VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNTION

摘要

在这篇文章中我们调查了大规模图片识别中卷积网络的深度对于精度的效果。我们的主要贡献是用一个有着 3*3 的卷积核的结构对深度增加的网络进行评估。它说明了把网络的深度提高到 16 到 19 层后会在先前最好的组态上性能有明显的提升。这些发现是我们队伍能在 2014 年 ImageNet 挑战中定位和分类赛中分别取得第一名和第二名的基础。对其它的数据库，我们的表现也非常好，达到了最好的记录。我们把表现最好的两个卷积网络模型公布于众，以供在计算机视觉的深度视觉表示方面更深的研究。

1 介绍

最近卷积网络在大规模图片和视频识别中获得了很大的成功，这是因为大规模的图片库（例如 ImageNet）和高性能的计算机系统（例如 GPU 或者大规模分布式集群），特别的，在深度视觉识别结构进步中扮演重要角色的是 ILSVRC，它给一些大规模的图片识别系统提供了一个测试的平台。例如 high_dimensional shallow feature encoding 和 deepConvNet

随着卷积网络成为计算机视觉领域的大部分模型，一些人尝试着去改进 AlexNet 的结构以获得更好的精度。例如在 2013 年 ILSVRC 比赛中表现最好的在第一个卷积层使用了更小的窗口尺寸和步进值，另一个在训练网络过程中是用来全部的图片和各种尺寸。在这篇文章中我们强调卷积网络结构中另一个重要的方面——卷积网络的深度。在最后我们发现这个结构的参数，并没有因为增加更多的卷积层提升深度而急剧增加，这可能是因为我们在所有的卷积层都使用了非常小的卷积核（3*3）.

最终我们得到了更加精确的卷积网络结构，它不仅在 ILSVRC 定位和分类比赛中达到了很好的精度，而且对其他的图片识别数据库也有很好的表现。甚至当我们使用相对简单的结构它也表现的很好。我们将表现最好的两个模型公开以供更深的研究。

本篇文章的其余部分如下所示组织。在第二部分，描述了卷积网络的组态。第三部分提供了图片分类和评估训练的细节。第四部分各种组态在 ILSVRC 分类任务上进行比较。第五部分总结了这篇文章。为了补充，我们在附录 A 中描述和评估了我们的 ILSVRC-2014

2 卷积组态

为了公正地测量由增加卷积网络深度带来的提升，我们所有的卷积网络层组态以同样的原则设计。在这个部分，我们首先描述了卷积网络组态的大体分布，然后详细描述在评估中用到的特殊的组态。在 2.3 部分我们讨论了设计的选择和与之前的较好的成果进行比较。

2.1 结构

在训练期间，卷积网络的输入是 224*224*3 的 RGB 图片。我们做的唯一的预处理就是在训练集上每个像素都减去 RGB 值的均值。图片通过一系列的卷积层，这些卷积层有着小的

感知域：3*3（这是捕捉到左右，上下，中心这些点的最小尺寸）。我们在一种组态中是用了 1*1 的卷积核，它可以被看做是对输入通道的一个线性转换器（后面跟着非线性）。卷积的步进值被设置为一个像素。卷积层的空间填充是指使得在卷积操作后保留原空间的分辨率，比如使用 3×3 卷积核，就填充 1 个像素。空间池化是由五个最大池化层完成的，每个池化层前面都会有若干个卷积层（并非所有的卷积层后都使用最大池化层）。 最大池化是以 2×2 像素窗口上执行，步幅为 2。

一堆卷积层（在不同的结构中有不同的深度）之后跟着三个全连接层，前两个全连接层每个都有 4096 个通道，第三个执行 1000ILSVRC 分类功能因此有一千个通道（每个通道代表一类）。最后一层是 softmax 层。所有网络的全连接层设置都是一样的。

所有隐藏层都有 ReLU 非线性层。我们注意到我们的网络没有一个装备有本地响应归一化，这种归一化在 ILSVRC 数据中没有提升网络的表现。但是却导致了内存消耗和计算时间的增加。在使用的情况下，LRN 的参数和 Krizhevsky 在 2012 年的一致。

2.2 组态

本文评估的卷积网络组态，在表一中列出。每列一种。下文中我们讲用网络的名字来代表他们（A 到 E）。所有网络的总体设计都在 2.1 部分提供，他们仅仅在深度上有所不同：从 A 的十一层的网络到 E 的十九层。卷积网络的宽度（频道数）非常小，从第一层的 64 在每个最大池化层后以二倍的因子增加到 512.

