

基于深度学习的图像识别进展： 百度的若干实践

都大龙 余轶南 罗 恒 等
百度公司

关键词：深度学习 图像分类 物体检测

概述：深度学习和图像识别

近年来在人工智能领域最受关注的，非深度学习莫属。自2006年吉奥夫雷·辛顿(Geoffery Hinton)等在《科学》(Science)杂志发表那篇著名的论文^[1]开始，深度学习的热潮从学术界席卷到了工业界。2012年6月，《纽约时报》披露“谷歌大脑(Google Brain)”项目，由著名的斯坦福大学机器学习教授吴恩达(Andrew Ng)和大规模计算机系统世界顶级专家杰夫·迪恩(Jeff Dean)共同主导，用1.6万个CPU核的并行计算平台训练深度神经网络(Deep Neural Networks, DNN)的机器学习模型，在语音和图像识别等领域获得巨大成功。

国内方面，2013年1月，百度成立深度学习研究院，公司CEO李彦宏担任院长。短短两年时间，深度学习技术被应用到百度的凤巢广告系统、网页搜索、

语音搜索、图像识别等领域，涵盖几十项产品。今天，用户在百度平台上的几乎每个服务请求，都被深度学习系统所处理。

人工智能的特征之一是学习的能力，即系统的性能是否会随着经验数据的积累而不断提升。所以，大数据时代的到来给人工智能的发展提供前所未有的机遇。在这个时代背景下，深度学习在包括图像识别等方面所取得的突破性进展并非偶然。

在百度的实践中，我们认识到深度学习主要在以下三个方面具有巨大优势：

1. 从统计和计算的角度看，深度学习特别适合处理大数据。在很多问题上，深度学习是目前我们能找到的最好方法。它集中体现了当前机器学习算法的三个大趋势：用较为复杂的模型降低模型偏差(model bias)，用大数据提升统计估计的准确度，用可扩展(scalable)的梯度下降算法求解大规模优化问题。

2. 深度学习不是一个黑箱系统。它像概率模型一样，提供一套丰富的、基于联接主义的建模语言(建模框架)。利用这套语言系统，我们可以表达数据内在的丰富关系和结构，比如用卷积处理图像中的二维空间结构，用递归神经网络(Recurrent Neural Network, RNN)处理自然语言等数据中的时序结构。

3. 深度学习几乎是唯一的端到端机器学习系统。它直接作用于原始数据，自动逐层进行特征学习，整个过程直接优化某个目标函数。而传统机器学习往往被分解为几个不连贯的数据预处理步骤，比如人工抽取特征，这些步骤并非一致地优化某个整体的目标函数。

让计算机识别和理解图像，是人工智能最重要的目标之一。尤其是在移动互联网时代，智能手机上的摄像头将人们日常看到的世界捕捉下来，图像和视频数据暴增，造就了图像大数据时代。

计算机视觉的主要内容就是图像识别：一方面，这个技术使得计算机像人类视觉系统一样，具有“看懂”世界的的能力，从而能自主适应环境、改造环境；另一方面，依靠识别图像内容，可以帮助我们更好地了解人，比如，通过用户产生的拍照内容了解用户的行为和喜好，或者通过识别用户手势理解用户的意图。借助图像识别让互联网服务更好地理解世界、洞察用户，也是百度深度学习研究院重点投入的技术研究方向之一。

有意思的是，深度学习研究的初衷主要就是应用于图像识别。迄今为止，尽管深度学习已经被应用到语音、图像、文字等方面，但深度学习领域发表的论文中大约70%是关于图像识别的。从2012年的ImageNet^[2]竞赛开始，深度学习在图像识别领域发挥出巨大威力，在通用图像分类、图像检测、光学字符识别(Optical Character Recognition, OCR)、人脸识别等领域，最好的系统都是基于深度学习的。前面所述深度学习的三大优势，在最近图像识别的进展中体现得淋漓尽致：模型结构越来越复杂，训练数据规模也不断增加；各种关于数据结构的先验知识被体现到新的模型结构中；端到端学习让我们越来越摒弃基于人工规则的中间步骤。

