指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 01

姓名：刘选

学号：2013302524

班号：10011303

深入与卷积

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed,
Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich
Google Inc. University of North Carolina, Chape Hill
University of Michigan, Ann Arbor Magic Leap Inc.
fszegedy,jiayq,sermanet,dragomir,dumitru,vanhouckeg@google.com
wliu@cs.unc.edu, reedscott@umich.edu, arabinovich@magicleap.com

摘要

我们提出了一个代号为开端（Inception）的深入卷积神经网络的体系结构，在2014年的大规模视觉识别挑战（ilsvrc14）的ImageNet项目的分类与检测上达到了新的行业状态。这种架构的主要特点是提高了网络上计算资源的利用率。通过精心设计的设计，我们增加了网络的深度和宽度，同时保持了计算预算常数。为了优化质量，架构上的决定是基于赫布原理和多尺度处理的直觉。被我们用于提交给ilsvrc14的一个特殊的结构叫做googlenet，一个22层的网络，其质量是在分类和检测的背景下进行评估的。

1、简介

在过去的三年里，由于在深度学习和卷积网络[10]方面的进步，我们的目标分类和检测能力有显著高。一个令人鼓舞的消息是，大部分的进步是不只是更强大的硬件，更大的数据集和更大的模型的结果，而主要是因为一个新的想法，即算法和改进的网络体系结构。没有新的数据来源，例如，通过顶部的条目在Ilsvrc 2014竞争除了分类用于检测目的的同一竞争数据集。我们提交给 Ilsvrc 2014的 googlenet 实际使用的参数比krizhevsky等人[9]在两年前的获奖架构少12倍，同时显的更准确。在目标检测方面，最大的收获不是来自越来越大的深层网络方面的天真应用，而是从深层的协同效应架构和计算机视觉中的经典,就像Girshick等人[6]的r-cnn算法。

另一个值得注意的因素是，随着正在进行的牵引力移动和嵌入式计算，我们的效率算法，特别是它们的功率和内存的使用，获得了重要地位。值得注意的是，在本文中提出的领导

深层结构的设计的想法包括这个因素，而不是完全固定在准确的数字上。对于大多数的实验，模型旨在保持乘加在推理时间上的 1.5 亿计算预算，以至于他们不结束纯粹是因为学术上的好奇心，但可以说，即使在大数据集上，投入真实世界中的也是一个合理的成本。

在本文中，我们将专注于一个有效的深层神经计算机视觉的网络架构，代号为起始，这个名字派生于网络上林等人的论文[12]以及著名的“我们需要更加深入“的互联网热潮[1]。在我们的研究中，“深”一词被用于两种不同的含义：首先，就意义而言，我们在“起始模块”的形成中引入了一个新的组织层次，更多是用于直接增加网络深度。一般来说，一方面可以认为成立模型作为一个逻辑的顶点[12]，同时从理论上对灵感和指导 Arora 等人的工作[2]。这种结构的好处是被 ilsvrc 2014 分类和检测的挑战验证过的，在那里它显着优于当前的行业状态。

2、相关工作

从传统意义上来说[10]，卷积神经网络（CNN）通常有一个标准的结构，堆叠的卷积层（任意地通过对比规范化和最大空间池化）后面跟随着单个或更充分的连接层。这种基本设计的变体普遍存在于图像分类的文献并且迄今为止已经在 MNIST，CIFAR 以及最值得注意的 ImageNet 项目的分类挑战赛[9，21]上取得了最好的结果。对于较大的数据集，如 ImageNet，近年的趋势一直是增加层数[12]和层大小[21，14]，同时使用脱落[7]来解决这个问题过度拟合。

尽管关注的最大池层导致损失精确的空间信息，相同的卷积网络结构[9]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]。