在表二中我们报告了每个网络组态的参数数目，尽管有着很大的深度，权重的数量并没有比有着更大卷积层和感知域的浅层网络多很多。

2.3 讨论

我们的网络组态和 ILSVRC2012 和 ILSVRC2013 中表现最好的模型有很大的不同。相比于在第一个卷积层用相对较大的感知域（例如 11*11 步进为 4 或者 7*7 步进为 2），我们在整个网络使用了非常小的 3*3 的感知域，它和输入的每个像素进行卷积（步进为 1）。很容易看出两个 3*3 的卷积层（中间没有池化）和 5*5 的卷积层有着同样大的感知域；三个 3*3 的卷积层和 7*7 的卷积层有着相同的感知域。所以我们用小的卷积核有什么收益呢？例如说，一堆 3*3 的卷积核来代替一个 7*7 的卷积层。第一，我们用三个非线性层代替了一个，这使得辨识度更高。第二我们降低了参数的数量，假设输入和输出都使用一个三层的 3*3 的卷积核有 C 个通道，参数数目是 3（3C*3C）=27C2.如果使用一个 7*7 的卷积层，这一层需要 7*7*C*C=49C2 的参数，参数增加 81％。 这可以被看作是在 7×7 卷积中实施正规化， 迫使他们通过 3×3 卷积核进行分解（两者之间注入非线性）。

使用 1×1 卷积核（配置 C，表 1）是一种增加决策函数的非线性而不影响卷积层感受野的方法。 尽管在我们的例子中，1×1 卷积本质上是一个线性投影到相同维度的空间上（输入和输出通道的数目是相同的），但激活函数引入了一个额外的非线性。 应该指出，最近 Lin 等人在其“网络中的网络”结构中使用了这种 1×1 卷积层（2014）。

Ciresan 等人以前曾使用过小尺寸的卷积滤波器 （2011 年），但他们的网络明显不如我们的深，并且他们没有对大规模 ILSVRC 数据集进行评估。 Goodfellow 等人（2014）将深度 ConvNets（11 个权重层）应用于街道号识别任务，并表明增加深度能获得更好的性能。GoogLeNet 是 ILSVRC-2014 分类任务中性能最好的一个入门版本，它的开发与我们的工作相独立，但相似的地方是都是基于非常深的卷积网络（22 个加权层）和小卷积滤波器（除 3×3 外，他们也使用了 1×1 和 5×5 卷积）。但是，它们的网络拓扑结构比我们的要复杂，并

且在第一层中，特征映射的空间分辨率减少的更加剧烈以减少计算量。正如将在 4.5 节中所显示的那样，我们的模型超过了 Szegedy 等人的模型（2014 年）的单网分类准确性。

3 分类框架

在前面的部分我们呈现了网络组态的细节，这个部分骂我们描述分类网络训练和评估的细节。

3.1 训练

卷积网络的训练过程大体上仿照 Krizhecsky2010（除了对不同尺寸的输入图片进行切片取样，接下来会解释到）。训练是通过使用小批量梯度下降（基于反向传播）的动量优化多项逻辑回归目标来实现的。batch 大小设置为 256，momentum 为 0.9。训练通过权值衰减（L2 惩罚系数设置为 $5*10^{-4}$）和前两个完全连接层（dropout 设置为 0.5）的 dropout 正则化来调整。 学习率最初设置为 0.01，然后在验证集精度停止提升时再降低 10 倍。 总的来说，学习率一共降低了 3 次，并且在 370K 个迭代（74 代）后停止了学习。 我们推测，尽管与 Krizhevsky2012 相比，网络的参数数量更多，网络深度也更大，但能用更少的迭代次数来实现收敛，这是因为：（a）更大深度和更小卷积核所带来的隐式正则化；（b）某些图层的预初始化。

