VGG-Net论文翻译——《Very Deep Convolutional Networks for Large-Scale Image Recognition》[[1]](#footnote-1)

Karen Simonyan & Andrew Zisserman

信息与电子工程学院 李斌 11531041

**学生的话：**

近些年来，随着卷积神经网络在计算机视觉方面的成功，深度学习如一股风潮席卷了包括检测、分割等诸多子领域。自2012年Alexnet推出后，世界各地的研究者都在思考如何能够得到更好的效果。其中，牛津大学的VGG Net就成功地利用了“更深的网络”和“更小的感受野”达到了这一目的。本文既为《计算机应用数学》课程之作业，也是学生课余自学的笔记。

Abstract

本文中我们探究了在大规模图像识别任务中，卷积神经网络的深度对于最终效果的影响。本文的主要贡献是：通过一种非常小的卷积核（3x3），有效地增加了网络的深度。同之前的方法相比，由于深度增加到了16~19层，在ImageNet竞赛中，效果得到了极大地提升。正是这些发现，帮助我们在2014年ImageNet竞赛中的定位（Localization）和分类（Classification）分别取得了第一和第二的位置。我们的模型还展示了对于其他数据集的泛化能力。我们将我们的两个最好的卷积网络模型公开，以促进计算机视觉方面近一步的研究。

1. Introduction

近年来，在大规模图像视频识别领域，卷积神经网络（ConvNet）取得了长足的进步。这一方面是由于一些大规模数据集的公开，如ImageNet。另一方面也得益于高性能的计算系统，如GPU和大规模分布式集群。特别要说明的一点是，ILSVRC竞赛在这其中扮演了一种重要的角色，它提供了一个重要的平台，从高维稀疏特征的编码，到今天的深度卷积神经网络。

随着卷积神经网络变得越来越普遍，无数的人都希望能够对原始的AlexNet进行改进，从而得到更好的效果。譬如说，在ILSVRC2013竞赛中，Zeiler和Fergus就利用了更小的感受野和更小的滑动距离（stride）。此外，还有人集中在了深度学习中的训练和测试的过程。在本文中，我们集中在了卷积神经网络另外一个重要的方面——深度。为达到这一目的，我们固定了其余的参数。并使用了非常小的卷积滑动窗（3x3）。将他们堆叠在一起有效地增加了网络的深度。

结果是，我们得到了更加犀利的卷积网络架构。它不仅仅创造了新的ILSVRC的记录，同时易于扩展到其他数据集。我们已经将我们的模型公开。

1. ConvNet Configurations

在本小节，我们首先介绍我们的卷积神经网络总的设计。然后在深度到具体的细节中去。我们设计的思考将在2.3小结中具体地讨论。

* 1. Architecture

在训练过程中，卷积网络的输入是一幅224x224固定大小的图片。唯一的预处理是，我们减去了训练集图像的均值。图像经过了一系列堆叠的卷积层。这些卷积层都应用了很小的感受野（3x3），这样小的感受野可能是已知的最小的，因为它只考虑了像素的八邻域。在某个模型变种中，我们甚至使用了1x1的卷积核，其实就是对于输入的一种线性变换。卷积的步长都设置为固定值1。而填充的设置则是基于卷积前后分辨率不变的考量。譬如说，填充为1像素。空间池化层应用最大池化（max pooling）。池化层使用2x2的滑动窗，步长为2像素。

一系列的卷积层堆叠后， 是3个全连接层（fully connected layer）。头两个全连接层拥有4096个通道。第三个全连接层为1000个通道（为了匹配ILSVRC竞赛中1000类的分类任务）。最后一层为Softmax层。

所有的激活层都是用ReLU非线性进行实现。需要说明的是，在我们的实现中，我们并没有加入局部响应归一化（Local Response Localisation, LRN）。据我们的观察， LRN并不会提升模型的效果，反而会内存开销和计算时间。

* 1. Configurations

所有卷积网络的细节设置，均以列在表一，每一列都代表模型的一种变化。在后文中，我们将使用A~E来指代这五类模型。所有的模型都遵从2.1.节所阐述的设计理念。他们的差别仅仅在与深度不同。从11层深度的网络A（8个卷积层和3个全连接层）以及19层深度的网络E（16个卷积层和3个全连接层）。卷积层的维度（个数）相当小，从开始的64，每隔一次池化就加一倍，知道达到512。

