人工智能技术分论坛

计算机视觉和深度学习现状和未来

邓志东教授/博士生导师

清华大学智能技术与系统国家重点实验室
清华大学计算机科学与技术系
北京信息科学与技术国家研究中心
michael@tsinghua.edu.cn

2018.12.04 · 深圳

提纲 OUTLINES

1、深度学习赋能的计算机视觉

2、探索具有认知理解能力的人工智能视觉

提纲 OUTLINES

1、深度学习赋能的计算机视觉

2、探索具有认知理解能力的人工智能视觉

1、深度学习赋能的计算机视觉

大数据驱动的深度学习已成为计算机视觉的主流方法

- 无论是单目/双目/红外摄像机，还是激光雷达/毫米波雷达成像，环境感知与自主导航，本质上可归纳为场景/目标的计算机视觉问题，主要涉及视觉感知（检测、定位、分割、跟踪、识别）与认知理解。

Hand-crafted/Fixed Feature Extractor → Trainable Classifier → SOPHIA

Trainable Hierachical Feature Extractor → Trainable Classifier → SOPHIA

Deep Learning = Learning Hierachical Feature Representation

1、深度学习赋能的计算机视觉

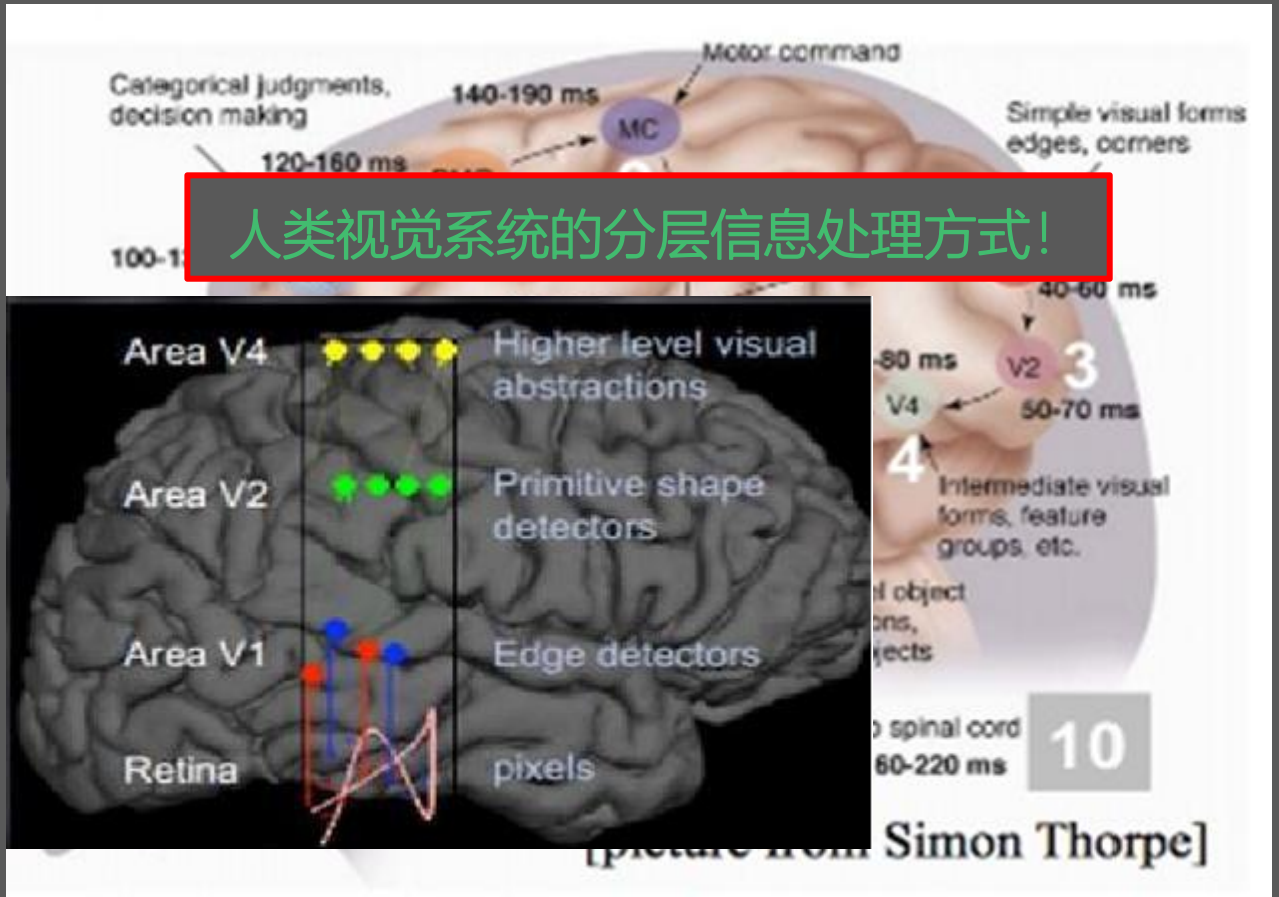
大数据驱动的深度卷积神经网络确实带来了更加接近于人类的视听觉感知能力（主要是目标检测、分割与识别能力）。

深度强化学习带来了超人类水平的棋类决策能力。

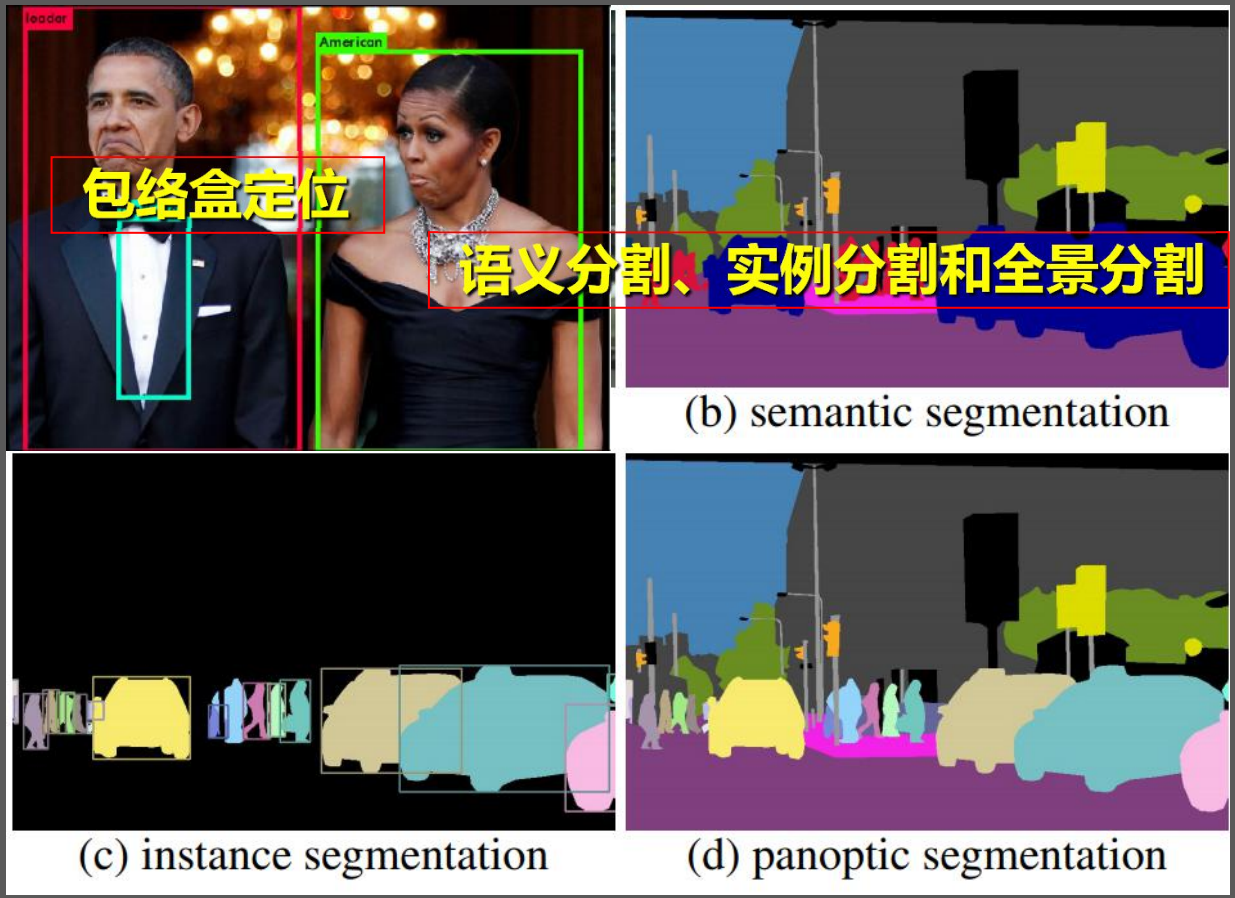
☆ 微小目标检测、目标/图像增强、目标分割/识别，动作/行为识别，超真实感合成，场景/文本分析，AutoML，信息融合，学习控制，语音交互，时序预测，决策辅助，对抗博弈，脑电融合