网络的初始权重是很重要的，因为不好的初始化会导致深度网络的梯度的不稳定。为了解决这个问题，我们从组态 A 开始训练，它足够浅所以可以用随机初始化来训练。当训练更深的结构时，我们用 A 网络的值来初始化前四个卷积层和后三个全连接层（其它层随机初始化）。对于预初始化的层，我们不减少其学习率，让他们随着训练过程改变。对于随机初始化的层，我们用均值为 0 方差为 0.01 的正太分布来初始化。值得注意的是，在提交论文后我们发现可以用 Glorot&Bengio2010 年的随机初始化过程来初始化权重值而不用预训练。

为了保证 224*224 的卷积网络输入。他们从缩放的训练图片中随机裁剪（每个随机梯度下降迭代每张图片裁剪一次）。为了更好的扩充训练集，被裁剪的图像经过随机水平翻转和随机 RGB 颜色偏移处理（Krizhevsky 2012）。下面将介绍训练图像缩放。

训练图像尺寸。设 S 是等比例缩放的训练图像的最小边，ConvNet 基于这些图像的裁剪作为输入（我们也称 S 为训练尺度）。 虽然裁剪大小固定为 224×224，但原则上 S 可以取不小于 224 的任何值：对于 S = 224，裁剪图将捕获整幅图像统计数据，完全跨越训练图像的最小边; 对于 S>224，裁剪图将对应于图像的一小部分，包含一个小物体或物体的一部分。

我们考虑设定训练尺度 S 的两种方法。第一种方法是固定 S，这对应于单尺度训练（注意采样作物中的图像内容仍然可以表示多尺度图像统计）。 在我们的实验中，我们评估了以两个固定尺度训练的模型：S = 256（已被广泛用于现有技术（Krizhevsky 等，2012; Zeiler＆Fergus，2013; Sermanet 等，2014））和 S = 384。给定一个 ConvNet 配置，我们首先使用 S = 256 来训练网络。为了加速 S = 384 网络的训练，它被初始化为具有 S = 256 的预训练权重，并且我们使用较小的学习率 0.001。

设定 S 的第二种方法是多尺度训练，其中通过从特定范围[Smin，Smax]（我们使用 Smin = 256 和 Smax = 512）随机采样 S 来独立缩放每张训练图像。 由于图像中的物体可能具有不同的大小，因此在训练时考虑到这一点是有益的。这也可以看作是通过尺寸抖动来增强训练集，其中单个模型被训练以识别多种类别的物体。出于速度的原因，我们通过对具有相同配置的单尺度模型的所有层进行微调来训练多尺度模型，并使用固定的 S = 384 进行预训练。

3.2 测试

在测试时，给定一个训练过的 ConvNet 和一个输入图像，它按以下方式分类。首先，将其等比例缩放到预定义的最小边，表示为 Q（我们也将其称为测试尺度）。我们注意到，Q 不一定等于训练尺度 S（如我们将在第 4 部分中所示，对每个 S 使用几个 Q 值可以提高性能）。然后，网络以类似于（Sermanet 等人，2014）的方式被密集地应用在重新缩放的测试图像上。也就是说，完全连接的层首先被转换成卷积层（第一个 FC 层转为 7×7 的卷积层，后两个 FC 层转为 1×1 卷积层）。然后将所得的全卷积网络应用于整个（未裁剪的）图像。其结果是一个类别得分映射，其通道数等于任务的目标分类数，以及一个可变的空间分辨率，取决于输入图像的大小。最后，为了获得固定大小的图像类别分数的向量，类别得分映射会被空间平均（加总池化）。我们还通过水平翻转图像来增强测试集；对原始图像和翻转图像的 softmax 分类概率进行平均以获得图像的最终分数。

由于全卷积网络应用于整个图像，因此不需要在测试时间对多个裁剪图像进行采样（Krizhevsky et al。，2012），这样效率较低，因为它需要网络对每个裁剪图像进行重新计算。同时，使用大量的裁剪图像，如 Szegedy 等人（2014）所做的那样可以提高准确性，因为与全卷积网络相比，它可以更精细地对输入图像进行采样。此外，由于卷积边界条件不同，多裁剪图像评估与密集评估是互补的：将 ConvNet 应用于裁剪图像时，卷积后的特征映射用零填充，而在密集评估的情况下，同一裁切图像的填充天然地来自于图像的相邻部分（由于卷积和空间池化），这大大增加了整个网络的感受野，因此捕获更多的上下文信息。尽管我们认为在实践中增加多裁切图像的计算时间并不能证明潜在的准确度增加，但我们对于每种尺寸规模（5×5 个常规栅格和 2 种翻转）都使用 50 个裁切图像来评估我们的网络，总共 150 个裁切图像、超过 3 个尺度，这与 Szegedy 等人使用的 4 种尺度、144 个裁切图像相当（2014）。