受到灵长类神经视觉系统的启发，Serre 等人[15]使用不同尺寸的 Gabor 滤波器来处理不同尺寸的图片。我们在这里也采用了类似的战略。然而，相对于混合两层深度模型，在开端结构中所有的滤波器都被学会了。此外，起始层被重复了许多次，导致了在 GoogLeNet 模型情况下的 22 层深度模型。

Network-in-Network 是林等人为了增加网络表达能力提出的一种方法。

在他们的模型中，另外的 1 X 1 卷积层被添加到网络中，增加了其深度。在我们的架构中，我们大量使用这种方法。然而在我们的网络中这种结构有双重作用，主要用于维数约减模块来移除计算瓶颈，否则这个瓶颈会限制我们网络的大小。这样不仅允许我们增加深度，而且允许我们增加宽度，而不会带来重大损失。

当前最流行的检测算法就是 Girshick 提出的 Regions-CNN；R-CNN

把检测任务分解为两个子任务：利用“底层的”线索（颜色或者像素的一致性）提取潜在的物体和使用 CNN 来分类潜在的物体。这样的一个两阶段方法，通过低纹理线索促使包围盒分割精度改变，成为一个强大分类能力的神经结构网络。在我们的检测意见书中，我们采用了类似的方法，但在两个方面都探讨了增强阶段，如多盒高对象预测包围盒召回，更好包围盒分类提案的集成方法。

3、动机和高度的考虑

提高深层神经网络性能的最直接的方法是增加其大小。这包括增加深度：净工作水平的数量以及宽度：各层次的单位数量。这是一个简单的和安全的训练更高质量模型的方式，特别是提供一个大的可用性标记训练数据量。然而，这个简单的解决方案有 2 个主要的缺点。

图 1：来自 ILSVRC 2014 1000 多类中的两个不同的类。需要领域知识去区分这些类。

大的网络需要更多的参数，较多的参数在固定的数据集下，容易造成网络过拟合，特别是如果标记的例子在训练集有限。高质量的大数据集获得起来是非常费力和昂贵的，经常需要专家评委区分不同可视化类，如 ImageNet，（即使在 1000 级 ilsvrc 子集）如图 1 所示。

大的网络需要更多的计算资源。例如连个相互连接的卷积网络，一致的增加网络的卷积数目，导致计算量二阶增涨。此外，如果额外增加的网络没有得到有效的利用（很多权值接近 0），会造成计算资源浪费。作为计算，预算总是有限的，计算效率很高资源是首选的不加选择地增加大小，甚至当主要目标是提高质量性能。

解决上了两个问题最基本的方式是：从全连接转到稀疏连接结构。稀疏连接除了模仿生物系统外，Arora 还提出了很好的理论基础。Arora 强调：“数据集的概率分布被一个大的稀疏的网络代表，那么理想的网络拓扑结构应该这样建立（这样的意思是：通过一层一层的统计层间激活值的相关性，然后通过对高度相关的输出进行聚类）”。尽管严格的数学证明需要很多的前提条件，但是这个观点和 Hebbian 准则（fire together，wired together）相互呼应，这也意味着在弱前提条件下，这个想法也是适用的。

此外，涉及到非一致稀疏数据数值计算时，当前的计算设施还很不高效。尽管算数操作的数量减少了 100 倍，查找和缓存丢失是如此显著以至于转到稀疏矩阵没有取得成功。通过利用经过稳定改进的，高度调节的数

值计算库来快速计算密集矩阵相乘，或者详尽地调节 CPU 和 GPU，会使得这种 gap 扩大。非一致稀疏矩阵的计算需要更加复杂的计算设施。当前的视觉机器学习系统只是通过卷积的方式来实现一种空间的局部稀疏性。然而，卷积在基础小的 patches（感觉应该每个 filter 或者 kernal）上是全连接的。传统的卷积网络在特征空间使用随机和稀疏的连接方式；因为他们（LeCun）想打破对称性和提升学习能力；然而为了更好的优化并行计算系统，这种趋势有转回到全连接方式。结构的一致性，大量的过滤器和更大的批量允许使用高效的密集计算。