百度深度学习研究院在基于深度学习的图像识别课题上开展了大量工作，并取得丰硕成果。

在将基于深度学习的图像识别应用于图像搜索、网页搜索、百度魔图、涂书笔记、作业帮、百度街景等互联网产品以及百度眼镜(BaiduEye)、自动驾驶等创新性研究项目方面，也积累了丰富经验。下面与大家分享若干个技术实践。

基于深度学习的图像分类和物体检测算法

图像分类(image classification)和物体检测(object detection)是图像识别的两个核心问题。前者主要对图像整体的语义内容进行类别判定，后者则定位图像中特定物体出现的区域并判定其类别。与图像分类相比，物

户分析、商品推荐等互联网应用中大有用武之地。

传统图像分类算法中具有代表性的是杨(Yang)等人^[3]在2009年提出的采用稀疏编码(sparse coding)表征图像、通过大规模数据训练支持向量机(support vector machine)进行图像分类的方法。这类方法在2010年和2011年的ImageNet^[1]图像分类竞赛中取得了最好成绩，其主要缺陷在于稀疏编码和分类模型是在不同目标函数的监督下分开训练得到的，两者无法有效地联合训练。变革发生于2012年，辛顿等人^[4]采用卷积神经网络(Convolutional Neural Network, CNN)将ImageNet图像Top5分类识别错误率从之前的25%降低

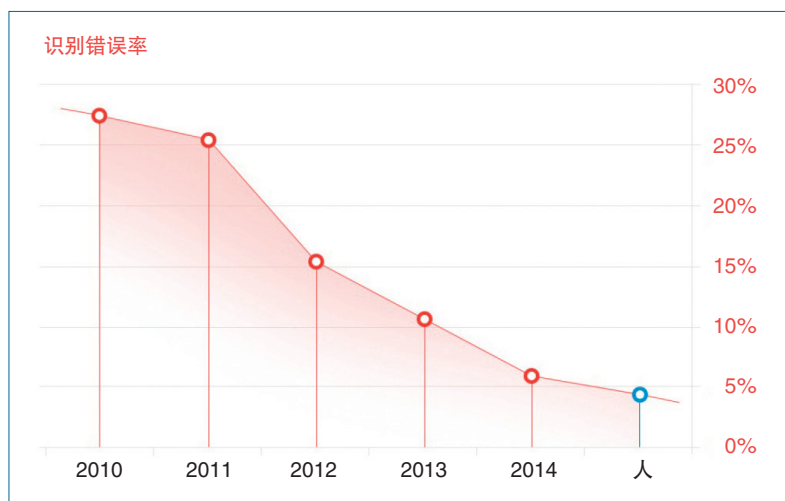


图1 2010年到2014年ImageNet竞赛的识别错误率变化以及人的识别错误率

体检测更加关注图像的局部区域和特定的物体类别集合，被视为更加复杂的图像识别问题。两项技术在信息检索、广告投放、用

到15%。随后，以卷积神经网络为代表的各种深度学习算法被广泛应用于传统的图像识别中，不断刷新纪录。截至2014年，Ima-

geNet 图像 Top5 分类的识别错误率已经降低到 6.73%^[5]。斯坦福大学的安德烈·卡帕西 (Andrej Karpathy) 等人^[6]对比了卷积神经网络和人在 ImageNet 数据库上的性能,发现目前最好的卷积神经网络模型距离人的识别率仅一步之遥(见图1)。而目前在较小的 CIFAR-10 数据库上,卷积神经网络的性能已经超过了人^[7]。

10 层卷积神经网络模型,结合图像的上下文信息,平均精度达到 40.3%。

近几年,深度学习在图像识别中的发展主要有以下几个趋势:

1. 模型层次不断加深。

2012 年,艾利克斯 (Alex) 获得当年 ImageNet 竞赛冠军时用的神经网络使用了 5 个卷积层(另外包括 3 个 pool 层和 2 个 norm

的条件)。

3. 海量的标注数据和适当的数据扰动。ImageNet 2012 分类竞赛的训练数据包含 120 万左右的标注样本,而 ImageNet 全库目前已经收集将近 2.2 万个类别共约 1420 万图像。但仅有这些数据仍不足以避免参数规模庞大的深度学习模型的过训练现象。结合图像数据的特点,包括平移、水平翻转、旋转、缩放等数据扰动方式被用于产生更多有效的训练数据,能够普遍提高识别模型的推广能力。

