VGG读书笔记

[1409.1556] Very Deep Convolutional Networks for Large-Scale Image Recognition (arxiv.org)

这篇论文介绍了一种深度卷积神经网络架构，称为VGG网络，可以在大规模图像分类和定位的任务上实现出色的性能。

Q1 论文试图解决什么问题?

这篇论文试图解决的问题是在大规模图像识别任务上提高准确率。在以往的研究中，其它人都是采用更小的感受野或者第一层卷积层的更小步幅、抑或是采用多尺度图像的训练，而在该篇文章中从网络的深度出发，来设计更深层次的卷积神经网络来提高识别准确率。

Q2 这是否是一个新的问题?

深层卷积神经网络用于图像分类和物体检测的问题已经存在了，比如2012年的深度ConvNets（Krizhevsky et al., 2012）（获得ILSVRC-2012冠军）。但是，在2014年之前，设计更深层次的卷积神经网络的尝试还比较有限。这篇论文提出的VGG网络是第一个达到16或更深层次的卷积神经网络，而且只使用了3×3卷积核，可以说这不是一个新问题，但VGG的深度思想和采用小卷积核更好的解决了图像分类和物体检测的的问题（而且在泛化能力也为语义分割这些任务提供了思想）。

Q3 这篇文章要验证一个什么科学假设?

这篇文章的主要目的是验证一个假设：设计更深层次、小卷积核(3×3)的卷积神经网络可以提高图像识别的性能。作者在在文章的主体部分（分类部分）实验中通过使用LRN、使用1×1卷积核、逐步增加卷积层数、对图像的多尺度裁剪抖动选择等操作，展示了VGG网络在大规模图像识别任务上的出色表现，从而验证了设计更深层次、小卷积核(3×3)的卷积神经网络确实可以提高图像识别的性能。

Q4 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员?

（论文中Introduction第一、二段回答了这个问题）

在VGG网络之前和同期，也有很多关于深度卷积神经网络的相关研究，如AlexNet、ZFNet、OverFeat等。

AlexNet是VGG网络之前提出的一个经典的深度卷积神经网络，由Alex Krizhevsky、Ilya Sutskever和Geoffrey Hinton提出，获得ILSVRC-2012冠军，这个可以归类成深度卷积方面有了很大的进步。

ZFNet是由Matthew Zeiler和Rob Fergus在AlexNet的基础上提出的改进版，其采用了更小的卷积核和更深的网络结构（论文中说的是利用更小的感受野大小和第一卷积层的更小步幅。），ILSVRC-2013的最佳提交，这个可以归类成在卷积核上做了改进。

OverFeat是由Pierre Sermanet和Yann LeCun提出的一个深度卷积神经网络，其采用了多尺度卷积和跨尺度连接的方式来提取图像中的目标特征（论文中说的是在整个图像和多个尺度上密集训练和测试网络），这个可以归类成在训练核测试的方法上做了改进。

这些网络都属于深度卷积神经网络领域的经典研究，在这些研究中，个人感觉，Geoffrey Hinton、Yann LeCun等都是深度学习领域的三巨头之二，而Pierre Sermanet在ZFNet和OverFeat都有贡献，AlexNet的作者Krizhevsky（在论文中多处出现这个名字），都值得我们关注。

Q5 论文中提到的解决方案之关键是什么?

论文提出的解决方案的关键：

- 增加卷积层数
- 使用小的3x3卷积核。
- 使用了抖动尺度来降低识别误差（这个没有在论文中没有说的很重要）

作者发现，增加卷积层数可以让网络获得更强的表达能力（也就说可以提取深度的特征），而使用小的卷积核不仅可以减少网络参数还增加了激活函数的从而增加了决策函数的非线性（论文中P3 Table2下方就举了用三个3×3的卷积层来代替一层7×7的例子，把7×7的卷积层通过非线性映射到成3×3的卷积层上增加网络的表达能力）。又可以从这个角度说为了达到大卷积核的效果（或者希望更好），使用小卷积核就必须增加卷积层数。

通过这种方式，论文中配置最好的VGG网络可以在大规模图像识别任务上实现出色的性能。

Q6 论文中的实验是如何设计的?