这就提出了一个问题，是否存在一个中间的步骤：一个可以利用额外稀疏性的结构，甚至在 filter 层，但是通过利用在密集矩阵上的计算来探索当前的硬件。稀疏矩阵计算研究建议：把稀疏矩阵聚类到密集子矩阵能够很好的解决问题。这种方法也可以应用在未来自动建立非一致深度结构上。

本文的 inception 结构起初提出用来评估一个复杂网络拓扑构建算法的假设输出，这个算法尝试近似一个稀疏结构，通过密集，容易获得的成分来覆盖假设输出。尽管是一个高度投机的事业，但相比参考网络的基础，[12]适度的收益在观察早期得到。随着调整差距扩大和开端网络结构在定位和目标检测的背景下被证明是特别有用的的基础网络。有趣的是，虽然大多数原来的结构选择受到质疑彻底的偏离，他们原来却是接近最优的局部。一个必须谨慎：虽然“开端网络结构”已成为一个成功的计算机视觉，仍然值得怀疑的是它的构建是否可以归因于指导原则。确保这将需要更多的彻底分析和验证。

4、结构细节

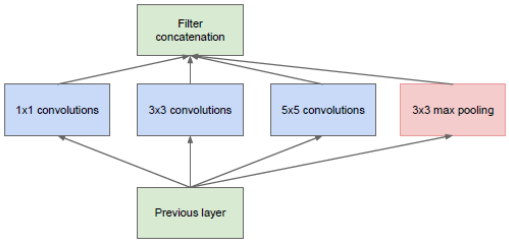
开端网络结构的主要思想是找到卷积网络中理想的局部稀疏结构怎样通过容易获得的密集成分来拟合和覆盖。假设通过卷积块来建立平移不变性，这样我们需要做的就是找到理想的局部构造，然后在空间上拓展重复。Arora 建议要逐层构建：需要分析后一层的统计相关性；然后把高度相关的单元聚类在一起。下一层聚集到一起的单元，连接到上一层的单元。

我们假设前几层的单元对应输入图像的一些区域，这些单元被聚集到 filter bank 中；这样前几层的相关单元聚集到局部区域。这意味着，许多的聚类被聚集到一个单一区域，他们可以被下一层的 1*1 的卷积覆盖。然而，我们也希望存在少数的空间拓展的聚类，这些聚类可以被在大的 patch 上的卷积覆盖，随着 patch 区域的增大，数量应该减小。

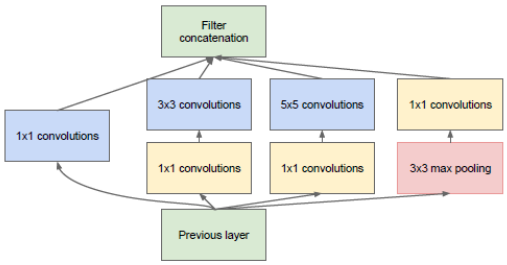
为了避免 patch alignment（区域布局）问题，当前的 inception 结构被限制在卷积尺寸 1*1，3*3 和 5*5，这种设计是基于更加的方便而不是必

须的。这种结构也意味着所有层 filter 的输出被聚集到一个单一的输出向量，从而形成下一层的输入。此外，由于 pooling 结构是当前卷积网络成功的关键，在每个阶段增加一个可选择的并行 pooling 结构。

上面 inception 模块的一个问题是计算问题，因为一个中等的 5*5 卷积运算在有很多的特征存在情况下都是被禁止的。这个问题更加的突显出来当 pooling 层加入其中时：输出 filter 的数量等于等于输入 filter 的数量。融合 pooling 层和卷基层层的输出会导致一个不可避免的增加输出的数量。尽管这个结构能够覆盖理想的稀疏结构，在计算上非常不高效。