值得一提的是,百度利用并行分布式深度学习平台 (PARALLEL Distributed Deep LEARNING, PADDLE),收集建立起规模更大、更符合个人电脑和移动互联网特点的图像数据仓库,这些数据结合深度学习算法产出的各种图像分类和物体检测模型,已经广泛服务于许多与图像有关的百度产品线。以互联网色情图片过滤为例,我们的训练数据囊括了 1.2 亿幅色情图像,分类精度达 99.4%。

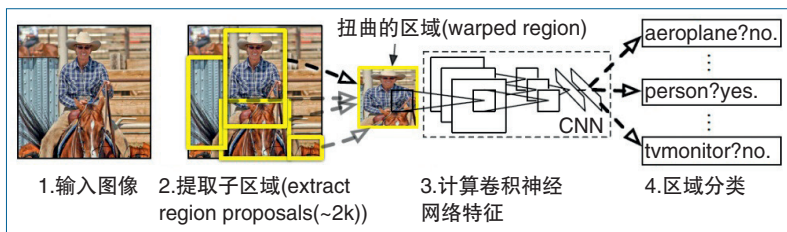


图2 区域卷积神经网络流程图^[9]

在物体检测方面,如图2所示,目前主流的算法大都采用扫描窗或是候选窗方法^[8],选取图像中许多大小位置不同的子区域进行分类(某种物体或是背景),最终得到感兴趣的物体出现的位置区域。扫描窗方法能够在相邻窗口之间共享特征,可以快速地扫描较大面积的图像;候选窗方法能够高效地在图像候选区域内进行识别,更为灵活地处理物体长宽比的变化,从而获得较高的交并比覆盖率。扫描窗和候选窗都是将物体检测问题归结为图像分类问题予以解决,因此,卷积神经网络同样可以在物体检测中大放异彩。在 ImageNet 2014^[2]的物体检测竞赛中,百度研发的物体检测算法在采用优化的候选框产生算法基础上,加上一个

层)。而到 2014 年,获得冠军的 GoogleNet^[5]使用了 59 个卷积层(另外包括 16 个 pool 层和 2 个 norm 层)。第二名的 VGG^[9]也使用 19 个卷积层,并获得较好的性能。模型深度的重要性不言而喻。

2. 模型结构日趋复杂。传统的卷积神经网络模型多使用简单的 conv-pool-norm 结构进行堆砌,GoogleNet^[5]的结果表明,并行多分辨率的 inception 结构能够融合图像在不同尺度上的有效信息,而 NIN(network-in-network)^[10]结构则通过低秩分解对较大参数规模的卷积层进行参数压缩,大大减小模型参数规模。这样做,一方面能够降低过拟合程度,提高模型的推广能力,另一方面则为大规模并行训练提供非常有利

基于端到端的序列学习:对传统光学字符识别框架的改造

光学字符识别的概念早在 20 世纪 20 年代便被提出,一直是模式识别领域研究中极具代表性的重要课题。近些年,随着移动互联网的发展,光学字符识别技术的应用场景也从传统的办公领域(例如邮政编码、书籍扫描和

文档传真) 逐渐渗入日常生活, 产生出许多以手机拍照光学字符识别作为入口的文字信息录入及查询类应用。

者包括基于类方向梯度直方图 (Histogram of Oriented Gradient, HOG) 特征的单字识别引擎^[12] 和基于 N-gram 的语言模型, 用于

杂度显著增大 (版面缺失、艺术字手写体常见、文字周边背景复杂), 而拍摄图像的条件又得不到很好的控制 (拍摄角度、距离导致的形变, 摄像头品质性能存在巨大差异, 光照和阴影变化复杂), 经典的光学字符识别技术架构难以满足实际应用的需求。究其原因, 是这一技术架构的处理流程繁琐冗长导致错误不断传递, 以及过分倚重人工规则却轻视大规模数据训练所致。

