用深度卷积神经网络进行ImageNet分类

Alex Krizhevsky
多伦多大学
kriz@cs.utoronto.ca

伊利亚-苏茨基弗
多伦多大学
ilya@cs.utoronto.ca

Geoffrey E. Hinton
多伦多大学
hinton@cs.utoronto.ca

摘要

我们训练了一个大型的深度卷积神经网络，将ImageNet LSVRC-2010竞赛中的120万张高分辨率图像分类到1000个不同的类别。在测试数据上，我们取得了37.5%和17.0%的前1名和前5名的错误率，大大优于之前的最先进水平。该神经网络有6000万个参数和65万个神经元，由5个卷积层组成，其中一些是最大集合层，还有3个全连接层，最后是1000路softmax。为了使训练速度更快，我们使用了非饱和神经元和非常有效的GPU实现卷积操作。为了减少全连接层中的过度拟合，我们采用了最近开发的名为 "dropout "的正则化方法，该方法被证明是非常有效的。我们还将这个模型的一个变体参加了ILSVRC-2012比赛，并取得了15.3%的前五名测试错误率，而第二好的作品则达到了26.2%。

1 简介

目前的物体识别方法主要使用机器学习方法。为了提高它们的性能，我们可以收集更大的数据集，学习更强大的模型，并使用更好的技术来防止过拟合。直到最近，标记图像的数据集相对较小--在数万张图像的数量级上（例如NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]）。简单的识别任务可以通过这种规模的数据集得到很好的解决，特别是如果它们得到了标签保护性转换的增强。例如，目前MNIST数字识别任务的最佳错误率（<0.3%）接近人类的表现[4]。但是现实环境中的物体表现出相当大的差异性，所以要学会识别它们，就必须使用更大的训练集。事实上，小型图像数据集的缺点已经被广泛认可（例如，Pinto等人[21]），但直到最近才有可能用数百万的图像来标注数据集。新的大型数据集包括LabelMe[23]，它由数十万张完全分割的图像组成，以及ImageNet[6]，它由超过22000个类别的1500万张标记的高分辨率图像组成。

为了从数以百万计的图像中学习成千上万的物体，我们需要一个具有较大学习能力的模型。然而，物体识别任务的巨大复杂性意味着即使像ImageNet这样大的数据集也不能说明这个问题，所以我们的模型也应该有很多先验知识来弥补我们没有的数据。卷积神经网络（CNN）就是这样一类模型[16, 11, 13, 18, 15, 22, 26]。它们的能力可以通过改变其深度和广度来控制，而且它们还对图像的性质做出了强有力的、大部分正确的假设（即统计的静止性和像素依赖的位置性）。因此，与具有类似大小的层的标准前馈神经网络相比，CNN的连接和参数要少得多，因此它们更容易训练，而其理论上的最佳性能可能只是略差。

尽管CNN具有吸引人的特质，尽管它们的局部结构相对高效，但它们在大规模应用于高分辨率图像时的成本仍然过高。幸运的是，目前的GPU与高度优化的二维卷积实现相搭配，足以促进有趣的大型CNN的训练，而且最近的数据集，如ImageNet，包含足够的标记实例来训练这样的模型，而没有严重的过拟合。

本文的具体贡献如下：我们在ILSVRC-2010和ILSVRC-2012比赛中使用的ImageNet子集上训练了迄今为止最大的卷积神经网络之一[2]，并取得了迄今为止在这些数据集上报告的最佳结果。我们编写了一个高度优化的二维卷积和训练卷积神经网络的所有其他固有操作的GPU实现，我们公开提供了[1]。我们的网络包含一些新的和不寻常的功能，这些功能提高了它的性能并减少了它的训练时间，这些将在第3节中详细介绍。我们的网络规模使得过拟合成为一个重要的问题，即使有120万个标记的训练例子，所以我们使用了几种有效的技术来防止过拟合，这些将在第4节中描述。我们最终的网络包含5个卷积层和3个全连接层，这个深度似乎很重要：我们发现去除任何卷积层（每个卷积层包含不超过模型参数的1%）都会导致性能下降。

最后，网络的规模主要受限于当前GPU上的可用内存量和我们愿意容忍的训练时间量。我们的网络在两个GTX580 3GB GPU上需要五到六天的时间来训练。我们所有的实验表明，只要等待更快的GPU和更大的数据集出现，我们的结果就可以得到改善。

