2014

该技术揭示了在模型的任何一层激发单个特征图的输入刺激。它还允许我们在训练期间观察特征的演变，并诊断模型的潜在问题。我们提出的视觉化技术使用Zeiler等人[29]提出的多层去卷积网络（deconvnet），将特征激活投射回输入像素空间。我们还通过遮挡输入图像的部分内容对分类器的输出进行了敏感性分析，揭示了场景的哪些部分对分类很重要。

利用这些工具，我们从Krizhevsky等人[18]的架构开始，探索不同的架构，发现在ImageNet上的结果优于他们。然后，我们探索该模型对其他数据集的泛化能力，只是在上面重新训练softmax分类器。因此，这是一种有监督的预训练，与Hinton等人[13]和其他人[1,26]所推广的无监督预训练方法形成鲜明对比。

1.1 相关工作

可视化。将特征可视化以获得对网络的直觉是常见的做法，但主要限于第一层，在那里可以投射到像素空间。在较高的层中，必须使用其他方法。[8]通过在图像空间中进行梯度下降，找到每个单元的最佳刺激，使单元的激活最大化。这需要仔细的初始化，并且不提供任何关于单元不变性的信息。受后者缺点的启发，[19]（扩展了[2]的想法）展示了如何围绕最优响应数值计算给定单元的Hessian，给出了一些关于不变性的洞察力。问题是，对于较高的层，不变性是非常复杂的，所以很难被简单的二次近似所捕获。相比之下，我们的方法提供了一个非参数化的不变性观点，显示了训练集的哪些模式激活了特征图。我们的方法类似于Simonyan等人[23]的当代工作，他们展示了如何通过从网络的全连接层向后投射，而不是我们使用的卷积特征，从convnet中获得显著性地图。Girshick等人[10]展示了在数据集中识别出对模型中较高层的强激活负责的斑块的可视化。我们的可视化的不同之处在于，它们不仅仅是输入图像的裁剪，而是自上而下的投影，揭示了每个斑块内刺激特定特征图的结构。

特征泛化。Donahue等人[7]和Girshick等人[10]在同时进行的工作中也探讨了我们对convnet特征泛化能力的证明。他们使用convnet特征在Caltech-101和太阳场景数据集上获得了最先进的性能，而在后者，则是在PASCAL VOC数据集上进行物体检测。

2 办法

我们在本文中使用标准的完全监督的Convnet模型，如LeCun等人[20]和Krizhevsky等人[18]所定义的。这些模型将一个颜色

二维输入图像 x_i ，通过一系列的层，到一个概率向量 \hat{y}_i 在 C 不同的类别。每一层包括：(i) 将前一层的输出（或在第一层的情况下，输入图像）与一组学习过的滤波器进行卷积；(ii) 将响应通过一个整流的线性函数（ $\text{relu}(x) = \max(x, 0)$ ）。

(iii)[可选择]局部邻域的最大集合和(iv)[可选择]局部对比操作，使各特征图的反应正常化。关于这些操作的更多细节，见[18]和[16]。网络的前几层是传统的全连接网络，最后一层是一个softmax分类器。图3显示了我们许多实验中使用的模型。

我们使用一大组有标签的图像 $\{x, y\}$ 来训练这些模型，其中标签 y_i 是一个表示真实类别的离散变量。一个适用于图像分类的交叉熵损失函数，用来比较 \hat{y}_i 和 y_i 。网络的参数（卷积层的滤波器、全连接层的权重矩阵和偏置）是通过反向传播损失对整个网络参数的导数来训练的，并通过随机梯度下降更新参数。训练的细节在第3节给出。

2.1 用一个Deconvnet进行可视化

理解convnet的运行需要解释中间层的特征激活。我们提出了一种新的方法，将这些活动映射到输入像素空间，显示什么输入模式最初导致了特征图中的特定激活。我们用一个去卷积网络（deconvnet）Zeiler等人[29]来进行这种映射。去卷积网络可以被认为是一个使用相同组件（排序、集合）的卷积网络模型，但是是反向的，所以不是把像素映射到特征，而是做相反的事情。在Zeiler等人[29]中，去神经网络被提议作为一种执行无监督学习的方式。在这里，他们不以任何学习的身份使用，只是作为一个已经训练好的convnet的探测。

为了检查一个convnet，一个deconvnet被连接到它的每一层，如图1（顶部）所示，提供一个回到图像像素的连续路径。开始时，一个输入图像被提交给Convnet，并在各层中计算出特征。为了检查一个给定的convnet激活，我们将该层的所有其他激活设置为零，并将特征图作为输入传递给附属的deconvnet层。然后，我们依次(i)解池，(ii)整顿和(iii)过滤，以重建下面一层中产生所选激活的活动。然后重复这一过程，直到达到输入像素空间。

