中国科学院大学网络空间安全学院/夏季学期/高级强化课

大数据与人工智能技术

第2章 计算机系统和人工智能赋能数据处理分析

第3节 基于人工智能的多媒体分析及应用

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

授课老师：岳银亮

2021年7月16日

本节内容大纲

目录

1 多媒体大数据

2 图像分析及应用

3 语音分析及应用

4 视频分析及应用

5 多模态融合及应用

6 深度伪造

多媒体大数据

- 目前非结构化的多媒体数据被快速和广泛的使用，充斥着日常生活的各个方面，并且多媒体资源易于访问、可用性强
- 多媒体数据主要包括文本、图像、音频和视频这四种

文本

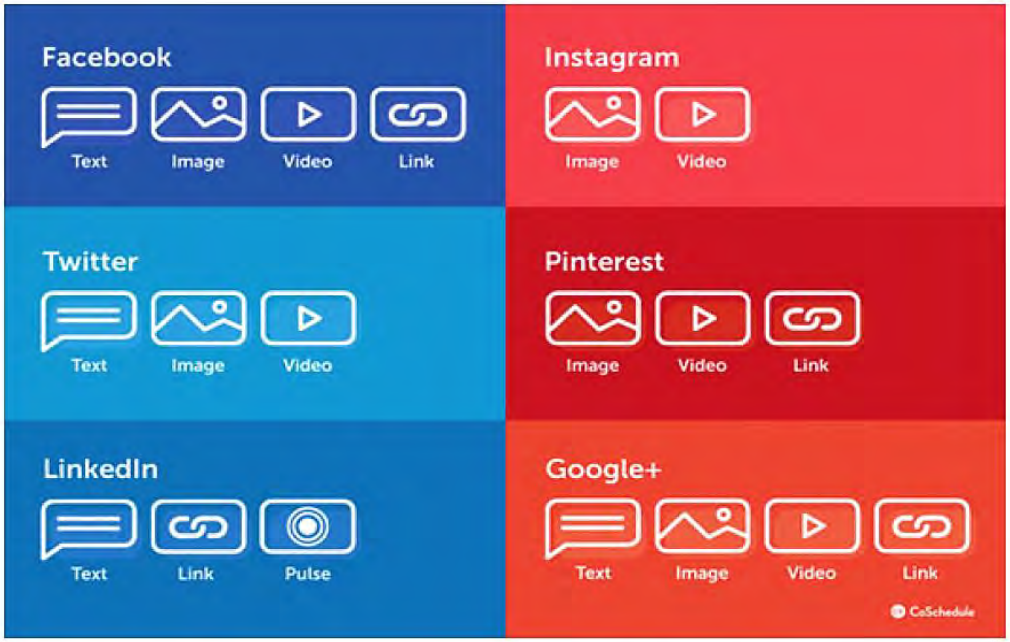
图像

音频

视频

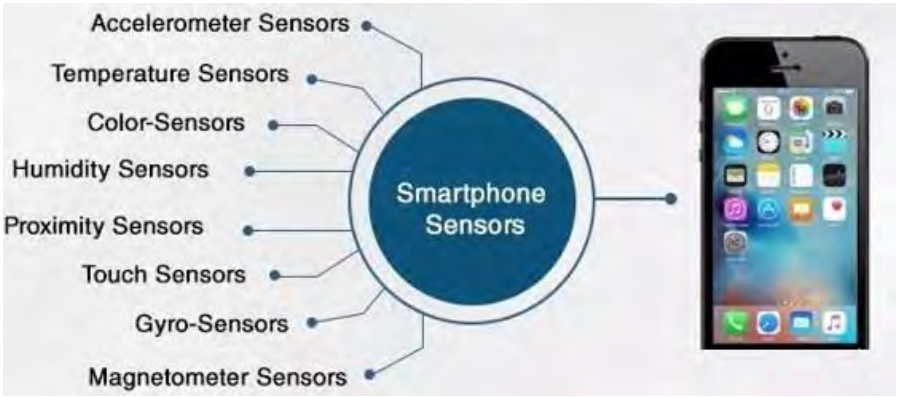
多媒体大数据

- 互联网和社交网络用户创建的多媒体数据是当前最常见、最大规模的、来源于人的 (human-sourced) 多媒体数据
  - **多媒体分享网站**: YouTube、iCloud、Flickr抖、音、快手…
  - **社交网络**：Facebook 、Instagram 、Twitter、微信、微博…



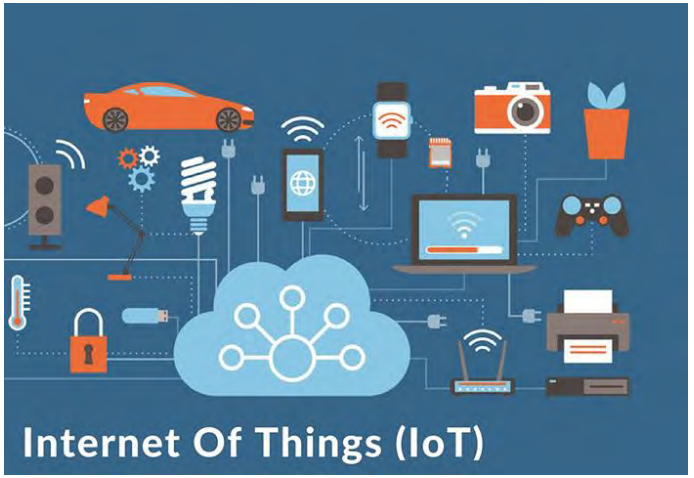
多媒体大数据

- **智能手机**：近年来，智能手机在人们的生活中已经超过了笔记本电脑等其他电子设备，数十亿人几乎随时随地携带智能手机
  - **智能手机的先进功能和技术**，如蓝牙、GPS、相机、强大的 CPU、网络连接等，可以访问和操作所有多媒体数据格式（例如，音频、图像、视频或文本）
  - **手机中的创新应用**的爆炸式增长使智能手机成为多媒体大数据的重要来源



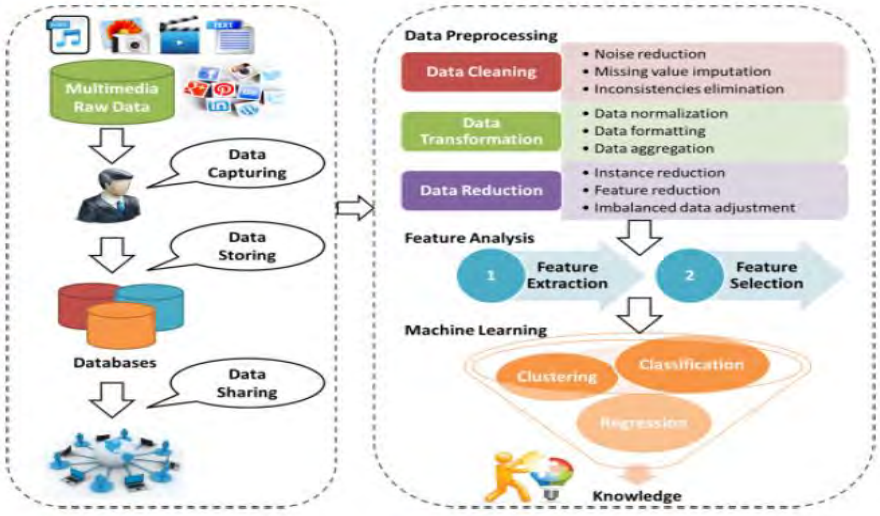
多媒体数据

- **物联网**：包括大量传感器和机器生成的数据，这些传感器和机器用于测量和记录现实世界中的事件
  - **固定传感器的数据**：交通传感器、网络摄像头、监控视频和图像、天气或污染传感器、家庭自动化设备数据等
  - **移动传感器跟踪的数据**：手机定位、汽车、卫星图像或移动设备中计算机系统的数据等