2 数据集

ImageNet是一个由超过1500万张贴有标签的高分辨率图像组成的数据集，属于大约22000个类别。这些图像是从网上收集的，并由人类标签员使用Ama-zon的Mechanical Turk众包工具进行标注。从2010年开始，作为Pascal视觉对象挑战赛的一部分，每年都会举办一个名为ImageNet大规模视觉识别挑战赛（ILSVRC）的比赛。ILSVRC使用ImageNet的一个子集，在1000个类别中的每一个都有大约1000张图片。总的来说，大约有120万张训练图像，5万张验证图像和15万张测试图像。

ILSVRC-2010是唯一一个有测试集标签的ILSVRC版本，所以这也是我们进行大部分实验的版本。由于我们的模型也参加了ILSVRC-2012的比赛，在第6节中，我们也报告了我们在这个版本的数据集上的结果，因为测试集的标签是不可用的。在ImageNet上，通常报告两个错误率：前1名和前5名，其中前5名错误率是指正确标签不在模型认为最可能的五个标签中的测试图像的比例。

ImageNet由不同分辨率的图像组成，而我们的系统需要一个恒定的输入尺寸。因此，我们将图像下采样为256 × 256的固定分辨率。给定一个矩形图像，我们首先重新缩放图像，使短边的长度为256，然后从得到的图像中裁剪出中央的256 × 256个补丁。除了从每个像素上减去训练集的平均活动外，我们没有以任何其他方式预处理图像。所以我们在像素的（居中）原始RGB值上训练我们的网络。

3 建筑

我们的网络结构在图2中得到了总结。它包含八个学习层--五个卷积层和三个全连接层。下面，我们将描述我们的网络结构的一些新颖或不寻常的特点。第3.1-3.4节是根据我们对其重要性的估计进行排序的，最重要的放在前面。

[1] http://code.google.com/p/cuda-convnet/

3.1 ReLU非线性

将神经元的输出f作为其输入x的函数的标准方法是$f(x) = \tanh(x)$ 或$f(x) = (1 + e^{-x})^{-1}$。就梯度下降的训练时间而言，这些饱和非线性要比非饱和非线性$f(x) = \max(0, x)$慢得多。按照Nair和Hinton[20]的说法，我们把具有这种非线性的神经元称为整流线性单元（ReLU）。深度卷积神经网络-
使用ReLU的作品比使用tanh单元的作品训练速度快几倍。图1证明了这一点，它显示了在CIFAR-10数据集上，一个特定的四层卷积网达到25%的训练误差所需的迭代次数。
作。这张图显示，如果我们使用传统的饱和神经元模型，我们将无法为这项工作实验如此大的神经网络。

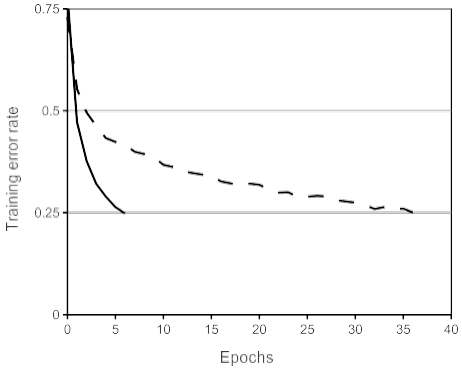

图1：一个带有ReLU的四层卷积神经网络（**实线**）在CIFAR-
10上达到25%的训练错误率，比带有tanh神经元的同等网络（**虚线**）快六倍。每个网络的学习率是独立选择的，以使训练尽可能快。没有采用任何形式的正则化。这里展示的效果的大小随网络结构的变化而变化，但具有ReLU的网络一致地比具有饱和神经元的同等网络快几倍。

我们并不是第一个考虑在CNN中替代传统神经元模型的人。例如，Jarrett等人[11]声称非线性$f(x)=\tanh(x)$与他们的对比度nor-x类型特别好。
在Caltech-
101数据集上，先进行恶化，再进行局部平均汇集。然而，在这个数据集上，首要问题是防止过度拟合，所以他们观察到的效果与我们在使用ReLU时报告的加速拟合训练集的能力不同。更快的学习对在大型数据集上训练的大型模型的性能有很大影响。