解除集合。在convnet中，最大集合操作是不可逆转的，但是我们可以通过在一组开关变量中记录每个集合区域内的最大值的位置来获得一个近似的逆转。在deconvnet中，unpooling操作使用这些开关，将上面一层的重建置于适当的位置，保留了刺激的结构。请看图1（底部）对该程序的说明。

矫正。convnet使用非线性，对特征图进行矫正，从而确保特征图始终为正值。

为了获得有效的

在每一层的特征重构（也应该是正的），我们将重构的信号通过一个非线性的 *relu*¹。

过滤。convnet使用学习过的filters对上一层的特征图进行卷积。为了近似地反转这一点，deconvnet使用相同的滤波器的转置版本（如其他自动编码器模型，如RBMs），但应用于矫正的地图，而不是下面一层的输出。在实践中，这意味着每一个滤波器都要在垂直和水平方向上翻转。

注意，在这个重建路径中，我们不使用任何对比度归一化操作。从高层往下投射时，使用的是上层convnet中的最大集合所产生的开关设置。由于这些开关设置是特定的输入图像所特有的，因此从单一的激活中得到的重建结果类似于原始输入图像的一小部分，其结构是根据其对特征行为的贡献而加权的。由于模型的训练是鉴别性的，它们隐含地显示了输入图像的哪些部分是鉴别性的。请注意，这些投影不是模型的样本，因为没有涉及生成过程。整个过程类似于反推一个单一的强激活（而非通常的梯度），即计算 $\frac{\partial z}{\partial x_n}$ ，其中 h 是特征图的元素具有强激活作用， x_n 是输入图像。然而，它在以下方面存在差异

(i) *relu*是独立施加的；(ii) 不使用对比度正常化操作。我们的方法有一个普遍的缺点，就是它只对单一的激活进行可视化，而不是对一个层中存在的联合活动进行可视化。尽管如此，正如我们在图6中所示，这些可视化是刺激模型中给定特征图的输入模式的准确表示：当原始输入图像中与该模式相对应的部分被遮挡时，我们看到特征图中的活动明显下降。

3 培训详情

我们现在描述一下大型的convnet模型，它将在第4节中被可视化。图3所示的结构与Krizhevsky等人[18]用于ImageNet分类的结构相似。其中的区别是，Krizhevsky的第3、4、5层使用的稀疏连接（由于该模型被分割在两个GPU上）在我们的模型中被密集连接所取代。其他与第1层和第2层有关的重要差异是在检查了图5中的可视化后得出的，如第4.1节所述。

该模型是在ImageNet 2012训练集（130万张图像，分布在1000个不同的类别中）[6]上训练的。每张RGB图像都经过预处理，将最小维度调整为256，裁剪中心256x256区域，减去每像素平均值（所有图像），然后使用10个大小为224x224的子裁剪（角+中心有（无）水平褶皱）。随机梯度下降法的小型批次大小为128，用于更新参数，开始时的学习率为 10^{-2} ，同时动量项为0.9。我们

¹我们还试着用前馈回流施加的二进制掩码进行整流。

操作，但产生的可视化效果明显不那么清晰。

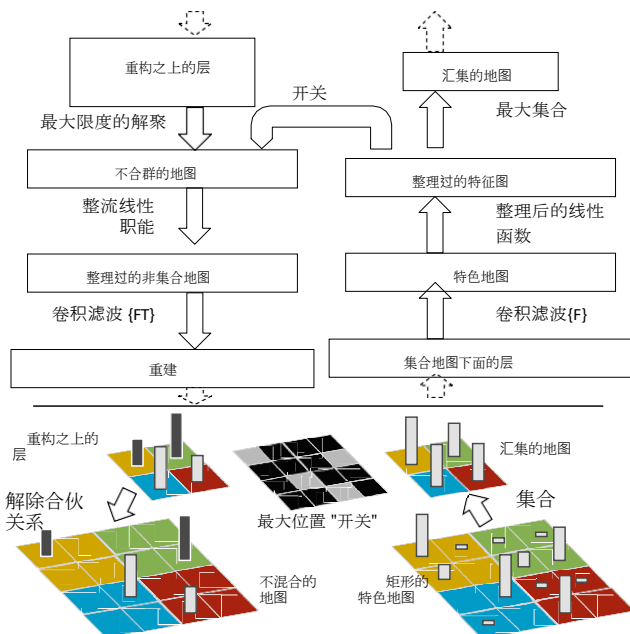


图1.上图：一个连接到convnet层（右）的deconvnet层（左）。deconvnet将从下面的层中重建一个近似版本的convnet特征。底部。解集网中的解集操作说明，使用开关记录了在convnet中解集时每个解集区域（彩色区域）的局部最大值的位置。黑/白条是特征图中的负/正激活。

