|  |
| --- |
| ARXIV:1409.4842v1[cs.cv]2014年9月17日 |

卷积法向纵深发展

魏刘洋清贾

北卡罗来纳大学教堂山分校。

皮埃尔・塞尔马内特・里德德拉戈米尔・安盖洛夫杜米特鲁・埃尔汉

谷歌公司密歇根大学谷歌公司谷歌公司

文森特・范豪克安德鲁・拉比诺维奇

谷歌公司

摘要

我们提出了一种代号为INCEP-TION的深层卷积神经网络体系结构，它负责在ImageNet大规模视觉识别挑战2014(ILSVRC14)中设置分类和检测的新技术水平。 这种体系结构的主要特点是提高了网络内部计算资源的利用率。 这是通过精心设计实现的，它允许增加网络的深度和宽度，同时保持计算预算不变。 为了优化质量，架构决策基于Hebbian原则和多尺度处理的直觉。 我们为ILSVRC14提交的材料中使用的一个具体体现是GoogleNet，这是一个22层的深层网络，其质量是在分类和检测的背景下评估的。

一 Introduction

在过去的三年中，主要由于深度学习，更具体地说是卷积网络[10]的进步，图像识别和目标检测的质量一直在以惊人的速度进步。 一个令人鼓舞的消息是，这一进展的大部分不仅是更强大的硬件、更大的数据集和更大的模型的结果，而且主要是新的思想、算法和改进的网络架构的结果。 例如，除了同一竞赛的分类数据集用于检测之外，ILSVRC2014竞赛的顶级条目没有使用新的数据源。 我们提交给ILSVRC2014的GoogleNet实际上比Krizhevsky等人两年前的获奖架构少用了12个参数[9]，同时更加精确。 目标检测的最大收益并不是单独使用深度网络或更大的模型，而是深度体系结构和经典计算机视觉的协同作用，如Girshick等人的R-CNN算法[6]。

另一个值得注意的因素是，随着移动和嵌入式计算的持续发展，我们算法的效率——尤其是它们的能力和内存使用——变得越来越重要。 值得注意的是，本文提出的深层体系结构设计考虑因素包括这一因素，而不是仅仅关注精度数字。 在大多数实验中，这些模型被设计成在推理时保持1:5亿的计算预算乘以加法，这样它们就不会最终成为纯粹的学术好奇心，而是能够以合理的成本投入到现实世界中，即使是在大型数据集上。

一

在本文中，我们将重点介绍一种高效的计算机视觉深层神经网络体系结构，代号为Inception，它的名字来源于Lin等人在网络论文[12]中的网络，以及著名的“我们需要更深入”的因特网模因[1]。 在我们的例子中，“深度”一词有两种不同的含义:首先，我们以“初始模块”的形式引入了一个新的组织层次，同时也以更直接的方式增加了网络深度。 一般而言，人们可以将初始模型看作是[12]的逻辑顶点，同时从Arora等人的理论工作中得到启发和指导[2]。 该架构的优势在ILSVRC2014分类和检测挑战中得到了实验验证，其性能明显优于当前技术水平。

2相关工作

从Lenet-5[10]开始，卷积神经网络(CNN)典型地具有标准结构-在卷积层（任选地随后是对比度归一化和最大池）之后是一个或多个完全连接的层。 这种基本设计的变体在图像分类文献中普遍存在，并且在MNIST和CIFAR上，尤其是在ImageNet分类挑战赛上取得了迄今为止最好的结果[9，21]。 对于ImageNet等较大的数据集，最近的趋势是增加层数[12]和层大小[21，14]，同时使用辍学[7]来解决过拟合问题。

尽管最大池层会导致精确空间信息的丢失，但与[9]相同的卷积网络结构也已成功地用于定位[9，14]、对象检测[6，14，18，5]和人体姿态估计[19]。 受灵长类视觉皮层神经科学模型的启发，Serre等人。 [15]使用一系列不同大小的固定Gabor滤波器来处理多尺度，类似于初始模型。 然而，与固定的两层深度模型[15]相反，初始模型中的所有滤波器都是学习的。 此外，初始层被重复多次，在GoogleNet模型的情况下导致22层深度模型。

Network-in-Network是Lin等人提出的一种方法。 [12]为了提高神经网络的表示能力。 当应用于卷积层时，该方法可以被看作是附加的11卷积层，典型地接着是整流线性激活[9]。 这使得它可以轻松地集成到当前的CNN管道中。 我们在体系结构中大量使用这种方法。 然而，在我们的设置中，11卷具有双重用途:最关键的是，它们主要用作降维模块来消除计算瓶颈，否则这些瓶颈将限制我们网络的大小。 这样不仅可以增加网络的深度，还可以增加网络的宽度，而不会造成严重的性能损失。

Girshick等人提出的基于卷积神经网络（R-CNN）的目标检测方法是当前的主流方法。 [6]见第4段。 R-CNN将整个检测问题分解为两个子问题:首先以类别不可知的方式利用低级线索（如颜色和超像素一致性）进行潜在对象建议，然后使用CNN分类器来识别这些位置处的对象类别。 这种两阶段的方法利用低级别线索的边界框分割的精确性，以及最先进的CNN的高度强大的分类能力。 我们在我们的检测提交中采用了类似的管道，但是在两个阶段都进行了探索增强，例如多框[5]预测用于更高的对象边界框召回，以及集成方法用于更好地对边界框建议进行分类。