数据驱动方法已被视为继实验科学、理论模型、模拟仿真之后的第四科学研究范式！

1、深度学习赋能的计算机视觉



计算机视觉：检测、定位与像素级分割问题



计算机视觉：检测、定位与像素级分割算法

深度学习方法

相关研究

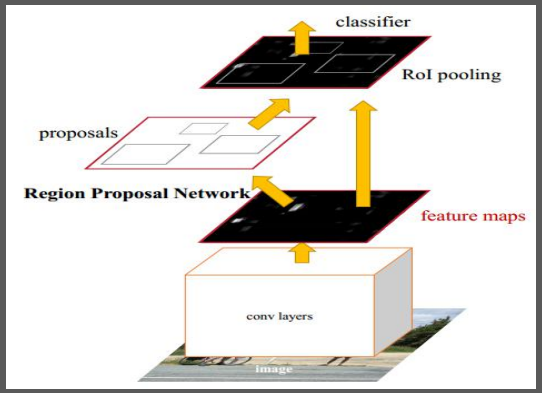
R-CNN (Girshick et al., 2014); Fast R-CNN (Girshick, 2015); Faster R-CNN (Ren, He, Girshick, & Sun, 2015); YOLO (Redmon et al., 2016); SSD (Liu et al., 2016); R-FCN (Dai, Li, He, & Sun, 2016); MS-CNN (Cai et al.,2016) ; RetinaNet (Lin, Goyal, Girshick, He, & Dollár, 2017); YOLOv3 (Redmon and Farhadi, 2017) ; Mask R-CNN (He et al., 2017); Mask^X R-CNN (Hu et al., 2017) ; PointNet (Qi et al., 2017); Complex-YOLO (Simon et al.,2018); Panoptic Segmentation (Kirillov et al.,2018)

1、深度学习赋能的计算机视觉

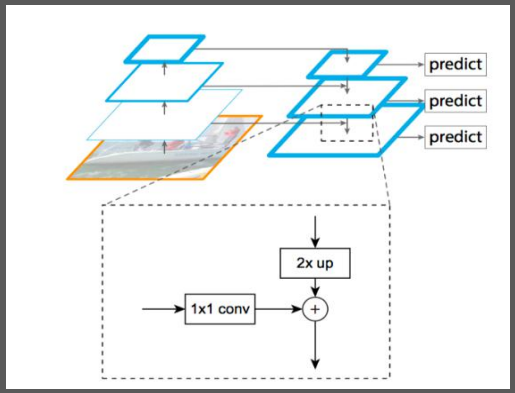
目标检测与分割模型

基于区域-卷积神经网络的检测与识别算法：

R-CNN (CVPR-2014) ➡ Fast R-CNN (ICCV-2015) ➡ Faster R-CNN (NIPS-2015) ➡ YOLO (CVPR-2016) ➡ SSD (ECCV-2016) ➡ R-FCN (NIPS-2016) ➡