在整个训练过程中，当验证误差趋于平稳时，手动退火学习率。Dropout[14]被用于全连接层（6和7），速率为0.5。所有的权重被初始化为 10^{-2} ，偏置被设置为0。

在训练过程中，第一层薄膜的可视化显示，少数的它们占主导地位。为了解决这个问题，我们将卷积，其RMS值超过10的固定半径的每个滤波器重新归一到这个固定半径。这一点至关重要，尤其是在模型的第一层，输入的图像大致在 $[-128, 128]$ 范围内。正如Krizhevsky等人[18]一样，我们制作了多个不同的作物和每个训练实例的flips，以提高训练集的大小。我们在70个历时后停止训练，使用基于[18]的实现，在单个GTX580 GPU上花费了大约12天。

4 Convnet可视化

使用第3节中描述的模型，我们现在使用deconvnet来可视化ImageNet验证集上的特征激活。

特征可视化。图2显示了训练完成后我们模型的特征可视化。对于一个给定的特征图，我们显示了前9个行为，每个行为都分别投射到像素空间，揭示了不同的

特征。

激发该地图的结构，并显示其对输入变形的不变性。在这些可视化的过程中，我们显示了相应的图像斑块。与只关注每个斑块内的判别结构的可视化相比，这些斑块的变化更大。例如，在第5层第1行第2列，这些斑块似乎没有什么共同点，但可视化显示，这个特定的特征图关注的是背景中的草，而不是前景物体。

各层的投影显示了网络中特征的层次性。第2层对角落和其他边缘/颜色的结点作出反应。第3层有更复杂的不变性，捕捉类似的纹理（如网状图案（第1行，第1列）；文字（R2,C4））。第4层显示了显著的变化，而且更具有阶级特征：狗脸（R1,C1）；鸟腿（R4,C2）。第5层显示了具有明显姿势变化的整个物体，例如，键盘（R1,C11）和狗（R4）。

训练期间的特征演变 on:图4显示了在训练过程中，在一个给定的特征图中最强的激活（跨越所有的训练例子）投射到像素空间的进展。外观上的突然跳动是由于最强激活的图像发生了变化。可以看到模型的下层在几个历时内就会收敛。然而，上层只有在相当多的历时（40-50）后才会发展起来，这表明需要让模型训练到完全收敛。

4.1 建筑选择

训练过的模型的可视化能让人深入了解其运行情况，它也能帮助人们首先选择好的架构。通过对Krizhevsky等人的架构的第一层和第二层的可视化（图5（a）和（c）），各种问题是显而易见的。第一层薄膜混合了极高和极低的频率信息，几乎没有对中频的覆盖。此外，第二层的可视化显示了由第一层卷积中使用的大跨度4引起混叠伪影。为了解决这些问题，我们(i)将第一层滤波器的尺寸从11x11减少到7x7，(ii)将卷积的步长定为2，而不是4。这个新的结构在第一和第二层特征中保留了更多的信息，如图5（b）和（d）所示。更重要的是，它还提高了分类性能，如第5.1节所示。

4.2 遮蔽敏感度

对于图像分类方法，一个很自然的问题是，该模型是真正识别图像中物体的位置，还是仅仅利用周围的环境。图6试图回答这个问题，通过系统地将输入图像的不同部分用灰色方块遮挡起来，并监测分类器的输出。这些例子清楚地表明，该模型正在对场景中的物体进行定位，因为当物体被遮挡时，正确类别的概率明显下降。图6还显示了顶级卷积层的最强特征图的可视化，以及该图中的活动（空间位置的总和）与遮挡者位置的关系。当



图2.完全训练好的模型中的特征的可视化。对于第2-5层，我们显示了整个验证数据集中的随机特征图子集的前9个激活点，使用我们的去卷积网络方法投射到像素空间。我们的重构不是来自模型的样本：它们是来自验证集的重构模式，在给定的特征图中引起高激活。对于每个特征图，我们也显示了相应的图像斑块。注意：(i)每个特征图中的强分组，(ii)在较高的层中有更大的不变性，(iii)图像中可判别的部分被夸大，例如狗的眼睛和鼻子（第4层，第1行，第1列）。最好以电子形式观看。压缩假象是30Mb提交限制的结果，而不是重建算法本身。