3动机和高层次的考虑

提高深层神经网络性能的最直接的方法是增加其规模。 这包括增加网络的深度（级别数）和宽度（每个级别的单元数）。 这是一种简单而安全的训练高质量模型的方法，特别是考虑到大量标记的训练数据的可用性。 然而，这个简单的解决方案有两个主要缺点。

更大的尺寸通常意味着更多的参数，这使得放大的网络更容易过拟合，特别是在训练集中标记的示例的数量有限的情况下。 这可能成为一个主要的瓶颈，因为创建高质量的培训集可能很棘手

二



(a)西伯利亚哈士奇(B)爱斯基摩犬

图1:ILSVRC2014分类挑战赛1000个类别中的两个不同类别。

而且代价高昂，特别是如果需要专业的人类评分员来区分像ImageNet（甚至是1000级ILSVRC子集）中的细粒度视觉类别，如图1所示。

均匀增加网络规模的另一个缺点是计算机资源的使用急剧增加。 例如，在深度视觉网络中，如果两个卷积层被链接，则其滤波器数目的任何均匀增加导致计算量的二次增长。 如果添加的容量使用效率不高（例如，如果大多数权重最终接近于零），则会浪费大量的计算时间。 因为在实践中计算预算总是有限的，所以优选计算资源的有效分配而不是大小的不加区分的增加，即使当主要目标是提高结果的质量时也是如此。

解决这两个问题的根本方法是最终从完全连接的体系结构转移到稀疏连接的体系结构，甚至在卷积中也是如此。 除了模仿生物系统外，由于Arora等人的开创性工作，这还具有更坚实的理论基础的优势。 [2]同上。 他们的主要结果是，如果数据集的概率分布可以用一个大的、非常稀疏的深层神经网络来表示，那么通过分析最后一层激活的相关统计量和对具有高度相关输出的神经元进行聚类，就可以逐层构建最优的网络拓扑。 尽管严格的数学证明需要非常严格的条件，但这一陈述与众所周知的赫比原理（神经元一起激发，一起连接）产生共鸣的事实表明，即使在不太严格的条件下，这一基本思想在实践中也是适用的。

缺点是，今天的计算基础设施在对非均匀稀疏数据结构进行数值计算时效率很低。 即使算术运算的数量减少了100，查找和缓存未命中的开销仍然占主导地位，因此切换到稀疏矩阵将不会得到回报。 通过使用稳定改进的、高度优化的数值库，可以利用底层CPU或GPU硬件的细微细节来实现极快的密集矩阵乘法，从而进一步扩大了差距[16，9]。 此外，非均匀稀疏模型需要更复杂的工程和计算基础设施。 目前大多数面向视觉的机器学习系统都是利用空间域中的稀疏性来构造卷积。 但是，卷积实现为到早期层中修补程序的密集连接的集合。 自[11]以来，ConvNet一直在特征维上使用随机稀疏连接表，为了打破对称性和提高学习效率，为了更好地优化并行计算，又恢复到与[9]完全连接的趋势。 该结构的均匀性和大量过滤器以及更大的批大小允许利用高效密集计算。

这就提出了一个问题:是否有希望采取下一个中间步骤:即使在理论建议的过滤器级别，也要使用额外的稀疏性，但要利用我们的

三

利用稠密矩阵的计算来实现现有的硬件。 关于稀疏矩阵计算的大量文献（如[3])表明，将稀疏矩阵聚类成相对稠密的子矩阵往往能给出稀疏矩阵乘法的最新实用性能。 在不久的将来，类似的方法将被用于非统一深度学习体系结构的自动构建，这似乎并不牵强。

Inception Architecture最初是第一作者的一个案例研究，该案例研究评估了复杂网络拓扑结构构建算法的假设输出，该算法试图逼近[2]中所暗示的视觉网络的稀疏结构，并用密集、可读的组件覆盖假设的结果。 尽管这是一项高度推测性的工作，但只有在对拓扑的确切选择进行两次迭代之后，我们才会看到相对于基于[12]的参考体系结构的适度收益。 在进一步调整学习速率、超参数和改进的训练方法之后，我们确定所得到的初始体系结构在作为[6]和[5]的基础网络的定位和对象检测方面特别有用。 有趣的是，尽管大多数最初的架构选择都受到了彻底的质疑和测试，但它们至少在本地是最优的。

然而，我们必须谨慎:尽管所提出的体系结构已成为计算机视觉的一个成功范例，但它的质量是否能归功于导致其构建的指导原则，仍是一个值得怀疑的问题。 确保这一点将需要更彻底的分析和验证:例如，如果基于下面描述的原则的自动化工具能够找到类似的、但更好的视觉网络拓扑结构。 最有说服力的证明是，如果一个自动化系统使用相同的算法创建网络拓扑，那么在其他领域也会获得类似的收益，但全局架构却截然不同。 至少，Inception体系结构的初步成功为今后朝着这个方向开展工作提供了坚实的动力。