针对复杂场景的特点和经典技术框架的不足, 我们对光学字符识别的系统流程和技术框架进行了大幅改造 (见图 4)。在系统流程方面, 引入文字检测概念, 和行分割合并成新的预处理模块, 任务是检测图像中包含文字的区域并生成相应文字行; 将字分割和单字识别合并成新的整行识别模块; 基于 N-gram 的语言模型解码模块予以保留, 但将主要依赖人工规则的版面分析和后处理模块从系统中删除。6 个步骤减少到 3 个步骤, 降低了传递误差造成的不良影响。作为预处理步骤, 新引入的文字行检测模块需要在复杂的自然图像中准确地提取长短不一的文字行区域。我们摒弃传统的二值化和连通域等基于规则的方法, 采用基于学习的 Boosting、卷积神经网络结合图模型 (graphic model) 的思路解决这一问题, 并在权威的公开评测中大幅超越之前最好的文字检测方法。此外, 由于整行文字识别是一个序列学习

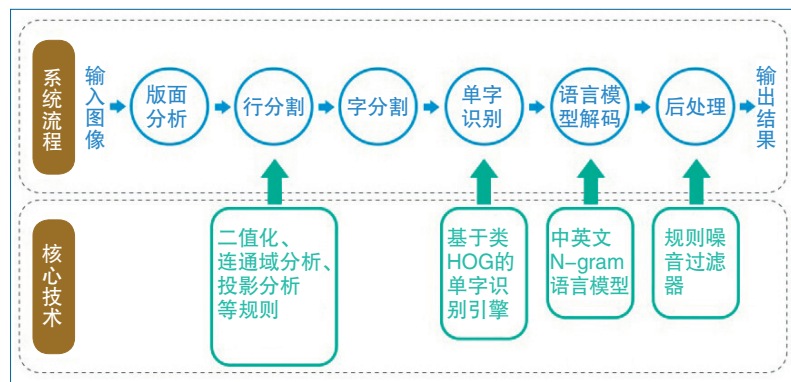


图3 经典的光学字符识别系统流程和技术框架

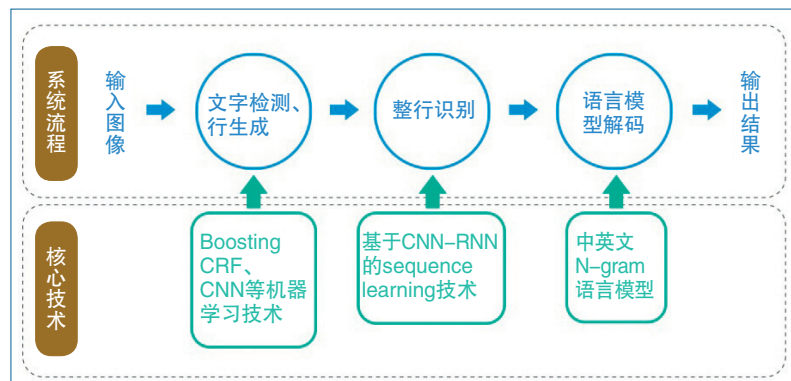


图4 基于CNN-RNN的序列光学字符识别流程

经典的字符识别系统的流程和技术框架如图 3 所示, 从输入图像到输出最终的文字识别结果, 历经版面分析、行分割、字分割、单字识别、语言模型解码和后处理。涉及的技术分为基于经验制定的规则和基于统计学习的模型^[11] 两大类。前者包括系统预处理阶段 (版面分析、行分割、字分割) 的二值化、连通域分析、投影分析等, 以及后处理阶段的规则噪声过滤器; 后

单字识别和语言模型解码阶段。在以印刷体文档扫描识别为代表的字符识别传统应用场景中, 版面结构的规则性较强, 字形、字体的一致性较高, 而文字同背景的区别性又较好。在数据简单、条件可控的情况下, 经典的字符识别技术架构通过细致的人工规则制定和适量的模型参数学习, 便可以达到比较理想的识别精度。但在广泛的自然场景中, 文字呈现出的图像信息复