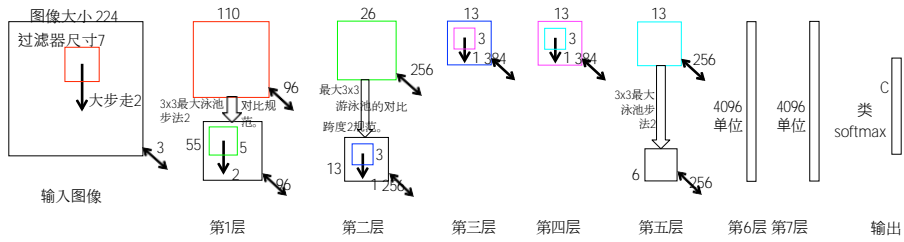


图3.我们的8层convnet模型的结构。一幅 224×224 的图像（有3个颜色平面）被作为输入。然后，所得到的特征图：(i) 通过一个矩形线性函数（未显示），(ii) 汇集（在 3×3 区域内最大，使用stride 2）和(iii) 对比度归一化的特征图，得到96个不同的 55×55 元素特征图。类似的操作在第2、3、4、5层重复。最后两层是完全连接的，从以下几方面获取特征

在顶层卷积层中，以矢量形式输入（ $6-6-256=9216$ 维）。最后一层是一个C-way softmax函数，C是类别的数量。所有卷积层和特征图的形状是方形的。

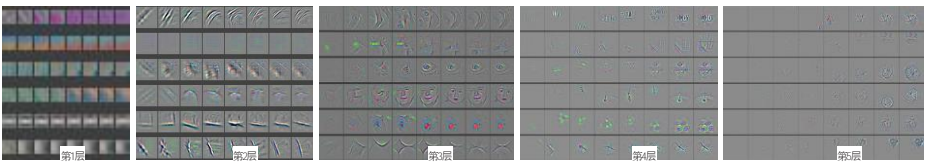


图4.通过训练，随机选择的模型特征子集的演变。每个层的特征都显示在不同的区块中。在每个区块内，我们显示了在历时[1,2,5,10,20,30,40,64]随机选择的特征子集。可视化显示了一个给定的特征图的最强激活（在所有训练实例中），使用我们的解网络方法投影到像素空间。色彩对比度是，该图最好以电子形式观看。

遮挡者覆盖了出现在可视化中的图像区域，我们看到特征图的活动强烈下降。这表明，可视化真正对应于刺激该特征图的图像结构，因此验证了图4和图2所示的其他可视化。

5 实验

5.1 ImageNet 2012

这个数据集由1.3M/50K/100K训练/验证/测试实例组成，分布在1000个类别上。表1显示了我们在这个数据集上的结果。

使用Krizhevsky等人[18]的确切架构，我们试图在验证集上复制他们的结果。在ImageNet 2012验证集上，我们的错误率在0.1%以内。

接下来，我们分析了我们的模型在第4.1节中概述的架构变化（第1层的 7×7 滤

波器和第2层的跨步卷积)下的性能。

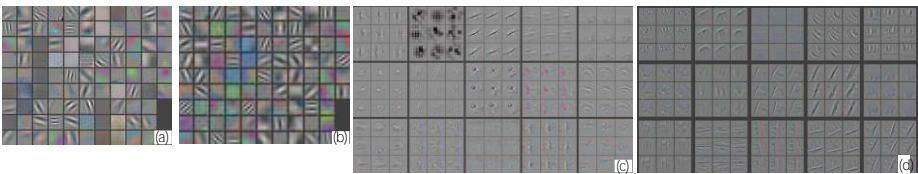


图5 (a): 第一层的特征没有进行特征尺度剪切。请注意，一个特征占主导地位。(b): 第一层特征来自 Krizhevsky 等人[18]。(c):我们的第一层特征。较小的步长（2 vs 4）和滤波器尺寸（7x7 vs 11x11）导致了更独特的特征和更少的 "死 "特征。(d):Krizhevsky 等人的第二层特征的可视化[18]。(e):我们第二层特征的可视化。这些更干净，没有(d)中可见的混叠伪影。

1 & 2).这个模型，如图3所示，明显优于Krizhevsky 等人[18]的架构，比他们的单一模型结果高出1.7%（测试前5名）。当我们结合多个 模型时，我们得到的测试误差为14.8%，提高了1.6%。这个结果接近于Howard[15]的数据增强方法所产生的结果，它可以很容易地与我们的结构相结合。然而，我们的模型与2013年Imagenet分类竞赛的冠军相比还有一定差距[28]。

表1.ImageNet 2012/2013分类错误率。*表示在ImageNet 2011和2012训练集上训练的模型。

误差百分比	瓦尔 顶-1	瓦尔 前五 名	测试 前五 名
Gunji 等人[12]。	-	-	26.2
DeCAF [7]	-	-	19.2
Krizhevsky 等人[18], 1 convnet	40.7	18.2	--
Krizhevsky 等人[18], 5个 convnets	38.1	16.4	16.4
Krizhevsky 等人* [18], 1个 convnets	39.0	16.6	--
。Krizhevsky 等人* [18], 7个 convnets	36.7	15.4	15.3
我们复制的			
Krizhevsky 等人, 1厥网	40.5	18.1	--
按图3所示, 1个定罪网	38.4	16.5	--
根据图3-(a), 有5个 convnets。	36.7	15.3	15.3
按图3, 1个定罪网, 但有			
第3,4,5层 : 512,1024,512地图 - (b)	37.5	16.0	16.1
6个信念网, (a)和(b)相结合	36.0	14.7	14.8
霍华德[15]	-	-	13.5
克拉里费[28]	-	-	11.7