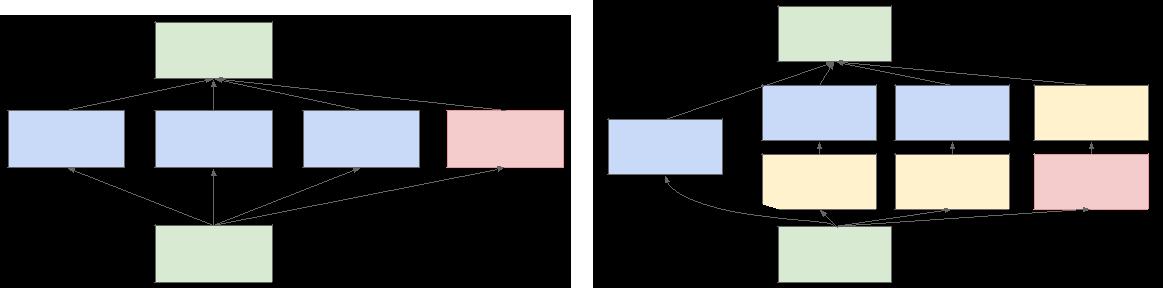
四 Architectural Details

Inception体系结构的主要思想是基于发现卷积视觉网络中的最优局部稀疏结构如何被容易获得的稠密组件近似和覆盖。 注意，假设翻译不变性意味着我们的网络将由卷积构建块构建。 我们所需要的只是找到最优的局部构造并在空间上重复它。 阿罗拉等人 [2]提出了一种逐层分析的方法，即对最后一层的相关统计量进行分析，并将其聚类成相关性较高的单元组。 这些集群构成下一层的单元，并连接到前一层的单元。 我们假设来自前一层的每个单元对应于输入图像的某个区域，并且这些单元被分组到滤波器组中。 在较低的层（接近输入的层），相关单元将集中在局部区域。 这意味着，我们最终会有很多团集中在一个单一的区域，它们可以被下一层的11个卷积层覆盖，如[12]中所建议的。 然而，人们也可以预期，在较大的斑块上可以被卷积覆盖的在空间上更分散的簇的数量将会较少，在越来越大的区域上的斑块的数量将会减少。 为了避免补丁对齐问题，初始体系结构的当前体现被限制为滤波器尺寸11、33和55，然而，该决定更多地基于便利性而不是必要性。 这还意味着，建议的架构是所有这些层的组合，它们的输出滤波器组级联到单个输出向量中，形成下一级的输入。 此外，由于池操作对于当前技术水平的卷积网络的成功至关重要，因此建议在每个这样的阶段添加一个替代并行池路径也应该具有额外的有益效果（参见图2(a))。

随着这些“初始模块”相互叠加，它们的输出相关统计信息必然会发生变化:随着更高抽象层的特征被更高抽象层捕获，它们的空间集中性预计会降低，这意味着随着我们向更高抽象层移动，3×3和5×5卷积的比率应该增加。

上述模块的一个大问题是，至少在这种朴素的形式中，即使是少量的55卷卷积，在具有大量滤波器的卷积层上也会非常昂贵。 一旦将池单元添加到组合中，这个问题就变得更加突出:它们的输出过滤器数量等于前一阶段的过滤器数量。 合并池层的输出和卷积层的输出将导致不可避免的

四



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 过滤器 |  |  |  |
|  | 过滤器 |  |  | 级联 |  |  |  |
|  | 级联 |  |  |  |  |  |  |
|  |  |  |  | 3x3卷积 | 5x5卷积 | 1x1卷积 |  |
| 1x1卷积 | 3x3卷积 | 5x5卷积 | 3x3最大池 | 1x1卷积 |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  | 1x1卷积 | 1x1卷积 | 3x3最大池 |  |
|  | 上一层 |  |  | 上一层 |  |  |  |
| (a)Inception Module，Naéve版本 | | | | (b)具有尺寸缩减的初始模块 | | |  |



图2:初始模块

逐级增加产出数量。 即使这种体系结构可以覆盖最优稀疏结构，它也会非常低效，导致计算在几个阶段内崩溃。

这就引出了所提出的体系结构的第二个想法:在计算需求增加太多的地方明智地应用维度缩减和投影。 这是基于嵌入的成功:即使是低维的嵌入也可能包含大量关于相对较大的图像补丁的信息。 然而，嵌入以密集的、压缩的形式表示信息，并且压缩的信息更难建模。 我们希望在大多数地方保持我们的表示稀疏（如[2]的条件所要求的），并且只在必须集体聚集信号时压缩信号。 也就是说，在昂贵的3×3和5×5卷积之前，用11卷积来计算约化。 除了用作还原外，它们还包括使用整流线性激活，这使它们具有双重用途。 最终结果如图2(b)所示。

一般而言，初始网络是由上述类型的模块彼此堆叠而成的网络，偶尔使用跨距为2的最大池层来将网格分辨率减半。 出于技术原因（培训期间的内存效率），只在较高层开始使用Inception模块，同时以传统的卷积方式保留较低层似乎是有益的。 这并不是绝对必要的，只是反映了我们目前执行中的一些基础设施效率低下。

这种体系结构的一个主要好处是，它允许显著增加每个阶段的单元数量，而不会在计算复杂度方面产生不可控的膨胀。 维度缩减的普遍使用允许将最后一级的大量输入滤波器屏蔽到下一层，首先缩减它们的维度，然后以较大的贴片大小卷积它们。 该设计的另一个实际有用的方面是，它与直觉一致，即视觉信息应该在不同的尺度上处理，然后聚合，以便下一阶段可以从不同的尺度上同时提取特征。

改进计算资源的使用允许增加每个级的宽度以及级的数量，而不会陷入计算困难。 另一种利用Inception体系结构的方法是创建稍微低劣的版本，但在计算上更便宜。 我们发现，所有包括的旋钮和杠杆都允许对计算资源进行受控的平衡，这可以使网络比使用非初始体系结构的类似执行网络快23倍，但是这需要在这一点上进行仔细的手动设计。

