ImageNet Classification with Deep Convolutional Neural Networks

摘要

我们训练了一格大型的，深度的卷积神经网络去辨别 ImageNet LSVRC-2010 比赛中的一千种不同种类的十二万张高分辨率的图片。在测试集上，我们达到了 top-1，top-5 错误率分别为 37.5%和 17.0%的成绩，这被认为比之前最好的成绩还要好。这个有着六千万参数和 650,000 的神经元的神经网络，由五个卷积层和三个全连接层组成，一些卷积层后跟着 max-pooling 层，全连接层的最后一层是 1000-way softmax。为了使训练更快，我们用了不饱和的神经元和卷积运算非常高效的 GPU。为了减少全连接层的过拟合，我们使用了最近发展出的正则化方法 dropout,这个被证明非常有效。我们也在 ILSVRC-2012 比赛中加入了这个模型，和第二名的 26.2%相比，我们的 top-5 错误率达到了 15.3%。

1 介绍

如今物体识别的方法对于机器学习非常重要。为了提高他们的性能，我们可以收集更大的数据库，学习更加强大的模型，用更好的技术防止过拟合。直到最近，有标记的图片数据集仍然非常小，只有 10k 的图片。在这种尺寸的数据集中简单的识别任务可以被很好的解决，特别是他们通过提供标签转换增强，例如说在 MNIST 数据集数字识别任务的最好的错误率已经比人类的表现还要好了（<0.3%），但是真实世界的物体呈现更加多样化的不同，所以为了识别他们需要更大的数据集，实际上，小的图片数据集的缺点已经被广泛的承认，但是直到最近收集百万张图片才成为可能。新的更大的数据库包括 LableMe,是有十万张被完全分割的图片组成，ImageNet 由超过 22，000 种类的 1500 万张被标记的高分辨率的图片组成。

为了从百万张图片中识别这几千种物体，我们需要一个有着强大学习能力的模型，然而，物体识别任务极高的复杂度意味着即使在 ImageNet 这么大的数据集上也很难被具体化，所以我们的模型应该有大量的先验知识去补偿我们没有的数据。卷积神经网络组成这种模型。他们的能力可以通过控制其深度和广度改变，通过神经网络他们也可以做出强大且大部分正确的假设，因此和有着相近尺寸的层的标准的前项神经网络比较，CNNs 有着更少的连接和参数，因而更容易训练，理论上最好的性能只差了一点点。尽管 CNNs 的具有吸引力的质量，尽管其高效的本地结构，但是将其应用于如此庞大规模的高分辨率的图片仍然是非常昂贵的。幸运的是如今的 GPUs，有着高度优化的 2d 卷积运算能力，足够强大来训练如此大的 CNNs，最近的数据库例如 ImageNet 也有足够用于训练的标记样本没有严重的过拟合。

这篇论文详细的贡献如下所示：我们在 ImageNet 的 2010 和 2010 比赛数据集上训练了一格最大的卷积神经网络，达到了迄今为止在此数据集上最好的效果。我们编写了基于 GPU 的高度优化的 2D 卷积实现并且将其免费提供给公众。我们的网络包括了一些新的不同寻常的特点来提升性能和减少训练时间，详细细节在第三部分。如此大的尺寸的网络即使有 120

万的训练样本也使得过拟合成为重要的问题，所以我们使用了一些高效的技术来防止过拟合，这部分内容在第四部分详细描述。最终的网络包括五个卷积层和三个全连接层，网络的深度似乎很重要，我们发现移去任意一个卷积层都会使得性能下降。

最后，网络的尺寸主要受限于当前 GPU 的内存和我们能忍受的训练时间。我们的网络在两块 3GB 的 GTX 580 GPU 训练了了 5 到 6 天，所有的实验表明我们的结果仍可以提升，只需等待更快的 GPU 和更打的数据集。

2 数据集

ImageNet 是由 22000 种类的 1500 万的高分辨率的图片组成的数据库，图片由人们在网上用亚马逊的工具进行标记得到。从 2010 年开始，ILSVRC 作为 pascal 物体识别的一部分每年一次举行。ILSVRC 使用共一千种类，每种一千张图片的来自 ImageNet 的数据集，总的来说，共有 120 万训练图片，50000 张验证图片。15 万张测试图片。