不同的ImageNet模型大小。在表2中，我们首先探索了Krizhevsky 等人[18]的架构，调整了层的大小，或完全删除了它们。在每一种情况下，模型都是用修改后

的结构从头开始训练。去掉全连接层 (6,7) 只会使误差略有增加 (在

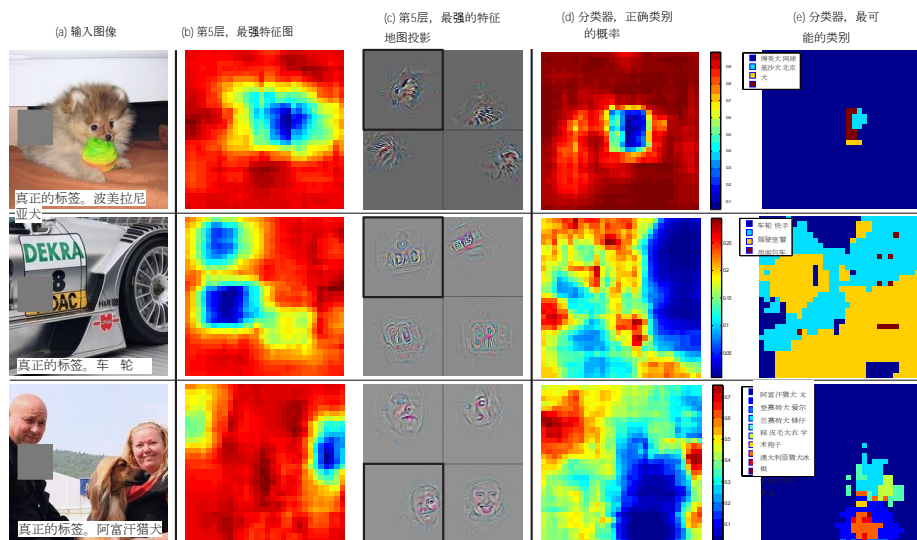


图6.三个测试例子，我们系统地用灰色方块遮盖场景的不同部分（第一列），看看顶部（第5层）特征图（（b）和（c））和分类器输出（（d）和（e））如何变化。（b）：对于灰度的每个位置，我们记录一个第5层特征图的总激活度（在未排除的图像中反应最强的那个）。（c）：这个特征图投射到输入图像（黑色方块）的可视化，以及这个图在其他图像中的可视化。第一行的例子显示最强的特征是狗的脸。当这一特征被掩盖时，该特征图的活动就会减少（（b）中的蓝色区域）。（d）：正确类别概率图，作为灰色方块位置的一个函数。例如，当狗的脸被遮住时，“博美犬”的概率明显下降。（e）：最可能的标签是遮挡者位置的一个函数。例如，在第一行中，对于大多数位置，它是“博美”，但如果狗的脸被遮住了，但球没有被遮住，那么它就预测为“网球”。在第2个例子中，汽车上的文字是第5层中最强的特征，但分类器对车轮最敏感。第三个例子包含多个物体。第5层中最强的特征是挑选出人脸，但分类器对狗（（d）中的蓝色区域）很敏感，因为它使用了多个特征图。

以下，我们指的是前5名的验证误差）。这是令人惊讶的，因为它们包含了大部分的模型参数。去掉中间卷积层中的两个，也会对错误率产生相对较小的影响。然而，去掉中间卷积层和全连接层后，得到的模型只有4层，其性能就会大大降低。这表明，模型的整体深度对于获得良好的性能是很重要的。然后我们修改我们的模型，如图3所示。改变全连接层的大小对性能没有什么影响（Krizhevsky等人的模型也是如此[18]）。然而，增加中间卷积层的大小会给性能带来有益的提高。但是，在增加这些层的同时，也扩大了全连接层的规模，导致了过度罚款。