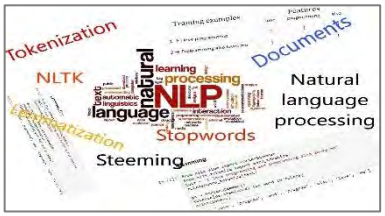


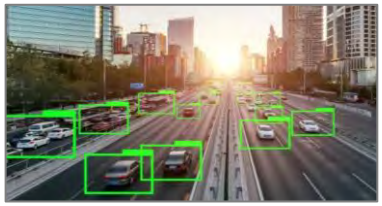
多媒体数据

- 多媒体数据符合“**大数据**”的特点，**数量大、种类多并具有价值**
- **多媒体大数据每天在不断的迅速产生**，比如2019年Facebook每天约有14亿活跃用户，日均上传照片约有3亿张，平台日均视频播放量达到了80亿
- **多媒体数据通常是非结构化且嘈杂的**，需要以有效和高效的方式处理、管理、挖掘、理解，并进行应用



多媒体数据





文本分析和挖掘

图像分析和挖掘

视频分析和挖掘

语音分析和挖掘

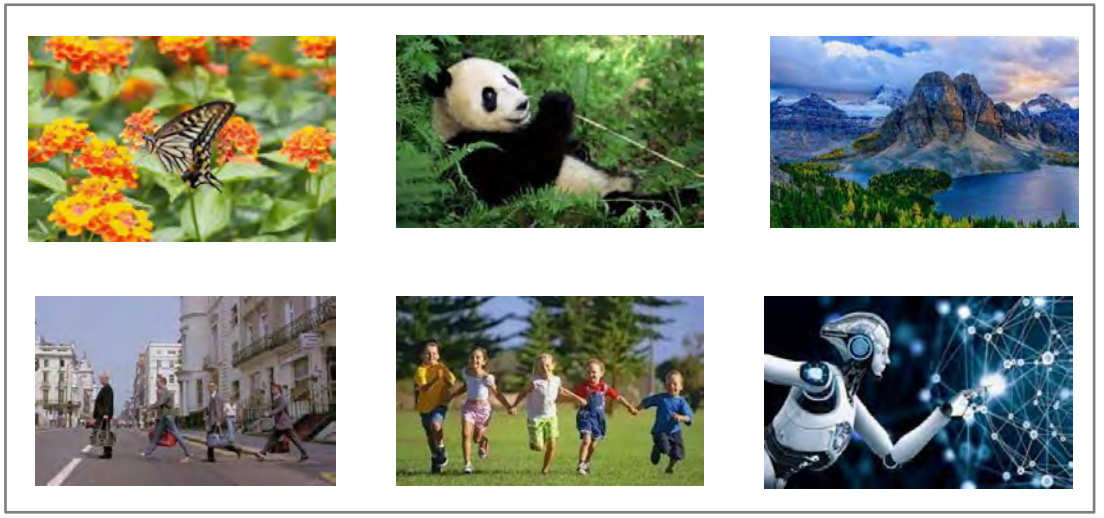
多模态数据融合

目录

- 多媒体大数据
- 图像分析及应用
- 语音分析及应用
- 视频分析及应用
- 多模态融合及应用
- 深度伪造

图像分析及应用

- 视觉是人类获取信息最主要的渠道，我们通过**视觉感知**着一切。人的大脑皮层， 有差不多 70% 都是在处理视觉信息
- 图像是最主要的非结构化数据类型，出现在生活的各个方面，是我们主要的信息来源，并且图像数据在爆炸式的增长

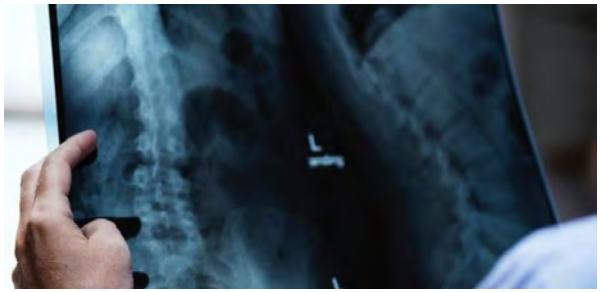


图像分析及应用

- 基于人工智能的图像分析属于计算机视觉领域(Computer Vision, CV)
- 作为人类，我们能够理解和描述如下图像中的场景
  - 这不仅涉及检测到前景中的四个人，一条街道和几辆汽车等基本信息。还能够了解人们正在行走，甚至知道他们是谁
  - 可以合理地推断出他们没有被车撞的危险，白色汽车停放位置不佳等
- **基于人工智能的图像分析**：**使计算机首先能够处理非结构化的图像数据，然后像人一样理解分析图像中的内容信息，从而实现各项应用**

图像分析及应用

- 基于人工智能的图像分析实际中有着广泛的应用
  - 门禁、支付宝上的人脸识别
  - 停车场、收费站的车牌识别
  - 上传图片到网站时的风险识别
  - 自动驾驶中的环境感知
  - 医疗保健领域中医学图像分析
  - 制造业中零件产品的故障发现
  - 农业领域预测疾病或虫害