模型

在实验中，作者使用了六个版本的深度神经网络，分别是A，A-LRN，B，C，D，E一共六个模型（模型的差异在于深度、是否用LRN、是否用conv1这三个方面），见下图。

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					

数据集

实验的数据集是ILSVRC 2012-2014 challenges中的数据集（1000个类别），有1.3M的训练集、50K的验证集和100K的测试集（而论文中提及说这里把验证集当作测试集来用了）。在评估方面，作者采用了top1和top5 误差（在论文中都有具体的定义）。

训练

在训练的时候，作者采用的是小批量梯度下降的方法，batch size是256，momentum设置为0.9，L2系数设置成5×10^{-4}，还对前两个全连接层进行了dropout（比率为0.5），学习率为10^{-2}，调了三次学习率，每次缩小十倍。

在参数初始化方面，作者一开始使用的是用配置A训练好的参数来初始化后面层数更深的网络，而后面发现其实随机初始化也一样可以。

在数据增广方面，作者借鉴了前人的一系列技术进行了重新缩放、随机裁剪、随机水平翻转和随机 RGB 颜色偏移等等操作，作者详细介绍了在训练图像尺寸的时候是怎么做的（多尺度训练，文中说的scale jitterring自己还是觉得翻译成尺度抖动比较合适）。

测试

由于测试输入图像的缩放后的尺寸不一定的等于我们在训练时候的训练图像尺寸大小，所以这里作者又借鉴了Sermanet等人在2014年提出的方法来将全连接层替换成卷积层、做测试集图像翻转、最后的分数翻转的和没翻转的做平均化，也通过多裁剪图像和密集评估的方法来测试网络的最终效果。

评估

在评估方面，作者分别评估了单尺度、多尺度、是否抖动训练尺度等等因素的效果，也评估了密集卷积网络评估和多裁剪的评估，并在不同的深度和模型大小之间进行了比较（也和之前其它人的模型效果进行了比较）。

Q7 用于定量评估的数据集是什么？代码有没有开源？

用于定量评估的数据集是ImageNet数据集，它包含超过100万张图像和1000个类别。论文中开源了VGG网络的最好的两个模型的代码，也提供了训练和测试VGG网络的详细说明。

> https://github.com/pytorch/vision/blob/main/torchvision/models/vgg.py Pytorch的vgg代码
>
> https://github.com/tensorflow/models/blob/master/research/slim/nets/vgg.py Tensorflow的vgg代码

Q8 论文中的实验及结果有没有很好地支持需要验证的科学假设？

论文作者在ImageNet数据集上展示了VGG网络在比以往更深的、卷积核为3×3的卷积神经网络中获得更好的性能，从而验证了设计更深层次的卷积神经网络可以提高图像识别性能的假设。

Q9 这篇论文到底有什么贡献？

1. 证明了在视觉表示中网络深度的重要性（更小的卷积核、更深的网络优于大卷积核浅层的网络）
2. 提出了一种深度卷积神经网络VGG，可以用于大规模图像分类和定位任务（而且有很好的性能）。
3. 通过实验证明，可以通过传统的卷积神经网络结构（LeCun等人，1989；Krizhevsky等人，2012）在显著增加深度的情况下实现在ImageNet挑战数据集上的最先进性能。
4. 通过开源代码，使其他研究人员可以构建自己的深度卷积神经网络，并在ImageNet数据集上进行评估。
5. 证明了VGG模型对于各种任务和数据集具有良好的泛化能力，为后续学术人员做迁移学习、语义分割等等一系列其它的研究提供了很大的帮助。

Q10 下一步呢？有什么工作可以继续深入？

VGG网络的提出使得深度卷积神经网络在图像识别领域得到广泛应用。在此基础上，未来可以通过以下几个方面继续深入：

1. 设计更深层次的卷积神经网络，探索更高的识别准确率。
2. 研究更有效的网络结构和训练策略，以提高训练速度和识别准确率。
3. 将深度卷积神经网络应用于更广泛的深度学习问题，如物体检测、语义分割等（比如在原文中Other Recognition Tasks中所提到的图像标题生成、纹理和材料识别）。
4. 进一步研究深度卷积神经网络的可解释性，以便更好地理解网络的决策过程。
5. 深度卷积神经网络的梯度消失和梯度爆炸问题的工作可以深入研究（事实上之后的ResNet和DenseNet就是分别引入了残差块和更加紧密的连接方式来解决这一问题）
6. 迁移学习，事实上，在pytorch官方文档中对迁移学习的介绍就是直接用的VGG的ConvNet，而冻结前面的ConvNet对后面的全连接层进行训练又或者冻结后面对前面进行微调，VGG给我们提取的深度特征在迁移学习中很有作用（在原文中的 GENERALISATION OF VERY DEEP FEATURES也说了在小数据集上部署大模型的情况）。
7. 自己一个奇怪的想法（不成熟）
    - 可不可以将VGG和ResNet做好的深度网络整合到autoEncoder中去，或者只是整合到encoder，让网络可以学习到更为复杂的特征存储起来，以后进行文生图或者图文生图等操作的时候就可以通过更少的prompts生成更丰富的图像？（想象力更丰富，个人觉得VGG提取深度特征应该可以做到）