（a）开端模块，幼稚版



（b）降维的开端模块

图 2：开端模块

这就引起了本文的第二个考量：当计算需求增加的时候，应用维数约减和 projection（投影）；这个是基于低维的嵌入可能包涵许多大的图像区域的信息。然而以一个密集，压缩的方法嵌入代表的信息和压缩的信息是非常困难的。我们只是想在大多数地方保持特征稀疏，压缩信息仅仅在特征不得不融合在一起时；这就是为什么 1*1 的卷积在 3*3 和 5*5 卷积前使用的原因；1*1filter 出了维数约减外，还起到增加非线性的作用。

一般滴，inception 网络是一个有 inception 模块堆叠组成的网络；偶尔使用 s=2 的 max-pooling 来使特征减半。由于技术的原因（训练阶段存储效率），似乎在高层使用 inception 模块更加有益，让底层的保持原始的卷积网络形状。

这个结构的主要贡献是：在可控的计算复杂度增量下，增加每个阶段的单元个数。无处不在的维数约减，允许上一层的大量的输入 filter “shield”到下一层。在一个大的 patch 卷积之前，先进行维数约减。

另一个有用的方面就是这种设计灵感来源于：视觉信息应该在不同尺度上处理，然后融合；以便下一层能够从不同的规模上提取特征。

5. GoogLeNet

由“googlenet”的名称，我们提及特殊的化身在我们提交的初始架构，被用于 ilsvrc 2014 赛事。我们也使用了一个深度和更广泛的初始网络与略优质量，但加入到合奏似乎仅

轻微提高了结果。我们省略了这个网络的细节，作为经验证据表明，确切影响结构的参数相对较小。表 1 所示最常见的实例中使用的竞争。这个网络（在我们的合奏中，不同的图像块训练 7 个模型中使用了 6 个采样方法。所有的部分，包括在本文开端网络模

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

表一：开端结构的典型----GoogleNet

备，包括那些有有限的计算资源，特

块内部的，都使用线性激活。它的小

在我们的网络中 RGB 颜色零均值空间可接受的范围是 224X224。“# 3X3 减少”和“# 5X5 减少”代表减少层中的 1X1 个过滤器的数目以前的 3X3 和 5X5 的卷积。一个可以看到在投影层中的 1X1 个过滤器的数目内置最大池的池项目列。所有这些减少投影层也使用整流线性激活。

该网络的设计与计算效率思维的实用性，使推理可以在个体上运行设

别是低内存占用。网络是 22 层深，当仅计算有参数的层（或 27 个层，如果我们也算池）。这个层（独立大厦块）使用的整体数量对于网络的建设是约 100。准确数字取决于通过机器学习基础设施如何计算层。平均池前使用分类是基于[12]，虽然我们的实施具有附加的线性层。线性层使我们能够很容易适应我们的网络到其他标签集，但它是大多是为了方便，我们不

希望它有重大影响。我们发现，从完全连接的移动层平均池 2 精度提高约 0.6%，但使用脱落仍然是必不可少的，即使删除完全连接的层。

鉴于网络的深度比较大，以一个有效的方式传播通过所有的层能力是一个值得关注的问题。在这项任务上较浅的网络表现强劲表明，在网络的中间层产生该功能的应该是很有判别力。通过增加备用分类连接到这些中间层，在分类的较低阶段的判别力被预期。这被认为在提供正规化的同时对抗消失的度问题。这些分类采取的形式规模较小的卷积网络的输出端的起始（4A）和（4D）模块。在训练过程中，他们的损失增加了网络的总损失具有折扣的重量（辅助分类的损失被加权 0.3）。在推论时期，这些备用网络被丢弃。后来控制实验示出的备用网络的效果是相对少数的（约 0.5%），他们只需要达到一个同样的效果。