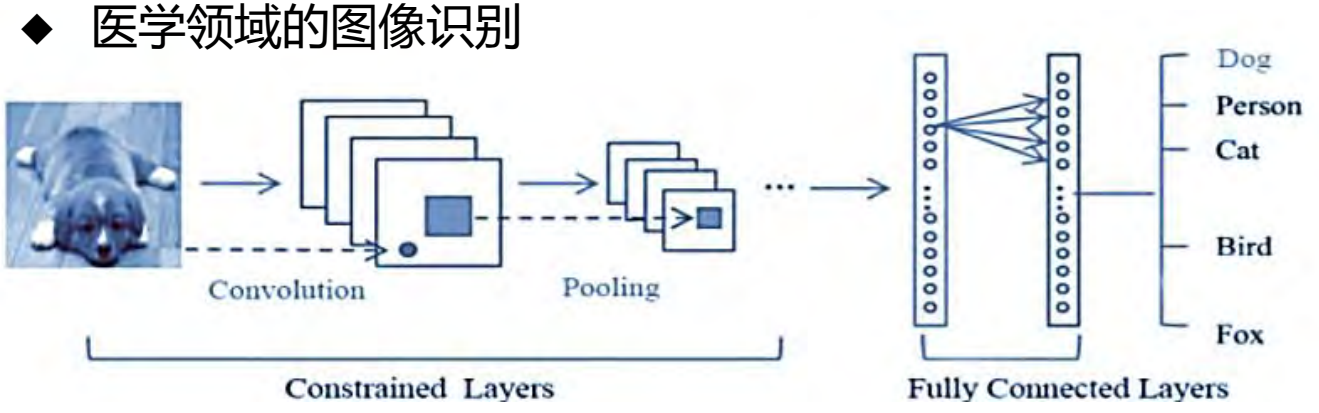


图像分析及应用

- **图像处理分析中的常见任务**
  - **图像分类**
  - 目标检测
  - 语义分割
  - 人脸识别
  - 文字识别
  - 图像生成

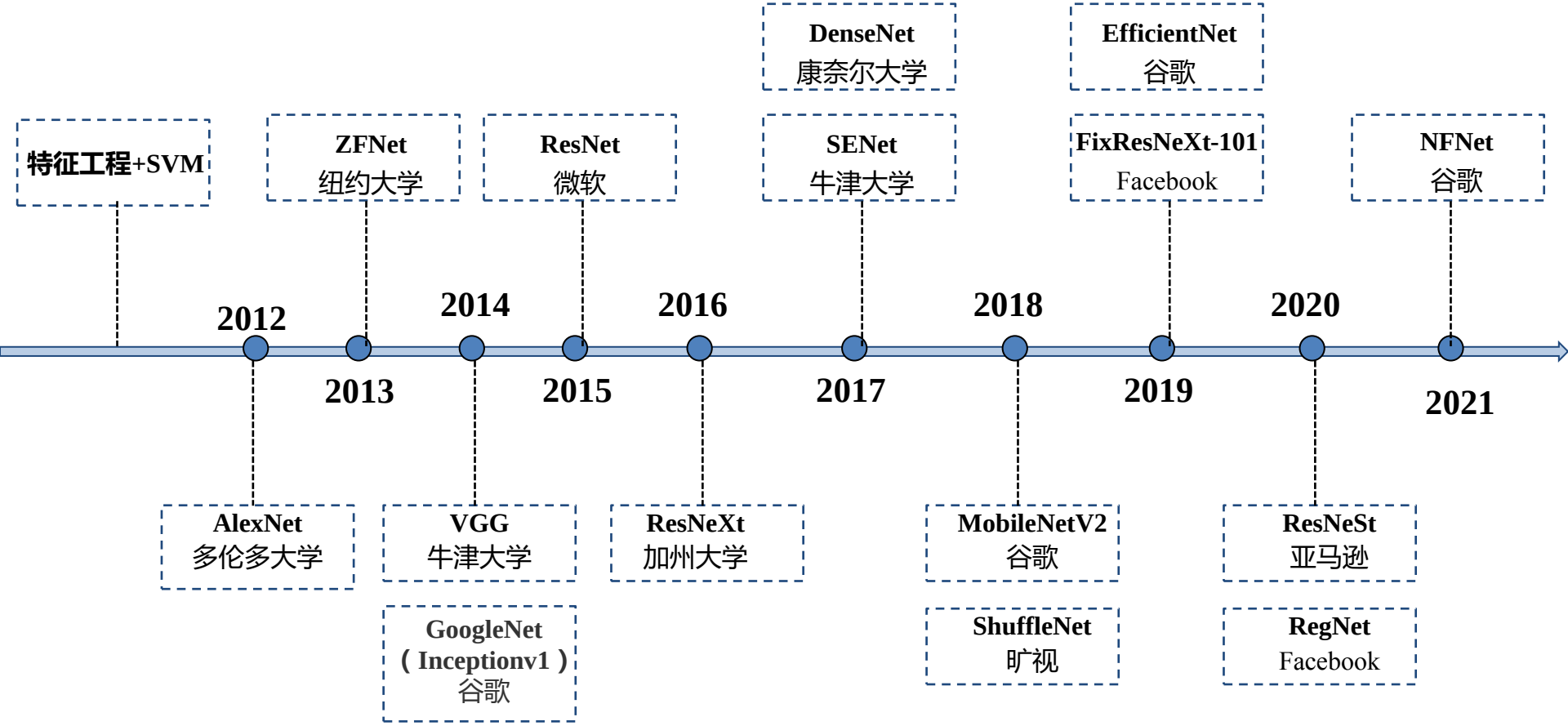
图像分析及应用

- **图像分类（Image Classification）**
  - **图像分类是根据图像的语义信息将不同类别图像区分开来**，是计算机视觉中重要的基本问题，也是图像检测、图像分割、物体跟踪、行为分析等其他高层视觉任务的基础
  - 图像分类在很多领域有广泛应用
    - 安防领域的人脸识别和智能视频分析
    - 交通领域的交通场景识别
    - 互联网领域基于内容的图像检索和相册自动归类
    - 医学领域的图像识别