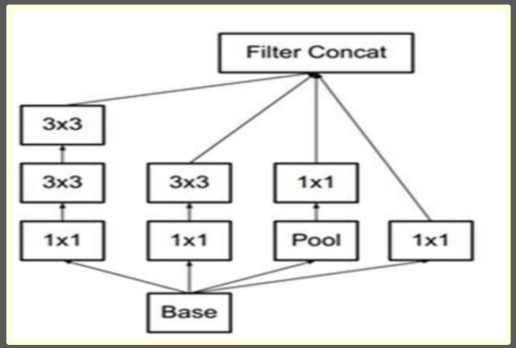


Faster R-CNN

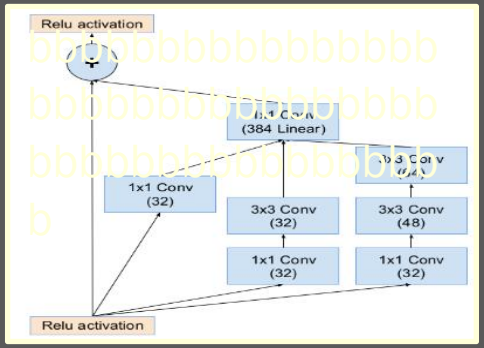


FPN

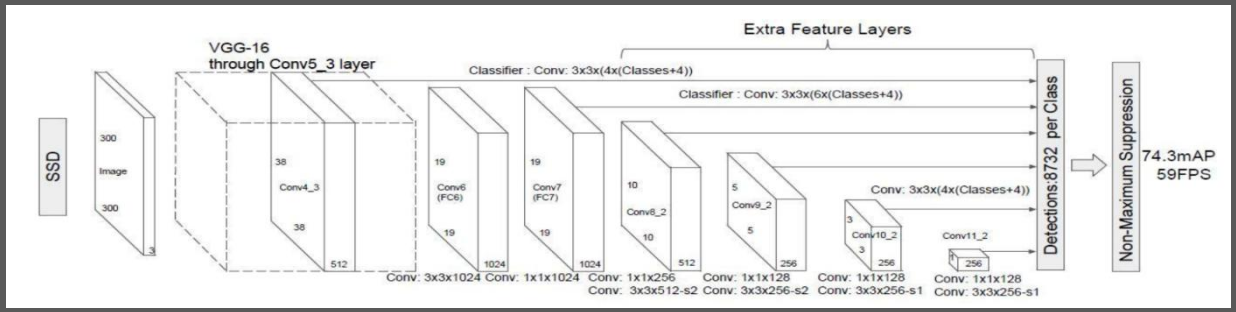
目标检测与分割模型



Inception-v2

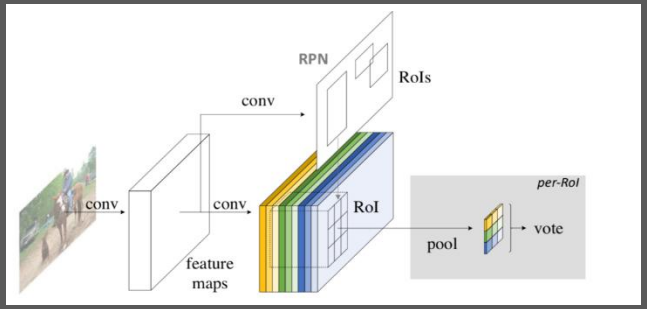


Inception-ResNet-v2



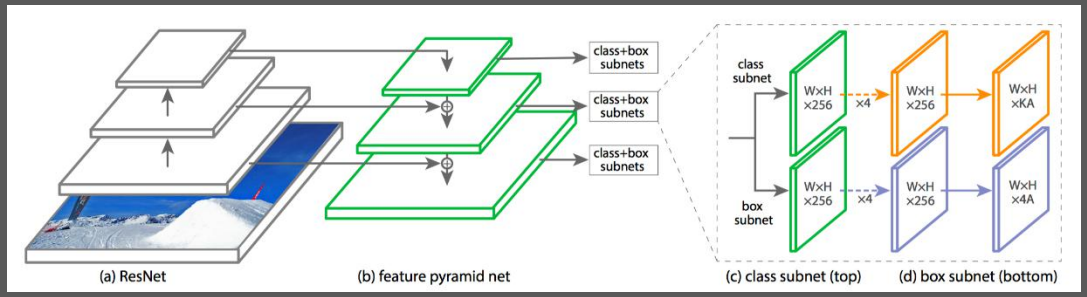
SSD

目标检测与分割模型



R-FCN

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t).$$



RetinaNet

计算机视觉：识别/分类问题

计算机视觉：识别/分类算法

深度学习方法

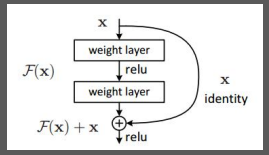
相关研究

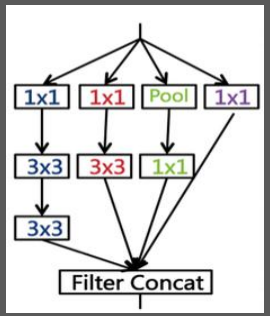
AlexNet (Krizhevsky, Sutskever , and Hinton, 2012); NIN (Lin, et al., 2014); VGG (Simonyan and Zisserman, 2014); Inception (Szegedy, *et al.*, 2015); GoogLeNet (Szegedy, et al., 2015); Batch Normalization (Ioffe and Szegedy, 2015); ResNet (He, et al., 2015); Dilated Convolutions (Yu and Koltun, 2015); DenseNet (Huang, et al., 2016); FPN (Lin et al., 2017); ResNext (Xie et al., 2017)

目标识别或分类模型

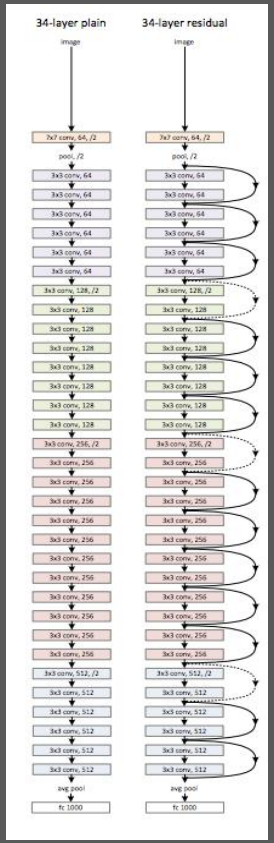
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