ILSVRC-2010 是 ILSVRC 测试集标签唯一公开的版本，所以我们大部分的实验基于这个版本。因为我们也用我们的模型参加了 ILSVRC-2012 的比赛，所以在第六部分也会报告我们在这个版本的数据集的结果，因为 测试集标签不公开。在 ImageNet,通常报告两种错误率，top-1 和 top-5，top-5 是在测试图片部分正确的标签不在模型认为最有可能的五个标签下的概率。

ImageNet 由很多不同分辨率的图片组成。我们的系统需要固定的输入维度。因此我们对图片进行采样将其转换为固定的 256*256 的分辨率，对于一个给定的矩形图片，我们首先对其进行缩放使其短边尺寸为 256，然后从中间裁剪出 256*256 的图片。除了减去训练集中图片每个像素点的均值外我们不对图片做其他任何预处理，所以我们的模型是在 RGB 中间值中训练的

3 结构

网络的结构总结在表二。包括八个学习层，五个卷积层和三个全连接层。除此之外我们描述了网络结构中一些创新的特点。3.1 到 3.4 根据重要性排序，最重要的排在前面。

3.1ReLU 非线性

输入标准的神经元模型输出函数输出函数应该是 f(x)=tanh(x)或者$(1+e^{-x})^{-1}$,在梯度下降法的训练过程中，这些饱和函数比不饱和函数 f(x)=max(0,x)要慢得多。根据 Nair 和 Histon 的说法，我们在神经元上使用这种非线性的 ReLU 单元，深度卷积神经网络的训练速度因此快了几倍。这个在表一中说明，表一展示了四层卷积网络在 CIFAR-10 数据库达到 25%的错误率时的迭代次数，这个表说明了在如此大的神经网络中我们不能使用传统的饱和神经元模型来进行实验。

我们不是第一个考虑在卷积神经网络中寻找传统神经元代替品的人。例如，Jattett 声明不饱和函数 f(x)=|tanhx|和他们的对比正则化和本地平均池化在 Caltech-101 数据集上工作地非常好。但是这个数据集主要考虑的问题是防止过拟合。因此他们观察到的效果和我们所报

告 ReLU 的对训练集的加速能力是不一样的。更快的学习对大型模型在大的数据集上的性能有更大的影响。

3.2 在多个 GPUs 上训练

一个 GTX 580 GPU 只有 3GB 的内存，限制了能够在其上训练的网络的最大尺寸。经计算得 120 万的训练样本足够来训练网络，但对于一个 GPU 来说太大了。所以我们把整个网络分布在两个 GPU 上。如今的 GPU 特别适合交叉并行运算，因为他们能够直接从其他 GPU 内读和写数据，不用通过主机的内存，我们使用的并行运算方案是在每个 GPU 上放一半的神经元，另一个小技巧是 GPU 之间的通信只在某些层内进行。这意思是，例如，第三层的卷积核的输入全部来自第二层，然而第四层的卷积核输入只来自第三层位于同一个 GPU 上的卷积核。选择这种连接模式是为了交叉验证的问题。但是这使我们精确的控制通信的数量直到它对计算量来说是可以接受的部分。

最终的结构是类似于柱状 CNN，除了我们的桶不是相对独立的。这个方案使我们的 top-1 和 top-5 错误率减少了 1.7%和 1.2%。另一方面，相比于在一个 GPU 上训练每个卷积层有着一半卷积核的网络，两块 GPU 比一块 GPU 训练花费更少的时间。

3.3 本地响应正则化

ReLU 有着很好的属性，并不需要输入正则化来防止饱和。如果一些样本对 ReLU 输入正激励，学习就发生在这些神经元上。然而我们发现下面的本地正则化方案确实有效。定义 $a^i_{x,y}$ 为神经元的活跃度，它有第 i 个神经元在（x.y）的位置上应用 ReLU 计算得来，本地响应归一化活跃度 $b^i_{x,y}$ 由下式给出

和是 n 个相邻的在相同的空间方向的卷积核映射的累加，N 是这一层总共的核的个数。这种本地响应归一化来源于真实神经中的侧抑制。让神经元在不同内核进行计算的活动中产生竞争。常数 k,n,a，b 都是超参数，他们的值在用验证集决定。我们取 k=2,n=5,阿尔法等于 10^{-4}，贝塔等于 0.75.在某些层的 Relu 后应用局部响应归一化。