侧面的额外网络的确切结构，包括备用分类，如下：

•一个平均池层有 5X5 个过滤器的大小和进展 3，导致 4X4X512 的输出为（4A），和 4X4X528 的（4D）阶段。

•用 128 个 1X1 个过滤器进行降维线性激活。

•一个完全连接的层，有 1024 个单位和整流线性激活。

•一个有 70%个比例的下降输出层。

•与 Softmax 损失作为分类器线性层（预测相同的 1000 个类作为主要的分类，但在推理时删除。由此产生的网络的示意图图 3。

6、训练方法

使用 google 的 DistBelief 分布式机器学习系统；本文只在 CPU 上部署，主要的限制来自于内存。本文使用异步的 SGD，动量项=0.9；每迭代 8 次，减小 4%的学习率。

对于图像采样方法一直在改变，并没固定的方法；已经收敛的网络是在其他的选择下训练的，所以给出一个有效的训练网络的方法是困难的。更为复杂的是，一些网络在相对小的 patch 上训练，一些网络在相对大的 patch 上训练，；还有一个技巧被证明很有效：提取各种各样的 patches，patches 的 size 均匀地分布在图像区域 8%和 100%之间，比例也在 3/4 和 4/3 之间随机选择。还有一些其他数据增益技术。

7、 ILSVRC2014 分类

2014 个挑战涉及的 ilsvrc 分类在 ImageNet 层次将图像分成 1000 个叶节点类别中的任务。大约有 120 万训练图像，50000 用于验证和 100000 测试图像。每个图像都与一个地面真的类别，和性能测量为基础得分最高的分类预测。双号码通常报道的的准确率，

其中：对第一个预测类的地面真理进

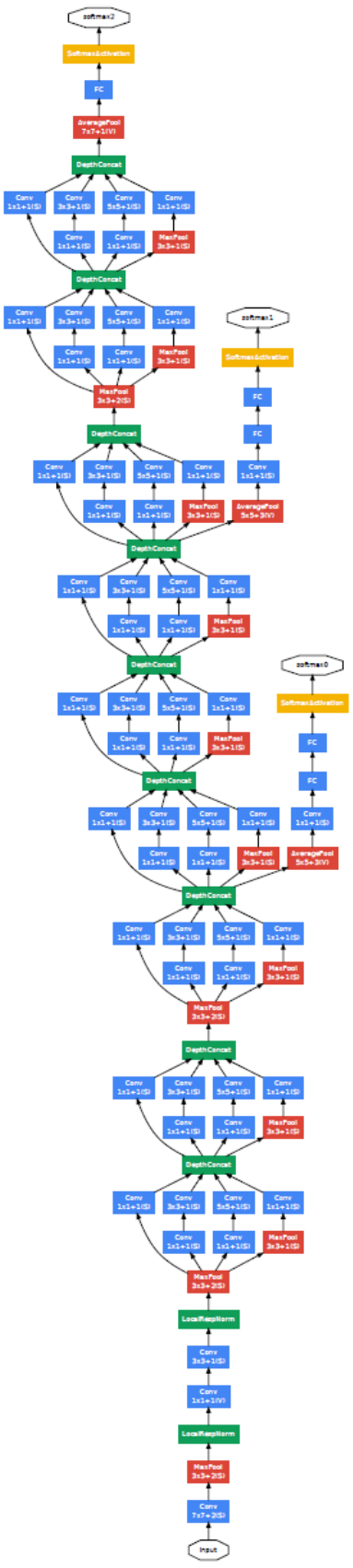


图 3:拥有所有钟声和口哨声的 GoogLeNet network

行比较，与前 5 的误差率，比较真实对前 5 个预测类：一个图像被认为正确分类如果地面真理是在 5，不管它的地位在他们。挑战采用前五错误率的目的排名。

我们参加了没有外部数据的挑战用于训练。除了训练技巧在本文中，我们采用了一套技术在测试过程中，以获得更高的性能，这是我们接下来要描述的。

1、我们独立训练的 7 个版本相同 googlenet 模型（包括一个更大的版本），和与他们进行整体预测。这些模型进行了训练与相同的初始化（即使由于相同的初始权重，由于一个监督）和学习率政策。他们不同的只是抽样方法与随机输入图像序列。

