DOI:10.3969/j.issn.1009-6868.2017.04.009
网络出版地址:http://kns.cnki.net/kcms/detail/34.1228.TN.20170705.1454.002.html

当深度学习遇到大视频数据

When Deep Learning Meets Big Video Data

中图分类号:TN929.5 文献标志码:A 文章编号:1009-6868 (2017) 04-0044-003

摘要: 视频信号是大数据中的大数据,这种海量视频数据带来了存储、传输、处理、管理等方面的挑战,同时也提供了大量有价值的信息和商业机会。认为深度学习颠覆了视觉理解的进程,从图像分类到物体检测、语义分割等更细更复杂的任务,从视频里物体的检测与跟踪到物体属性和行为的分析,特别是关于人和车的理解技术。指出随着计算能力和大数据持续快速增长,加上深度学习、主动学习、迁移学习、无监督学习、强化学习等强大机器学习技术继续发展,让机器可以像人一样看到并理解世界的前景是乐观的。

关键词: 深度学习;大视频数据;人工智能

Abstract: As the biggest big data, big video data imposes significant challenges to storage, transmission, processing, and management, but it also provides an enormous amount of rich information and huge business opportunities. Deep learning has fundamentally changed the way we understand visual information, from image classification to object detection and semantic segmentation, and from object detection and tracking to the analysis of object attributes and behaviors in video. This is especially true in understanding humans and vehicles. With the rapid growth of big compute, big data, and advanced machine learning techniques such as deep learning, active learning, transfer learning, unsupervised learning, and reinforcement learning, machines will soon be able to see and understand the physical world in the same way as humans do.

Key words: deep learning; big video data; artificial intelligence

曾文军/ZENG Wenjun
罗翀/LUO Chong
(微软亚洲研究院,北京 100080)
(Microsoft Research Asia, Beijing 100080, China)

- 视频数据已渗透到人类日常生活的方方面面,视频分析的应用也因此是多方面的
- 深度学习颠覆了视觉数据理解的进程
- 基于深度学习的计算机视觉市场争夺日趋白热化。这种激烈竞争反过来将会进一步刺激加快计算机视觉和视频分析技术的发展

1 人工智能离不开视觉计算

人工智能(AI)是当今科技世界炙手可热的词语,每个人都在谈论。在过去2~3年里,AI击败人类的新闻不断,从Facebook的面部识别技术DeepFace达到和人类一样的识别精度[1],微软深度学习系统在图像识别上打败人类[2],机器在智商测试中击败人类[3],到AlphaGo击败围棋世界冠军李世石,AI的高热度在继续。

那么这些突破有哪些共性呢?第一,他们都归因于大数据的到来,例如数千个小时有标注的语音数据,数千万有标签的图像等;第二,离不开巨大的计算资源支撑,包括图形处理器(GPU)和云集群的到来和普及。在此基础上机器学习技术才取得显著进展,特别是深度学习的飞速发展。我们正处在AI的黄金时代。

AI离不开感知,而视觉是我们最主要的感知手段。研究结果表明:人的感知、学习、认知和活动有80%~85%是通过视觉介导的[4]。如果不能获取并处理视觉信息,就没法研究真实世界的人工智能,由此可见计算机视觉对人工智能发展的重要性。

视频信号在大数据中占很大比重,现在网络上70%~80%的流量是由视频信号所组成的,可以说它是大数据中的大数据。这些数据可能在几年前还不太容易得到,但是随着各种摄像头的普及,视频数据得以更详细的记录物理世界发生的一切。由此产生了海量的大视频数据,这种大数据给我们带来了存储、传输、处理、管理等方面的挑战,同时也提供了很大的机会,让机器帮助分析理解视频大数据就成了我们观察了解物理世界的一条捷径。现在我们通过分析这个大数据,提取有价值的信息,从而去支持新的产品或者服务,所以这里

收稿时间:2017-05-28
网络出版时间:2017-07-05

面蕴藏了巨大的商业机会。视频数据已渗透到人类日常生活的方方面面，视频分析的应用是多方面的，包括居家、企业、零售、公共安全、交通、制造等，市场巨大。比如：预计全球家居安防解决方案市场将以8.7%的复合年增长率增长，到2020年将达到475亿美元[5]，半自动车市场预计到2018年将达到214亿美元[6]。

2 深度学习颠覆了视觉理解的进程

视觉信号分析的发展起起伏伏，每到一定阶段都会出现“瓶颈”，其中很大的瓶颈就是没有足够量的数据，所以模型或算法的发展都受到一定的限制。直到大约2009年，ImageNet产生了。它是迄今为止最大的有标记的图像数据库，根据WordNet的层次结构组织，有超过10万个概念，每个概念有数百到数千幅附属的图像。ImageNet在过去几年大大促进了计算机视觉和图像分析的发展。

在ImageNet的基础上，近几年有一些与图像识别相关的挑战赛，如业界熟知的图像分类比赛就是利用100多万标注图像，进行1 000种分类方法准确性比较的挑战赛。还有一些如物体检测、场景检测、场景分析和语义分割等基于ImageNet的比赛。

