深度学习笔记 | 第6讲：CNN图像分类发家史之从LeNet5到ResNet

louwill 狗熊会 2018-09-10

新朋友点蓝色字免费订阅

大家好！又到了每周一狗熊会的深度学习时间了。在上一期的笔记分享中，小编和大家直观展示卷积神经网络在学习过程中的特征可视化，明确了神经网络是逐层学习图像特征的过程。从本节内容开始，小编将用连续三期给大家详细介绍下 CNN 在计算机视觉领域的三大应用任务：图像分类、目标检测和图像分割。CNN 在短短的十几年的快速发展历程中，为更好的解决上述三大视觉任务不断的做出了努力和创新。本期小编就和大家一起来看深度学习计算机视觉的首要任务——图像分类。

1

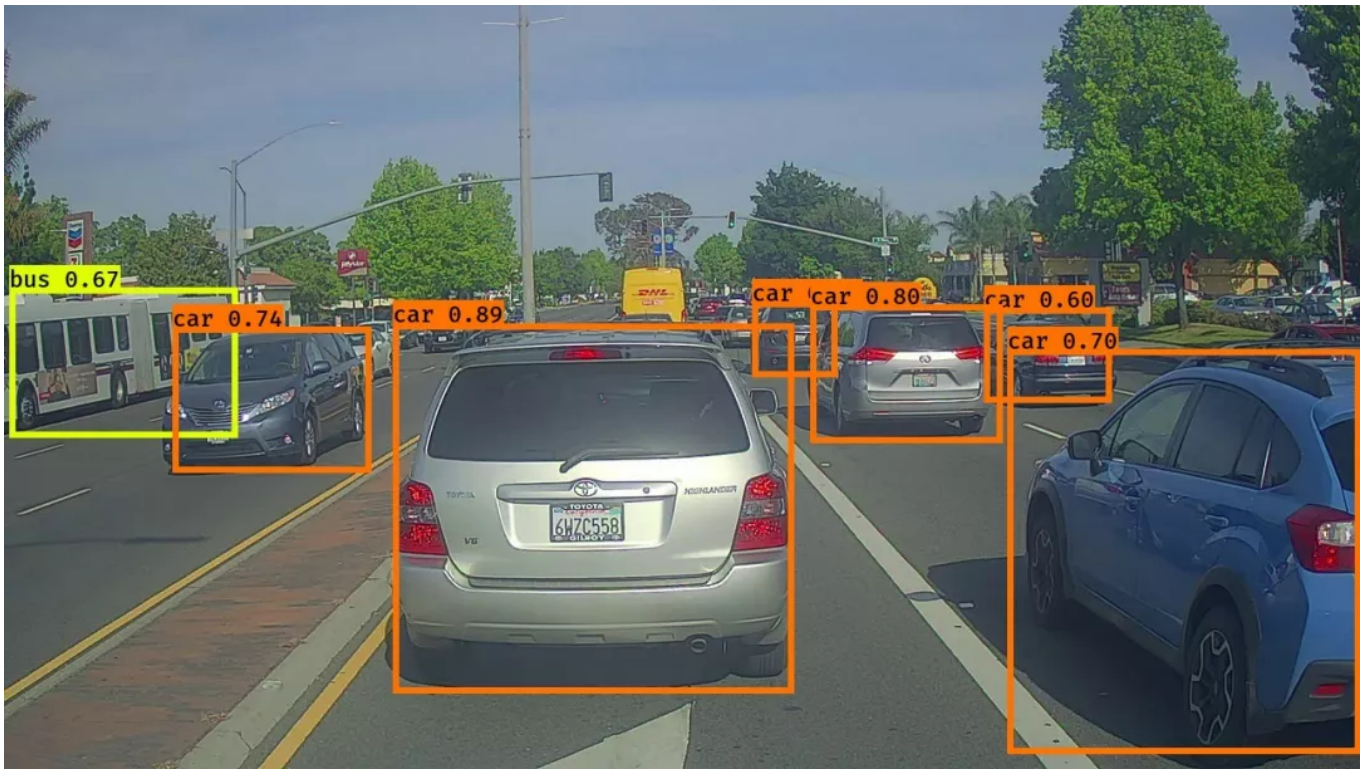
计算机视觉的三大任务

自从神经网络和深度学习方法引入到图像领域，经过近些年来的发展，从一开始的图像分类逐渐延伸到目标检测和图像分割领域，深度学习也逐渐在计算机视觉领域占据绝对的主导地位。如果要想利用深度学