在表2中， 我们报告了所有设置的参数数量。尽管我们的模型深度很高，我们的模型参数比之前那些更加稀疏的模型的参数居然还要少！

* 1. Discussion

我们这个卷积网络模型的设计与ILSVRC2012和2013年冠军队伍的设计完全不同。与其采用相对较大的感受野（如11x11的卷积核，4的步长或者7x7的卷积核，2的步长）。我们在整个网络中都应用了非常小的卷积核（3x3的卷积核，1的步长）。容易证明，两层3x3的卷积层，与一层5x5所起到的效果是一样的。而三层3x3的卷积层，与7x7所起到的效果是一致的。那么问题来了，使用3层3x3的卷积层比一层7x7的卷积，到底好在哪儿呢？第一，我们应用了三层非线性操作，而不是一层。这使得最终的决策函数区分度更高。第二，我们有效地减少了参数的格式。假设输入输出都是通道数都是3C，那么级联3x3的参数为3\*(32C2)=27C2。而对于单一7\*7的卷积核，需要的参数为72C2=49C2。参数提高了81%还多！采用级联3x3的卷积核，可以视作强迫7\*7的卷积核分解为3个独立的卷积核。

对于1\*1的卷积核的应用可以视作，在不改变感受野大小的情况下， 增加非线性的一种有效手段。即使，在我们的例子中，1\*1等同于在同等维度下的线性变换，另一种非线性变换ReLU也同时被引入。

小尺度卷积核也曾经被应用过，但是他们的网络相对于我们的来说，还是太浅了一点。并且，他们并没有应用到ILSVRC这样的数据集中去。Goodfellow等人将深度卷积网络应用到了街景任务识别的工作中去，并证明了深度越深，表现越好。GoogLeNet，ILSVRC2014年的冠军，与我们的工作同时独立地进行，但都是基于非常深的网络以及非常小的感受野。他们的工作与我们相比较更加复杂，并且特征图分辨率的降低更加快，这一定程度上降低了计算负担。

1. Classification Framework

在之前的小节中，我们主要介绍了网络的设计，在这一小节中，我们将关注网络训练和评估的一些细节。

* 1. Training

总体的训练流程是参照AlexNet在2012年的文献（除了没有对输入图片做裁剪，这点后面会详细叙述）。明确的说，整个训练过程，是以多项逻辑斯蒂回归为目标函数，使用mini-batch with momentum梯度下降法作为优化方法进行训练的。训练中加入了正则化项，其中的weight decay（即L2范数惩罚因子为5·10-4）。以及dropout方法应用到了头两个全连接层中（dropout因子为0.5）。学习速率初始被设置为10-2，当验证集准确率不再降低后，再变成其十分之一。总的来说，学习率在整个过程中会降低3次，当循环迭代370k次后（74次迭代样本集），完全停止。我们认为，尽管深度很大，参数很多，但网络并不需要迭代很多次。这主要是因为：1. 显示的正则化项；2. 对于特定层的初始化工作。

对于参数的初始化是非常重要的，不好的初始化过程可能会使学习过程收敛到局部最优点去。要解决这个方法，我们首先训练网络A，网络A是一个浅层网络，这个网络直接使用随机初始化即可完成训练。然后，我们再训练更深层的网络，我们将深层网络的前四层卷积层和三层全连接层初始化为A的网络参数，而其它层进行随机初始化。我们没有降低预训练层数的学习率，允许他们在训练过程中改变。对于随机初始化，我们使用了正态分布进行，零均值，方差为10-2。偏置均被初始化为0。值得一提的是，在论文发表后，我们才发现可以使用Glorot和Bengio在2010年发表的随机初始化方法。

得到224x224的卷积网络输入图片，图片需要随机裁剪变成训练图片。为了进一步地扩大训练数据集，有可能还会使用镜像翻转、RGB颜色转换等方法。

附：caffe算法实现

Caffe使用Google的protobuffer标准。建立一个深度学习模型，只需要给出模型的定义（train.prototxt、deploy.prototxt）。训练的定义（solver.prototxt）即可。

此处给出，模型的定义代码。见deploy.prototxt文件。

1. 本文翻译有删改。 [↑](#footnote-ref-1)