VGG





Inception-v2



ResNet

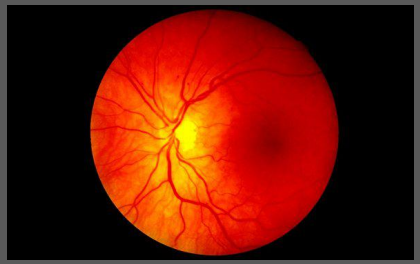
1、深度学习赋能的计算机视觉

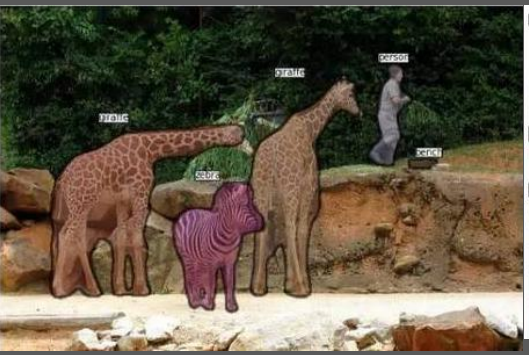
场景分割与解析

★ 利用区域-全卷积神经网络，基于图像样本中每一个像素的分类标签进行监督学习，完成像素级别的场景分类

目标检测、分割与识别

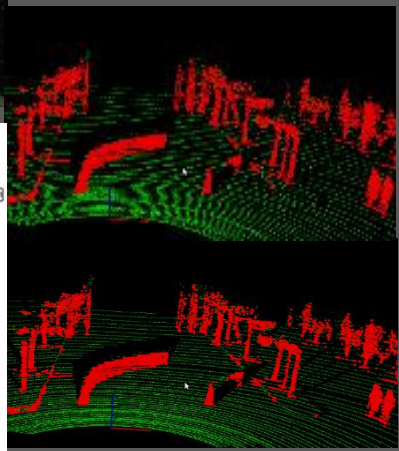
★ 基于区域-卷积神经网络的实例分割



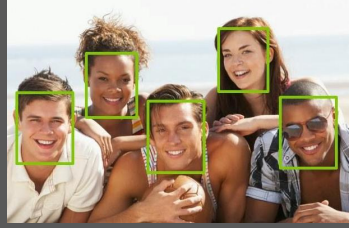


KITTI：

基于DFFA (2017)的可行驶路面与车道线检测



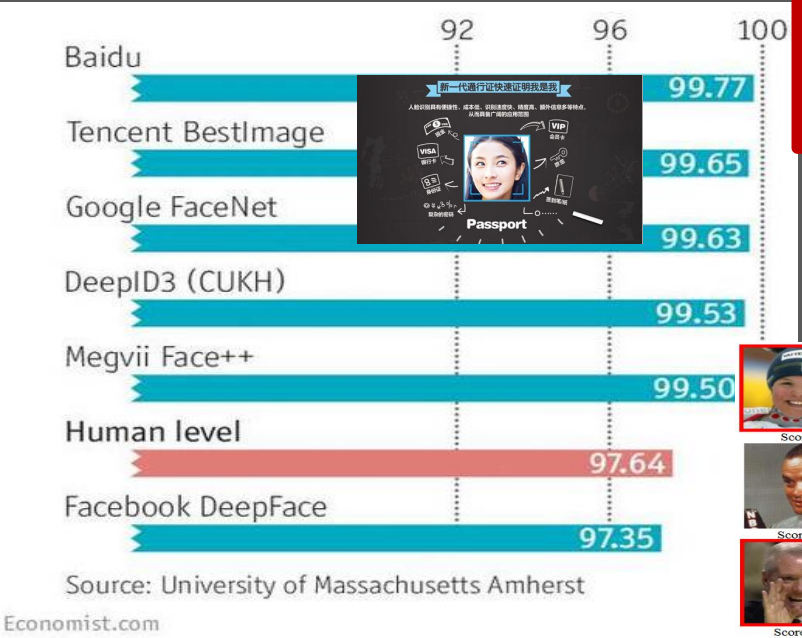
人脸检测、定位与识别



★ 针对LFW人脸识别库，深度卷积神经网络超过了人类的识别能力

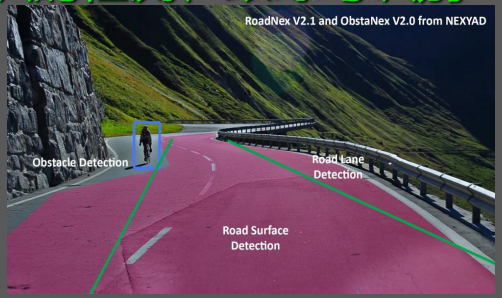
2017年

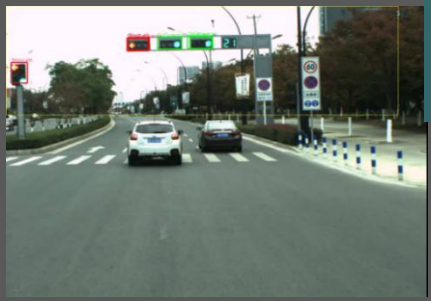
1. 阅面科技 ReadSense 99.82%;
2. 平安PingAn AI Lab 99.80%;
3. 腾讯YouTu Lab 99.80%

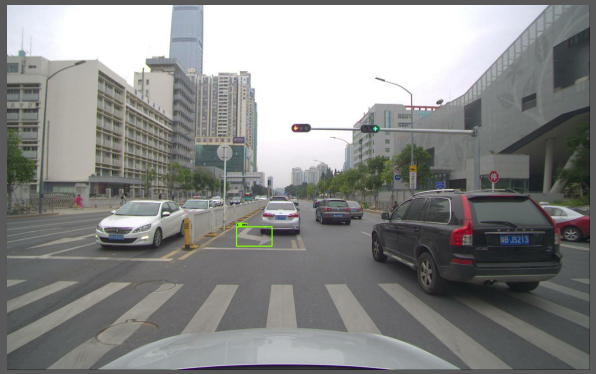


目标检测、跟踪与识别

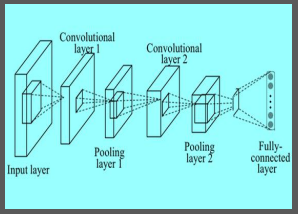
★ 对视频图像，基于深度学习的自动唇读，可行驶路面，交通信号灯、汽车尾灯与地面交通标识的检测、跟踪与识别