习技术开启计算机视觉领域的研究，明确并深刻理解计算机视觉的三大任务非常关键。如下图所示：

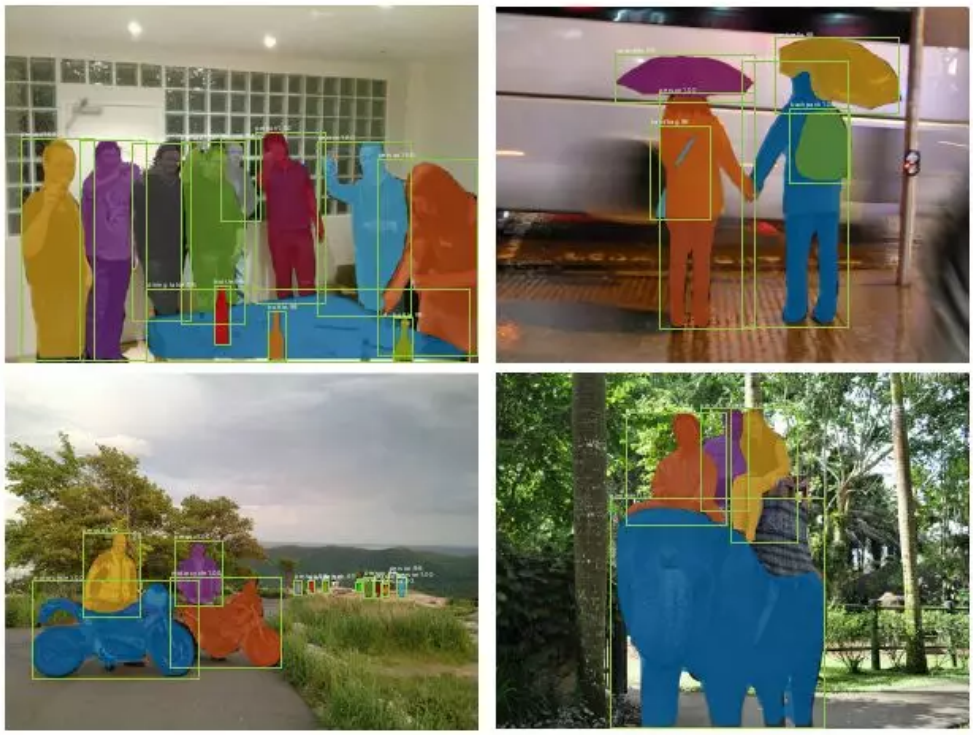
从图中我们可以简单描述计算机视觉三大任务的要义。图像分类就是要回答这张图像是一只猫的问题，跟传统的机器学习任务并无区别，只是我们的输入由数值数据变成图片数据。本节的内容就是介绍 CNN 在图像分类的发展历史上出现的一些经典的网络。

而目标检测则不仅需要回答图像中有什么，而且还得给出这些物体在图像中位置问题，以图中为例就是不仅要识别图中的阿猫阿狗，还得给出阿猫阿狗的具体定位。所以目标检测的任务简单而言就是分类+定位。在无人驾驶的应用中，我们的目标是训练出一个具有极高准确率的物体检测器，在工业产品的瑕疵检测中，我们的目标是能够快速准确的找出产品中的瑕疵区域，在医学肺部结节的检测中，我们的任务是能够根据病人肺部影像很好的检测出结节的位置。