关于ImageNet图像分类比赛，在2012年前由于分类错误率很高，从而限制了它的实用。2012年，Hinton的实验室第1次把深度神经网络应用到图像分类任务上，其性能才得以大幅提升[7]，充分展示了深度神经网络对视觉研究的极大潜能，也掀起了视觉研究的新高潮，让人们看到了计算机视觉实用化的希望。短短几年后的今天，深度神经网络技术发展迅猛，在ImageNet图像分类上的性能已超过人类，人们研究的重点也从图像分类转移到图像物体检测、语义分割等更细、更复杂的任务。

图像分析已经有了很大的进步，视频分析和理解方面进展则稍显缓慢。视频信号相比于图像信号有更大的挑战，因为它是一个更高维的信号，所含内容的多样性也很复杂，所以要去判断它、理解它都很困难，当然数据量很大也是另外一个原因。除此之外，在很多情况下视频是提供实时监测控制的，因此对处理速度等指标也有很高的要求，加之标注视频数据时每1帧都要标注，费时、费力且成本高昂，视频发展相比图像来讲还是落后一些。当然，如何获得足够多训练数据也是必须解决的难点。

前面谈到视频分析的一些应用场景，尽管不同应用场景有不同技术要求，但有些基本技术是共享的，比如物体的检测与跟踪。人是我们日常生活和工作的核心，因此也是大多数图像/视频的最主要实体。对人的分析是视频理解中的关键一步。因此很多研究团队包括微软亚洲研究院最近几年都专注于以人为中心的视频分析，例如检测与识别人、人的属性、人的行为，甚至是人的意向。由于近年来大数据、计算能力和深度学习技术的进步，对视觉数据中人的理解技术已取得了很大的进步。机器检测和识别人脸的性能已经达到了和人相仿的程度，并在身份验证、安全、智能零售、智能媒体管理等领域得到广泛应用。人体检测的性能也有了显著提高，在一些基准数据集上达到超过80%的准确度。人的各种属性，如性别、年龄、情感、手势与身体姿势，以及衣服颜色类别等也可以很好地提取，以帮助更好地了解一个人的状态。人体姿态估计技术的性能也达到了数年前不可想象的水平，并极大地方便了人的动作识别。

3 车辆和车牌检测与识别案例

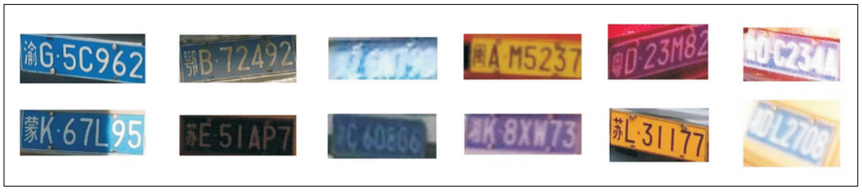
日常生活中，尤其是城市生活中，车辆是重要性仅次于人的目标类别。深度学习技术的飞速发展也大大的带动了与车辆相关的计算机视觉技术的发展，其中，最重要的就是车牌和车辆的图像检测与识别技术。

车牌是车辆的身份证，车牌自动识别技术有着非常广泛的应用，例如：车辆进入管控区域时的权限验证，进入停车场或高速公路时的收费管理，或者道路车辆违章摄像。目标通常分为合作目标和非合作目标。合作目标的图像检测和识别技术已经相当成熟，在大量应用的车牌识别系统中包含图像采集、车牌检测、字符抽取和字符识别4个步骤，其中图像采集环节是可控的，即图像采集对象是合作目标。比如车辆进入停车场时，需要车辆在低速甚至完全静止状态下完成图像采集，而且拍照时车牌的位置相对固定，这就在最大程度上保证了图像清晰，同时有效限制了车牌检测时的搜索范围。

然而，在大数据时代出现了一些车牌检测的新应用需求。这些应用面对的是不可控的图像采集对象，即非合作目标。图像质量良莠不齐，车牌类别、大小、出现位置、光照条件等都有很大的不确定性，如图1所示。例如：交通管理部门希望能够从公交车摄像头获取的图像数据中自动提取违章占用公交专用道车辆的号牌信息，然而公交车摄像头获取的是非合作目标的图像，基于合作目标的图像检测和识别技术显然无法满足应用需求。利用我们在车牌检测技术上最新的基于深度学习的研究成果，可以准确、高效的解决这一难题。

图2展示了一些我们获得的车牌检测结果示例。我们的方案可以在不同的光照条件下准确定位到大小、视角不同的各类车牌的4个角点。

视频车牌模糊也是一个非常典型的新应用需求。用户在视频网站分享视频时，如果镜头中有车辆出现，用户希望能够模糊掉车牌信息以免侵犯他人隐私，就需要利用视频车牌模糊技术，其中车牌检测和跟踪是关键所在。基于深度学习的图像车牌检测和跟踪可大大提高视频中车牌的召回率，提升车牌模糊的性能。