这个方案和 Jarrett 的对比归一化方案有些类似。但是我们的更类似于亮度归一化，因为我们没有减去平均活跃度。响应归一化分别减少了我们的 top-1 和 top-5 错误率 1.4%和 1.2%。

我们也在 CIFAR-10 数据集上验证了效果，没有正则化四层的 CNN 达到了 13%的错误率，有正则化的错误率为 11%

3.4 重叠池化

CNN 中的池化层对邻近组的在同一个核映射中的神经元的输出求和。一般来说，被邻近池化单元总结的邻居节点是没有重叠的。为了更精确，一个池化层可以被认为由相隔 s 像素的池化单元组成，每个由本地 z*z 的池化单元总结。如果我们设置 s 和 z 相等，我们遵守了常用的传统的 CNN 池化。如果我们设置 s 小于 z。我们使用了重叠池化，这也是我们在我们的网络中使用的，我们设置 s=2,z=3.这个方案减少了我们的 top-1 和 top-5 错误率 0.4%和 0.3%。和不重叠池化的方案相比，s=z=2，我们观察到在训练期间重叠池化更不容易过拟合。

3.5 总体结构

现在我们准备来描述我们的 CNN 的总体结构，像表二描述的那样，网络包括八个带权重

的层；前五个是卷积层其余的是全连接层。最后一个全连接层输出 1000-way 的 softmax 代表对一千种分类的预测。网络采取最大值的多标量的客观衰减模型，…….

第二层，第四层和第五层的卷积核只和前一层位于同一个 GPU 的卷积核连接。第三层的卷积核和第二层所有的卷积核连接。响应归一层在第一层和第二层之后。最大池化层在每个响应归一层和第五个卷积层后。ReLU 被应用于每个卷积层和全连接层的输出。

第一个卷积层利用 96 个大小为 11×11×3、步长为 4 个像素（这是同一核映射中邻近神经元的感受野中心之间的距离）的核，来对大小为 224×224×3 的输入图像进行滤波。第二个卷积层需要将第一个卷积层的（响应归一化及池化的）输出作为自己的输入，且利用 256 个大小为 5×5×48 的核对其进行滤波。第三、第四和第五个卷积层彼此相连，没有任何介于中间的 pooling 层与归一化层。第三个卷积层有 384 个大小为 3×3×256 的核被连接到第二个卷积层的（归一化的、池化的）输出。第四个卷积层拥有 384 个大小为 3×3×192 的核，第五个卷积层拥有 256 个大小为 3×3×192 的核。全连接层都各有 4096 个神经元。

4 减少过拟合

我们的神经元网络结构有 6000 万的参数，尽管 ILSVRC 的一千种类使得每个训练样本在图片到标记的映射增加了 10 比特的约束，学习如此多的参数而不带相当大的过拟合是不够的，因此，我们描述了两种主要的防止过拟合的方法。

4.1 数据增强

减少过拟合最简单和最常用的方法是用标签保护转换法人为的扩大数据。我们用了两种直接的数据增强方法，每一种方法都用很少的计算量从原始图像中产生转换后的图片. 在我们的实现中，转换图像是由 CPU 上的 Python 代码生成的，而 GPU 是在之前那一批图像上训练的。所以这些数据增强方案实际上是免费计算的。

数据增强的第一种形式由生成图像转化和水平反射组成。为此，我们从 256×256 的图像中提取随机的 224×224 的碎片（和它们的水平反射），并在这些提取的碎片上训练我们的网络（这就是图 2 中输入图像是 224×224×3 维的原因）。这使得我们的训练集规模扩大了 2048 倍，尽管由此产生的训练样例一定高度依赖。如果没有这个方案，我们的网络会有大量的过拟合，这将迫使我们使用小得多的网络。在测试时，该网络通过提取五个 224×224 的碎片（四个边角碎片和中心碎片）和它们的水平反射（因此总共是十个碎片）做出了预测，并在这十个碎片上来平均该网络的 softmax 层做出的预测。