3.3 实现细节

我们的实现源自公开发布的 C ++ Caffe 工具箱（Jia，2013）（2013 年 12 月推出），但包含许多重大修改，使得我们能在装有多个 GPU 的单系统中执行训练和评估，以及能够对多种规模的全尺寸（未裁剪）图像（如上所述）进行训练和评估。多 GPU 训练利用数据并行性，并且通过将每批训练图像分成几个 GPU 批次并在每个 GPU 上并行处理来执行。计算 GPU 批梯度后，计算它们的均值以获得完整批次的梯度。梯度计算在 GPU 中是同步的，因此结果与在单个 GPU 上进行训练时完全相同。尽管最近提出了更加复杂的加速 ConvNet 训练的方法（Krizhevsky，2014），它们针对网络的不同层使用模型和数据并行性，但我们发现，与使用单个 GPU 相比，我们概念更简单的方案在现成的 4 GPU 系统上已经提供了 3.75 倍的加速。在配备四个 NVIDIA Titan Black GPU 的系统上，根据架构的不同，训练一个网络需要 2-3 周的时间。

4 分类实验

数据库 在这个部分我们提供了卷积网络结构在 ILSVRC2014 数据库上图片分类的结果（被用于 ILSVRC2012-2014 挑战）。数据库包含 1000 类图片，分为三个部分，训练集（1.3M 图片），验证集（50K 图片），测试集 （100K 带分类标记的图片）。分类性能用两个方法来评估。Top-和 top-5 错误率。前者是前者是多分类误差，例如错误分类的图片比例，后者是 ILSVRC 主要的评估方法。按照 Top-5 预测分类中不存在真实分类的图像所占比例计算。对于大多数实验，我们使用验证集作为测试集。还对测试集进行了一些实验，并将其作为 ILGVRC-2014

竞赛“VGG”团队的参赛作品（Russakovsky 等，2014）提交给官方 ILSVRC 服务器。

4.1 单尺寸评估

我们以评估在 2.2 部分描述的单尺寸独立卷积网络模型网络组态开始，图片尺寸设置如下: 对于固定的 S，Q=S，对于 S 属于[Smin,Smax]之间的，Q=0.5（Smin+Smax）.结果在表 3 中展示。

首先我们注意到在模型 A 使用本地响应归一化并没有比没有使用 LRN 有所提高。因此我们没有在更深的网络结构使用正则化。

第二，我们观察到随着深度增加分类错误减少，从 11 层的 A 到 19 层的 E。值得注意的是，尽管在同样深度，组态 C（包含有三个 1*1 的卷积核）表现比使用 3*3 的卷积核的组态 D 要差。这说明额外的非线性并没有用（C 比 B 表现好）。当深度达到 19 层后结构的错误率达到饱和，但是更深的网络结构可能适合更大的数据库。我们也把网络 B 与一个带有 5*5 卷积核的浅层网络进行了比较，这个网络来源于网络 B，其将网络的一对 3*3 卷积核换成了一个 5*5 的卷积核（他们的有着相同的感受野）， 浅层网络的 top-1 错误率要比网络 B 的错误率高 7%（在中心裁剪的条件下），这使我们确信小滤波器的深层网络比大滤波器的浅层网络表现更好。

最后，在训练时加入尺寸抖动（S 属于[Smin,Smax]）比训练时用固定的小边（S=256 或者 S=384）导致了更好的结果，甚至在测试时只使用单一的尺寸。这证明了使用尺寸抖动来扩充训练集确实对捕捉多尺寸图片数据有所帮助。