表2.ImageNet 2012 在对Krizhevsky 等人[18]的模型和我们的模型进行各种架构改变后的分类错误率（见图3）。

误差百分比	火车 顶-1	瓦尔 顶-1	瓦尔 前五 名
我们对Krizhevsky 等人[18]的复制, 1 convnet	35.1	40.5	18.1
删除了第3,4层	41.8	45.4	22.1
删除了第7层	27.4	40.0	18.4
删除了第6,7层	27.4	44.8	22.4
删除了第3、4、6、7层	71.1	71.3	50.1
调整图层6,7: 2048个单位	40.3	41.7	18.8
调整图层6,7: 8192单位	26.8	40.0	18.1
我们的模型（如图3）	33.1	38.4	16.5
调整图层6,7: 2048个单位	38.2	40.2	17.6
调整图层6,7: 8192单位	22.0	38.8	17.0
调整图层3、4、5 : 512、1024、512地图	18.8	37.5	16.0
调整图层6,7 : 8192单位和 第3、4、5层 : 512、1024、512地图	10.0	38.3	16.9

5.2 特征归纳

上述实验表明，我们的ImageNet模型的卷积部分对获得最先进的性能非常重要。图2的可视化显示了在卷积层中学习到的复杂不变性，这一点得到了支持。我们现在探索这些特征提取层对其他数据集的概括能力，即Caltech-101[9]、Caltech-256[11]和PASCAL VOC 2012。为了做到这一点，我们保持ImageNet训练模型的第1-7层不变，并使用新数据集的训练图像在上面训练一个新的softmax分类器（针对适当数量的类）。由于softmax包含的参数相对较少，它可以从相对较少的例子中快速训练出来，对于某些数据集就是如此。

实验将我们从ImageNet获得的特征表示与其他方法使用的手工制作的特征进行了比较。在我们的方法和现有的方法中，Caltech/PASCAL的训练数据只被用来训练分类器。由于它们的复杂程度相似（我们的是softmax，其他的是linear SVM），因此特征表示对性能至关重要。值得注意的是，这两种表征都是用Caltech和PASCAL训练集以外的图像建立的。例如，HOG描述符中的超参数是通过对行人数据集的系统实验确定的[5]。

我们还尝试了从头开始训练模型的第二种策略，即把第1-7层重设为随机值，在PASCAL/Caltech数据集的训练图像上训练它们以及softmax。

一个复杂的问题是，Caltech的一些数据集有一些图像也在ImageNet的训练数据中。使用归一化的相关度，我们

识别了这些少数的 "重叠 "图像²并将它们从我们的Imagenet训练集中删除，然后重新训练我们的Imagenet模型，从而避免了训练/测试污染的可能性。

Caltech-101：我们遵循[9]的程序，每类随机选择15或30幅图像进行训练和测试，每类最多50幅图像，在表3中报告每类的平均准确率，使用5个训练/测试折页。每类30张图像的训练时间为17分钟。预先训练的模型比[3]中30张图像ages/class的最佳re-reported结果多出2.2%。我们的结果与Donahue 等人[7]最近发表的结果一致，他们获得了86.1%的准确性（30张图片/类）。然而，从头开始训练的convnet模型表现得很糟糕，只达到了46.5%，这表明在这么小的数据集上训练一个大型convnet是不可能的。

表3.我们的Convnet模型与两种领先的替代方法相比，Caltech-101的分类精度

# # 火车	累计百分比 15/班	累计百分比 30/班
Bo 等人[3]。	-	81.4 ± 0.33
Yang 等人[17]。	73.2	84.3
非预科生	22.8 ± 1.5	46.5 ± 1.7
图像网训练的convnet	83.8 ± 0.5	86.5 ± 0.5

Caltech-256：我们按照[11]的程序，每类选择15、30、45或60张训练图像，在表4中报告每类的平均准确率。我们的ImageNet预训练模型比Bo 等人[3]取得的最先进的结果要好得多：60张训练图像/类，74.2%比55.2%。然而，与Caltech-101一样，从头开始训练的模型表现不佳。在图7中，我们探讨了 "一次性学习"[9]制度。使用我们预先训练好的模型，只需要6张Caltech-256的训练图像就可以打败使用10倍图像的领先方法。这显示了ImageNet特征提取器的力量。

PASCAL 2012。我们使用标准的训练图像和验证图像，在 ImageNet-retrained convnet的基础上训练一个20路softmax。这并不理想，因为PASCAL 图像，可能包含多个对象，而我们的模型只是为每张图像提供一个独家预测。表5显示了测试集的结果，与领先的方法进行了比较：比赛中的前两名作品和Oquab 等人[21]的同期工作，他们使用了一个具有更合适分类器的convnet。PASCAL和ImageNet图像在性质上有很大不同，前者是完整的场景，而后者则不同。这可能解释了我们的平均值

²对于Caltech-101，我们发现了44张共同的图像（在9,144张总图像中），任何特定类别的最大重叠量为10张。对于Caltech-256，我们发现了243张共同的图像（在30,607张