无人驾驶的目标检测示例

而图像分割则是需要实现像素级的图像分割，以图中为例就是要把每个物体以像素级的标准分割开来，这对算法的要求则更高。这其中包括语义分割和实例分割，至于这其中的区别和细节小编将在后面两期中再做阐述。



定位+实例分割

——2——

CNN图像分类发展史

在神经网络和深度学习领域，Yann LeCun 是不得不提的一位大佬。他于 1998 年在 IEEE 上发表了一篇 42 页 的长文（*Gradient-based learning applied to document recognition，1998*），文中首次提出卷积-池化-全连接的神经网络结构，由 LeCun 提出的七层网络命名为 **LeNet-5**，因而也为他赢得了卷积神经网络之父的美誉。

LeNet-5 的网络结构如下：

LeNet-5 共有 7 层，输入层不计入层数，每层都有一定的训练参数，其中三个卷积层的训练参数较多，每层都有多个滤波器，也叫特征图，每个滤波器都对上一层的输出提取不同的像素特征。所以 LeNet-5 的简略结构如下：

输入-卷积-池化-卷积-池化-卷积（全连接）-全连接-全连接（输出）

作为标准的卷积网络结构，LeNet-5 对后世的影响深远，以至于在16 年后，谷歌提出 Inception 网络时也将其命名为 GoogLeNet，以致敬 Yann LeCun 对卷积神经网络发展的贡献。然而 LeNet-5 提出后的十几年里，由于神经网络的可解释性问题和计算资源的限制，神经网络的发展一直处于低谷。

故事的转折发展在 2012 年，也就是现代意义上的深度学习元年。2012年，深度学习三巨头之一的 Geoffrey Hinton 的学生 Alex Krizhevsky 率先提出了**AlexNet** *(ImageNet Classification with Deep Convolutional Neural Networks，2012)*，并在当年度的 ILSVRC（ImageNet大规模视觉挑战赛）以显著的优势获得当届冠军，top-5 的错误率降至了 16.4%，相比于第二名 26.2% 的错误率有了极大的提升。这一成绩引起了学界和业界的极大关注，计算机视觉也开始逐渐进入深度学习主导的时代。

AlexNet 继承了 LeCun 的 Le-Net5 思想，将卷积神经网络的发展到很宽很深的网络当中，相较于 Le-Net5 的六万个参数，AlexNet 包含了 6 亿三千万条连接，6000 万个参数和 65 万个神经元，其网络结构包括 5 层卷积，其中第一、第二和第五层卷积后面连接了最大池化层，然后是 3 个全连接层。AlexNet 的网络架构如图所示：

AlexNet 不算池化层总共有 8 层，前 5 层为卷积层，其中第一、第二和第五层卷积都包含了一个最大池化层，后三层为全连接层。所以 AlexNet 的简略结构如下：

输入>卷积>池化>卷积>池化>卷积>卷积>卷积>池化>全连接>全连接>全连接>输出

AlexNet 在 ILSVRC 2010 八张测试集图片上的识别效果：

AlexNet 就像是打开了深度学习的潘多拉魔盒，此后不断有新的网络被提出，这些都极大的繁荣了深度学习的理论和实践，致使深度学习逐渐发展兴盛起来。在 2013 年的 ILSVRC 大赛中，Zeiler 和 Fergus 在 AlexNet 的基础上对其进行了微调提出了 **ZFNet**，使得 top5 的错误率下降到 11.2%，夺得当年的第一，鉴于和 AlexNet 的结构过于相似，小编这里就不再对 ZFNet 细述。

到了 2014 年，不断的积累实践和日益强大的计算能力使得研究人员敢于将神经网络的结构推向更深层。在 2014 年提出的 **VGG-Net** (*Very Deep Convolutional Networks for Large-Scale Image Recognition，2014*)中，首次将卷积网络结构拓展至 16 和 19 层，也就是著名的 VGG16 和 VGG19。相较于此前的 LeNet-5 和 AlexNet 的 5x5 卷积和 11x11 卷积，VGGNet 结构中大量使用 3x3 的卷积核和 2x2 的池化核。VGGNet 的网络虽然开始加深但其结构并不复杂，但作者的实践却证明了卷积网络深度的重要性。深度卷积网络能够提取图像低层次、中层次和高层次的特征，因而网络结构需要的一定的深度来提取图像不同层次的特征。