3.2 在多个GPU上进行训练

一台GTX580
GPU只有3GB内存，这限制了在其上训练的网络的最大规模。事实证明，120万个训练实例足以训练出太大的网络，而这些网络在一个GPU上是放不下的。因此，我们将网络分散在两个GPU上。目前的GPU特别适合于跨GPU并行化，因为它们能够直接从对方的内存中读出和写入，而不需要通过主机内存。我们采用的并行化方案基本上是将一半的内核（或神经元）放在每个GPU上，并有一个额外的技巧：GPU只在某些层进行通信。这意味着，例如，第3层的内核从第2层的所有内核图中获取输入。然而，第4层的内核只从驻扎在同一GPU上的第3层的内核图中获取输入。选择连接模式是交叉验证的一个问题，但这允许我们精确调整通信量，直到它是计算量的一个可接受的部分。

由此产生的架构与Cires¸an等人[5]采用的 "柱状"CNN有点类似，只是我们的柱子不是独立的（见图2）。与在一个GPU上训练的每个卷积层中只有一半内核的网络相比，这个方案将我们的前1名和前5名错误率分别降低了1.7%和1.2%。双GPU网络的训练时间比单GPU网络略少[2]。

[2]在最后的卷积层中，单GPU网络实际上与双GPU网络有相同数量的内核。这是因为该网络的大部分参数都在第一个全连接层中，该层将最后一个卷积层作为输入。因此，为了使两个网络的参数数量大致相同，我们没有将最后的卷积层（以及后面的全连接层）的大小减半。因此，这种比较偏向于单GPU网络，因为它比双GPU网络的 "一半大小 "还要大。

3.3 局部反应归一化

ReLU有一个理想的特性，即它们不需要输入归一化来防止它们饱和。如果至少有一些训练实例对ReLU产生了积极的输入，学习就会在该神经元中发生。
然而，我们仍然发现，以下的局部归一化方案有助于
泛化。用$a^i_{x,y}$
表示一个神经元的活动，该活动是通过在位置上应用内核i计算出来的。
(x, y)，然后应用ReLU非线性，响应归一化的活动b^i
表示 x,y 。

$$b^i_{x,y} = a^i_{x,y} \Big/ \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2\right)^{\beta}$$

其中，总和在同一空间位置的n个 "相邻
"内核图上运行，N是该层中内核的总数量。当然，内核图的排序是任意的，并在训练开始前确定。这种反应正常化实现了一种横向抑制的形式，其灵感来自于真实神经元中的类型，在使用不同内核计算的神经元输出之间产生了大活动的竞争。常数k、n、α和β是超参数，其值通过验证集确定；我们使用k=2，n=5，α=10^{-4}，β=0.75。我们
在某些层中应用ReLU非线性后，应用了这种归一化（见第3.5节）。

这个方案与Jarrett等人[11]的局部对比度归一化方案有些相似，但我们的方案更正确的说法是
"亮度归一化"，因为我们不减去平均活动。响应归一化使我们的前1名和前5名错误率分别降低了1.4%和1.2%。我们还在CIFAR-
10数据集上验证了这个方案的有效性：一个四层的CNN在没有归一化的情况下取得了13%的测试错误率，在归一化后取得了11%的测试错误率[3]。

3.4 重叠集合

CNN中的汇集层总结了同一核图中相邻的神经元组的输出。传统上，相邻的汇集单元所总结的邻域是不重叠的（例如，[17，11，4]）。更准确地说，一个集合层可以被认为是由一个间隔为s像素的集合单元网格组成，每个集合单元总结的邻域大小为z
z，以集合单元的位置为中心。如果我们设定s=z，我们就会得到通常采用的传统的局部集合法
在CNN中。如果我们设定$s < z$，我们就会得到重叠池。这就是我们在整个
网络，s=2，z=3。与非重叠方案s=2 ，z=2 相 比 ，
这个方案将前1名和前5名的错误率分别降低了0.4%和0.3%，后者产生的输出维度相当。
在训练过程中，我们普遍观察到，采用重叠池的模型发现过拟合的难度略大。