5GoogleNet

在ILSVRC14竞赛中，我们选择了GoogleNet作为我们的团队名称。 这个名字是对Yann Lecun开创的Lenet5网络的敬意[10]。 我们还使用GoogleNet来指代我们提交比赛时所使用的初始架构的具体体现。 我们还使用了一个更深、更广的初始网络，其质量略差，但将其添加到整体中似乎略有改善。 我们省略了该网络的细节，因为我们的实验已经表明，确切的体系结构参数的影响是相对的

五

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类型 | 修补程序大小/ | 产量 | 深度 | 第一名 |  | 一 | #33 | 第三名 |  | 三 | #5和5 | 第五名 |  | 五 | 游泳池 | 参数 | 行动 |  |
| 大步走 | 大小 |  |  | 减少 |  |  | 减少 |  |  | 项目 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 卷积 | 77=2 | 11211264 | 一 |  |  |  |  |  |  |  |  |  |  |  |  | 2.7千瓦 | 34米 |  |
| 最大池 | 33=2 | 56-5664 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 卷积 | 33=1 | 5656192 | 二 |  |  |  | 64岁 | 192页 | |  |  |  |  |  |  | 112K | 360米 |  |
| 最大池 | 33=2 | 2828192 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 初期(3A) |  | 2828256 | 二 | 64岁 | |  | 九十六 | 128名 | |  | 十六 | 三十二 | |  | 三十二 | 159K型 | 128米 |  |
| 初期(3B) |  | 2828480 | 二 | 128名 | |  | 128名 | 192页 | |  | 三十二 | 九十六 | |  | 64岁 | 380K | 304m |  |
| 最大池 | 33=2 | 1414480 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 初始(4A) |  | 1414512 | 二 | 192页 | |  | 九十六 | 208 | |  | 十六 | 四十八 | |  | 64岁 | 364K型 | 73M型 |  |
| 初期(4B) |  | 1414512 | 二 | 160美元 | |  | 112 | 二百二十四 | |  | 24岁 | 64岁 | |  | 64岁 | 437K型 | 88米 |  |
| 初始(4C) |  | 1414512 | 二 | 128名 | |  | 128名 | 256人 | |  | 24岁 | 64岁 | |  | 64岁 | 463K型 | 100米 |  |
| 初始(4D) |  | 1414528 | 二 | 112 | |  | 144 | 288 | |  | 三十二 | 64岁 | |  | 64岁 | 580K | 1.19亿美元 |  |
| 初期(4E) |  | 1414832 | 二 | 256人 | |  | 160美元 | 三百二十 | |  | 三十二 | 128名 | |  | 128名 | 840K | 170米 |  |
| 最大池 | 33=2 | 77832 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 初期(5A) |  | 77832 | 二 | 256人 | |  | 160美元 | 三百二十 | |  | 三十二 | 128名 | |  | 128名 | 1072K型 | 54米 |  |
| 初期(5b) |  | 771024 | 二 | 三百八十四 | |  | 192页 | 三百八十四 | |  | 四十八 | 128名 | |  | 128名 | 1388K型 | 71米 |  |
| 平均池 | 77=1 | 111024 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 辍学率（40%） |  | 111024 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 线性 |  | 111000 | 一 |  |  |  |  |  |  |  |  |  |  |  |  | 1000K | 1米 |  |
| 软最大值 |  | 111000 | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

表1:GoogleNet是Inception架构的化身

次要的。 这里，为了演示起见，表1中描述了最成功的特定实例（名为GoogleNet）。 在我们的集成中，7个模型中有6个使用了完全相同的拓扑（用不同的采样方法训练）。

所有卷积，包括初始模块内的卷积，都使用整流线性激活。 在我们的网络中，接收场的大小是224，224，取RGB颜色通道的平均减法。 “#33reduce”和“#55reduce”表示在卷积33和55之前使用的缩减层中的11个筛选器的数目。 在Pool Proj列中的内置最大池之后，可以看到项目层中11个过滤器的数量。 所有这些缩小/投影层也使用整流线性激活。

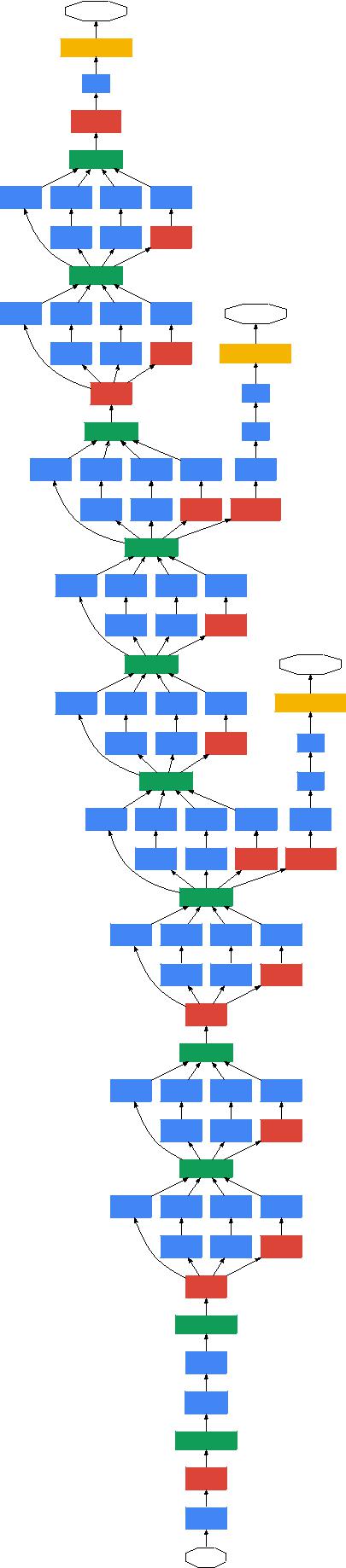
该网络的设计考虑了计算效率和实用性，使得推理可以在单个设备上运行，包括那些计算资源有限的设备，特别是内存占用较少的设备。 如果只计算具有参数的层，则网络深度为22层（如果同时计算池，则为27层）。 用于构建网络的层（独立构建块）总数约为100。 然而，这个数字取决于所使用的机器学习基础设施系统。 在分类器之前使用平均池是基于[12]的，尽管我们的实现不同之处在于我们使用了额外的线性层。 这使得我们的网络能够很容易地适应和微调其他标签集，但它大多是方便的，我们不希望它有一个主要的影响。 研究发现，从完全连接的层转移到平均汇集可将TOP-1的准确度提高约0.6%，但即使在移除完全连接的层之后，使用辍学仍然是必要的。