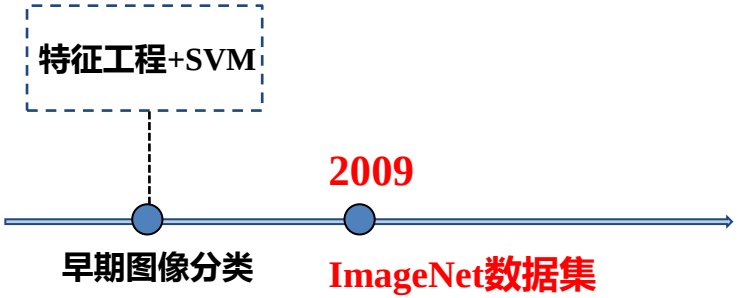
图像分析及应用

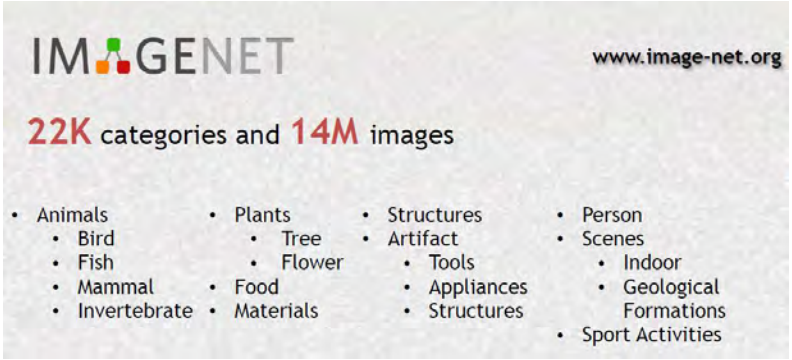
- **图像分类的发展**



图像分析及应用

- **图像分类的发展**



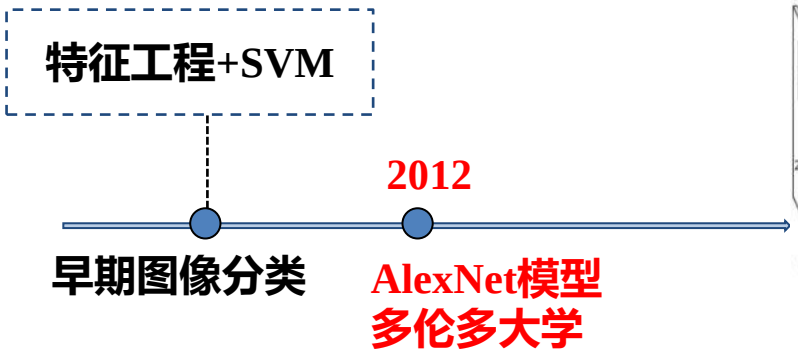


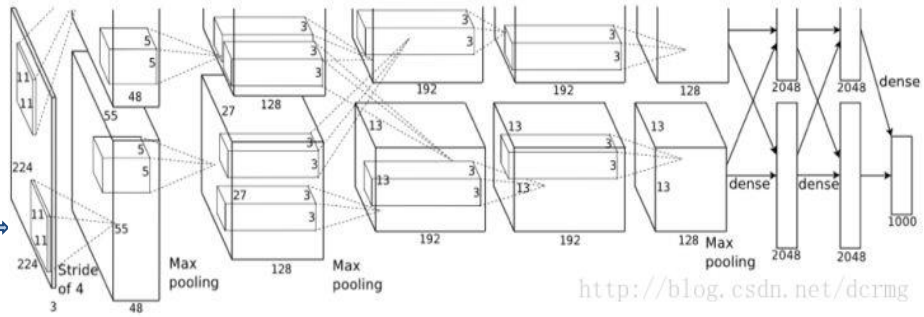
- 2009年斯坦福大学的李飞飞等人在 CVPR 2009 发表了**ImageNet数据集，是计算机视觉领域常用的数据集之一**。在图像分类、目标分割和目标检测中有着无法撼动的地位。这篇论文在 ImageNet 发布十周年之际，于 CVPR 2019 大会上获得了经典论文奖
- 从 2010 年起至2017，每年 ImageNet 官方会举办挑战赛**ILSVRC**，每年都会有层出不穷的图像分类模型

Deng J, Dong W, Socher R, et al, ". Imagenet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248-255, 2009.

图像分析及应用

- **图像分类的发展**



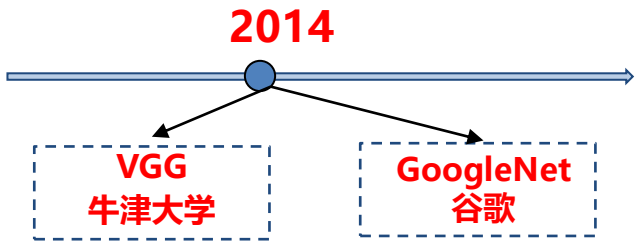


- Alexnet由多伦多大学的Alex Krizhevsky设计提出，是2012年ImageNet竞赛的冠军模型
- 该模型使用卷积神经网络（CNN）和GPU而赢得图像识别竞赛。**ALexnet使人们意识到卷积神经网络的优势，以及可以利用GPU加速卷积神经网络训练**
- **Alexnet是首次将深度学习用于大规模图像分类中，**从AlexNet之后，涌现了一系列CNN模型，不断地在ImageNet上刷新成绩。

Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017

图像分析及应用

- **图像分类的发展**



- **VGG**：牛津大学和Google DeepMind的研究员一起研发提出，取得了ILSVRC2014比赛分类项目的第二名
- **VGG的特点**
  - **优点**： VGG结构简单、容易扩展，并且很适合迁移学习，至今仍被广泛使用
  - **缺点**： VGG可以看成是加深版的AlexNet，缺点是：参数多，计算效率较低

- **GoogLeNet**：2014年由谷歌提出，是ILSVRC2014比赛的冠军模型
- **GoogLeNet的特点**
  - **优点**：增加模型的深度和网络宽度，但它在网络上做了更大胆的优化，避免了深度网络中存在的参数多、计算效率低、梯度消失等难以优化的问题
  - **缺点**：扩展迁移性没有VGG好

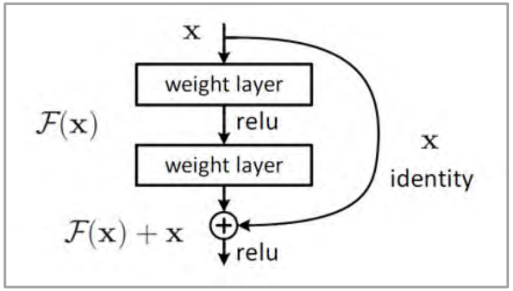
Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
Szegedy C, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

图像分析及应用

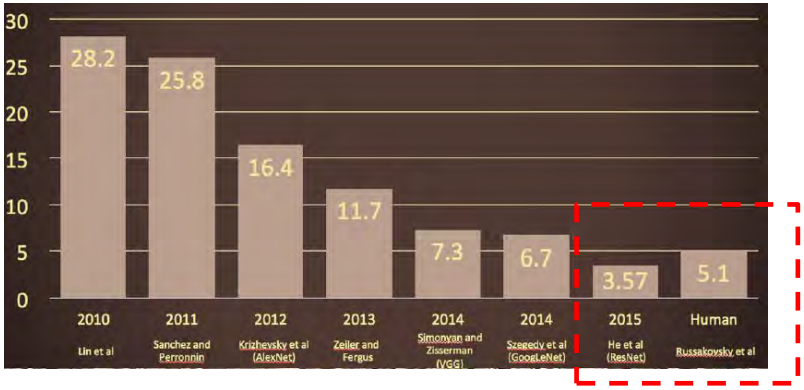
- **图像分类的发展**

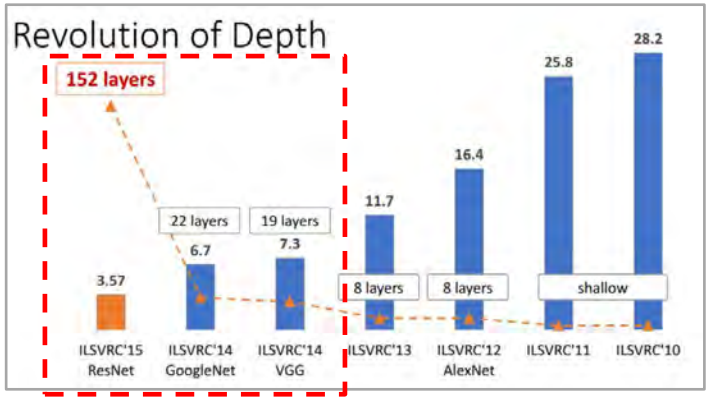
2015 ResNet-微软

- ResNet由微软何凯明等四位学者提出，是2015年ImageNet图像分类、图像物体定位和图像物体检测比赛的冠军，CVPR2016最佳论文
- **ResNet网络很深：**2014年的VGG才19层，而2015年的ResNet多达152层，通过**残差学习**等技巧解决了**深度网络的退化问题，**让我们**可以训练出更深的网络**



- Resnet凭借它优异的性能**成为了第一个图像分类精确度大于人类的神经网络**





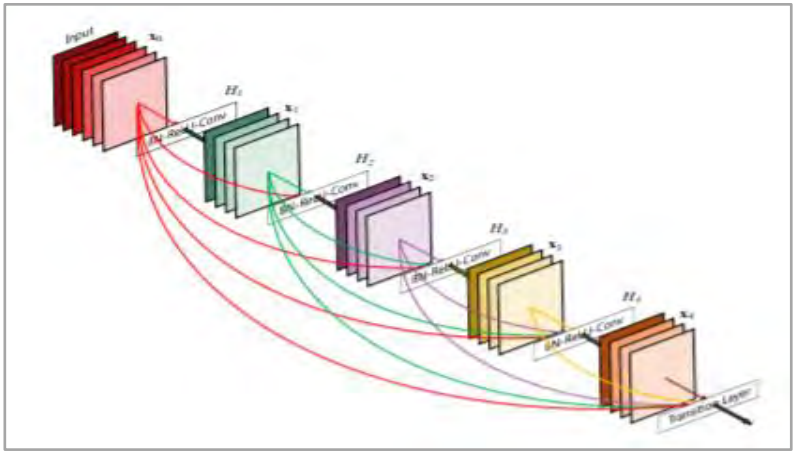
He K, Zhang X, Ren S, et al, ". Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770-778, 2016.