3.5 整体架构

现在我们准备描述我们的CNN的整体结构。如图2所示，该网络包含八个带权重的层；前五个是卷积层，其余三个是全连接层。最后一个全连接层的输出被送入一个1000路softmax，产生1000个类别标签的分布。我们的网络最大化了多叉逻辑回归的目标，这相当于最大化了预测分布下正确标签的对数概率在整个训练案例中的平均值。

第二、第四和第五卷积层的内核只与前一层中位于同一GPU上的内核图相连（见图2）。第三卷积层的神经元与第二层的所有内核图相连。完全连接层的神经元与前一层的所有神经元相连。响应正常化层紧随第一和第二卷积层。第3.4节中描述的那种最大集合层，紧随反应正常化层以及第五卷积层。ReLU非线性被应用于每个卷积层和全连接层的输出。

第一个卷积层用96个大小为11 11 3的核子过滤224 224 3的输入图像。
的跨度为4像素（这是相邻的接收场中心之间的距离）。

[3]由于篇幅限制，我们无法详细描述这个网络，但这里提供的代码和参数文件对其进行了精确的规定：http://code.google.com/p/cuda-convnet/。

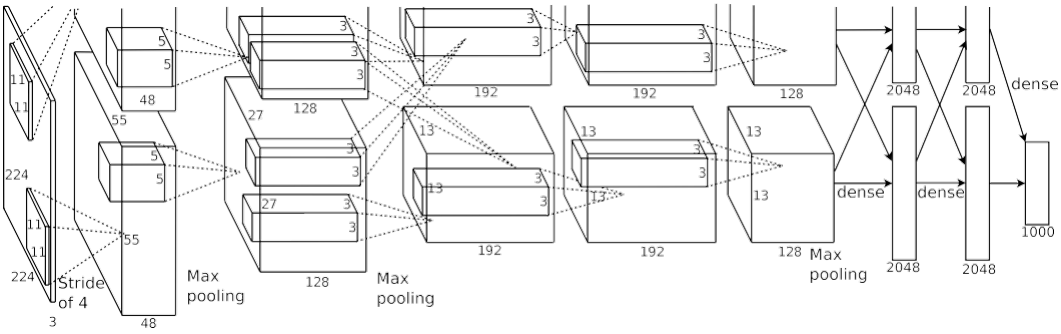

图2：我们的CNN的架构图，明确显示了两个GPU之间的职责划分。一个GPU运行图中顶部的层部分，而另一个则运行底部的层部分。GPU只在某些层进行通信。网络的输入是150,528维，网络其余层的神经元数量由253,440-186,624-64,896-64,896-43,264-4096-1000给出。

核心图中的神经元）。第二个卷积层将第一个卷积层的（响应归一化和池化）输出作为输入，并用大小为5548的256个核对其进行过滤。第三、第四和第五卷积层相互连接，没有任何中间的池化或规范化层。第三卷积层有384个大小为

3256

的核，与第二卷积层的（归一化、池化）输出相连。第四卷积层有384个大小为3×3×192的核，第五卷积层有256个大小为3×3×192的核。全连接层每个有4096个神经元。

4 减少过度拟合

我们的神经网络架构有6000万个参数。尽管ILSVRC的1000个类使每个训练实例对从图像到标签的映射施加了10比特的约束，但事实证明，这不足以在没有相当大的过拟合的情况下学习这么多的参数。下面，我们将描述我们对抗过拟合的两种主要方式。

4.1 数据扩增

减少图像数据过拟合的最简单和最常见的方法是使用标签保护的变换来人为地扩大数据集（例如，[25，4，5]）。我们采用了两种不同形式的数据放大，这两种方法都允许以很少的计算量从原始图像中产生转换图像，因此转换后的图像不需要存储在磁盘上。在我们的实施中，转换后的图像是在CPU上用Python代码生成的，而GPU正在对前一批图像进行训练。因此，这些数据增强方案实际上是无计算的。

数据增强的第一种形式包括生成图像的平移和水平反射。我们通过提取随机的224 个斑块（和它们的水平反射）来实现这一目的。
256 256张图像，在这些提取的斑块上训练我们的网络[4]。这增加了我们的
训练集的系数为2048，当然，由此产生的训练实例是高度相互关联的。
依赖性。如果没有这个方案，我们的网络就会出现严重的过拟合，这将迫使我们使用更小的网络。在测试时，网络通过提取五个224个斑块（四个角斑块和中心斑块）以及它们的水平反射（因此总共有十个斑块）来进行预测，并对网络的softmax层对这十个斑块的预测进行平均。