考虑到网络的相对较大的深度，以有效的方式将梯度传播回所有层的能力是一个值得关注的问题。 一个有趣的见解是，相对较浅的网络在这项任务上的强大性能表明，网络中间层产生的特征应该是非常有区别的。 通过添加连接到这些中间层的辅助分类器，我们期望鼓励分类器中较低阶段的区分，增加传播回来的梯度信号，并提供额外的正则化。 这些分类器采用较小的卷积网络的形式，该卷积网络置于初始(4a)和(4d)模块的输出之上。 在训练过程中，它们的损失以折扣权重被加到网络的总损失中（辅助分类器的损失以0.3加权）。 在推理时，这些辅助网络被丢弃。

包括辅助分类器在内的附加网络的具体结构如下:

* 平均汇集层具有55个滤波器大小和跨距3，导致(4A)级的输出为44512，(4D)级的输出为44528。

六



|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | SoftMax2软件 | |  |  |  |  |  |  |  |  |
|  | SoftMax激活 | | |  |  |  |  |  |  |  |  |
|  |  | 财务委员会 | |  |  |  |  |  |  |  |  |
|  |  | 平均池 | |  |  |  |  |  |  |  |  |
|  |  | 7x7+1(V) | |  |  |  |  |  |  |  |  |
|  |  | 深度 | |  |  |  |  |  |  |  |  |
| 康沃尔 | 康沃尔 | | 康沃尔 |  | 康沃尔 |  |  |  |  |  |  |
| 1x1+1(s) | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |  |  |  |  |
|  | 康沃尔 | | 康沃尔 |  | 最大池 | |  |  |  |  |  |
|  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |  |  |  |  |
|  |  | 深度 | |  |  |  |  |  |  |  |  |
| 康沃尔 | 康沃尔 | | 康沃尔 |  | 康沃尔 |  |  | softmax1软件 | |  |  |
| 1x1+1(s) | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |  |  |
|  |  |  |  |  |
|  | 康沃尔 | | 康沃尔 |  | 最大池 | | SoftMax激活 | | | |  |
|  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |
|  |  |  |  |  |  |
|  |  |  | 最大池 |  |  |  |  | 财务委员会 |  |  |  |
|  |  |  | 3x3+2(s) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  | 深度 | |  |  |  |  | 财务委员会 |  |  |  |
|  | 康沃尔 | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 |  |  |  |
| 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | | 1x1+1(s) | |  |  |
|  |  | 康沃尔 | | 康沃尔 | | 最大池 | | 平均池 | |  |  |
|  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | | 5x5+3(V) | |  |  |
|  |  |  | 深度 | | |  |  |  |  |  |  |
|  | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 | |  |  |  |
|  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |  |  |
|  |  |  | 康沃尔 | | 康沃尔 | | 最大池 | |  |  |  |
|  |  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |  |  |
|  |  |  | 深度 | | |  |  |  |  | 软MAX0 |  |
|  | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 | |  | SoftMax激活 |  |
|  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |  |
|  |  |  |  |
|  |  |  | 康沃尔 | | 康沃尔 | | 最大池 | |  | 财务委员会 |  |
|  |  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |  |
|  |  |  |  |  |  |
|  |  |  |  | 深度 | | |  |  |  | 财务委员会 |  |
|  |  |  | 康沃尔 | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 |  |
|  |  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | | 1x1+1(s) |  |
|  |  |  |  | 康沃尔 | | 康沃尔 | | 最大池 | | 平均池 |  |
|  |  |  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | | 5x5+3(V) |  |
|  |  |  |  |  |  | 深度 | |  |  |  |  |
|  |  |  | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 | |  |
|  |  |  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |
|  |  |  |  |  | 康沃尔 | | 康沃尔 | | 最大池 | |  |
|  |  |  |  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |
|  |  |  |  |  |  | 最大池 | |  |  |  |  |
|  |  |  |  |  |  | 3x3+2(s) | |  |  |  |  |
|  |  |  |  |  |  | 深度 | |  |  |  |  |
|  |  |  | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 | |  |
|  |  |  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |
|  |  |  |  |  | 康沃尔 | | 康沃尔 | | 最大池 | |  |
|  |  |  |  |  | 1x1+1(s) | | 1x1+1(s) | | 3x3+1(s) | |  |
|  |  |  |  |  |  | 深度 | |  |  |  |  |
|  |  |  | 康沃尔 | | 康沃尔 | | 康沃尔 | | 康沃尔 | |  |
|  |  |  | 1x1+1(s) | | 3x3+1(s) | | 5x5+1(s) | | 1x1+1(s) | |  |