数据增强的第二种形式是改变训练图像中 RGB 通道的强度。具体来说，我们在整个 ImageNet 训练集的 RGB 像素值中使用 PCA。对于每个训练图像，我们成倍增加已有主成分，比例大小为对应特征值乘以一个从均值为 0，标准差为 0.1 的高斯分布中提取的随机变量。这样一来，对于每个 RGB 图像 $I_{xy}=[I^r_{xy}, I^G_{xy}, I^B_{xy}]$,我们增加下面的值：

P_i 和 分别是 RGB 像素值的三维协方差矩阵的第 i 个特征向量和特征值。阿尔法 i 是上述的随机变量。每一个阿尔法 i 对于特定的训练图片值提取一次直到其再次被训练。这个方案大概抓住了自然图像一个重要的属性，光照强度和颜色变化不影响物体的识别。这个方案使得 top-1 错误率减少了 1%

4.2 Dropout

结合许多不同模型的预测是一种非常成功的减少测试误差的方式[1,3]，但它先前训练花了好几天时间，似乎对于大型神经网络来说太过昂贵。然而，有一个非常有效的模型组合版本，它在训练中只花费两倍于单模型的时间。最近推出的叫做“dropout”的技术，它做的就是以 0.5 的概率将每个隐层神经元的输出设置为零。以这种方式“dropped out”的神经元既不利于前向传播，也不参与反向传播。所以每次提出一个输入，该神经网络就尝试一个不同的结构，但是所有这些结构之间共享权重。因为神经元不能依赖于其他特定神经元而存在，所以这种技术降低了神经元复杂的互适应关系。正因如此，要被迫学习更为鲁棒的特征，这些特征在结合其他神经元的一些不同随机子集时有用。在测试时，我们将所有神经元的输出都仅仅只乘以 0.5，对于获取指数级 dropout 网络产生的预测分布的几何平均值，这是一个合理的近似方法。

我们在图 2 中前两个全连接层使用 dropout。如果没有 dropout，我们的网络会表现出大量的过拟合。dropout 使收敛所需的迭代次数大致增加了一倍。

5.学习的细节

我们使用随机梯度下降法和一批大小为 128、动力为 0.9、权重衰减为 0.0005 的样例来训练我们的网络。我们发现，这少量的权重衰减对于模型学习是重要的。换句话说，这里的权重衰减不仅仅是一个正则化矩阵：它减少了模型的训练误差。对于权重 w 的更新规则为

其中 i 是迭代指数，v 是动力变量，ε 是学习率， 是目标关于 w、对 求值的导数在第 i 批样例 上的平均值。

我们用一个均值为 0、标准差为 0.01 的高斯分布初始化了每一层的权重。我们用常数 1 初始化了第二、第四和第五个卷积层以及全连接隐层的神经元偏差。该初始化通过提供带正输入的 ReLU 来加速学习的初级阶段。我们在其余层用常数 0 初始化神经元偏差。

我们对于所有层都使用了相等的学习率，这是在整个训练过程中手动调整的。我们遵循的启发式是，当验证误差率在当前学习率下不再提高时，就将学习率除以 10。学习率初始化为 0.01，在终止前降低三次。我们训练该网络时大致将这 120 万张图像的训练集循环了 90 次，在两个 NVIDIA GTX 580 3GB GPU 上花了五到六天。

6 结果

我们在 ILSVRC-2010 测试集上的结果总结于表 1 中。我们的网络实现了 top-1 测试集误差率 37.5%，top-5 测试集误差率 17.0%。ILSVRC-2010 大赛中取得的最好表现是 47.1%与 28.2%，它的方法是用不同特征训练六个 sparse-coding 模型，对这些模型产生的预测求平均值，自那以后公布的最好结果是 45.7%与 25.7%，它的方法是从两类密集采样的特征中计算出费舍尔向量，用费舍尔向量训练两个分类器，再对这两个分类器的预测求平均值。

我们也在 ILSVRC-2012 大赛中输入了我们的模型，并在表 2 中报告结果。由于 ILSVRC-2012 测试集标签是不公开的，我们不能对试过的所有模型都报告测试误差率。在本段的其余部分，我们将验证误差率与测试误差率互换，因为根据我们的经验，它们之间相差不超过