数据增强的第二种形式包括改变训练图像中的RGB通道的强度。具体来说，我们对整个ImageNet训练集的RGB像素值的集合进行PCA。在每张训练图像中，我们添加所发现的主成分的倍数。

[4]这就是为什么图2中的输入图像是224×224×3维的原因。

幅度与相应的特征值乘以从平均数为零、标准差为0.1的高斯中抽取的随机变量成正比。因此，对每个RGB图像像素$I_{xy} = [I^R_{xy}, I^G_{xy}, I^B_{xy}]^T$我们添加以下数量。

$$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

其中$\mathbf{p}_i$和λ_i分别是RGB像素值的3×3协方差矩阵的第1个特征向量和特征值，α_i是前述的随机变量。每个α_i，对某一特定训练图像的所有像素只画一次，直到该图像再次被用于训练，这时点，它被重新绘制。这个方案近似地捕捉到了自然图像的一个重要特性，即物体身份对光照强度和颜色的变化是不变的。这个方案将top-1的错误率降低了1%以上。

4.2 辍学

结合许多不同模型的预测是减少测试错误的一个非常成功的方法[1, 3]，但对于已经需要几天时间来训练的大型神经网络来说，它似乎太昂贵了。然而，有一个非常有效的模型组合版本，在训练期间只需要花费大约2倍的成本。最近引入的技术被称为"丢弃"[10]，包括以0.5的概率将每个隐藏神经元的输出设置为零。以这种方式被"剔除"的神经元对前向传递没有贡献，也不参与反向传播。因此，每次有输入时，神经网络都会对不同的架构进行采样，但所有这些架构都共享权重。这种技术减少了神经元的复杂共同适应，因为一个神经元不能依赖特定的其他神经元的存在。因此，它被迫学习更强大的特征，这些特征与其他神经元的许多不同的随机子集一起使用。在测试时，我们使用所有的神经元，但将它们的输出乘以0.5，这是一个合理的近似值，即取指数型多辍学网络产生的预测分布的几何平均值。

我们在图2的前两个全连接层中使用了dropout。如果没有dropout，我们的网络会出现严重的过拟合。弃权使收敛所需的迭代次数大约增加了一倍。

5学习的细节

我们使用随机梯度下降法训练我们的模型，批次大小为128例，动量为0.9，权重衰减为0.0005。我们发现，这种少量的权重衰减对模型的学习很重要。换句话说，这里的权重衰减不仅仅是一个正则器：它减少了模型的训练误差。权重w的更新规则是

$$v_{i+1} := 0.9 \cdot v_i - 0.0005 \cdot \varepsilon \cdot w_i - \varepsilon \left\langle \frac{\partial L}{\partial w}\Big|_{w_i} \right\rangle_{D_i}$$

$$w_{i+1} := w_i + v_{i+1}$$

其中i是迭代指数，v是动量变量，ε是学习率，$\left\langle \frac{\partial L}{\partial w}\Big|_{w_i} \right\rangle_{D_i}$是是指在第1批$D_i$，目标相对于$w$的导数的平均数，评价于$w_i$。

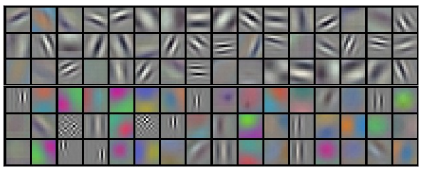

图3：大小为96个卷积核的第×卷积层在224×224×3输入图像上学习的11×11×3。前48个内核是在GPU 1上学习的，而底部48个内核是在GPU上学习的。2.详见第6.1节。

我们从标准差为0.01的零均高斯分布中初始化了每层的权重。我们将第二、第四和第五卷积层以及全连接隐藏层中的神经元偏置初始化为常数1。这种初始化通过为ReLU提供正的输入来加速学习的早期阶段。我们用常数0来初始化其余各层的神经元偏差。

我们对所有的层使用相同的学习率，在整个训练过程中手动调整。我们遵循的启发式方法是，当验证错误率不再随当前的学习率提高时，将学习率除以10。学习率被初始化为0.01，而