图像分析及应用

- **图像分类的发展**

2016 DenseNet 康奈尔大学

- DenseNet由康奈尔大学在2016年提出，是CVPR2017最佳论文
- ResNet模型的核心是残差连接，有助于训练过程中梯度的反向传播，能训练出更深的CNN网络。DenseNet模型的基本思路与ResNet一致，但做出了其他创新
  - DenseNet提出了一种前面所有层与后面层的密集残差连接方式
  - DenseNet通过特征在channel上的连接来实现特征重用，让DenseNet在参数和计算成本更少的情形下实现比ResNet更优的性能

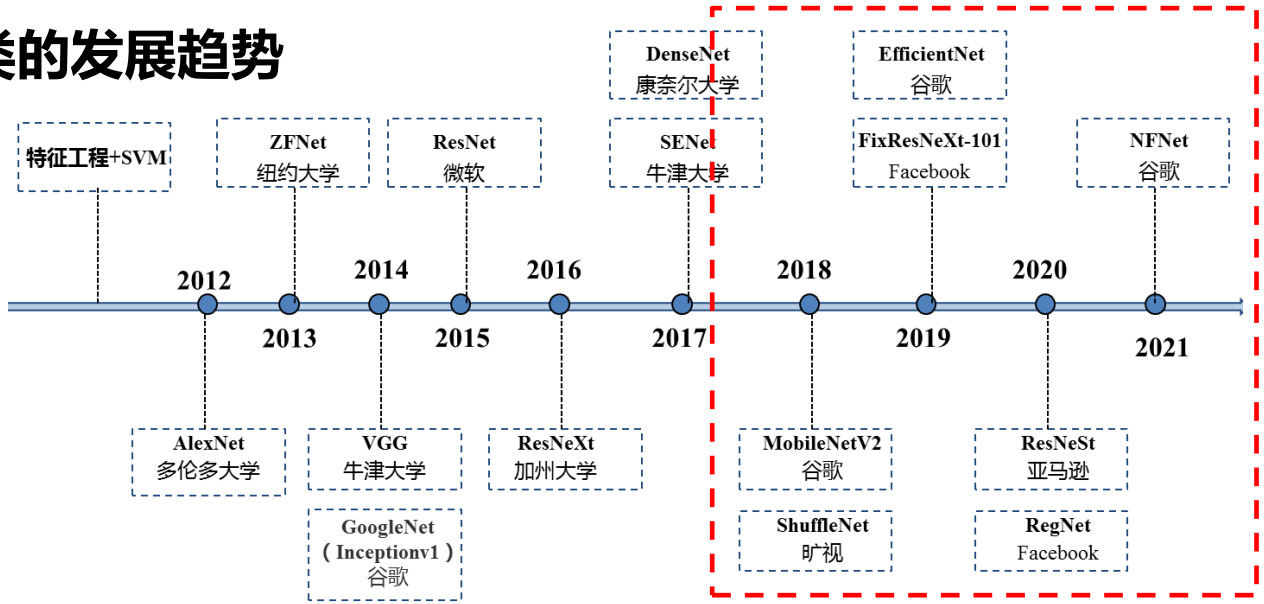


	ResNet	DenseNet
创新	• 残差学习 • shortcuts连接 • 加深网络不退化	• 密集shortcut连接 • 特征重用 • 引入过渡层
第L层输出	$x_\ell = H_\ell(x_{\ell-1}) + x_{\ell-1}$,	$x_\ell = H_\ell([x_0, x_1, \ldots, x_{\ell-1}])$,
拼接方式	Element-wise add	Concatenate
训练速度（相对）	快	慢，由于通道叠加原因，需要频繁读取内存，拖慢速度
参数量（相对）	多	少

Huang G, Liu Z, Van Der Maaten L, et al, “. Densely connected convolutional networks,” in Proc. CVPR, 2017. pp. 4700-4708, 2017.

图像分析及应用

- **图像分类的发展趋势**



- **趋势：模型越来越深，参数逐年增加**，从2012年有 60M 参数量的 AlexNet 到2019年有着 829M 的 FixResNeXt-101 32×48d，模型越来越大
- 从 2018 年开始，**缩小参数量、提高计算效率并保持很好的性能表现的研究逐渐增多**。2019最著名的谷歌提出的小型化模型是EfficientNet，仅有 66M 的 参数量，但已经接近当时的 SOTA 分数了。2021年谷歌的DeepMind团队发布了NFNet，性能和当前表现最佳的模型相当，训练速度却快了 8.7 倍