2、在测试过程中，我们比 krizhevsky 等人采取了更积极的截弃方法。[9]。具体地，我们调整了图像的 4 个尺度，较短的尺寸(高度或宽度)为 256，288，320 和 352，分别采取左，中心和右这些大小的图像方格（就图像，我们采取的顶部，中心和底部方格）。对于每一个方格，我们然后采取 4 个角落和中心 224X224 方格以及平方大小 224X224，和他们的镜像版本。这导致 4X3X6X2 = 144 截弃每张图。类似的方法在过去的霍华德安得烈[8]一年的记录中用过，我们的经验验证执行略差于拟议的计划。我们注意，这种侵略性的裁剪可能不是必要的在实际应用中，由于更多的受益，现在一个合

理的截弃量就变得很贫瘠（像我们稍后将展示的那样）。

3、SOFTMAX 概率的平均值在多个裁剪和超过所有的个人分类，以获得最后预测。在我们的实验中，我们分析验证数据的替代方法，如最大集中在裁剪和平均超过分类，但它们导致性能不如简单的平均。

在本文的其余部分，我们分析了多因素，有助于整体性能的最终提升。

我们最后提交的挑战获得前 5 的误差率为 6.67%的验证和测试数据，排在众多参与者中的第一个。这相对减少了 56.5%的相比于 2012 的有监督方法，减少了约 40%相对于以前的最佳方法（clarifai），它们都均采用外部数据训练分类器。表 2 显示了在过去的 3 年里顶级的方法。

我们还分析和报告多个性能测试选择，通过不同数量的模型和不同量的裁剪，用来预测表 3 中的图。当我们使用一个模型，我们选择了一个最低验证数据的误差率。为了不过度拟合的验证数据集测试数据统计，所有的数字都有报道在数据集中。

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

表:2：分类性能

Number of models	Number of Crops	Cost	Top-5 error	compared to base
1	1	1	10.07%	base
1	10	10	9.15%	-0.92%
1	144	144	7.89%	-2.18%
7	1	7	8.09%	-1.98%
7	10	70	7.62%	-2.45%
7	144	1008	6.67%	-3.45%

表 3：GoogleNet 分类性能降低

8、ILSVRC 2014 检测和结果

ILSVRC 检测任务是预测图像中大约 200 个类别的边框。检测方法类似 R-CNN，首先进行区域提取，通过结合 Selective Search 和 multi-box 预测来改进区域提取算法。为了减少错误 positives 的数量，像素增加 2 倍；这种方法把 selective search 提取的区域减半，然后又从 multi-box 结果中添加 200 个区域，最后大约使用 60%的区域，

Team	Year	Place	mAP	external data	ensemble	approach
UvA-Euvision	2013	1st	22.6%	none	?	Fisher vectors
Deep Insight	2014	3rd	40.5%	ImageNet 1k	3	CNN
CUHK DeepID-Net	2014	2nd	40.7%	ImageNet 1k	?	CNN
GoogLeNet	2014	1st	43.9%	ImageNet 1k	6	CNN

表 4：检测性能比较。未报告的值用问号标注

同时增加覆盖范围从 92%到 93%。削减区域最后使 mAP 提高 1%，最后使用 6 个网络来对每个区域进行分类。我们在没有使用额外预训练数据的情况下，没有使用 bounding boxregression

的情况喜爱，获得了单个网络获得38.02%;6个网络获得了43.9%的成绩;说明multi-module对于分类和检测都很重要。

我们首先报告的顶部检测结果和显示自检测任务的第一版进展。相比到2013个结果,准确率几乎翻了一番。最优秀的表演团队都使用了卷积网络。我们报告的官方分数在表 4 中并且对于每一个团队共同的策略：使用外部数据，集合模型或上下文模型。外部数据通常是预训练的模型，ilsvrc12 分类数据后来在检测数据上提炼。有些球队还提到定位数据的使用。自一个好的部分本地化任务边界框不包含在检测数据集，可以预先训练一个一般的边界这一数据相同的方式解释变量分类盒用于预训练。Googlenet 记录没有使用定位数据培训。