在论文中，作者使用了 A-E 五个不同深度水平的卷积网络进行试验，从A到E网络深度不断加深，网络的具体信息如下：

VGG 的网络结构非常规整，2-2-3-3-3的卷积结构也非常利于编程实现。卷积层的滤波器数量的变化也存在明显的规律，由64到128再到256和512，每一次卷积都是像素成规律的减少和通道数成规律的增

加。VGG16 在当年的 ILSVRC 以 7.32% 的 top5 错误率取得了当年大赛的第二名。这么厉害的网络为什么是第二名？因为当年有比 VGG 更厉害的网络，也就是前文提到的致敬 LeNet-5 的 **GoogLeNet**。

GoogLeNet *（Going Deeper with Convolutions，2014）*在借鉴此前 1x1 卷积思想的基础上，通过滤波器组合构建 Inception 模块，使得网络可以走向更深且表达能力更强。从 2014 年获得当届 ILSVRC 冠军的 Inception v1 到现在，光 Inception 网络就已经更新到 v4 了，而后基于 Inception 模块和其他网络结构的组合而成的网络就更多了，比如说 Inception Resnet。

通常在构建卷积结构时，我们需要考虑是使用 1x1 卷积、3x3 卷积还是 5x5 卷积及其是否需要添加池化操作。而 GoogLeNet 的 Inception 模块就是帮你决定采用什么样的卷积结构。简单而言，Inception 模块就是分别采用了 1x1 卷积、3x3 卷积和 5x5 卷积构建了一个卷积组合然后输出也是一个卷积组合后的输出。如下图所示：

对于 28x28x192 的像素输入，我们分别采用 1x1 卷积、3x3 卷积和 5x5 卷积以及最大池化四个滤波器对输入进行操作，将对应的输出进行堆积，即 32+32+128+64=256，最后的输出大小为 28x28x256。所以总的而言，Inception 网络的基本思想就是不需要人为的去决定使用哪个卷积结构或者池化，而是由网络自己决定这些参数，决定有哪些滤波器组合。

构建好 Inception 模块后，将多个类似结构的Inception模块组合起来便是一个Inception 网络，如下图所示：

GoogLeNet 在当年度激烈的 ILSVRC 大赛中以 6.67% 的 top5 错误率荣膺第一名，让同样出色的 VGG Net 只能屈居第二。

此前通过 VGG Net 和 GoogLeNet 中，我们了解到卷积神经网络也可以进行到很深层，VGG16 和 VGG19 就是证明。但卷积网络变得更深呢？当然是可以的。深度神经网络能够从提取图像各个层级的特征，使得图像识别的准确率越来越高。但在2014年和15年那会儿，将卷积网络变深且取得不错的训练效果并不是一件容易的事。

深度卷积网络一开始面临的最主要的问题是梯度消失和梯度爆炸。那什么是梯度消失和梯度爆炸呢？所谓梯度消失，就是在深层神经网络的训练过程中，计算得到的梯度越来越小，使得权值得不到更新的情形，这样算法也就失效了。而梯度爆炸则是相反的情况，是指在神经网络训练过程中梯度变得越来越大，权值得到疯狂更新的情形，这样算法得不到收敛，模型也就失效了。当然，其间通过设置 relu 和归一化激活函数层等手段使得我们很好的解决这些问题。但当我们将网络层数加到更深时却发现训练的准确率在逐渐降低。这种并不是由过拟合造成的神经网络训练数据识别准确率降低的现象我们称之为退化（degradation）。