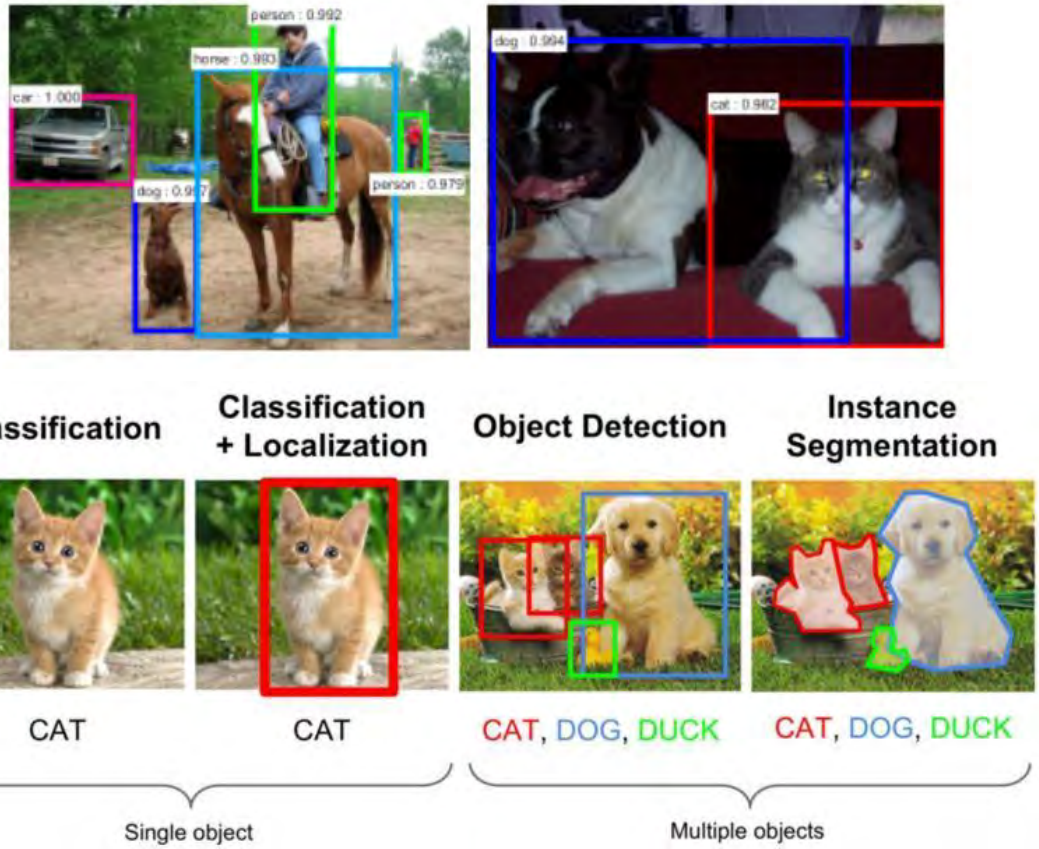
图像分析及应用

- **图像处理分析中的常见任务**
  - 图像分类
  - **目标检测**
  - 语义分割
  - 人脸识别
  - 文字识别
  - 图像生成

图像分析及应用

- **目标检测（Object Detection）**

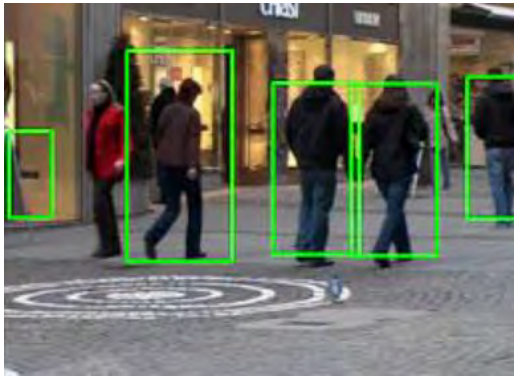
➢ 目标检测任务的目标是给定一张图像，让**计算机找出图像中所有物体目标的位置**，**并给出每个目标的具体类别**

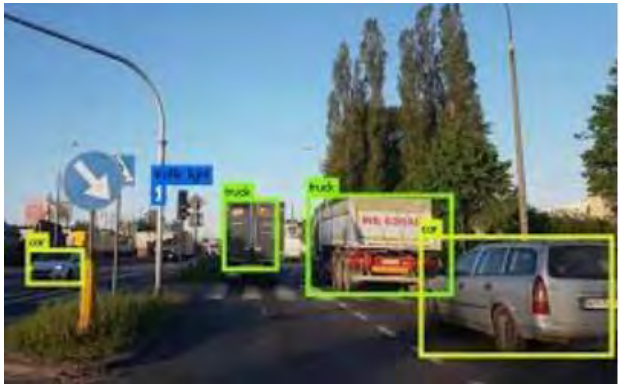


图像分析及应用

- **目标检测（Object Detection）**

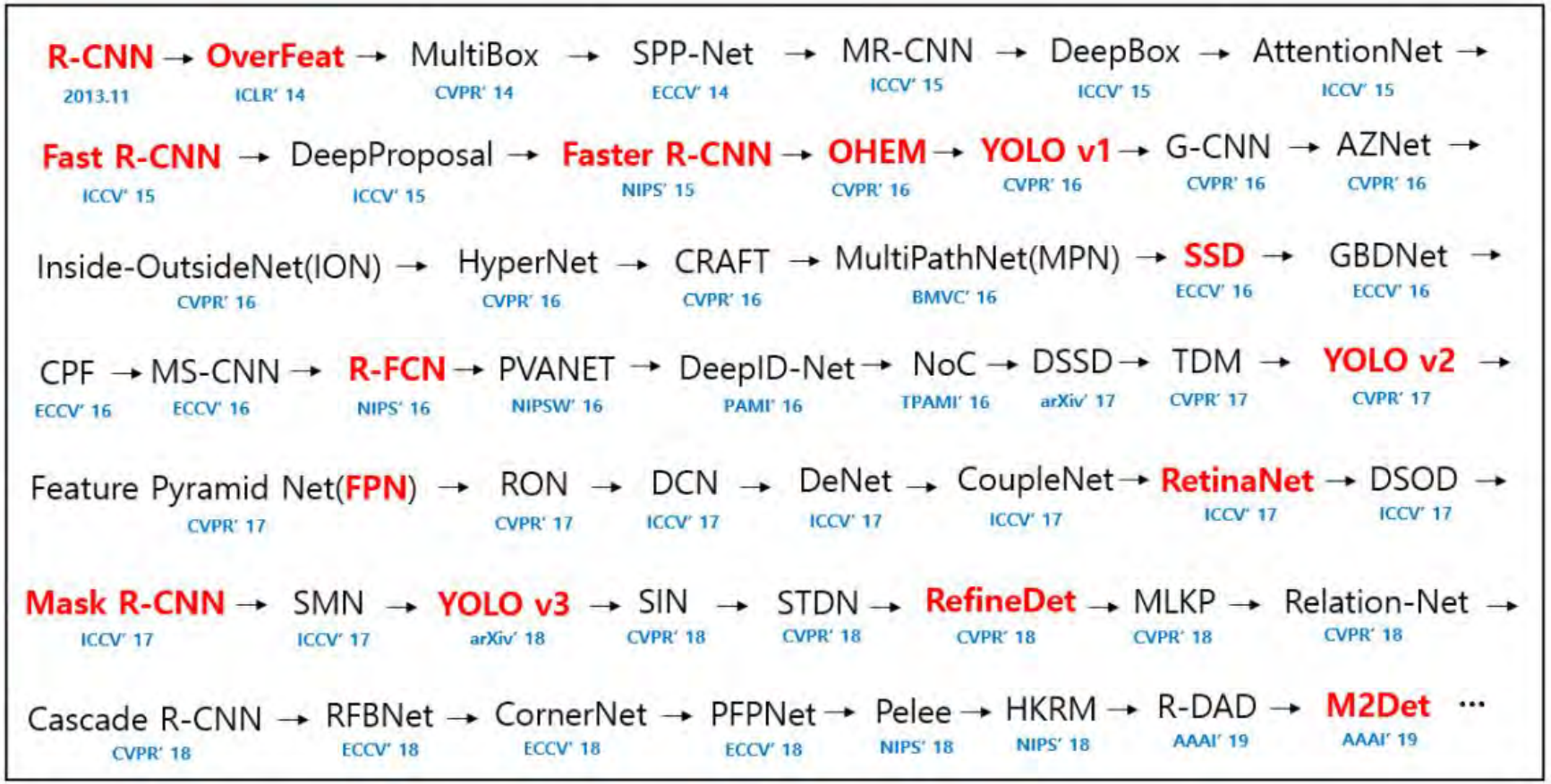
- 目标检测是计算机视觉和数字图像处理的一个热门方向，广泛应用于交通检测、机器人导航、智能视频监控、工业检测、航空航天等诸多领域，通过计算机视觉减少对人力资本的消耗，具有重要的现实意义
- 目标检测是泛身份识别领域的一个基础性的算法，对后续的人脸识别、步态识别、人群计数、场景语义理解、实例分割等任务起着至关重要的作用