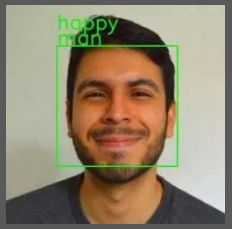


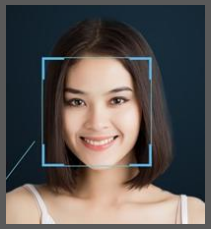


人脸表情检测与识别

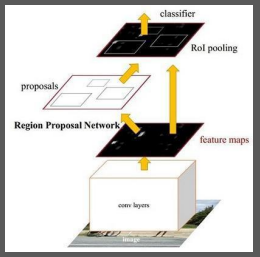


★ 基于深度学习的人脸表情识别



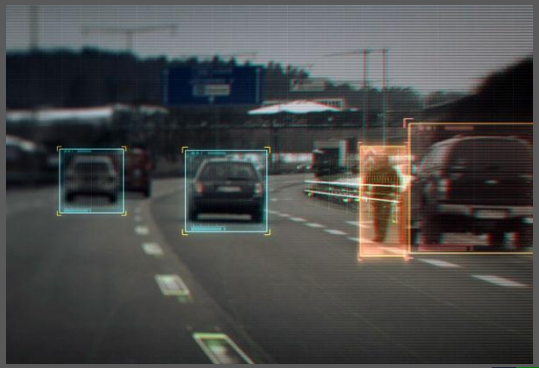


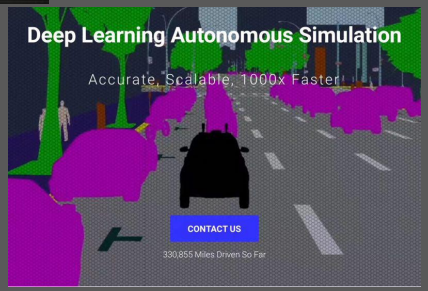
障碍物检测与识别



★ 基于区域-卷积神经网络的机动车、非机动车与行人的检测与识别

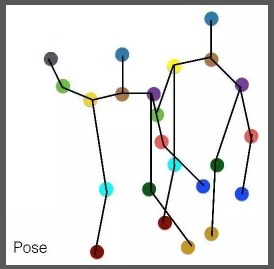
基于人工智能的行人检测

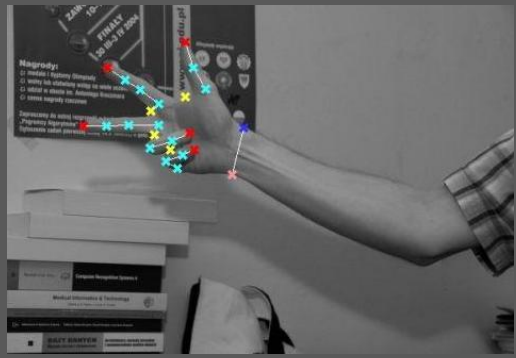


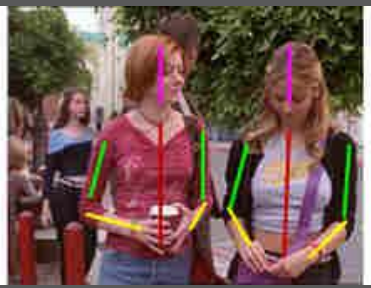


动作与行为意图的检测与识别

★ 基于深度神经网络的手势、动作识别与行为意图预测





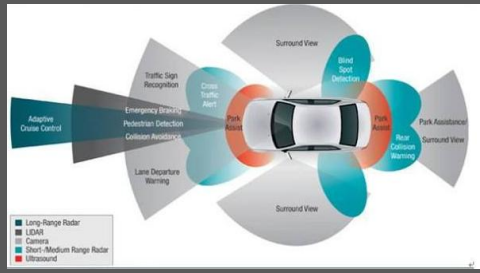


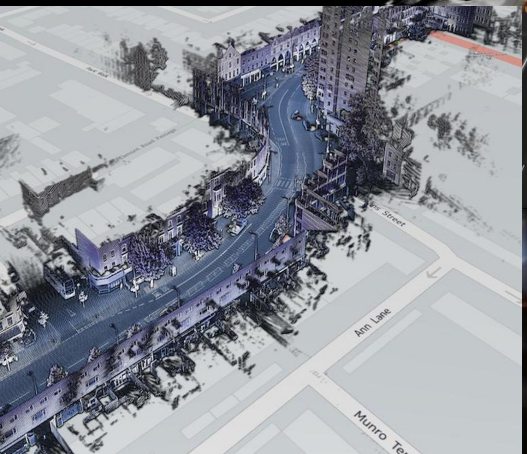
基于深度学习视觉的运动参数估计与行为意图预测

多模态视觉信息融合

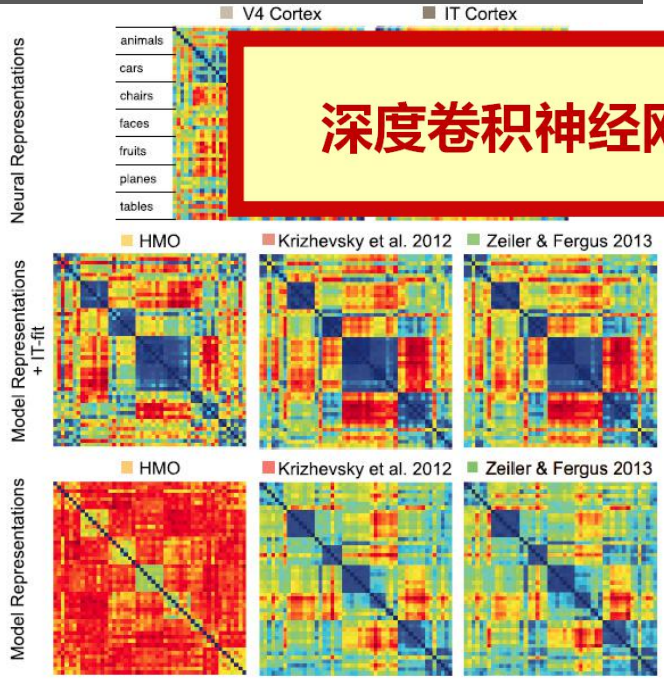
★ 基于信息融合的多模态深度学习视觉

- 与先验模型与领域专家知识结合

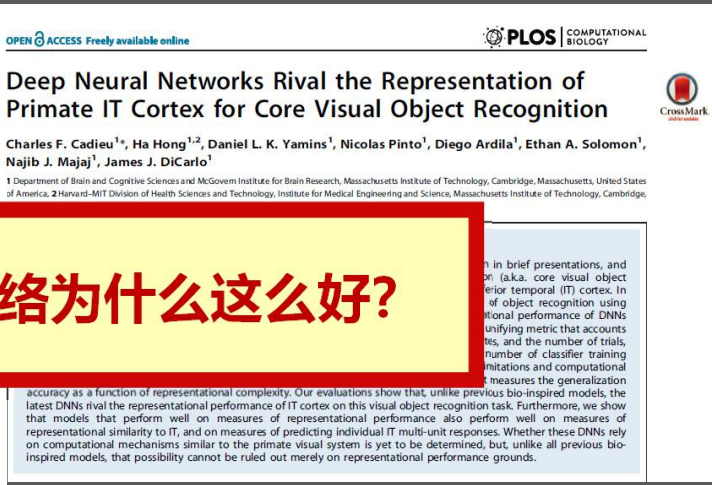




深度神经网络媲美灵长类动物的IT皮层
(美国MIT麦戈文脑科学研究所，*PLoS*, Dec. 2014)



深度卷积神经网络为什么这么好?



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition

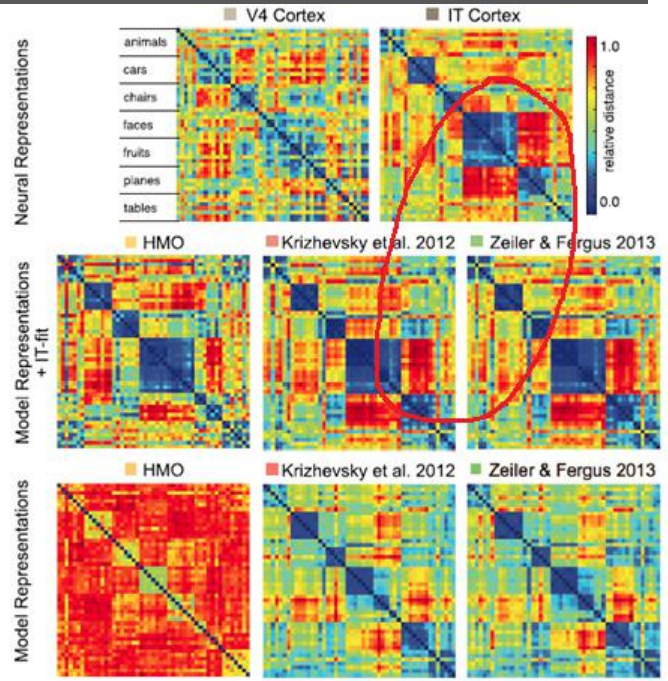
Charles F. Cadieu[1]*, Ha Hong[1,2], Daniel L. K. Yamins[1], Nicolas Pinto[1], Diego Ardila[1], Ethan A. Solomon[1], Najib J. Majaj[1], James J. DiCarlo[1]

1 Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 2 Harvard–MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge,

...n in brief presentations, and ...r (a.k.a. core visual object ...erior temporal (IT) cortex. In ... of object recognition using ...tional performance of DNNs ... unifying metric that accounts ...tes, and the number of trials, ... number of classifier training ...mitations and computational ... measures the generalization accuracy as a function of representational complexity. Our evaluations show that, unlike previous bio-inspired models, the latest DNNs rival the representational performance of IT cortex on this visual object recognition task. Furthermore, we show that models that perform well on measures of representational performance also perform well on measures of representational similarity to IT, and on measures of predicting individual IT multi-unit responses. Whether these DNNs rely on computational mechanisms similar to the primate visual system is yet to be determined, but, unlike all previous bio-inspired models, that possibility cannot be ruled out merely on representational performance grounds.

C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, et al., "Deep neural networks rival the representation of primate IT cortex for core visual object recognition", *PLoS Comput. Biol.*, vol. 10, no. 12, pp. e1003963, Dec. 2014. doi:10.1371/journal.pcbi.1003963

在多级多层特征的自动提取上，深度卷积神经网络与生物视觉通路具有某种相似性，同时也符合Hubel & Wiesel模型（1981年获诺贝尔生理或医学奖）



Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition

Charles F. Cadieu[1]*, Ha Hong[1,2], Daniel L. K. Yamins[1], Nicolas Pinto[1], Diego Ardila[1], Ethan A. Solomon[1], Najib J. Majaj[1], James J. DiCarlo[1]

1 Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 2 Harvard–MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

The primate visual system achieves remarkable visual object recognition performance even in brief presentations, and under changes to object exemplar, geometric transformations, and background variation (a.k.a. core visual object recognition). This remarkable performance is mediated by the representation formed in inferior temporal (IT) cortex. In parallel, recent advances in machine learning have led to ever higher performing models of object recognition using artificial deep neural networks (DNNs). It remains unclear, however, whether the representational performance of DNNs rivals that of the brain. To accurately produce such a comparison, a major difficulty has been a unifying metric that accounts for experimental limitations, such as the amount of noise, the number of neural recording sites, and the number of trials, and computational limitations, such as the complexity of the decoding classifier and the number of classifier training examples. In this work, we perform a direct comparison that corrects for these experimental limitations and computational considerations. As part of our methodology, we propose an extension of "kernel analysis" that measures the generalization accuracy as a function of representational complexity. Our evaluations show that, unlike previous bio-inspired models, the latest DNNs rival the representational performance of IT cortex on this visual object recognition task. Furthermore, we show that models that perform well on measures of representational performance also perform well on measures of representational similarity to IT, and on measures of predicting individual IT multi-unit responses. Whether these DNNs rely on computational mechanisms similar to the primate visual system is yet to be determined, but, unlike all previous bio-inspired models, that possibility cannot be ruled out merely on representational performance grounds.