CONVCONVMAXPOOL

1x1+1(s)1x1+1(s)3x3+1(s)

最大池

3x3+2(s)

局部范数

康沃尔

3x3+1(s)

康沃尔

1x1+1(V)

局部范数

最大池

3x3+2(s)

康沃尔

7x7+2(S)

输入

图3:GoogleNet网络中的所有花哨信息

七

* 一个包含128个滤波器的1-1卷积，用于降维和整流线性激活。
* 具有1024个单元和整流线性激活的完全连接层。
* 压降层，输出压降比为70%。
* 以SoftMax损失作为分类器的线性层（预测与主分类器相同的1000个类别，但在推理时移除）。

结果网络的示意图如图3所示。

6培训方法

我们的网络使用DistBelief[4]分布式机器学习系统进行训练，使用最少量的模型和数据并行性。 虽然我们只使用了基于CPU的实现，但粗略估计表明，GoogleNet网络可以在一周内通过少量的高端GPU进行收敛训练，主要限制是内存使用。 我们的训练采用0.9动量的异步随机梯度下降[17]，固定的学习速率计划（每8个周期降低4%的学习速率）。 Polyak平均[13]用于创建推理时使用的最终模型。

在比赛前的几个月里，我们的图像采样方法发生了很大的变化，已经收敛的模型与其他选项一起进行了训练，有时还结合了变化的超参数，如辍学率和学习率，因此很难对训练这些网络的最有效的单一方法给出明确的指导。 使问题进一步复杂化的是，一些模型主要是受[8]的启发，针对较小的相对作物进行训练，另一些则针对较大的相对作物进行训练。 尽管如此，在竞争之后被证实非常有效的一个处方包括对图像的各种大小的斑块进行采样，这些斑块的大小在图像面积的8%和100%之间均匀分布，并且其纵横比在3=4和4=3之间随机选择。 此外，我们还发现Andrew Howard[8]的光度畸变在一定程度上有助于克服过拟合现象。 此外，我们开始使用随机插值方法（双线性、面积、最近邻和三次插值，具有相等的概率）来调整相对较晚的值，并且与其他超参数变化结合使用，因此我们不能确定使用它们是否会对最终结果产生积极的影响。

7ILSVRC2014分类挑战赛设置和结果

ILSVRC2014分类挑战涉及将图像分类为ImageNet层次结构中的1000个叶节点类别之一。 大约有120万幅图像用于培训，5万幅图像用于验证，10万幅图像用于测试。 每个图像与一个背景真相类别相关联，并且基于最高评分分类器预测来测量性能。 通常报告两个数字:Top-1准确率，用于将地面真相与第一个预测类进行比较；Top-5错误率，用于将地面真相与前5个预测类进行比较:如果地面真相在前5个预测类中，则无论其在前5个预测类中的排名如何，图像都被视为正确分类。 该挑战使用前5个错误率进行排名。

我们参加了挑战，没有外部数据用于培训。 除了本文前面提到的培训技术之外，我们还在测试过程中采用了一套技术来获得更高的性能，我们将在下面进行详细说明。

1. 我们独立地训练了同一GoogleNet模型的7个版本（包括一个更宽的版本），并用它们执行集成预测。 这些模型使用相同的初始化（甚至使用相同的初始权重，主要是因为疏忽）和学习速率策略进行训练，它们仅在抽样方法和它们看到输入图像的随机顺序上不同。
2. 在试验中，我们采用了一种比Krizhevsky等人更积极的种植方法。 [9]见附件一。 具体地说，我们将图像调整为4个尺度，其中较短的维度（高度或宽度）分别为256、288、320和352，取这些调整大小的图像的左、中、右正方形（在肖像图像的情况下，取顶部、中心和底部正方形）。 然后，对于每个正方形，我们取4个角和中心224、224以及

八

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 团队 | |  | 年份 | 地点 | 错误（前5位） | | | 使用外部数据 | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | |  |  |  |  |  |  |  |
|  | 监管 | |  | 2012年 | 第1次 | 16:4% | |  | 否 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 监管 | |  | 2012年 | 第1次 | 15:3% | |  | ImageNet22K图像 | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 克拉费 | |  | 2013年 | 第1次 | 11:7% | |  | 否 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 克拉费 | |  | 2013年 | 第1次 | 11:2% | |  | ImageNet22K图像 | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | MSRA公司 | |  | 2014年 | 第3次 | 7:35% | |  | 否 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | VGG格式 | |  | 2014年 | 第2次 | 7:32% | |  | 否 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 谷歌网 | |  | 2014年 | 第1次 | 6:67% | |  | 否 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | 表2:分类性能 | | | | | |  |  |
|  | |  |  | | | |  |  | |  | |
| 型号数量 | |  | 作物数量 | | | | 成本 | 前5个错误 | | 与基数相比 | |
|  |  |  |  | |  |  |  |  | |  | |
| 一 |  |  | 一 | |  |  | 一 | 10:07% | | 基地 | |
|  |  |  |  | |  |  |  |  | |  |  |
| 一 |  |  | 十 | |  |  | 十 | 9:15% | | -0.92% |  |
|  |  |  |  | |  |  |  |  | |  |  |
| 一 |  |  | 144 | |  |  | 144 | 7:89% | | -2.18% |  |
|  |  |  |  | |  |  |  |  | |  |  |
| 七 |  |  | 一 | |  |  | 七 | 8:09% | | -1.98% |  |
|  |  |  |  | |  |  |  |  | |  |  |
| 七 |  |  | 十 | |  |  | 七十 | 7:62% | | -2.45% |  |
|  |  |  |  | |  |  |  |  | |  |  |
| 七 |  |  | 144 | |  |  | 1008年 | 6:67% | | -3.45% |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