总的图像中)，任何特定类别的最大重叠量为18。

表4.Caltech 256分类 cation 准确率

# # 火车	累计百分比 15/班	累计百分比 30/班	累计百分比 45/班	累计百分比 60/班
Sohn 等人[24]。	35.1	42.1	45.7	47.9
Bo 等人[3]。	40.5 ± 0.4	48.0 ± 0.2	51.9 ± 0.2	55.2 ± 0.3
非预科。	9.0 ± 1.4	22.5 ± 0.7	31.2 ± 0.5	38.8 ± 1.4
ImageNet-pretr.	65.7 ± 0.2	70.6 ± 0.2	72.7 ± 0.4	74.2 ± 0.3

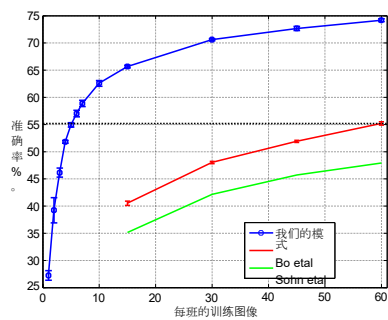


图7.Caltech-256的分类性能随着每类训练图像数量的变化而变化。使用我们预先训练的特征提取器，每类只用6个训练实例，我们超过了Bo 等人[3]的最佳报告结果。

表5.PASCAL 2012分类结果，将我们的Imagenet训练的con-vnet与领先的两种方法和Oquab 等人的最新方法进行比较[21]。

累计百分比	[22]	[27]	[21]	我们的	累计百分比	[22]	[27]	[21]	我们的
飞机	92.0	97.3	94.6	96.0	餐桌	63.2	77.8	69.0	67.7
自行车	74.2	84.2	82.9	77.1	犬类	68.9	83.0	92.1	87.8
鸟类	73.0	80.8	88.2	88.4	马	78.2	87.5	93.4	86.0
船只	77.5	85.3	60.3	85.5	摩托车	81.0	90.1	88.6	85.1
瓶子	54.3	60.8	60.3	55.8	个人	91.6	95.0	96.1	90.9
公交车	85.2	89.9	89.0	85.8	盆栽	55.9	57.8	64.3	52.2
汽车	81.9	86.8	84.4	78.6	绵羊	69.4	79.2	86.6	83.6
猫咪	76.4	89.3	90.7	91.2	沙发	65.4	73.4	62.3	61.1
椅子	65.2	75.4	72.1	65.0	火车	86.7	94.5	91.1	91.8
牛	63.2	77.8	86.8	74.4	电视机	77.4	80.7	79.8	76.1
平均值	74.3	82.2	82.8	79.0	# 赢得了	0	11	6	3

性能比领先的竞争结果[27]低3.2%，然而我们确实在5个级别上击败了他们，有时差距很大。

5.3 特征分析

我们探讨了在我们的Imagenet预训练模型的每一层中的特征有多大的辨别力。我们通过改变ImageNet模型的层数，，将线性SVM或softmax分类器放在上面。表6显示了Caltech-101和Caltech-256的结果。对于这两个数据集，随着我们对模型的提升，可以看到一个稳定的改进，使用所有层可以获得最好的结果。这支持了这样一个前提，即随着特征层次的加深，它们会学到越来越强大的特征。

表6.在我们的ImageNet-pretrained convnet中，每层特征图所包含的判别信息的分析。我们对来自Confnet不同层（如括号内所示）的特征进行线性SVM或softmax训练。较高的层数通常会产生更多的区分性特征。

	Cal-101 (30/班)	Cal-256 (60/班)
证券交易所 (1)	44.8 ± 0.7	24.6 ± 0.4
证券交易所 (2)	66.2 ± 0.5	39.6 ± 0.3
证券交易所 (3)	72.3 ± 0.4	46.0 ± 0.3
证券交易所 (4)	76.6 ± 0.4	51.3 ± 0.1
证券交易所 (5)	86.2 ± 0.8	65.6 ± 0.3
视觉识别系统 (7)	85.5 ± 0.4	71.7 ± 0.2
软体公司 (5)	82.9 ± 0.4	65.7 ± 0.5
软体动物 (7)	85.4 ± 0.4	72.6 ± 0.1

6 讨论

我们以多种方式探索了为图像分类训练的大型卷积神经网络模型。首先，我们提出了一种新的方法，将模型内的活动可视化。这揭示了这些特征远不是随机的、不可解释的模式。相反，它们显示了许多直观的理想属性，如构成性、不断增加的不变性和随着我们层层上升的类别区分。我们还展示了这些可视化如何被用来识别模型的问题，从而获得更好的结果，例如改进Krizhevsky等人的[18]令人印象深刻的ImageNet 2012结果。然后，我们通过一系列的闭塞实验证明，该模型在进行分类训练的同时，对图像中的局部结构高度敏感，而不仅仅是利用广泛的场景背景。对该模型的消融研究表明，网络的最小深度，而不是任何单独的部分，对该模型的性能至关重要。

