深度卷积网络

马倩 (11521052)

2016 年 5 月 16 日

1 引言

卷积神经网络 (CNN) 已经出现了 20 多年 [3]，它是一种适合以图像作为输入的深度神经网络。它采用类似生物视觉中接受野功能的卷积核来卷积图像得到图像多层特征，并且由于共享卷积核权重，使得这种结构的参数数量不至于过大，即使比较深的网络 [3] 依然可以用前向传播算法训练并得到很好的效果。虽然 CNN 提出了非常长的时间，但一直没有受到太大的关注，一方面是由于计算能力不足，这导致训练网络时间非常长，二是由于网络有很多参数，因此需要大量训练数据，而当时还没有足够多的数据。直到 [2] 的出现，[2] 的网络结构和 [1] 很像，但使用了一些最新的技术，比如 Dropout、ReLU 激活函数和 Max-pooling 等，更加重要的是巨大的数据量 (ImageNet) 和更强的计算能力 (GPU) 的应用。这使得网络能更好更快的被训练并使网络具有很好的泛化能力，这个网络在计算机视觉挑战 ILSVRC 的图片分类中取得了巨大成功，也使得 CNN 受到巨大关注，并逐渐成为了计算机视觉很多研究方向的主流工具。本次讨论的文章 [1]Very Deep Convolutional Networks for Large-Scale Image Recognition 是对 [2] 的改进，它通常被称为 VGG，这个网络现在已经得到非常广泛的应用。

此论文主要是研究了卷积网络的深度对网络性能的影响，并提出了一个和 AlexNet[2] 相似但更深的网络 VGG，提升网络深度到 16-19 层 (包括了 pooling 层) 可以使得图片分类正确率得到很大的提升。

2 方法概述

2.1 网络结构

与 AlexNet 相比，VGG 主要有以下的特点：

- 卷积核大小固定为 3×3
- 卷积核移动间隔 (stride) 固定为 1
- 没有采用局部响应归一化 (LRN[2])

Max-pooling 采用 2×2 窗口，stride 为 2。AlexNet 中卷积核的大小不是固定的，比如第一层是 7×7，第二层用了 5×5，而 VGG 中全部使用 3×3 卷积核。两个 3×3 的卷积层可以看层是一个 5×5 的卷积层, 如图 1 所示。三个

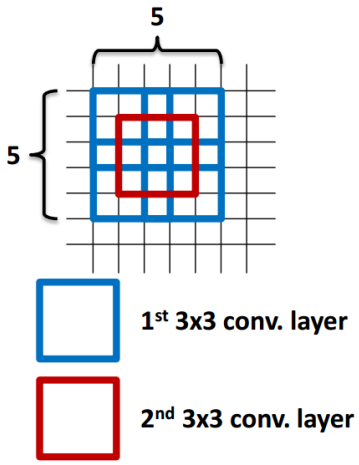

Figure 1: 用两个 3×3 卷积核与 5×5 卷积核

3×3 的卷积层可以看层一个 7×7。那么之所以用 3 个 3×3 卷积层而不使用一个 7×7 卷积层的原因有以下两点：

(1) 使用 3 个卷积层的同时也使用了 3 次非线性函数，因此可以使学习出的特征更可分。

(2) 可以降低参数个数，假设 3 层的输入和输出是 C 个通道，则一共有 $3(3^2C^2) = 27C^2$ 个参数，相应使用 7×7 的一个卷积层需要 $7^2C^2 = 49C^2$ 个参数，相比多了 81% 的参数。

本文提出了 VGG 的多种配置，每种配置的区别只在与深度不同，具体每一层结构都是统一的。作者将他们命名为 (A-E)。比如 A 有 11 个参数层 (不包括 pooling 层，8 个卷积层，3 个全连接层)，E 有 19 个参数层 (16 个卷积层和 3 个全连接层)。VGG 的输入层固定为 224×224 的 RGB 图像，这与 AlexNet

相同。网络的每一层的通道数比较小，从第一个卷积层 64 通道开始，每经过 pooling 后通道乘以 2，直到等于 512 时停止增长。三个全连接层神经元个数分别为 4096、4096 和 1000，最后是一个 soft-max 层。所有配置总结在表 1。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 1: VGG 不同配置的网络结构

2.2 训练

VGG 使用和 AlexNet 相似的训练步骤，只有少许改动。目标函数是多维逻辑回归对象，使用代冲量的 mini-batch 梯度下降，batch 为 256，冲量为 0.9。在

前两个全连接层使用 0.5 的 dropout 正则化。一共 37 万次迭代，对比 AlexNet 的训练，VGG 训练迭代次数更少，原因在于更好的正则化和预训练。训练中比较有趣的是网络权重初始化。深度学习的兴起就是由于无监督的权重初始化技术的开发，不过现在使用了很多正则化网络方法，所以不用预训练也可以得到很好的结果。但是本文有权重初始化步骤，具体是用随机初始化参数的方法训练比较浅的网络 A，然后用得到的权重初始化更深的网络然后再监督训练。

训练网络很重要的一点是训练数据，[2] 提出了几个增加数据的方法，比如水平翻转和 RGB 的随机变化等。本文提出了由于训练数据的尺度并不固定，因此不同图片的采样个数也应该不同，即多尺度训练。

