图像简史——程序员眼中的图像发展史

作者：图鸭科技　　微信公众号:tucodec

人，是感官的动物。

我们的大脑，像一块复杂度极高的 CPU，每天在接收着各种格式的数据，进行着无休止的计算。我们以各种感官接触着这个世界，抽取着不同感官下的信息，从而认知了世界。而图像作为承载信息最为丰富的一种媒介，在人类探索智慧的历史中，一直占据着重要的位置。人用这样一双肉眼如何识别不同类别的图像（image classification and pattern recognition），如何在图像中分割出形形色色的物体（semantic segmentation and object detection），如何从模糊的图像中想象出物体的轮廓（image super-resolution），如何创作出天马行空的图画（image synthesis），都是目前机器视觉图像处理领域关注的热点问题。全世界的研究者都希望有朝一日，计算机能代替人眼来识别这一幅幅图像，发现在图像中隐藏的密码。

图像分类

图像分类是图像处理中的一个重要任务。在传统机器学习领域，去识别分类一个一个图像的标准流程是特征提取、特征筛选，最后将特征向量输入合适的分类器完成特征分类。直到 2012 年 Alex Krizhevsky 突破性的提出 AlexNet 的网络结构，**借助深度学习的算法，将图像特征的提取、筛选和分类三个模块集成于一体**，设计 5 层卷积层加 3 层全连接层的深度卷积神经网络结构，逐层对图像信息进行不同方向的挖掘提取，譬如浅层卷积通常获取的是图像边缘等通用特征，深层卷积获取的一般是特定数据集的特定分布特征。AlexNet 以 15.4%的创纪录低失误率夺得 2012 年 ILSVRC（ImageNet 大规模视觉识别挑战赛）的年度冠军，值得一提的是当年亚军得主的错误率为 26.2%。AlexNet 超越传统机器学习的完美一役被公认为是深度学习领域里程碑式的历史事件，一举吹响了深度学习在计算机领域爆炸发展的号角。

（图为李飞飞博士和她的 ImageNet 数据集）

时间转眼来到了 2014 年，GoogleNet 横空出世，此时的深度学习，已经历 ZF-net，VGG-net 的进一步精炼，在网络的深度，卷积核的尺寸，反向传播中梯度消失问题等技术细节部分已有了详细的讨论，Google 在这些技术基础上引入了 Inception 单元，大破了传统深度神经网络各计算单元之间依次排列，即卷积层->激活层->池化层->下一卷积层的范式，将 ImageNet 分类错误率提高到了 6.7%的高水平。

在网络越来越深，网络结构越来越复杂的趋势下，深度神经网络的训练越来越难，2015 年 Microsoft 大神何恺明（现就职于 Facebook AI Research）为了解决训练中准确率先饱和后降低的问题，将 residual learning 的概念引入深度学习领域，其核心思想是当神经网络在某一层达到饱和时，利用接下来的所有层去映射一个 f(x)=x 的函数，由于激活层中非线性部分的存在，这一目标几乎是不可能实现的。

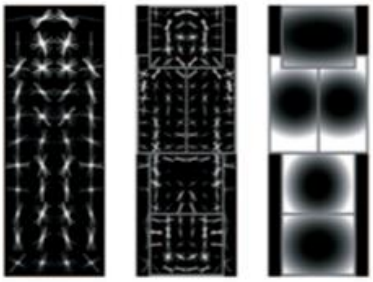
但 ResNet 中，将一部分卷积层短接，则当训练饱和时，接下来的所有层的目标变成了映射一个 f(x)=0 的函数，为了达到这一目标，只需要训练过程中，各训练变量值收敛至 0 即可。Resdiual learning 的出现，加深网络深度提高模型表现的前提下保证了网络训练的稳定性。2015 年，ResNet 也以 3.6%的超低错误率获得了 2015