4.2 多尺寸评估

使用单尺寸评估了卷积网络模型后，现在我们评估在测试时尺度抖动的作用。它由运行一个模型在一张测试图片的几个缩放版本上组成（对应于不同的 Q 值）。随后将分类结果平均。考虑到训练和测试尺寸的巨大不同会导致性能的下降，用固定训练的模型在三个尺寸的测试图片进行评估，测试尺寸接近训练尺寸，Q={S-32,S,S+32}.同时，在训练时的尺寸抖动使得网络在测试时适用于更宽范围的尺寸。所以用 S 在[Smin,Smax]之间训练的模型在一个更大范围的尺寸 Q={Smin,0.5（Smin+Smax），Smax}进行评估。

结果（呈现在表 4）说明在测试时使用尺寸抖动表现更好（在表 3 内展示了，相比于用固定尺寸来评估一个模型）。如前文所说，最深的组态（D 和 E）表现最好，尺寸抖动比用一个最小的边 S 训练表现好。我们最好的单网络在验证集上的表现是 top-1 和 top-5 的错误率分别是 24.8%和 7.5%，（在表 4 中突出显示），在测试集组态 E 达到了 top-5 错误率 7.3%。

4.3 多切片评估

在表 5 中我们比较了密集卷积网络和多切片评估进行比较（详见 3.2 部分），我们通过平均它们的 softmax 输入评估了两种技术的互补性，可以看出使用多切片比密集评估表现稍微好一点。事实上这两种方法是互补的，所以它们配合起来表现比每一种都要好。上面表明，我们才行这是因为卷积边界条件不同的处理。

4.4 卷积网络融合

到现在为止，我们评估的都是单独的卷积网络模型的表现。这部分实验，我们通过平均他

们的 softmax 分类概率均值综合了几种模型的输出。这提高了性能由于模型之间的互补，这种方法在 2012 年（Krizhevsky 等，2012）和 2013 年（Zeiler＆Fergus，2013; Sermanet 等，2014）的最优 ILSVRC 提交中使用过。

结果在表 6 中展示。但是本次 ILSVRC 比赛中我们只训练了单尺度的网络，和多尺度的模型 D（通过仅对完全连接层而不是所有层进行微调），结果由此产生的 7 个网络的融合模型具有 7.3％的 ILSVRC 测试错误率。 提交之后，我们考虑了只有两个表现最好的多尺度模型（配置 D 和 E）的集合，它使用密集评估将测试错误率降低到 7.0％，而使用组合密集和多裁切图像评估将测试错误率则降低到 6.8％。作为参考，我们表现最佳的单模型实现了 7.1％的误差（模型 E，表 5）。

4.5 与当前最先进技术比较

最后我们将我们的结果和表 7 中最先进的结果进行比较，在 ILSVRC-2014 挑战赛分类任务中，我们 VGG 队使用七个模型融合以 7.3%的错误率获得了第二名，提交后，我们用两个模型的融合将错误率降低到了 6.8%。

可以从表 7 看出，我们的深度卷积网络比之前在 ILSVRC-2012 和 ILSVRC2013 比赛的最好结果的模型表现好的多。我们的结果也对分类任务获胜者（GoogLeNet 有 6.7％的错误率）具有竞争力，并且大大优于 ILSVRC-2013 的获奖提交 Clarifai，其使用外部训练数据达到 11.2％，而不使用外部数据则为 11.7％。考虑到我们的最佳结果仅仅是通过两个模型融合来实现的，并显著地低于大多数 ILSVRC 提交中的使用的模型，这相当不简单。 就单网性能而言，我们的架构实现了最好的结果（7.0％的测试错误），优于单个 GoogLeNet 0.9％。 值得注意的是，我们并没有偏离 LeCun 等人的经典 ConvNet 架构（1989），但通过大幅增加深度来改善它。

5 总结

在这项工作中，我们评估了用于大规模图像分类的深层卷积网络（多达 19 个权值层）。 已经证明，表示层的深度有利于分类准确性，并且通过大幅增加网络深度便可以使用传统的 ConvNet 架构来实现 ImageNet 挑战数据集上的最新性能（LeCun 等，1989; Krizhevsky 等，2012）。 在附录中，我们还展示了我们的模型能很好地泛化应用于其他的任务和数据集，不亚于甚至性能优于那些深度略浅、更复杂的识别流水线。 我们的结果再一次证实了视觉表示中深度的重要性。