2.3 测试

[2] 采用的是测试时将图片切割多个 244×244 输入网络并取平均，本文的方法是将全连接层转化成卷积层 [4]，如果都为卷积层的化输入就可以不固定 (因为卷积核可以作用在任何位置)，最后使用 sum-pooling 得到每一类的得分结果。

3 实验结果

在 ImageNet 大挑战 2014 中，VGG 在目标定位中得到第一名，在图片分类中得到第二名。

作者使用 Caffe 来构建和训练网络，并公布了网络结构和训练后的参数，其中 19 层网络参数大小为 578MB，以下是模型中一些网络层的描述：左边是卷积层，中间是 pooling 层，右边是全连接层。可以看到参数是完全按照论文中讨论设置的。

```
layers {
  bottom: "data"
  top: "conv1_1"
  name: "conv1_1"
  type: CONVOLUTION
  convolution_param {
    num_output: 64
    pad: 1
    kernel_size: 3
  }
}
```

```
layers {
  bottom: "conv5_4"
  top: "pool5"
  name: "pool5"
  type: POOLING
  pooling_param {
    pool: MAX
    kernel_size: 2
    stride: 2
  }
}
```

```
layers {
  bottom: "fc7"
  top: "fc7"
  name: "drop7"
  type: DROPOUT
  dropout_param {
    dropout_ratio: 0.5
  }
}
```

采用作者提供的模型和训练好的数据，用 Caffe 的 Python 接口可以来做预测和可视化参数 [5]。以下是重要代码片段：

```
model_def = 'vgg.prototxt' #网络结构定义
model_weights = 'VGG_ILSVRC_19_layers.caffemodel' #训练好的网络模型
net = caffe.Net(model_def,
                model_weights,
                caffe.TEST) #得到 caffe 内部的网络结构

net.blobs['data'].data[...] = transformed_image #数据输入网络
output = net.forward() #运行网络得到结果(概率)

filters = net.params['conv1_1'][0].data #第一个卷积层的数据
feat = net.blobs['conv1_1'].data[0, :64] #第一层的特征图
feat = net.blobs['pool1'].data[0] #第一个 pooling 层
```

将文件夹下的小猫的图片作为输入，程序会预测图片的类别，并输出属于每一类的概率。可以将参数和中间结果可视化，可视化的结果如图 2,3,4 所示。

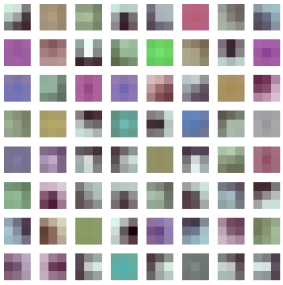

Figure 2: 第一个卷积层的参数可视化

Figure 3: 第一个卷积层的特征图

图 2 是参数的可视化，可以看到第一个卷积层有 64 个卷积核，每个核有 3×3 的参数，这些卷积核作用于输入图片后，得到的效果如右图。图 3 显示图 2 的卷积核的功能，即用每一个卷积核卷积输入图像得到的结果，比如第四行第列就基本上是得到了原图片的灰度图，第二行第五列的卷积核功能可能是将图片做了分段。还有好几个得到的是图像的边缘。图 4 是第一个 pooling 层，采用最大值采样。

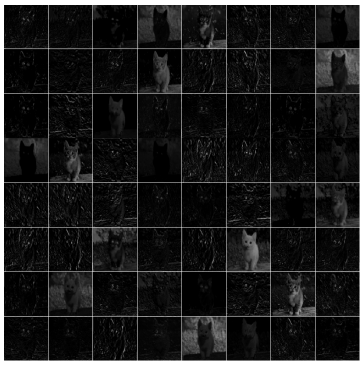

Figure 4: 第一个 pooling 层特征图

4 小结与讨论

VGG 是一个在 AlexNet 基础上改进的网络结构，主要的特点在于卷积核大小固定不变并且很小 (3×3)，使得加深网络的同时可以保证网络的参数不会大量增长，让网络具有很好的泛化能力。VGG 提出了多种深度的网络的配置。在网络的训练，VGG 也有所改进，主要在于采用浅层网络的参数作为更深网络的参数的初始值，这种预训练可以加快训练速度并得到更好的结果。VGG 也提出了新的训练数据预处理方法，同时测试时图片只用一次通过网络，提高了测试速度。

VGG 目前应用已经非常广泛。虽然目前有更新、分类结果跟好的网络，但是 VGG 依然有很多应用，比如学习出来的特征比手工设计的特征可能要好，因此可以用 VGG 的中间层的特征图做下一步分析的基础。VGG 网络也说明提升网络的深度是改进卷积网络性能的一个很重要的方面，不过也不有无限制的提升，始终有限制网络深度提升的问题 (比如由于梯度消失使得训练更困难)。

References

[1] Simonyan. K. & Zisserman A.. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR. 2015.

[2] Krizhevsky A. Sutskever & Hinton G. E. ImageNet classification with deep convolutional neural networks. NIPS. 2012.

[3] LeCun Y. Boser B. Denker J. S. Henderson D. Howard R. E. Hubbard W. & Jackel L. D. Backpropagation applied to handwritten zip code recognition. Neural Computation. 1989.

[4] Sermanet P. Eigen D. Zhang X. Mathieu M. Fergus R. & LeCun Y. OverFeat: Integrated Recognition. Localization and Detection using Convolutional Networks. ICLR. 2014.

[5] http://nbviewer.jupyter.org/github/BVLC/caffe/blob/master/examples/00-classification.ipynb