年 ImageNet 挑战赛的冠军，这一技术也超越了人类的平均识别水平，意味着人工智能在人类舞台中崛起的开始。

图像中的物体检测

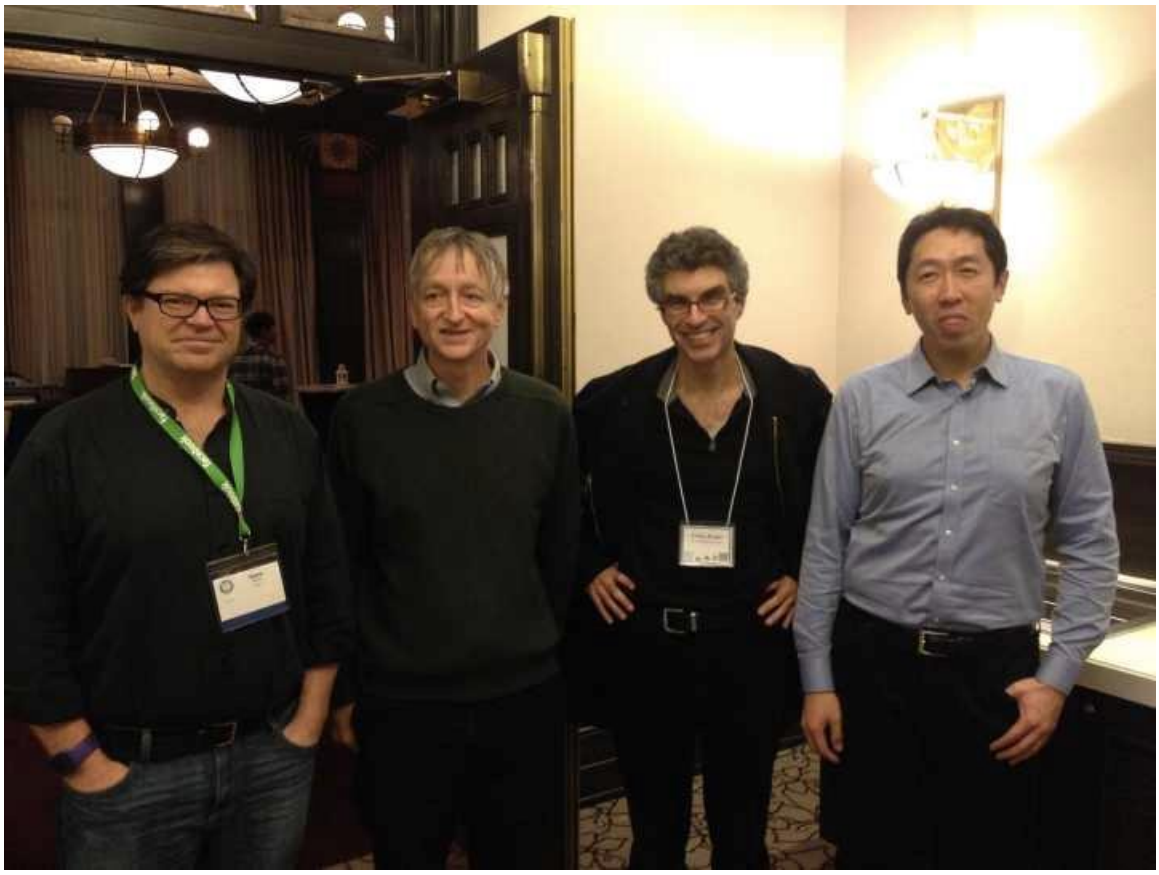
图像分类任务的实现可以让我们粗略的知道图像中包含了什么类型的物体，但并不知道物体在图像中哪一个位置，也不知道物体的具体信息，在一些具体的应用场景比如车牌识别、交通违章检测、人脸识别、运动捕捉，单纯的图像分类就不能完全满足我们的需求了。

这时候，需要引入图像领域另一个重要任务：物体的检测与识别。在传统机器领域，一个典型的案例是利用 HOG（Histogram of Gradient）特征来生成各种物体相应的“滤波器”，HOG 滤波器能完整的记录物体的边缘和轮廓信息，利用这一滤波器过滤不同图片的不同位置，当输出响应值幅度超过一定阈值，就认为滤波器和图片中的物体匹配程度较高，从而完成了物体的检测。这一项工作由 Pedro F. Felzenszalb，Ross B. Girshick， David Mcallester 还有 Deva Ramanan 以 Object Detection with Discriminatively Trained Part-Based Models 共同发表在 2010 年 9 月的 IEEE Transactions on Pattern Analysis and Machine Interlligence 期刊上。



（传统机器学习典型案例，HOG 特征滤波器完整的记录了人的整体轮廓以及一些如眼睛、躯干、四肢等特征部位的细节信息）

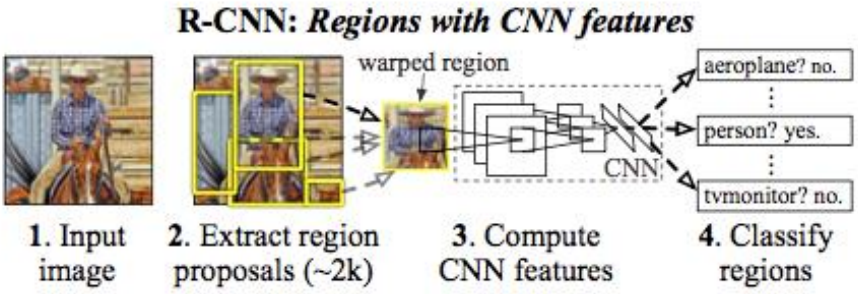
时间如白驹过隙，惊鸿一瞥， 四年过去，Ross B. Girishick 已由当年站在巨人肩膀上的 IEEE Student Member 成长为了 AI 行业内独当一面的神级人物，继承了深度学习先驱的意志，在 2014 年 CVPR 会议上发表题为 Rich Feature Hirarchies for Accurate Object Detection and Semantic Segmentation 文章。RCNN，一时无两，天下皆知。