最后，我们展示了ImageNet训练的模型如何能够很好地推广到其他数据集。对于Caltech-101和Caltech-256，这些数据集足够相似，以至于我们可以击败最好的报告结果，在后者的情况下，我们有相当大的优势。我们的convnet模型对PASCAL数据的概括性较差，可能是

尽管没有对该任务进行调整，但我们的性能仍然在3.2%的最佳报告结果范围内，这与数据集偏差有关[25]。例如，如果使用一个不同的损失函数，允许每幅图像有多个目标，我们的性能可能会有所提高。这自然会使网络也能解决物体。

Acknowledgments. 作者要感谢Yann LeCun的有益讨论和，感谢NSERC、NSF grant #1116923和Microsoft Research的支持。

参考文献

1. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: NIPS, 第153-160页(2007)
2. Berkes, P., Wiskott, L.: On analysis and interpretation of inhomogeneous quadratic forms as receptive fields. 神经计算 (2006)
3. Bo, L., Ren, X., Fox, D.: 使用分层匹配追求的多路径稀疏编码。 In: CVPR (2013)
4. Ciresan, D.C., Meier, J., Schmidhuber, J.: 用于图像分类的多列深度神经网络。 In: CVPR (2012)
5. Dalal, N., Triggs, B.: 用于行人检测的定向梯度直方图。 In: CVPR (2005)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: 一个大规模的分层图像数据库。 In: cvpr 2009 (2009)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. arXiv: 1310.1531 (2013)
8. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: 深度网络的高层特征的可视化。 技术报告，蒙特利尔大学(2009)
9. Fei-fei, L., Fergus, R., Perona, P.: 物体类别的一次性学习。 IEEE Trans. PAMI (2006)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv: 1311.2524 (2014)
11. Griffin, G., Holub, A., Perona, P.: Caltech 256. 加州理工学院技术报告 (2006)
12. Gunji, N., Higuchi, T., Yasumoto, K., Muraoka, H., Ushiku, Y., Harada, T., Kuniyoshi, Y.: 分类条目。 Imagenet竞赛 (2012)
13. Hinton, G.E., Osindero, S., Teh, Y.: 深度信念网的快速学习算法。 神经计算 18, 1527-1554 (2006)
14. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: 通过防止特征检测器的共同适应来改进神经网络。 In: arXiv:1207.0580 (2012)
15. Howard, A.G.: 基于深度卷积神经网络的图像分类的一些改进。 arXiv 1312.5402 (2013)
16. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: 什么是物体识别的最佳多阶段结构？ In: ICCV (2009)

17. Jianchao, Y., Kai, Y., Yihong, G., Thomas, H.: 利用稀疏编码进行图像分类的线性空间金字塔匹配。In:CVPR (2009)

18. Krizhevsky, A. , Sutskever, I., Hinton, G.: 用深度对话神经网络进行图像分类。 In:NIPS (2012)
19. Le, Q.V., Ngiam, J., Chen, Z., Chia, D., Koh, P., Ng, A.Y. 。 Tiled convolutional neural networks.在。 NIPS (2010)
20. LeCun, Y., Bottou, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D. 。 应用于手写邮政编码识别的反向传播法。 Neural Comput.1(4), 541-551 (1989)
21. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: 使用卷积神经网络学习和转移中层图像表征。 In:CVPR (2014)
22. Sande, K., Uijlings, J., Snoek, C., Smeulders, A.: Hybrid coding for selective search.In:2012年PASCAL VOC分类挑战赛(2012)
23. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks:Visualising图像分类模型和显著性地图。 arXiv 1312.6034v1 (2013)
24. Sohn, K., Jung, D., Lee, H., Hero III, A.: 用于物体识别的稀疏、分布式、卷积feature表征的古代学习。 In:ICCV (2011)
25. Torralba, A., Efros, A.A.: Unbiased look at dataset bias.In:CVPR (2011)
26. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: 用去噪自动编码器提取和合成稳健特征。 In:ICML, 第1096-1103页(2008)
27. Yan, S., Dong, J., Chen, Q., Song, Z., Pan, Y., Xia, W., Huang, Z., Hua, Y., Shen, S.: 用于子类别感知的物体分类的通用层次匹配。 In:PASCAL VOC Classification Challenge 2012 (2012)
28. Zeiler, M.: Clarifai (2013), <http://www.image-net.org/challenges/LSVRC/2013/results.php>
29. Zeiler, M., Taylor, G., Fergus, R.: 用于中高级特征学习的自适应去卷积网络。 在。 ICCV (2011)