(sequence learning) 问题, 我们有针对性地研发出基于双向长短期

识别技术应用于许多用户产品中。比如, 百度“涂书笔记”能够帮

户, 以及由此产生更多的反馈, 让我们能够大量收集数据, 高效地












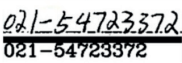
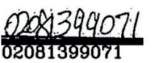
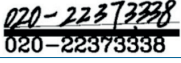
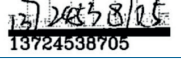
 menthol	 please	 international	
 you	 contents	 borel	 together
 huanghelou	 thinking	 founder	 serum
 021-54723372		 02081399071	
 020-22373338		 13724538705	

图5 基于CNN-RNN的序列光学字符识别结果

记忆神经网络 (Bidirectional Long Short-term Memory, BLSTM)^[13]

的递归神经网络序列模型学习算法, 结合卷积神经网络模型提取出的图像特征, 不考虑每个字符出现的具体位置, 只关注整个图像序列对应的文字内容, 使得单字分割和单字识别问题融为一体, 最终实现深度学习理论追求的理想——端到端训练。这样做能够充分利用文字序列上下文关联进行消歧, 避免传统方法中字符分割造成的不可逆转的错误。如图5所示, 这一序列学习模型极其擅长识别字分割比较困难的文字序列, 甚至包括潦草的手写电话号码。此外, 这一序列学习模型还使得训练数据的标注难度大为降低, 便于收集更大规模的训练数据。不同语言 (即便字词、句子的长短结构迥异) 光学字符识别问题也可以纳入同一个技术框架内统一解决, 大幅降低系统维护成本。

目前, 百度已经将光学字符

应用中扮演重要角色, 体现出自然图像中文字识别的特殊价值。

并行分布式深度学习平台

深度学习近年在语音识别、图像识别、机器翻译等领域取得的突破性进展, 引发工业界尤其是互联网行业的广泛兴趣, 谷歌、百度、脸谱纷纷成立专门的深度学习技术研究部门。深度学习技

术应用, 以及由此产生更多的反馈, 让我们能够大量收集数据, 高效地使用这些数据使得我们有机会训练高度复杂的模型来处理更具挑战的人工智能任务。为了实现这种产品、用户、数据的正反馈, 应用深度学习需要解决三个不同维度的挑战。首先, 底层计算维度。相对于每时每刻都在飞速增长的数据, 计算机单机的计算能力远远无法满足需要, 超大规模的并行计算势在必行。其次, 算法模型开发维度。随着越来越广泛的关注、大量研究机构的投入, 深度学习成为人工智能最活跃的领域。新的算法模型不断涌现, 新的、好的结果不断刷新, 需要迅速开发、迭代新的方法模型。第三, 一线业务部门应用维度。互联网每天都在产生新的产品、新的应用, 将深度学习高效、便捷地整合到不同的产品和应用中面临新的挑战。百度深度学习研究院开发并行分布式深度学习平台 (见图6) 的初衷就是为了

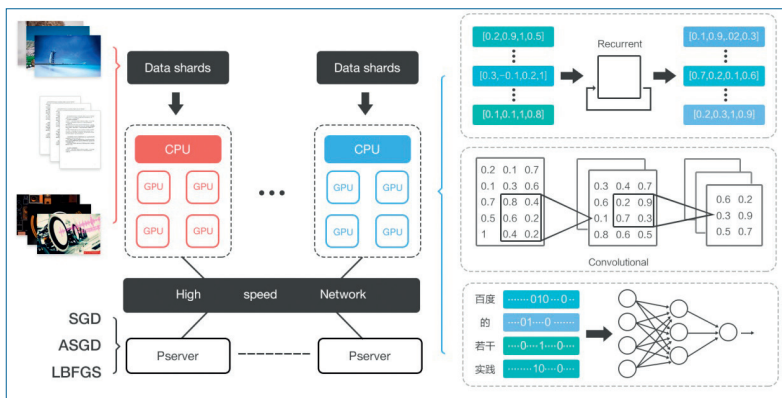


图6 并行分布式深度学习平台

术应用于互联网产品, 可大大增强用户体验, 进一步吸引更多用

户应对这些挑战。它支持超大规模并行深度学习优化 (数据的分布、

模型的分布),支持CPU/GPU混合计算、对不同类型数据(如文本、图像、声音等)的优化、丰富灵活的网络类型(如卷积神经网络、递归神经网络等)、各类主流多机优化算法(如SGD,ASGD,LBFGS等)。