在终止前减少三次。我们通过120万张图像的训练集对网络进行了大约90个周期的训练，这在两个NVIDIA GTX 580 3GB GPU上花费了五到六天的时间。

6 结果

我们在ILSVRC-2010上的结果总结于表1。我们的网络实现了前1名和前5名测试集的错误率为**37.5%**和**17.0%**[5]。在ILSVRC-2010的比赛中，用一种方法平均了六个在不同特征上训练的稀疏编码模型所产生的预测结果，所取得的最好成绩是47.1%和28.2%[2]，此后，用一种方法平均了两个在Fisher Vectors（FVs）上训练的分类器的预测结果，从两种密集取样的特征中计算出来，所发表的最好成绩是45.7%和25.7%[24]。

模型	顶-1	前五名
稀疏编码[2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
有线电视新闻网	**37.5%**	**17.0%**

表1：关于ILSVRC-2010年的测试集的结果比较。*斜体字*是其他人取得的最佳结果。

我们还在ILSVRC-2012的competition中输入了我们的模型，并在表2中报告了我们的结果。由于ILSVRC-2012测试集的标签没有公开，我们无法报告所有我们尝试的模型的测试错误率。在本段的其余部分，我们使用验证和测试的错误率可以互换，因为根据我们的经验，它们的差异不超过0.1%（见表2）。本文描述的CNN实现了18.2%的前5名错误率。预测结果的平均值五个类似的CNN的错误率为16.4%。训练一个CNN，在最后一个集合层上增加第六个卷积层，对整个ImageNet 2011秋季版（15M图像，22K类别）进行分类，然后在ILSVRC-2012上进行"微调"，得到的错误率为16.6%。将在整个2011年秋季版上预训练的两个CNN的预测结果与上述五个CNN的预测结果平均起来，错误率为**15.3%**。第二好的测试项目取得了26.2%的错误率，其方法是对从不同类型的密集采样特征计算的FV上训练的七个分类器的预测进行平均[7]。

最后，我们还报告了2009年秋季版本的ImageNet的错误率，该版本有10,184个类别和890万张图片。在这个数据集上，我们遵循文献中的惯例，使用一半的图像进行训练，一半进行测试。由于没有在这个测试集上，我们的分割方式与以前的作者所使用的分割方式有所不同，但这并不明显影响结果。在这个数据集上，我们的前1名和前5名的错误率分别为**67.4%** 和**40.9%**，由上述网络达到，但在最后一个集合层上有一个额外的第六卷积层。在这个数据集上发表的最佳结果是78.1%和60.9%[19]。

模型	TOP-1 (VAL)	前五名(VAL)	前五名(测试)
SIFT + FVs [7]	-	-	*26.2%*
1 CNN	40.7%	18.2%	-
5个CNN	38.1%	16.4%	**16.4%**
1个CNN*	39.0%	16.6%	-
7个CNNs*	36.7%	15.4%	**15.3%**

表2：ILSVRC-2012验证和测试集的错误率比较。*斜体字*是其他人取得的最佳结果。带星号的模型*是为了对整个ImageNet 2011秋季版进行分类而 "预训练 "的。详见第6节。

6.1 定性评价

图3显示了该网络的两个数据连接层所学习的卷积核。该网络已经学会了各种频率和方向选择的内核，以及各种彩色斑点。请注意两个GPU表现出的专业性，这是第3.5节中描述的限

制性连接的结果。GPU 1上的内核在很大程度上是不分颜色的，而GPU 2上的内核在很大程度上是特定颜色的。这种特殊化发生在每次运行期间，并且与任何特定的随机权重初始化无关（除GPU的重新编号外）。