图像分析及应用

- **目标检测研究发展**



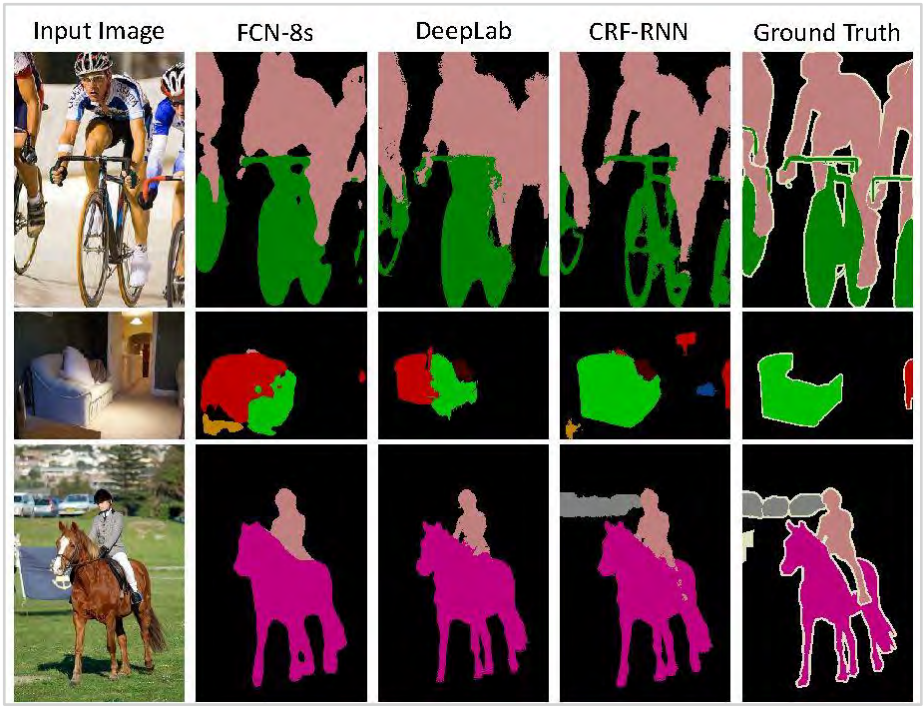
图像分析及应用

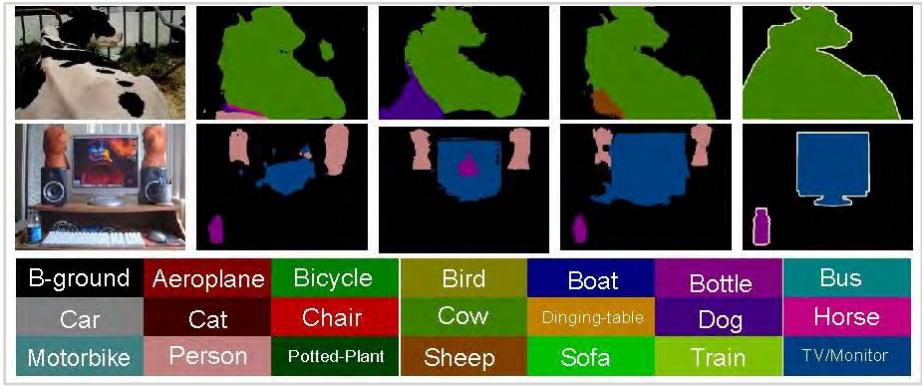
- **图像处理分析中的常见任务**
  - 图像分类
  - 目标检测
  - **语义分割**
  - 人脸识别
  - 文字识别
  - 图像生成

图像分析及应用

- **语义分割（Semantic Segmentation）**

➢ 图像语义分割从字面意思上理解就是**让计算机根据图像的语义来进行分割**。具体是指**将标签或类别与图片的每个像素关联**的一种深度学习算法。它用来识别构成可区分类别的像素集合





图像分析及应用

- **语义分割（Semantic Segmentation）**

- 语义分割**结合了图像分类、目标检测和图像分割**，通过一定的方法将图像分割成具有一定语义含义的区域块，并识别出每个区域块的语义类别，实现从底层到高层的语义推理过程，最终得到一幅具有**逐像素语义标注**的分割图像
- 语义分割一般是针对图像进行**像素级分类**。具体而言，语义图像分割就是**将每个像素都标注上其对应的类别**。
  - **图像分类**：判别图中物体是什么，比如是猫还是狗；
  - **目标检测**：寻找图像中的物体并进行定位（通过边界框）；
  - **语义分割**：对图像进行像素级分类，预测每个像素属于的类别，不区分个体；

图像分析及应用

- **语义分割的应用**

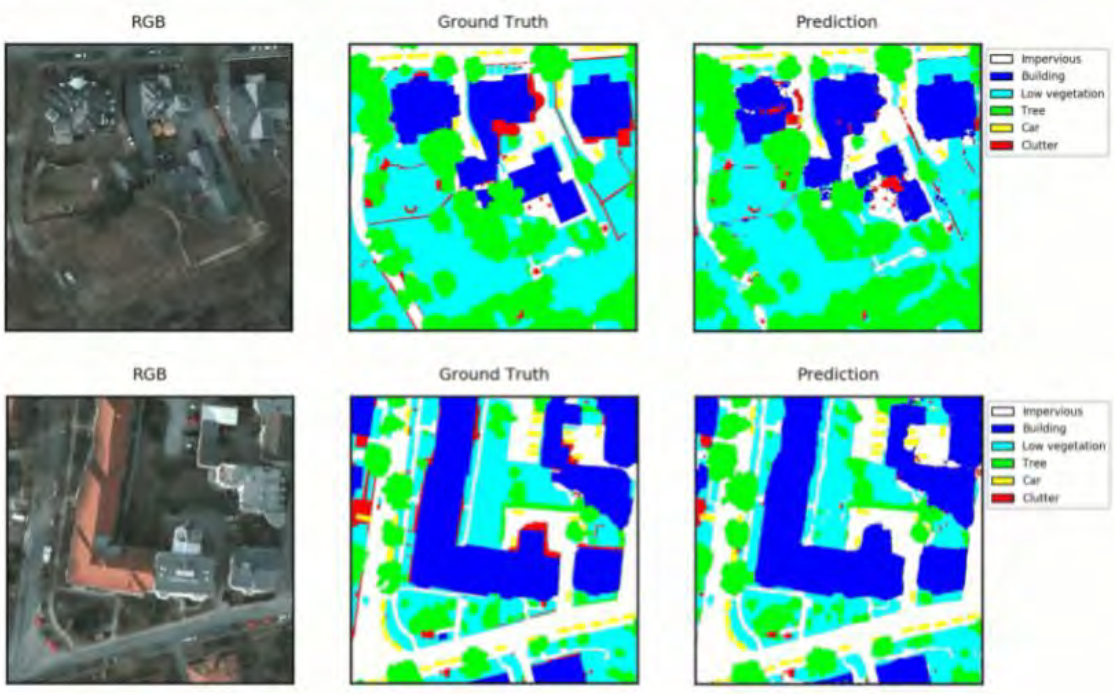
  - **无人车驾驶**：语义分割是无人车驾驶的核心算法技术，车载摄像头，或者激光雷达探查到图像后输入到神经网络中，后台计算机可以自动将图像分割归类，通过区分道路与障碍物，比如行人、人行道、电线杆和其他汽车，让汽车识别可行驶的路径。