表3:GoogleNet分类性能细分

方块大小调整为224224，以及它们的镜像版本。 结果是每幅图像有4362=144个作物。 Andrew Howard[8]在前一年的条目中也使用了类似的方法，我们通过经验验证，该方法的性能略差于所提议的方案。 我们注意到，在实际应用中，这种积极的种植可能没有必要，因为在有了合理数量的作物之后，更多作物的好处就变得微不足道了（我们将在后面说明这一点）。

1. 对多个作物和所有单个类筛选器上的SoftMax概率进行平均，以获得最终的预测。 在我们的实验中，我们分析了验证数据的替代方法，例如作物上的最大池和分类器上的平均，但是它们导致了比简单平均更差的性能。

在本文的其余部分，我们分析了影响最终提交的总体性能的多种因素。

我们在挑战中的最终提交在验证和测试数据上获得6.67%的前5名误差，在其他参与者中排名第一。 这与2012年的监督办法相比相对减少了56.5%，与上一年的最佳办法（CLARIFAI）相比相对减少了约40%，这两种办法都使用外部数据来培训分类器。 下表显示了一些性能最佳的方法的统计信息。

我们还通过改变下表中预测图像时使用的模型数量和作物数量来分析和报告多个测试选项的性能。 当我们使用一个模型时，我们选择了验证数据上错误率最低的一个模型。 所有数字都在验证数据集中报告，以避免与测试数据统计数据过度匹配。

8ILSVRC2014检测挑战设置和结果

ILSVRC检测任务是在200个可能的类中产生围绕图像中的对象的边界框。 如果检测到的对象与GroundTruth的类匹配，并且它们的边界框重叠至少50%（使用Jaccard索引），则被视为正确。 无关的检测被视为假阳性并受到处罚。 与分类任务相反，每个图像可以包含

九

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 团队 | | 年份 | |  | 地点 | 地图 | 外部数据 | | 合奏 | 方法 | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 紫外-可见分光光度计 | | 2013年 | |  | 第1次 | 22:6% | 无 | | 是吗？ | Fisher向量 | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 深刻的洞察力 | | 2014年 | |  | 第3次 | 40:5% | ImageNet1K图像 | | 三 | 有线电视新闻网 | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 中大深网 | | 2014年 | |  | 第2次 | 40:7% | ImageNet1K图像 | | 是吗？ | 有线电视新闻网 | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 谷歌网 | | 2014年 | |  | 第1次 | 43:9% | ImageNet1K图像 | | 六 | 有线电视新闻网 | |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | 表4:检测性能 | | | |  |  |  |
|  |  |  |  |  |  |  | |  | | |  |
|  | 团队 |  |  |  | 地图 | 上下文模型 | | 包围盒回归 | | |  |
|  |  |  |  |  | |  |  |  |  |  |  |
|  | 特里普斯-搜神 |  |  | 31:6% | |  | 否 |  | 是吗？ |  |  |
|  |  |  |  |  | |  |  |  |  |  |  |
|  | 伯克利愿景 |  |  | 34:5% | |  | 否 |  | 是的 |  |  |
|  |  |  |  |  | |  |  |  |  |  |  |
|  | 紫外-可见分光光度计 |  |  | 35:4% | |  | 是吗？ |  | 是吗？ |  |  |
|  |  | |  |  | |  |  |  |  |  |  |
|  | 中大深网2 | |  | 37:7% | |  | 否 |  | 是吗？ |  |  |
|  |  |  |  |  | |  |  |  |  |  |  |
|  | 谷歌网 |  |  | 38:02% | |  | 否 |  | 否 |  |  |
|  |  |  |  |  | |  |  |  |  |  |  |
|  | 深刻的洞察力 |  |  | 40:2% | |  | 是的 |  | 是的 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

表5:单型号检测性能

许多物体或没有物体，它们的规模可能从大到小不等。 使用平均平均精度（MAP）报告结果。

GoogleNet采用的检测方法类似于文献[6]中的R-CNN，但增加了初始模型作为区域分类器。 此外，通过将选择性搜索[20]方法与多框[5]预测相结合来改进区域提议步骤，以用于更高的对象边界框召回。 为了减少误报率，超像素尺寸增加了2。 这使来自选择性搜索算法的建议减半。 我们将来自多方框[5]的200个区域提案相加，总共产生了[6]所使用的建议的大约60%，同时将覆盖率从92%增加到93%。 在扩大覆盖面的情况下减少建议书数量的总体效果是，单个模型的平均精度提高了1%。 最后，在对每个区域进行分类时使用了6个ConvNet集合，使分类结果的准确率从40%提高到43.9%。 注意，与R-CNN相反，由于时间不够，我们没有使用包围盒回归。

我们首先报告顶级检测结果，并显示自第一版检测任务以来的进展。 与2013年的结果相比，精确度几乎翻了一番。 表现最好的团队都使用卷积网络。 我们在表4中报告官方分数和每个团队的共同策略:使用外部数据、集成模型或上下文模型。 外部数据通常是用于预训练模型的ILSVRC12分类数据，该模型随后在检测数据上被细化。 一些团队还提到了本地化数据的使用。 由于本地化任务边界框的很大一部分不包括在检测数据集中，因此可以用该数据预先训练一般边界框回归器，就像将分类用于预先训练一样。 GoogleNet条目没有使用本地化数据进行预培训。