C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, et al., "Deep neural networks rival the representation of primate IT cortex for core visual object recognition", *PLoS Comput. Biol.*, vol. 10, no. 12, pp. e1003963, Dec. 2014. doi:10.1371/journal.pcbi.1003963

计算机视觉：大数据的支撑

- 公开评测数据集是完备大数据，算法性能仅反映了深度神经网络本身达到甚至超过人类水平的感知能力

★公开评测数据集：

视觉物体检测、识别与分割 - ImageNet，MS COCO，PASCAL VOC-2007/VOC-2012，Caltech-101，Caltech-256，CIFAR-10，CIFAR-100，MNIST，US-PS，SVHN等；

人脸识别 - LFW， PubFig ，MTFL，Caltech人脸数据库，FDDB，CelebA，CK+，FER-2013， JAFFE等；

交通标识识别 - GTSRB，TRoM等

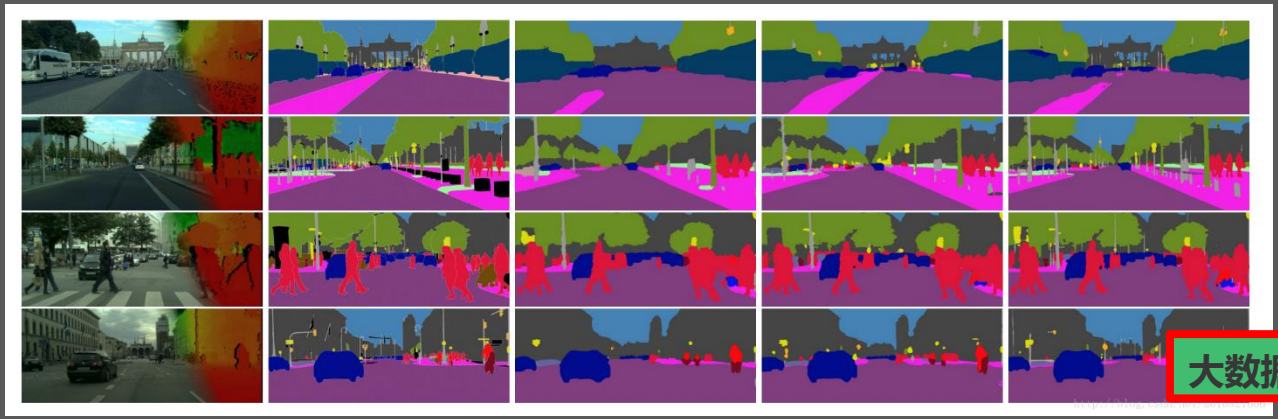
大数据

计算机视觉：大数据的支撑

- 落地应用中，开放环境下不存在完备大数据，但须尽可能多地积累大数据，采集与喂食的大数据越多，越能获得更好的感知直觉。

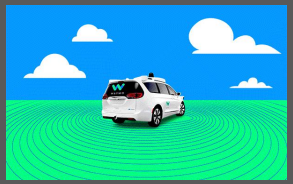
★专有大数据资源：

其重要性如同原油一样，巨头企业视之为战略资源！！



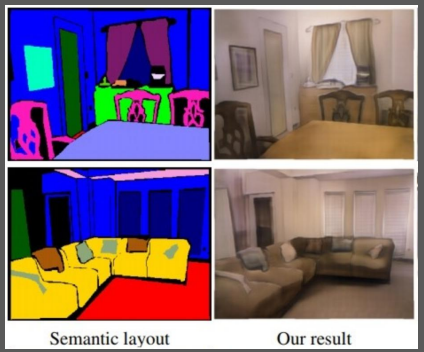
计算机视觉：大数据的支撑

预训练/迁移学习；合成/GAN或仿真数据

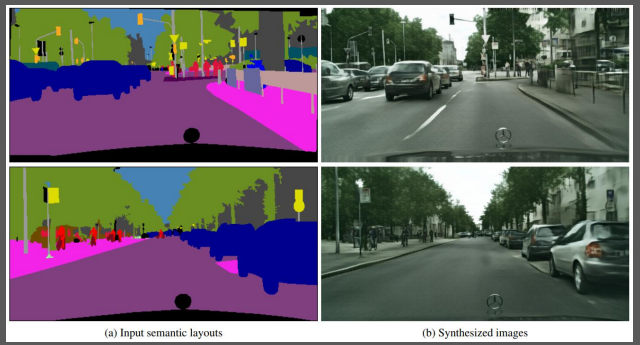


https://arxiv.org/pdf/1809.11096.pdf

利用GAN合成超分辨率真实感图像



https://arxiv.org/abs/1707.09405v1



计算机视觉：大数据的支撑

大数据燃料：喂食越多，越能获得更好的感知直觉

应用落地需要追逐大数据

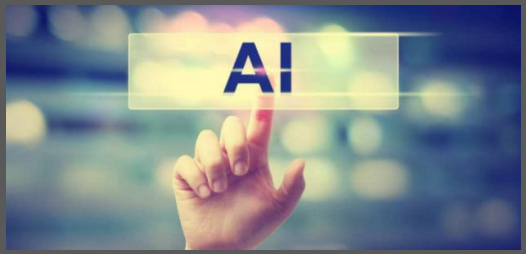
谁拥有与利用的标签大数据越多，谁的技术成熟度就越高；

在开放环境下，如何解决大数据的完备性问题？

众集&众包或寻求专业数据标注公司合作

长尾效应：

比如说识别率从 99.999% 提高到 99.99999%，需要的是指数级增长的大数据，需要极大的资源付出！



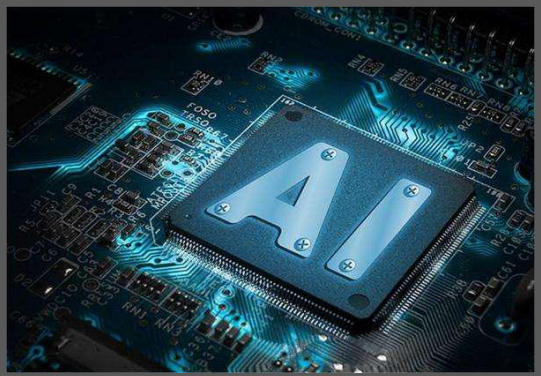
大数据：真实条件下有标签的巨量数据

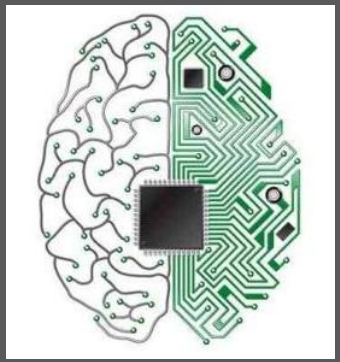
大数据

计算机视觉：算力的支撑

算力引擎：针对深度神经网络的加速能力大幅提高

- 离线训练；
- 基于云平台的在线推断应用；
- 终端推断应用（边缘计算）



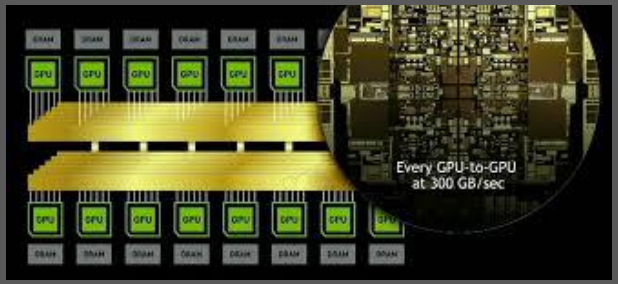


★ 通用人工智能芯片

超级GPU集群AI训练服务器

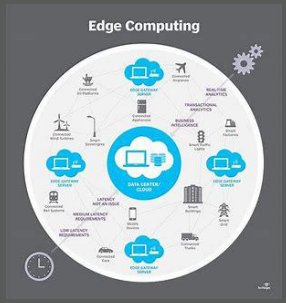
- 英伟达Tesla P100深度学习芯片具有150亿个晶体管，运算速度达到21.2万亿次，研发预算超过20亿美元；最新推出的Tesla V100的晶体管个数超过210亿；
- 英伟达最新DGX-2超级深度学习服务器，利用16块Tesla V100进行并联，显存512GB，提供最高达2,000TPLOP的深度学习计算能力。