在并行分布式深度学习平台上,为了应对计算上的挑战,我们在多个层次上(多线程、单机多GPU、CPU/GPU混合、CPU/GPU集群)实现的并行计算,针对不同类型的数据(文本、语音、图像、视频)采用不同策略优化模型,使我们能够最大限度地为各种计算任务优化计算资源。同时,为了应对不断涌现的新模型、新算法,我们实现了灵活的系统框架,开发者可以方便地复用以前的代码灵活地增加新算法、新模型,并且以近乎透明的方式使用各种计算资源以及并行分布式深度学习平台的优化策略。

互联网每天都在产生海量数据,既有语音、图像、视频这种稠密的自然数据,也有文本、社交关系这种稀疏的人为数据。尤其是后者,通常是高维稀疏且不断增加、变化的(譬如新的概念、词语、人物),给深度学习的应

用带来巨大挑战。一方面,为了更好地处理海量稠密数据,并行分布式深度学习平台支持使用多机多GPU卡对大规模神经网络进行快速优化,通过计算和通讯的并行以及流化大块数据的多级通讯(GPU到主机,主机直接网络传输,主机到GPU),充分降低了通讯开销,有效提升了训练速度。另一方面,根据高维稀疏数据的特点,并行分布式深度学习平台还提出并实现了许多非常具有针对性的体系结构和算法:

1. 由于海量的高维数据需要规模极大的模型与之匹配,因此模型和数据只能分布式地存储在大量的节点上。稀疏的数据与随之而来的稀疏梯度一起,使调度节点间的通信变得十分复杂。并行分布式深度学习平台针对这种复杂的场景进行了精巧的优化,可以不断地扩大模型和数据的规模。

2. 尽管有海量的数据,但是由于数据的稀疏性,过拟合仍然是需要时刻警惕的问题。并行分布式深度学习平台在实践中摸索出一套在多机并行稀疏数据情况下,控制模型规模和复杂度的算法。在提高模型泛化能力

的同时,减小模型规模,减轻给线上系统性能带来的压力。

3. 并行分布式深度学习平台对同时需要稠密矩阵运算和稀疏矩阵运算的场景进行了优化。在一个复杂网络里,针对不同层的特点,灵活地配置、使用CPU或GPU进行计算,为在多模态(文本、图像、视频)下大规模应用深度学习奠定基础。

并行分布式深度学习平台取得的成果以及未来

并行分布式深度学习平台高效的性能,尤其是对于稀疏数据的特别优化,使得深度学习应用到工业级别的广告点击预估、网页搜索排序,大大提高数据的规模、速度、泛化结果。同时,并行分布式深度学习平台灵活的系统框架大大降低了开发使用的门槛,让深度学习技术在百度知道、百度杀毒等产品上得到迅速推广。

随着深度学习在物体识别、自然语言处理领域的应用不断取得突破,未来的深度学习不仅会像人一样去听、去读、去看、去感受,更将会不断地在更大的规模上处理海量的数据;新的计算

