147 | 计算机视觉领域的深度学习模型（二）：VGG & GoogleNet

2018-09-10 洪亮劼

在上第一期的分享中，我们通过一篇经典论文讲了 AlexNet 这个模型。可以说，这篇文章是深度学习在计算机视觉领域发挥作用的奠基之作。

AlexNet 在 2012 年发表之后，研究界对这个模型做了很多改进工作，使得这个模型得到了不断优化，特别是在 ImageNet 上的表现获得了显著提升。今天我们就来看看针对 AlexNet 模型的两个重要改进，分别是 VGG 和 GoogleNet。

VGG 网络

我们要分享的第一篇论文题目是《用于大规模图像识别的深度卷积网络》（Very Deep Convolutional Networks for Large-Scale Image Recognition）[1]。这篇文章的作者都

来自于英国牛津大学的“视觉几何实验室”（Visual Geometry Group），简称 VGG，所以文章提出的模型也被叫作 **VGG 网络**。到目前为止，这篇论文的引用次数已经多达 1 万 4 千次。

首先，我们简单来了解一下这篇论文的作者。

第一作者叫卡伦·西蒙彦（Karen Simonyan），发表论文的时候他在牛津大学计算机系攻读博士学位。之后，西蒙彦加入了谷歌，在 DeepMind 任职，继续从事深度学习的研究。

第二作者叫安德鲁·兹泽曼（Andrew Zisserman），是牛津大学计算机系的教授，也是计算机视觉领域的学术权威。他曾经三次被授予计算机视觉最高荣誉“马尔奖”（Marr Prize）。

这篇论文的主要贡献是什么呢？一个重要贡献就是研究**如何把之前的模型（例如 AlexNet）加深层次，从而能够拥有更好的模型泛化能力，最终实现更小的分类错误率。**

为了更好地理解这篇文章的贡献，我们来回忆一下 AlexNet 的架构。AlexNet 拥有 8 层神经网络，分别是 5 层卷积层和 3 层全联通层。AlexNet 之所以能够有效地进行训练，是因为这个模型利用了“线性整流函数”（ReLu）、数据增强（Data Augmentation）以及 Dropout 等手段。这些方法让 AlexNet 能够达到 8 层。

但是，学术界一直以来都认为，从理论上看，神经网络应该是层数越多，泛化能力越好。而且在理论上，一个 8 层的神经网络完全可以加到 18 层或者 80 层。但是在现实中，梯度消失和过拟合等情况让加深神经网络变得非常困难。在这篇论文中，VGG 网络就尝试从 AlexNet 出发，看能否加入更多的神经网络层数，来达到更好的模型效果。

那 VGG 是怎么做到加深神经网络层数的呢？总体来说，VGG 对卷积层的“过滤器”（Filter）进行了更改，达到了 19 层的网络结构。从结果上看，和 AlexNet 相比，VGG 在 ImageNet 上的错误率要降低差不多一半。可以说，**这是第一个真正意义上达到了“深层”的网络结构。**

VGG 在“过滤器”上着手更改，那么具体的改变细节究竟有哪些呢？简单来说，就是在卷积层中仅仅使用“3*3”的“接受域”（Receptive Field），使得每一层都非常小。我们可以从整个形象上来理解，认为这是一组非常“瘦”的网络架构。在卷积层之后，是三层全联通层以及最后一层进行分类任务的层。一个细节是，VGG 放弃了我们之前介绍的 AlexNet

中引入的一个叫“局部响应归一化”（Local Response Normalization）的技术，原因是这个技巧并没有真正带来模型效果的提升。

VGG 架构在训练上的一个要点是先从一个小的结构开始，我们可以理解为首先训练一个 AlexNet，然后利用训练的结果来初始化更深结构的网络。作者们发现采用这种“**初始训练**”（Pre-Training）的办法要比完全从随机状态初始化模型训练得更加稳定。

GoogleNet

我们要分享的第二篇论文题目是《更深层的卷积》（Going deeper with convolutions）[2]。因为这篇论文的作者基本都来自于谷歌，所以文章提出的模型有时候又被叫作**GoogleNet**。这篇论文拥有 8 千多次的引用数。