▲图1 不可控的图像采集对象

◀图2
典型的车牌检测结果

图像和视频中的车辆检测因其在自动驾驶、道路监测控制中的应用受到了广泛的关注。然而，深度神经网络的出现使车辆检测的精度有了质的提升。KITTI是车辆检测领域一个著名的公开数据集。在深度神经网络被大规模应用到物体检测领域之前，Regionlets[8]曾作为一个标杆方案，获得了较高的检测准确率。其在简单、中等难度和较难数据类别上的准确率分别为86.5%，76.56%和59.82%。然而近年来，随着Faster R-CNN[9]模型的提出，Regionlets在KITTI车辆检测排行榜上已退居到第50名的位置。截至目前，在中等难度的车辆检测上已有超过10种方案可以获得超过90%的准确率。另外，深度神经网络的出现也推动了车型车款识别（定位到车型车款），车辆精细化识别（定位到具体车辆）等方向的发展，使得智能城市的构想不再遥远。

总之，随着计算能力的持续快速增长，加上深度学习、主动学习、迁移学习、强化学习等强大机器学习技术继续发展，让机器可以像人一样看到并理解世界的前景是乐观的。

4 结束语

再好的研究成果，最终只有在实际应用中得到验证才能体现它的真正价值。微软亚洲研究院研发的视频分析技术正在通过微软认知服务这个平台，以视频应用程序编程接口（API）的形式提供给广大人工智能领域的开发者，帮助大家方便而高效地开发和视频相关的人工智能应用系统。这些技术也已成为微软Azure云平台的媒体分析服务的重要组成部分，可提供企业级的智能服务。类似的，其它高科技公司如Google、Amazon、Facebook等也相继推出基于深度学习的计算机视觉API，从而使得市场争夺日趋白热化。这种激烈竞争反过来将会进一步刺激加快计算机视觉和视频分析技术的发展，最终使人工智能更快、更深入地渗透到人类日常生活和工作中去。

参考文献

[1] Sophos. Facebook's DeepFace facial recognition technology has human-like accuracy[EB/OL]. (2015-02-06)[2017-06-11]. https://nakedsecurity.sophos.com/2015/02/06/facebooks-deepface-facial-recognition-technology-has-human-like-accuracy/
[2] NOVET J. Microsoft Researchers Say Their Newest Deep Learning System Beats Humans — and Google[EB/OL]. (2015-02-09)[2017-06-11]. https://venturebeat.com/2015/02/09/microsoft-researchers-say-their-newest-deep-learning-system-beats-humans-and-google/
[3] MIT Technology Review. Deep Learning Machine Beats Humans in IQ Test[EB/OL]. (2015-06-12)[2017-06-11]. https://www.technologyreview.com/s/538431/deep-learning-machine-beats-humans-in-iq-test/
[4] Brainline. Vision Problems[EB/OL]. [2017-06-11]. http://www.brainline.org/landing_pages/categories/vision.html
[5] Markets and markets. Home Security Solutions Market - Global Forecast to 2020 [EB/OL]. (2017-03)[2017-06-11]. http://www.marketsandmarkets.com/Market-Reports/home-security-solutions-market-701.html
[6] Markets and markets. Semi Autonomous Market for Passenger Car-Global Trends & Forecast to 2018[EB/OL]. (2017-05)[2017-06-11]. http://www.marketsandmarkets.com/Market-Reports/near-autonomous-passenger-car-market-1220.html
[7] 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
[8] WANG X, YANG M, ZHU S, et al. Regionlets for Generic Object Detection[C]// IEEE International Conference on Computer Vision. USA: IEEE Computer Society, 2013: 17-24. DOI: 10.1109/ICCV.2013.10
[9] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149. DOI: 10.1109/TPAMI.2016.2577031

作者简介

曾文军，微软亚洲研究院副院长，IEEE Fellow，密苏里大学、中国科技大学、西安交大、天津大学等多所学校博士生导师，2003—2016任密苏里大学计算机科学系终身教授；目前负责微软亚洲研究院视频分析和理解的研发，为微软认知服务和Azure媒体分析服务提供技术；曾对国际标准ISO MPEG、JPEG2000和Open Mobile Alliance的发展作出重大贡献，担任多个IEEE期刊和杂志的副主编、多个IEEE会议共同主席或技术程序委员会主席；已发表160篇论文，并有2部关于多媒体安全和社交多媒体的著作。

罗翀，微软亚洲研究院网络多媒体组主管研究员，中国科学技术大学兼职教授、博士生导师，IEEE高级会员；主要研究方向为多媒体云计算、多媒体通信、计算机视觉等；自2011年起担任IEEE Infocom会议的技术委员会成员；已在ACM Mobicom、IEEE Infocom等顶尖学术会议以及多份IEEE期刊上发表论文30余篇，拥有10余项国际发明专利。