[5]如第4.1节所述，不对十个斑块进行平均预测的错误率为39.0%和18.3%。

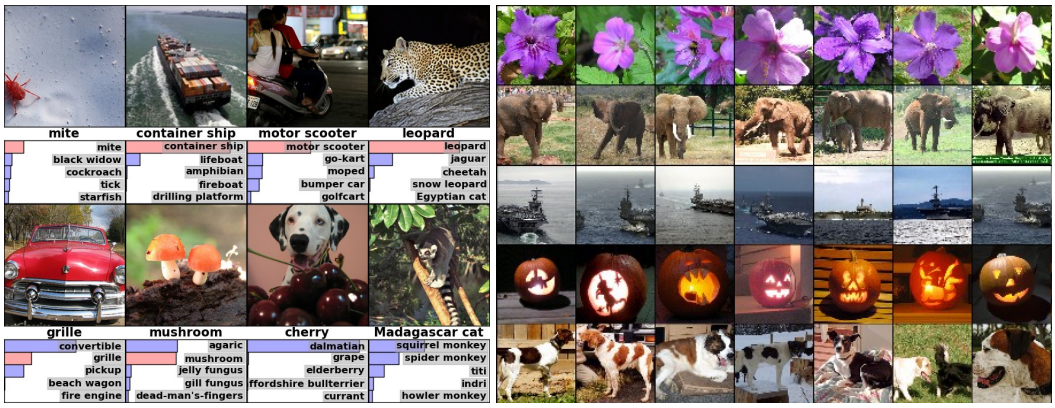

图4：（**左**）八张ILSVRC-2010测试图像和我们的模型认为最有可能的五个标签。正确的标签写在每张图片下面，分配给正确标签的概率也用红条显示（如果它恰好在前5名中）。(**右图**）第一列是五张ILSVRC-2010的测试图像。其余各列显示了六幅训练图像，这些图像在最后一个隐藏层产生的特征向量与测试图像的特征向量的欧氏距离最小。

在图4的左侧面板中，我们通过计算网络在8张测试图像上的前5名预测结果来定性地评估网络所学到的东西。请注意，即使是偏离中心的物体，如左上方的螨虫，也能被网络识别。大多数前五名的标签都显得很合理。例如，只有其他类型的猫被认为是豹子的合理标签。在某些情况下（格栅、樱桃），对于照片的预期焦点确实存在模糊性。

另一种探测网络视觉知识的方法是考虑最后一个4096维隐藏层的图像所引起的特征激活。如果两幅图像产生的特征激活向量具有较小的欧氏分离度，我们可以说，神经网络的高层认为它们是相似的。图4显示了来自测试集的五幅图像和来自训练集的六幅图像，根据这个衡量标准，它们各自最相似。请注意，在像素层面上，检索到的训练图像与第一列中的查询图像在L2上一般不接近。例如，检索到的狗和大象以各种姿势出现。我们在补充材料中介绍了更多测试图像的结果。

使用4096维实值向量之间的欧几里得距离来计算相似性是低效的，但可以通过训练一个自动编码器将这些向量压缩成短的二进制代码来使其高效。这应该会产生一个比对原始像素应用自动编码器更好的图像检索方法[14]，后者不利用图像标签，因此倾向于检索具有相似边缘模式的图像，无论它们在语义上是否相似。

7 讨论

我们的结果表明，一个大型的深度卷积神经网络能够在一个极具挑战性的数据集上使用纯粹的监督学习取得破纪录的结果。值得注意的是，如果去掉一个卷积层，我们网络的性能就会下降。例如，去掉任何一个中间层，都会使网络的最高性能损失约2%。因此，深度对于实现我们的结果真的很重要。

为了简化我们的实验，我们没有使用任何无监督的预训练，尽管我们预计它将有所帮助，特别是如果我们获得足够的计算能力来显著增加网络的大小而不获得相应的标记数据量的增加。到目前为止，我们的结果已经随着我们的网络规模的扩大和训练时间的延长而有所改善，但我们仍有许多数量级的工作要做，以便与人类视觉系统的近时路径相匹配。最终，我们希望在视频序列上使用非常大的深度卷积网络，因为时间结构提供了非常有用的信息，而这些信息在静态图像中是缺失的或不太明显。

参考文献

[1] R.M. Bell和Y. Koren。来自netflix奖挑战的教训。*ACM SIGKDD探索通讯*，9（2）：75-79，2007。

[2] A.Berg, J. Deng, and L. F e i -Fei.2010年大规模视觉识别挑战。www.image-net.org/challenges.2010.