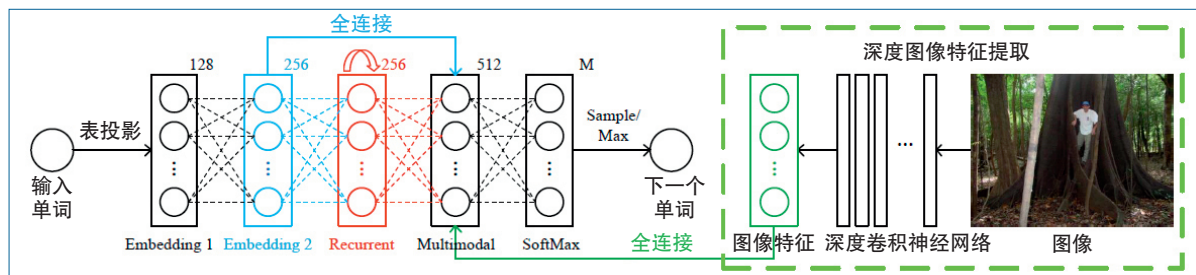


图7 图像生成语言算法中使用的递归神经网络模型结构

平台也将会不断涌现(我们相信会有更多为深度学习进行优化的硬件);新的深度学习算法赋予模型更多的意识和主动性,与增强学习的结合,让深度学习将不仅能够识别,而且能够获取高层的知识、进行推理、对外界产生反馈。并行分布式深度学习平台则将贯通数据、硬件、算法,不断推进人工智能的实践。作为并行分布式深度学习平台迈向未来人工智能的一步,我们开始用自然语言描述图片^[14],在没有任何人工干预的情况下,仅仅使用图文数据教机器描述图片(见图7、图8)。多伦多大学^[15]、斯坦福大

语言与感知的物理世界联系起来。这对于解决长期困扰人工智能的难题——“常识推理(common sense reasoning)”是重要的一步。

总结和展望

过去几年,得益于深度学习算法,图像识别技术的研究和应用飞速发展。图像标注、目标检测、物体分割、姿态估计、人脸识别、光学字符识别,几乎所有经典的图像识别技术都在深度学习算法的帮助下取得突破性进展。谷歌、脸谱、微软、亚马逊、百度都投入巨资收购和建设以图

点进行了大量实践,获得许多颇具价值的经验和知识:

丰富的图像扰动是我们将关于图像的先验知识用于深度学习输入端的有效手段 不同于许多其他数据,图像和视频在时间、空间维度上具有良好的连续性和结构性,且包含大量冗余信息。无论使用平移和翻转,还是旋转、缩放、高斯和椒盐噪声、错切等图像处理变换,都能够产生大量有效的训练数据,增强深度学习模型的鲁棒性。

结构化损失函数是我们将模型化知识用于深度学习输出端的有效方式 无论是序列解码还是图模型预测,采用人工模型对深度学习模型输出进行后处理时,具有针对性的结构化损失函数往往能够帮助深度学习过程更快地收敛到更加理想的状态。

参数的稀疏化、图像的多分辨率通道、多任务的联合学习是我们将关于问题的认知和理解注入到深度学习模型结构中的有效方式 全卷积模型中的低秩约束和全联通层中的L1正则约束已经在许多大模型训练中获得很好的效果,而多分辨率的卷积模型也在图像分类、目标检测和物体分割等问题中展现出传统单分辨率模型不具备的优势,多任务的联合学习更是使各种任务在深度学习模型中不同层面上相互帮助和约束。

从没有感知域(receptive field)的深度神经网络,到固定感知域的卷积神经网络,再到可



Tourists are sitting at a long table with a white table cloth and are eating;
(一群游客坐在一张铺着白色桌布的长桌旁用餐)



A dry landscape with green trees and bushes and light brown grass in the foreground and reddish-brown round rock domes and a blue sky in the background;
(一片背景为红褐色岩石圆顶和蓝天,前景为一些绿树、灌木和浅棕色小草的干燥的景观)

图8 根据图像生成语言描述的模型输出结果

学^[16]、谷歌^[17]、微软^[18]也纷纷发表了类似的工作成果,我们在后续工作^[19]中与这些成果进行了比较,我们的模型在相关任务(句子生成、句子检索、图片检索)中均有更加优秀的表现。这项工作把人工智能的两大分支——自然语言处理和计算机视觉无缝地连为一体,使计算机能够真正将

像识别为主要课题的人工智能技术团队,各种以图像识别技术为卖点的初创公司更是如雨后春笋般涌现;拍照搜索、视频监控、智能家居、机器人、增强现实,图像识别技术以前所未有的速度与广度向日常生活渗透,不断孕育令人印象深刻的新科技产品。在这股大潮中,百度结合自身特

变感知域的递归神经网络，深度学习模型在各种图像识别问题中不断演进。曾经爆炸式增长的参数规模逐步得到有效控制，人们将关于图像的先验知识逐渐用于深度学习，大规模并行化计算平台愈加成熟，这些使我们能够从容应对大数据条件下的图像识别问题。展望未来，基于深度学习的图像识别问题可围绕以下几个重点展开：