0.1%（见表 2）。本文所描述的 CNN 实现了 18.2%的 top-5 误差率。对五个相似 CNN 的预测求平均值得出了 16.4%的误差率。训练一个在最末 pooling 层之后还有一个额外的第六个卷积层的 CNN，用以对整个 ImageNet 2011 年秋季发布的图像（15M 张图像，22K 种类别）进行分类，然后在 ILSVRC-2012 上“微调”它，这种方法得出了 16.6%的误差率。用在整个 2011 年秋季发布的图像上预训练的两个 CNN，结合先前提到的五个 CNN，再对这七个 CNN 作出的预测求平均值，这种方法得出了 15.3% 的误差率。比赛中的第二名实现了 26.2%的误差率，用的方法是从不同类密集采样的特征中计算 FV，用 FV 训练几个分类器，再对这几个分类器的预测求平均值。

最后，我们还报告在 ImageNet 2009 年秋季版本上的误差率，该版本有 10,184 种类别与 890 万张图像。在这个数据集上，我们按照文献惯例，用一半图像来训练，用另一半图像来测试。由于没有确定的测试集，我们的划分必然不同于以前的作者使用的划分，但这并不会明显地影响到结果。我们在该数据集上的 top-1 误差率和 top-5 误差率分别为 67.4%和 40.9%，这是通过上述的网络得到的，但还有个附加条件，第六个卷积层接在最后一个 pooling 层之后。该数据集上公布的最佳结果是 78.1%和 60.9%。

6.1 定性评价

图 3 显示了通过该网络的两个数据连接层学习到的卷积核。该网络已经学习到各种各样的频率与方向选择核，以及各种颜色的斑点。注意两个 GPU 显现出的特性，3.5 节中描述了一个结果是限制连接。GPU1 上的核大多数颜色不明确，而 GPU2 上的核大多数颜色明确。这种特性在每一次运行中都会出现，且独立于所有特定的随机权重初始化（以 GPU 的重新编数为模）。

在图 4 左边面板上，通过计算该网络在八个测试图像上的 top-5 预测，我们定性地判断它学到了什么。注意到即使是偏离中心的物体，比如左上角的一小块，也可以被网络识别。大多数的 top-5 标签似乎合情合理。例如，只有其他类型的猫科动物被认为是对豹貌似合理的标签。在某些情况下（铁栅、樱桃），对于图片意图的焦点存在歧义。

探测网络的视觉知识有另一种方法，就是考虑由位于最后的 4096 维隐层上的图像引起的特征激活。如果两个图像用小欧氏分离产生了特征激活向量，我们可以说，在神经网络的更高级别上认为它们是相似的。图 4 显示了测试集中的五个图像，以及训练集中根据这一标准与其中每一个最相似的六个图像。注意，在像素级别，检索到的训练图像一般不会接近第一列中的查询图像。例如，检索到的狗和大象表现出各种各样的姿势。我们会在补充材料里给出更多测试图像的结果。通过使用两个 4096 维实值向量之间的欧氏距离来计算相似性是低效的，但它可以通过训练一个自动编码器将这些向量压缩为短的二进制代码来变得高效。这应该会产生一个比应用自动编码器到原始像素要好得多的图像检索方法，它不利用图像标签，此后还有一种用相似边缘图案来检索图像的倾向，而不论它们在语义上是否相似。

7 讨论

我们的研究结果表明，大型深度卷积神经网络在一个非常具有挑战性的数据集上使用纯粹的监督学习，能够达到破纪录的结果。值得注意的是，如果有一个卷积层被移除，我们的网络性能就会降低。例如，除去任何中间层都将导致该网络的 top-1 性能有 2%的损失。所以该层次深度对于达到我们的结果确实是重要的。为了简化实验，我们没有使用任何无监督的预训练，即使我们预计它将带来帮助，特别是我们可以获得足够的计算能力来显著地扩大

网络规模，而不带来标记数据量的相应增加。到目前为止，我们的结果有所改善，因为我们已经让网络更大，训练时间更久，但是为了匹配人类视觉系统的 infero-temporal 路径，我们仍然有更高的数量级要去达到。最终我们想要在视频序列上使用非常大型的深度卷积网络，其中的瞬时结构会提供非常有用的信息，这些信息在静态图像中丢失了或极不明显。