在表 5 中，我们使用一个单一模型进行比较。顶级性能模型是令人瞠目的只提高了有着 3 个模型整体 0.3 点，而 googlenet 得到明显更强的整体效果。

9、结论

通过容易获得的密集 building block 来近似理想的稀疏结构是一个可行的方法。这种方法的主要贡献是在微小增加计算量的情况下明显地提升效果。我们的方法有力地证明了转移到稀疏结构是一个灵活并且有用的想法。这个主要优点方法是相比于浅和较窄的体系结构，在一个温和的增长计算方面有显著的质量增益。

我们的目标检测工作是有竞争力的，尽管没有利用上下文，也没有进行包围盒回归，这意味着还进一步证明了“开端结构”的优势。对于这两种分类和检测，人们认为类似质量的结果可以实现更昂贵类似深度的非初始类型网络宽度。然而，我们的方法产生了坚实的证据，移动稀疏的架构是可行的和有用的想法一般。这表明未来旨在努力创建稀疏和更精细的结构，在自动化的基础上[2]，以及关于应用的见解其他领域的体系结构。

Team	mAP	Contextual model	Bounding box regression
Trimps-Soushen	31.6%	no	?
Berkeley Vision	34.5%	no	yes
UvA-Euvision	35.4%	?	?
CUHK DeepID-Net2	37.7%	no	?
GoogLeNet	38.02%	no	no
Deep Insight	40.2%	yes	yes

表 5：单模型性能检测

参考文献

[1] Know your meme: We need to go deeper. http://knowyourmeme.com/memes/we-need-to-go-deeper. Accessed: 2014-09-15.

[2] S. Arora, A. Bhaskara, R. Ge,

and T. Ma. Provable bounds for learning some deep representations. CoRR, abs/1310.6343, 2013.
[3] U. V. C, ataly ¨urek, C. Aykanat, and B. Uc,ar. On two-dimensional sparse matrix partitioning: Models, methods, and a recipe. SIAM J. Sci. Comput., 32(2):656–683, Feb. 2010.
[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng. Large scale distributed deep networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, NIPS, pages 1232–1240. 2012.
[5] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In CVPR, 2014.
[6] R. B. Girshick, J. Donahue, T. Darrell, and J. 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] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580, 2012.
[8] A. G. Howard. Some improvements on deep convolutional neural network based image classification. CoRR, abs/1312.5402, 2013.
[9] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional 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, Dec. 1989.
[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition.

Proceedings of the IEEE, 86(11):2278 - 2324, 1998.
[12] M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013.
[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] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. CoRR, abs/1312.6229, 2013.
[15] T. Serre, L.Wolf, S. M. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell., 29(3):411 - 426, 2007.
[16] F. Song and J. Dongarra. Scaling up matrix computations on shared-memory manycore systems with 1000 cpu cores. In Proceedings of the 28th ACM International Conference on Supercomputing, ICS ' 14, pages 333 - 342, New York, NY, USA, 2014. ACM.
[17] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton. On the importance of initialization and momentum in deep learning. In ICML, volume 28 of JMLR Proceedings, pages 1139 - 1147. JMLR.org, 2013.
[18] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, NIPS, pages 2553 - 2561, 2013.
[19] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. CoRR, abs/1312.4659, 2013.
[20] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In Proceedings of the 2011 International Conference on Computer Vision, ICCV ' 11, pages 1879 - 1886, Washington, DC, USA, 2011. IEEE Computer Society.
[21] M. D. Zeiler and R. Fergus. Visualizing and understanding

convolutional networks. In D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, ECCV, volume 8689 of Lecture Notes in Computer Science, pages 818–833. Springer, 2014.