GoogleNet 不仅和 VGG 一样在把架构做“深”上下文章，而且在模型的效率上比 AlexNet 更加优秀。作者们利用了比 AlexNet 少 12 倍的参数，在更深的架构上达到了更好的效果。

GoogleNet 创新的重点是在网络架构上。和 AlexNet 以及 VGG 都不同的是，GoogleNet 的作者们认为更加合适的网络架构不是简单地把相同的卷积层叠加起来，然后再把相同的全联通层叠加。如果我们需要更深的架构，必须从原理上对网络架构有一个不同的理解。作者们认为，网络结构必须走向“稀疏化”（Sparsity），才能够达到更深层次、更高效的目的。

那么，能否直接用稀疏结构来进行网络的架构呢？过去的经验表明，这条路并不那么直观。第一，直接利用稀疏的结构所表达的网络结构效果并不好，第二，这样做就无法利用现代的硬件，特别是 GPU 的加速功能。现代的 GPU 之所以能够高效地处理视觉以及其他一系列类似的问题，主要的原因就是**快速的紧密矩阵运算**。所以，直接使用稀疏结构有一定的挑战。

这篇论文的核心思想就是希望**用一组“局部的”（Local）紧密结构来逼近理想中的最优的稀疏化结构**，从而能够在计算上达到高效率，同时在理论思想上，能够利用稀疏化结构来达到更深的网络架构。

这种局部模块被作者们称作是**Inception 模块**。什么意思呢？传统上，卷积层都是直接叠加起来的。而这篇论文提出的 Inception 模块，其实就是让卷积层能够在水平方向上排列

起来，然后整个模块再进行垂直方向的叠加。至于水平方向排列多少个卷积层，垂直方向排列多少 Inception 模块，都是采用经验试错的方式来进行实验的。

这篇论文最终提出的 GoogleNet 有 22 层网络结构。如果把所有的平行结构都算上的话，整个网络超过了 100 层。为了能够在这么深的结构上训练模型，作者们还采用了一种方法，那就是在中间的一些层次中**插入分类器**。相比之下，我们之前遇到过的网络结构都是在最后一层才有一个分类器。分类器层的特点就是最终的标签信息会在这里被利用，也就是说，分类的准确性，或者说是图片中物体究竟是什么，都会被这个标签信息所涵盖。在中间层加入分类器，其实就是希望标签信息能够正确引导中间层的目标，并且能够让梯度依然有效经过。

在实验中，GoogleNet 模型可以说是达到了非常好的效果。在 2014 年的 ImageNet 竞赛中，GoogleNet 和 VGG 分列比赛的第一名和第二名。两个模型之间的差距仅有不到 1 个百分点。

小结

今天我为你讲了两篇基于深度学习的经典论文，讨论了两个模型 VGG 和 GoogleNet。这两个模型在 AlexNet 的基础上做了不少的革新。

一起回顾一下要点：第一，VGG 模型对卷积层的“过滤器”进行了更改，实现了 19 层的网络结构，可以说是第一个真正意义上达到了“深层”的网络结构；第二，GoogleNet 模型的创新是在网络架构上，利用稀疏化结构达到了更深的网络架构。

最后，给你留一个思考题，总结和比较 VGG 和 GoogleNet 这两个模型，我们看到了深度模型研发的一个什么趋势？

欢迎你给我留言，我们一起讨论。

参考文献

1. K. Simonyan, A. Zisserman. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. International Conference on Learning Representations, 2015.
2. C. Szegedy et al. **Going deeper with convolutions**. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 1-9, 2015.

上一篇　146 | 计算机视觉领域的深度学习模型（一）：AlexNet

下一篇　148 | 计算机视觉领域的深度学习模型（三）：ResNet

精选留言 (4)

写留言

吴文敏
2018-09-26

1

更深的模型更少的参数

展开

sky
2018-09-11

1

第一是如何让网络越来越深，参数越来越多，第二是如何优化网络结构，优化网络结构需要对深度神经网络有更基础的理论性理解

sky
2018-09-11

第一是如何让网络越来越深，参数越来越多，第二是如何优化网络结构，优化网络结构需要对深度神经网络有更基础的理论性理解

Andy
2018-09-10

请问老师计算机视觉会被这些复杂的模型统治吗?

展开