（图为深度学习“上古四杰”，从左向右依次为[1]卷积神经网络的提出者 Yan Lecun，[2]被誉为“深度学习之父”、深度学习领路人，近期逆流而行提出深度网络 Capsule 概念的 Geoffery Hinton，[3]《Deep Learning》一书作者 Yoshua Bengio， [4]前斯坦福人工智能实验室主任 Andrew Ng（吴恩达））

RCNN 的核心思想在于将一个物体检测任务转化为分类任务，RCNN 的输入为一系列利用 selective search 算法从图像中抽取的图像块，我们称之为 region proposal。经过 warping 处理，region proposals 被标准化到相同的尺寸大小，输入到预先训练好并精细调参的卷积神经网络中，提取 CNN 特征。得到了每一个 proposal 的 CNN 特征后，针对每一个物体类别，训练一个二分类器，判断该 proposal 是否属于该物体类别。2015 年，为了缩短提取每一个 proposal 的 CNN 特征的时间，Girishick 借鉴了 Spatial Pooling Pyramid Network（SPPnet）中的 pooling 技术，首先利用一整幅图像提取 CNN 特征图谱，再在这张特征图谱上截取不同的位置的 proposal，从而得到不同尺寸的 feature proposals，最后将这些 feature proposals 通过 SPPnet 标准化到相同的尺寸，进行分类。这种改进，解决了 RCNN 中每一个 proposal 都需要进行 CNN 特征抽取的弊端，一次性在整图上完成特征提取，极大的缩短了模型的运行时间，因而被称作“Fast R-CNN”，同名文章发表于 ICCV 2015 会议。

2015 年，Girishick 大神持续发力，定义 RPN（region-proposal-network）层，取代传统的 region proposal 截取算法，将 region proposal 的截取嵌入深度神经网络中，进一步提高了 fast R-CNN 的模型效率，因而被称作“Faster R-CNN”，在 NIPS2015 上 Girishick 发表了题为“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”的关键文章，完成了 RCNN 研究领域的三级跳壮举。



(图为RCNN算法流程，最终可识别“马”以及骑在马背上的“人”)

图像生成

随着时代的发展，科学家们不仅仅是技术的研究者，更是艺术的创造者。

在人工智能领域的另一位新一代灵魂人物，Ian Goodfellow 在 2014 年提出了 Generative Adversarial Net 的概念，通过定义一个生成器（generator）和一个判别器（discriminator）来完成图像生成任务。其原理在于生成器的任务是从随机噪声中“创造”出接近目标图像的“假图像”去欺骗判别器，而判别器的任务是去甄别哪一些图像是来自于真实的数据集，哪一些图像是来自于生成器，在生成器和判别器的互相对抗中，通过合理的损失函数设计完成训练，最终模型收敛后，判别器的概率输出为常数 0.5，即一幅图像来自于生成器和真实数据集的概率相同，生成器生成的图像的概率分布无限趋近于真实数据集。

GAN 技术成为 2015，2016 年深度学习研究的热门领域，在图像恢复、降噪、超分辨重建等方向获得了极佳的表现，衍生出一系列诸如 WGAN，Info-GAN，DCGAN，Conditional-GAN 等技术，引领了一波风潮。

(图为利用 Cycle-GAN 技术，由一幅普通的照片生成莫奈、梵高等风格的油画)

图像的故事才刚刚开始。

当我们把一帧帧图像串联在一起，变成流动的光影，我们研究的问题就从空间维度上扩展到了时间维度，我们不仅需要关心物体在图像中的位置、类别、轮廓形状、语义信息，我们更要关心图像帧与帧之间的时间关系，去捕捉、识别一个物体的运动，去提取视频的摘要，去分析视频所表达的含义，去考虑除了图像之外的声音、文本标注，去处理一系列的自然语言，我们的研究一步一步，迈向了更广阔的星辰与大海。

图像和视频，都是虚拟的一串串数字，一个个字节，但却让这个世界更加真实。