在表5中，我们只使用一个模型比较结果。 表现最好的模型是由深刻的洞察力，令人惊讶的是，3个模型的集成只提高了0.3个点，而GoogleNet通过集成获得了显著更强的结果。

9结论

我们的结果似乎给出了一个确凿的证据，即用容易得到的密集构建块来近似期望的最优稀疏结构是改进计算机视觉的神经网络的一个可行方法。 与较浅和较窄的网络相比，该方法的主要优点是在适度增加计算要求的情况下获得显著的质量增益。 我还注意到，尽管我们的检测工作既没有利用上下文，也没有执行边界框，但仍然具有竞争力

十

这一事实进一步证明了Inception体系结构的强大。 尽管人们期望通过更昂贵得多的深度和宽度相似的网络可以获得类似质量的结果，但我们的方法产生了坚实的证据，证明迁移到更稀疏的体系结构总体上是可行和有用的。 这表明，在[2]的基础上，以自动化的方式创建更稀疏和更精细的结构是一项很有前途的未来工作。

10致谢

我们要感谢Sanjeev Arora和Aditya Bhaskara就[2]进行了富有成果的讨论。 此外，我们还要感谢“不相信”团队的支持，特别是Rajat Monga、Jon Shlens、Alex Krizhevsky、Jeff Dean、Ilya Sutskever和Andrea Frome。 我们还要感谢Tom Duerig和Ning Ye在光度畸变方面的帮助。 此外，如果没有查克·罗森伯格和哈特维格·亚当的支持，我们的工作是不可能完成的。

参考资料

1. 了解你的迷因:我们需要深入了解。 http://knowyourmeme.com/memes/

[我们需要再深入一点。 访问日期:2014-09-15。](http://knowyourmeme.com/memes/we-need-to-go-deeper)

1. Sanjeev Arora、Aditya Bhaskara、Rong Ge和Tengyu Ma。 学习一些深层表征的可证界。 Corr，ABS/1310.6343，2013年。

“

[3] Umit V. C¸atalyurek,¨ Cevdet Aykanat, and Bora Uc¸ar. On two-dimensional sparse matrix par-titioning: Models, methods, and a recipe. SIAM J. Sci. Comput., 32(2):656–683, February 2010.

[4] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc’aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc V. Le, and Andrew Y. Ng. Large scale distributed deep networks. In P. Bartlett, F.c.n. Pereira, C.j.c. Burges, L. Bot-tou, and K.q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1232–1240. 2012.

[5] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable ob-ject detection using deep neural networks. In Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on, 2014.

[6] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on, 2014.

[7] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhut-dinov. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580, 2012.

[8] Andrew G. Howard. Some improvements on deep convolutional neural network based image classification. CoRR, abs/1312.5402, 2013.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep con-volutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1106–1114, 2012.

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

[11] Yann LeCun, Leon´ Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[12]林敏，陈强，严水成。 网络中的网络。 Corr，ABS/1312.4400，2013年。

[13] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. SIAM J. Control Optim., 30(4):838–855, July 1992.

[14] Pierre Sermanet, David Eigen, Xiang Zhang, Michael¨ Mathieu, Rob Fergus, and Yann Le-Cun. Overfeat: Integrated recognition, localization and detection using convolutional net-works. CoRR, abs/1312.6229, 2013.

十一

1. Thomas Serre，Lior Wolf，Stanley M。Bileschi，Maximilian Riesenhuber和Tomaso Poggio。 使用类皮层机制的鲁棒对象识别。 IEEE传输。 肛门模式。 马赫。 Intell。，29（3）:411-426，2007。
2. 宋凤光和杰克·东加拉。 在具有1000个CPU核的共享内存多核系统上扩展矩阵计算。 第28届ACM国际超级计算会议论文集，ICS'14，第333-342页，纽约，纽约，美国，2014。 ACM。
3. Ilya Sutskever，James Martens，George E。Dahl和Geoffrey E。Hinton。 论初始化和动力在深度学习中的重要性。 《第30届机器学习国际会议论文集》，ICML2013，美国佐治亚州亚特兰大市，2013年6月16日至21日，第28卷，第1139-1147页。 jmlr.org，2013年。
4. Christian Szegedy Alexander Toshev和Dumitru Erhan 用于目标检测的深层神经网络。 在Christopher J.C。Burges，Leon'Bottou，Zoubin Ghahramani和Kilian Q。Weinberger主编的《神经信息处理系统进展》26:27，2013年神经信息处理系统年会。 2013年12月5日至8日在美国内华达州塔霍湖举行的会议记录，第2553-2561页。
5. 亚历山大·托舍夫和克里斯蒂安·塞格迪。 深度姿态:通过深度神经网络进行人体姿态估计。 Corr，ABS/1312.4659，2013年。
6. Koen E。A。van de Sande，Jasper R。R。Uijlings，Theo Gevers和Arnold W。M。Smeulders。 分割作为目标识别的选择性搜索。 2011年国际计算机视觉会议论文集，ICCV'11，1879-1886页，华盛顿特区，美国，2011年。 IEEE计算机协会。
7. Matthew D。Zeiler和Rob Fergus。 可视化和理解卷积网络。 在David J。Fleet，Tomas-Pajdla，BerntSchiele和Tinne Tuytelaars，编辑，计算机视觉-ECCV2014-第13届欧洲会议，2014年9月6日至12日，瑞士苏黎世，Pro-Seedings，Part I，第8689卷计算机科学讲义，第818-833页。 Springer，2014年。

十二