由上图我们可以看到 56 层的普通卷积网络不管是在训练集还是测试集上的训练误差都要高于 20 层的卷积网络。是个典型的退化现象。退化问题不解决，咱们的深度学习就无法 go deeper. 于是何恺明等一干大佬就提出了残差网络 **ResNet** *（Deep Residual Learning for Image Recognition，2015）*。

要理解残差网络，就必须理解残差块（residual block）这个结构，因为残差块是残差网络的基本组成部分。回忆一下我们之前学到的各种卷积网络结构（LeNet-5/AlexNet/VGG），通常结构就是卷积池化再卷积池化，中间的卷积池化操作可以很多层。类似这样的网络结构何恺明在论文中将其称为普通网络（Plain Network），何凯明认为普通网络解决不了退化问题，我们需要在网络结构上作出创新。

何恺明给出的创新在于给网络之间添加一个捷径（shortcuts）或者也叫跳跃连接（skip connection），可以让捷径之间的网络能够学习一个恒等函数，使得在加深网络的情形下训练效果至少不会变差。残差块的基本结构如下：

以上残差块是一个两层的网络结构，输入 `X` 经过两层的加权和激活得到 `F(X)` 的输出，这是典型的普通卷积网络结构。但残差块的区别在于添加了一个从输入 `X` 到两层网络输出单元的 shortcut，这使得输入节点的信息单元直接获得了与输出节点的信息单元通信的能力，这时候在进行 `relu` 激活之前的输出就不再是 `F(X)` 了，而是 `F(X)+X`。当很多个具备类似结构的这样的残差块组建到一起时，残差网络就顺利形成了。残差网络能够顺利训练很深层的卷积网络，其中能够很好的解决网络的退化问题。

或许你可能会问凭什么加了一条从输入到输出的捷径网络就能防止退化训练更深层的卷积网络？或是说残差网络为什么能有效？我们将上述残差块的两层输入输出符号改为和，相应的就有：

加入的跳跃连接后就有：

在网络中加入 L2 正则化进行权值衰减或者其他情形下，l+2 层的权值 `W` 是很容易衰减为零的，假设偏置同样为零的情形下就有 = 。深度学习的试验表明学习这个恒等式并不困难，这就意味着，在拥有跳跃连接的普通网络即使多加几层，其效果也并不逊色于加深之前的网络效果。当然，我们的目标不是保持网络不退化，而是需要提升网络表现，当隐藏层能够学到一些有用的信息时，残差网络的效果就会提升。所以，残差网络之所以有效是在于它能够很好的学习上述那个恒等式，而普通网络学习恒等式都很困难，残差网络在两者相较中自然胜出。

由很多个残差块组成的残差网络如下图右图所示：

ResNet 在 2015 年 ILSVRC 大赛上 top5 单模型的错误率达到了 3.57%，在其他数据集上也有着惊人的表现，结果当然就是收割各类奖项了。

以上便是本期的主要内容。

本篇笔记小编和大家介绍了以深度学习为主导的计算机视觉的三大任务：图像分类、目标检测和图像分割，并对每个任务进行了简单的介绍。然后重点花了大量篇幅介绍了 CNN 图像分类的发展历程，从上个世纪的 LeNet-5 到 开启深度学习元年的 AlexNet，以及此后的 VGG Net、GoogLeNet 和 ResNet等。对各个网络细节感兴趣和想要了解更多内容的朋友，不妨将这几篇论文一一找来仔细研读，想必定有一番收获。这一期到这里就结束啦，咱们下期见！

【参考资料】

https://www.deeplearning.ai/

Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[J]. 2013, 8689:818-833.

Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015:1-9.

He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015:770-778.

作者简介

鲁伟，狗熊会人才计划一期学员。目前在杭州某软件公司从事数据分析和深度学习相关的研究工作，研究方向为贝叶斯统计、计算机视觉和迁移学习。

识别二维码，查看作者更多精彩文章

识别下方二维码成为狗熊会会员！

友情提示：

个人会员不提供数据、代码，

视频only！

个人会员网址：http://teach.xiong99.com.cn

点击“**阅读原文**”，成为狗熊会会员！

阅读原文