增强学习 与卷积神经网络和递归神经网络相比，增强学习产生的模型能够根据数据特点更灵活地产生输入序列，并通过更加模糊的监督方式进行模型训练。这样可以精简模型的复杂度，提高预测速度，同时大幅降低训练数据的标注难度，使得学习和预测过程不需要过多的人工参与，形式上更接近真正智能的学习模式。

大规模弱标注和部分标注数据的应用 随着模型规模的不断增大，获取大规模带标注的训练数据成为一道难题。和传统的强标注数据不同，在互联网场景中，以用户点击数据为代表，我们很容易获取大量包含噪音的弱标注数据，以及只有部分相关信息被标注的训练数据。采用适当的网络模型和结构化损失函数，是充分利用这些带有瑕疵但规模惊人的标注数据的关键。

低层视觉和高层视觉的广泛结合 以深度信息、立体视觉、光流场、图像分割等为代表的底层视觉方法将在深度学习框架下同语义级别的高层视觉广泛

结合，大大提高图像识别系统的通用性和鲁棒性。

适合进行深度学习模型计算的硬件高速发展 最近几个月，英特尔、英伟达和高通都宣布其硬件产业布局将为更好地支持深度学习计算而服务，开发速度更快、体积更小、更省电的计算硬件单元，聚焦于智能汽车、无人机、智能家居、可穿戴式设备等新兴电子消费品市场。

毫无疑问，基于深度学习算法的图像识别技术已经为人工智能领域中“感知”这一核心问题开启全新局面。随着理论和实践的不断深入、硬件和产品的不断推动，以图像识别为首的各种感知技术将很快填平现实物理世界和虚拟网络世界之间的沟壑，迎来人工智能全面爆发的时代。■



都大龙
百度资深研发工程师。主要研究方向为计算机视觉、深度学习、OCR等。
dudalong@baidu.com



余轶南
百度资深研发工程师。主要研究方向为计算机视觉和机器学习等。
yuyinan@baidu.com



罗恒
百度深度学习实验室高级研究员。主要研究方向为深度学习、非监督学习。
luoheng@baidu.com

其他作者：张健 黄畅
徐伟 余凯

参考文献

- [1] Geoffrey E. Hinton, and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science* 313.5786 (2006): 504~507.
- [2] Olga Russakovsky, Jia Deng, Hao Su, and et al.. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575, 2014.
- [3] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification. *Computer Vision and Pattern Recognition*, 2009.
- [4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012.
- [5] Szegedy Christian, et al. Going deeper with convolutions. arXiv preprint arXiv:1409.4842 (2014).
- [6] <http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/>.
- [7] <http://blog.kaggle.com/2015/01/02/cifar-10-competition-winners-interviews-with-dr-ben-graham-phil-culliton-zygmunt-zajac/>.
- [8] R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.
- [9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
- [10] M. Lin, Q. Chen, S. Yan,

- Network In Network. arXiv preprint arXiv:1312.4400, 2013.
- [11] Guo Hong, Ding Xiaoqing, Zhang Zhong, and et al.. Realization of A High-Performance Bilingual Chinese-English OCR System, Proceedings of the Third International Conference on Document Analysis and Recognition, 978-981.
- [12] H. Liu, X. Ding, Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes, in: Proceedings of the 8th ICDAR, Seoul, Korea, 2005:19-23.
- [13] A. Graves. Supervised Sequence Labelling with Recurrent Neural Networks. PhD thesis.
- [14] J. Mao, W. Xu, Y. Yang, and et al.. Explain Images with Multimodel Recurrent Neural Networks. Deep Learning and Representation Learning Workshop: NIPS 2014.
- [15] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014a.
- [16] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In arXiv:1406.5679, 2014.
- [17] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.
- [18] Fang Hao, Saurabh Gupta, Forrest Iandola, et al. From captions to visual concepts and back. arXiv preprint arXiv:1411.4952, 2014.
- [19] Junhua Mao, Wei Xu, Yi Yang, and et al.. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). arXiv preprint arXiv:1412.6632, 2014.