* P40，计算能力12T，显存24GB；
* V100，计算能力14T，显存32GB



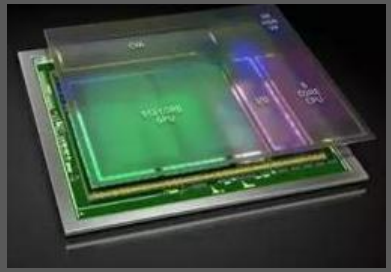
★ 通用人工智能芯片

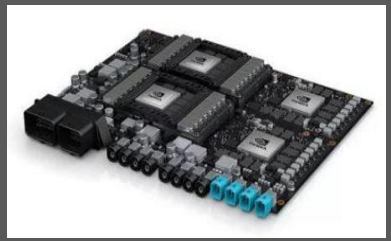
云端与终端AI推断芯片



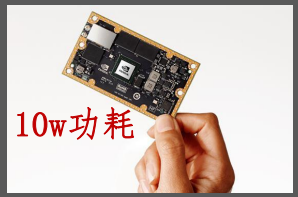
与物联网结合的边缘计算，对人工智能提出了特殊的要求：

轻量化的深度学习模型 + 超低功耗的微小型AI芯片





英伟达DRIVE PX系列



英伟达深度学习芯片 Tegra X1及Jetson TX2

★ 基于ASIC/FPGA的专用人工智能芯片

包括ASIC与FPGA异构混合的深度神经网络芯片

- 专用集成电路（Application Specific Integrated Circuit, ASIC）
- 现场可编程门阵列（Field-Programmable Gate Array, FPGA）

特点：❶低功耗，高性能；

❷通常用于深度学习的推断（inferernce）

例如，基于FPGA的深度学习处理器（DPU），其能耗比可提升至少1,000倍（相对传统GPU），且成本更低！

FPGA市场目前由Xilinx和Altera主导；

Altera被英特尔以167亿美元收购；Xilinx与IBM合作

- 基于云平台的在线应用或移动终端应用

★ 基于ASIC/FPGA的专用人工智能芯片

云端与终端AI推断芯片

■ 英特尔、谷歌、亚马逊等的AI芯片

- 谷歌TPU系列为专门为深度学习定制的ASIC芯片，完全支持谷歌的TensorFlow开源代码框架；
- TPU1，Cloud TPU（TPU2），TPU3

■ 中国AI芯片

- 华为海思（麒麟970/麒麟980），昇腾910/昇腾310；
- 中科院计算所的ASIC寒武纪DPU；
- 地平线的征程/旭日嵌入式AI处理器；
- 深鉴科技基于FPGA的低功耗DPU（被美国赛灵思收购）；
- 百度昆仑芯片；

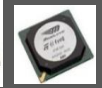







谷歌TPU系列





英特尔系列



百度昆仑


寒武纪芯片



征程2.0及Matrix 1.0

★ 类脑芯片

人脑的功耗仅20瓦左右！

皮层神经元个数140亿

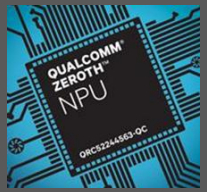
1) 基于传统CMOS工艺

❶ IBM真北（TrueNorth）

（*Science*, vol. 345, 8 Aug. 2014, pp. 668-673）

100万发放神经元；2.56亿突触连接；仅63毫瓦功耗！

❷ 高通Zeroth类脑处理器

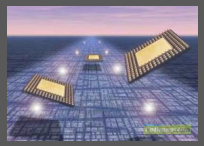


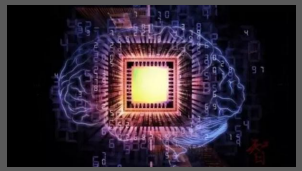
2) 基于新型忆阻器件

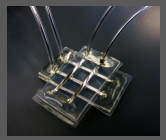
忆阻器，全称记忆电阻（Memristor），是硬件模拟突触的理想方式；

集成度更高、读写速度更快；

功耗更低！！







Nature，vol. 521, pp.61–64, May 2015

计算机视觉：应用场景

弱人工智能+

IoT, 5G, VR/AR，无人零售，智能安防，智能家居，智慧城市，自动驾驶，智能机器人，无人自主系统，并渗透更多垂直应用领域，…

大数据人工智能：将无处不在，可望替换更多依赖人类视觉功能的服务性工种和更多需要环境适应性及自主性的复杂体力劳动

提纲 OUTLINES

1、深度学习赋能的计算机视觉

2、探索具有认知理解能力的人工智能视觉

2、探索具有认知理解能力的人工智能视觉

基于深度学习的计算机视觉更接近于人的感知能力，但大数据人工智能视觉尚缺乏认知理解能力。

重大理论挑战：

视觉感知（基于小样本） + 认知理解

什么是认知智能?