[3] L.Breiman.随机森林。*机器学习*，45（1）：5-32，2001。

[4] D.Cires¸an, U. Meier, and J. Schmidhuber.用于图像分类的多列深度神经网络。*Arxiv预印本arXiv:1202.2745*，2012。

[5] D.C. Cires¸an, U. Meier, J. Masci, L.M. Gambardella, and J. Schmidhuber.用于视觉物体分类的高性能神经网络。*Arxiv预印本arXiv:1102.0183*，2011。

[6] J.Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.ImageNet:一个大规模的分层图像数据库。In *CVPR09*, 2009.

[7] J.Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. *ilsvrc-2012*, 2012.url `http://www.image-net.org/challenges/LSVRC/2012/`。

[8] L.Fei-Fei, R. Fergus, and P. Perona.从少数训练实例中学习生成性视觉模型。在101个物体类别上测试的增量贝叶斯方法。*Co mputer Vision and Image Understand- ing*, 106(1):59-70, 2007.

[9] G. Griffin, A. Holub, and P. Perona.Caltech-256物体类别数据集。技术报告7694，California技术研究所，2007。URL http://authors.l`ibrary.caltech.edu/7694`。

[10] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov.通过防止特征检测器的共同适应来改进神经网络工程。*arXiv预印本arXiv:1207.0580*，2012。

[11] K.Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun.什么是物体识别的最佳多阶段结构？*在国际计算机视觉会议上*，第2146-2153页。IEEE, 2009.

[12] A.Krizhevsky.从微小的图像中学习多层的特征。硕士论文，多伦多大学计算机科学系，2009年。

[13] A.Krizhevsky.cifar-10上的卷积深度信念网络.*未发表的手稿*，2010年。

[14] A.Krizhevsky and G.E. Hinton.使用非常深的自动编码器进行基于内容的图像检索。在*ESANN*, 2011.

[15] Y.Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, et al.手写数字识别的反向传播网络。在*神经信息处理系统的进展中*，1990年。

[16] Y.LeCun, F.J. Huang, and L. Bottou.具有姿势和光照不变性的通用物体识别的学习方法。在*计算机视觉和模式识别，2004年。CVPR 2004.2004年IEEE计算机学会会议论文集*，第2卷，第II-97页。IEEE, 2004.

[17] Y.LeCun, K. Kavukcuoglu, and C. Farabet.卷积网络和视觉中的应用。在*电路与系统（ISCAS），2010年IEEE国际研讨会论文集中*，第253-256页。IEEE, 2010.

[18] H.Lee, R. Grosse, R. Ranganath, and A.Y. Ng.卷积深度信念网络用于可扩展的非超视距学习分层表示。在第*26届国际机器学习年会论文集中*，第609-616页。ACM, 2009.

[19] T.Mensink, J. Verbeek, F. Perronnin, and G. Csurka.Metric Learning for Large Scale Image Classification:以接近零的成本推广到新的类别。在*ECCV -欧洲计算机视觉会议上*，意大利佛罗伦萨，2012年10月。

[20] V.Nair and G. E. Hinton.矫正的线性单元改善了限制性波尔茨曼机。In *Proc. 27th International Conference on Machine Learning*, 2010.

[21] N.Pinto, D.D. Cox, and J.J. DiCarlo.为什么现实世界的视觉物体识别是困难的？*PLoS computational biology*, 4(1):e27, 2008.

[22] N.Pinto, D. Doukhan, J.J. DiCarlo, and D.D. Cox。一种高通量筛选方法来发现生物启发的视觉表现的良好形式。*PLoS计算生物学*，5(11)：e1000579，2009。

[23] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman.Labelme：一个数据库和基于网络的图像注释工具。*国际计算机视觉杂志*，77（1）：157-173，2008。

[24] J.Sánchez and F. Perronnin.用于大规模图像分类的高维签名压缩。在*计算机视觉和模式识别（CVPR），2011年IEEE会议上*，第1665-1672页。IEEE, 2011.

[25] P.Y. Simard, D. Steinkraus, and J.C. Platt.应用于视觉文件分析的卷积神经网络的最佳实践。在*第七届国际文档分析和识别会议论文集*，第二卷，第958-962页，2003年。

[26] S.C. Turaga, J.F. Murray, V. Jain, F. Roth, M. Helmstaedter, K. Briggman, W. Denk, and H.S. Seung。Convolutional网络可以学习生成图像分割的亲和图。*Neural Computation*, 22(2):511-538, 2010.