☆ 认知智能，即对人类深思熟虑行为的模拟，包括记忆、常识、知识学习、推理、规划、决策、意图、动机与思考等高级智能行为

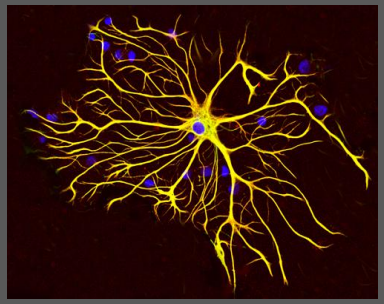
现状：追求看清、听清，有识别无理解

未来：要看懂、听懂、读懂！

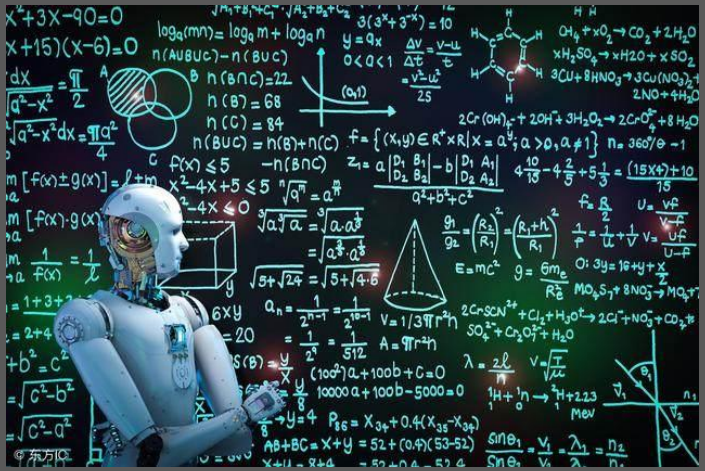
2、探索具有认知理解能力的人工智能视觉

连接主义、行为主义与符号主义的有机结合

核心是让AI有自己的语言

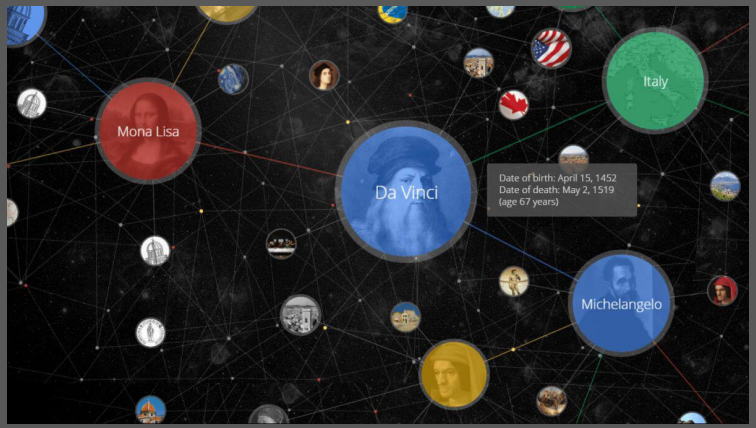


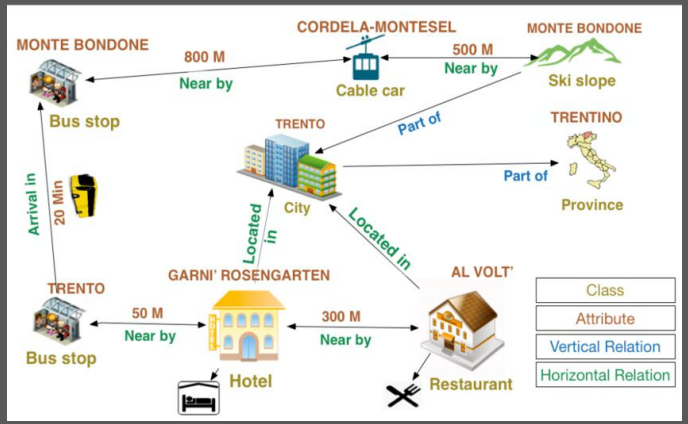
AI视觉引擎



2、探索具有认知理解能力的人工智能视觉

大数据感知智能+概率图模型/知识图谱





发展具有自主学习能力的知识图谱/Knowledge Graph

2、探索具有认知理解能力的人工智能视觉

从特征学习拓展到知识学习

概念结构图：以嵌入结构+概率图模型（知识图谱）的方式抽象、延伸概念，赋以其内涵与外延，以实现**认知理解**；

知识表达：概念结构图通过学习进行时空递归，获得**知识**的表达、记忆与学习能力；

规则学习：语言智能，核心是如何具有学习能力；

知识学习：规则+常识+经验

结语：

☆ 无论是单目/双目/红外摄像机，还是激光雷达/毫米波雷达成像，感知问题大多可归结为场景或目标的计算机视觉问题；

☆ 在大数据和超强计算能力的支撑下，基于深度学习的计算机视觉更加接近于人的视觉感知能力；

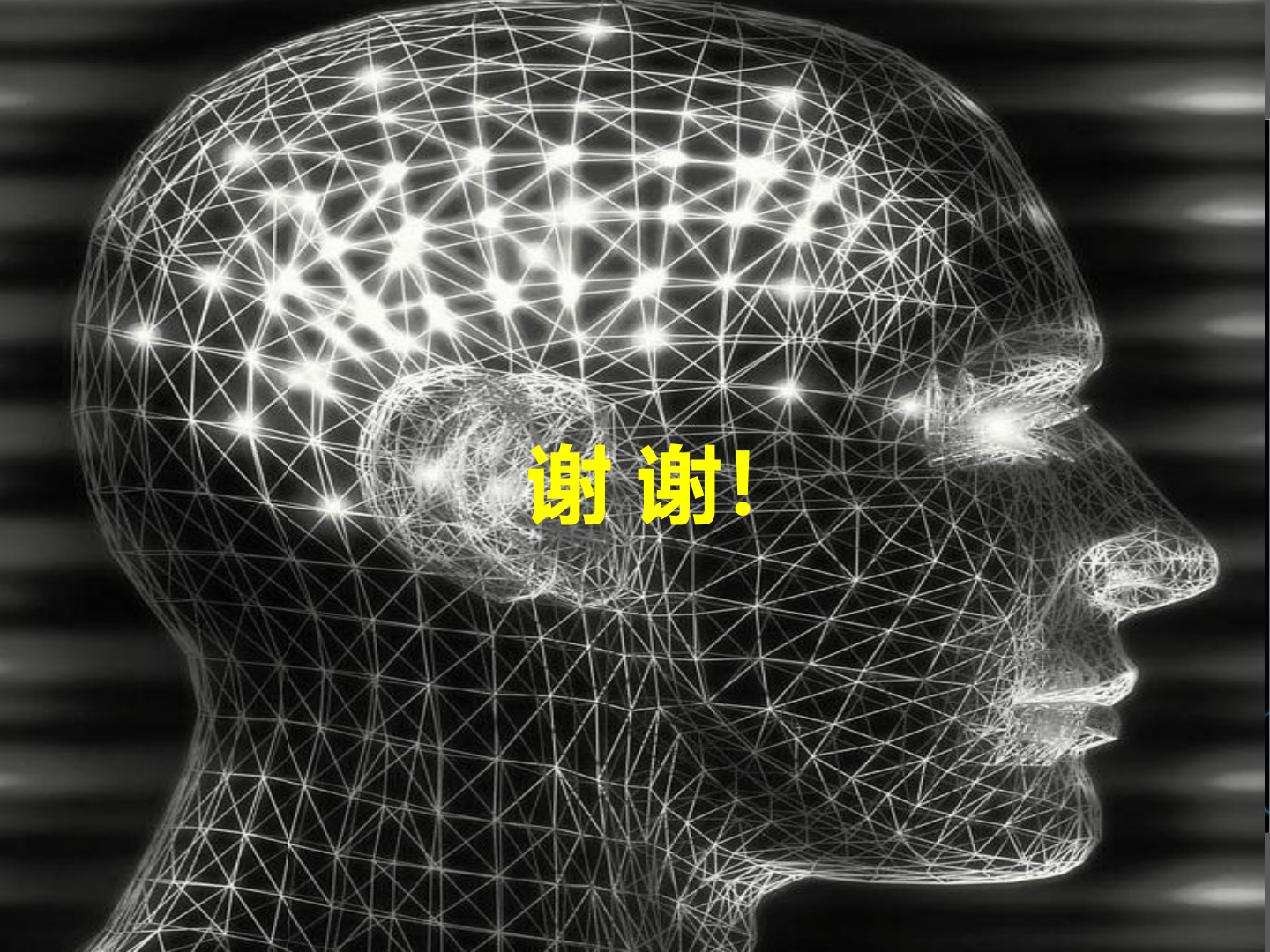
☆ 开放环境下不存在完备大数据，但须尽可能多地运用标签大数据，并且还要考虑到数据与性能的长尾效应，因此产业应用中亟需发展基于小样本的深度学习视觉方法；

☆ 须突破多模态视觉融合技术，进一步增强视觉感知的用户体验；

☆ 大数据人工智能视觉尚缺乏认知理解能力，已成为目前人工智能前沿探索中面临的一个重大技术挑战；

☆ 结合具有自主学习能力的概率图模型或知识图谱，发展具有认知理解能力的计算机视觉技术，可望成为人工智能技术与产业的核心赋能力量；

☆ 轻量化、低功耗、低成本人工智能必将成为发展智能物联网的基础。


谢 谢!