图像分析及应用

- **语义分割的应用**

➢ **地理信息系统**：可以通过训练神经网络让机器输入卫星遥感影像，自动识别道路，河流，庄稼，建筑物等，并且对图像中每个像素进行标注。

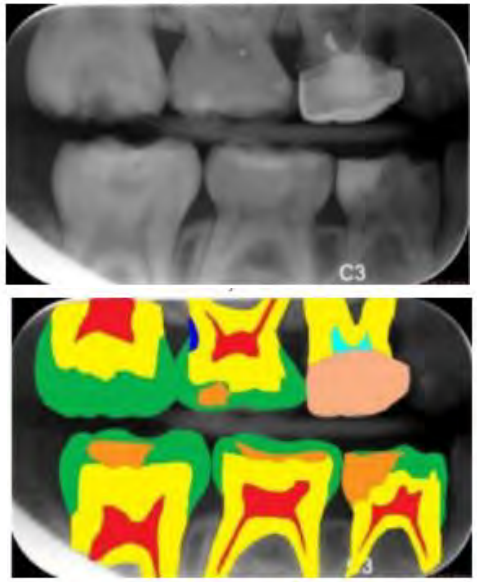
左边为卫星遥感影像，中间为真实的标签，右边为神经网络预测的标签结果

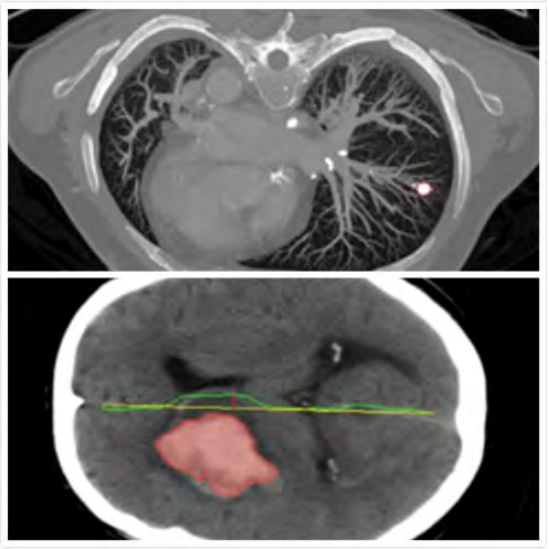


图像分析及应用

- **语义分割的应用**

➢ **医疗影像分析**：随着人工智能的崛起，将神经网络与医疗诊断结合也成为研究热点，智能医疗研究逐渐成熟。在智能医疗领域，语义分割主要应用有肿瘤图像分割，龋齿诊断等。(下图分别是龋齿诊断，头部CT扫描紧急护理诊断辅助和肺癌诊断辅助)





图像分析及应用

- **图像处理分析中的常见任务**
  - 图像分类
  - 目标检测
  - 语义分割
  - **人脸识别**
  - 文字识别
  - 图像生成

图像分析及应用

- **人脸识别（Facial recognition）**

➢ 人脸识别是基于人的脸部特征信息进行身份识别的一种生物识别技术。人脸识别的目的就是要**判断图片和视频（视频是由图片构成的）中人脸的身份**

门禁系统

安防监控等系统

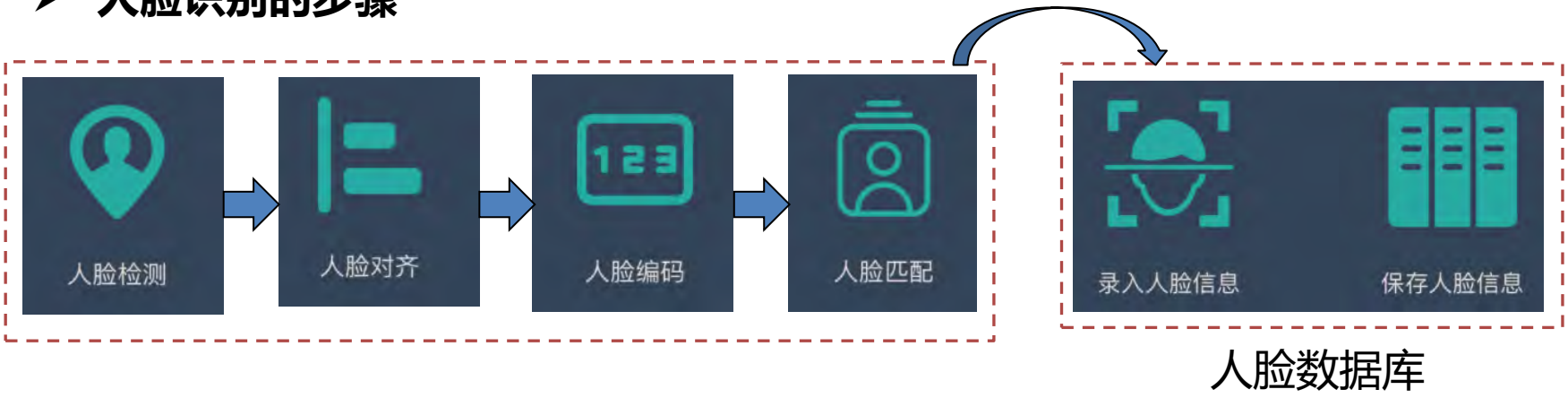
支付系统

自主服务系统（如ATM）

图像分析及应用

- **人脸识别（Facial recognition）**

➢ **人脸识别的步骤**

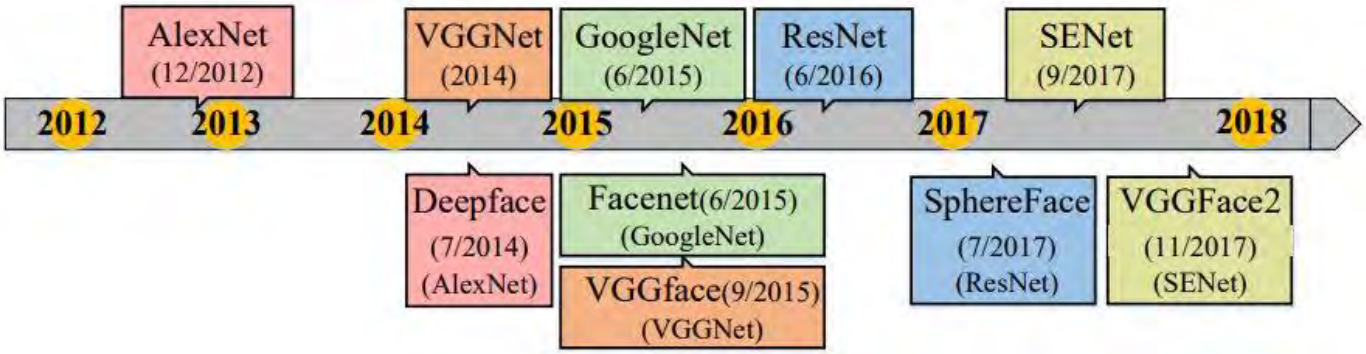


➢ **人脸识别的难点**

图像分析及应用

- **人脸识别的发展阶段**

- 在学术研究上基于深度学习的算法从最开始的 VGG 网络到 GoogLeNet 再到 ResNet，网络模型总体上呈现出更深，更宽的趋势
- 以旷视、商汤等为代表的科技厂商在学术公开竞赛中取得好成绩，开始发展实际业务，通过不断扩大实际数据集，算法性能也在逐渐的提升
- 与第一阶段相反，开始在不降低识别性能的基础上，研究网络的轻量化。轻量化的主要目的有两个：一个是提升算法的速度，甚至能够部署到移动端；另一个就是便于硬件实现，从而将人脸识别算法直接做成一个硬件模块

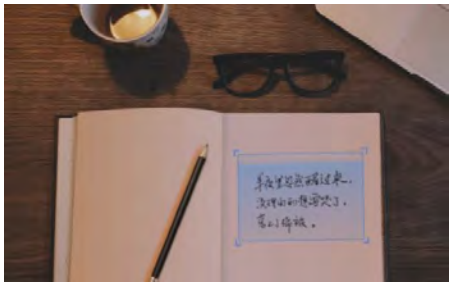


图像分析及应用

- **图像处理分析中的常见任务**
  - 图像分类
  - 目标检测
  - 语义分割
  - 人脸识别
  - **文字识别**
  - 图像生成

图像分析及应用

- **场景文字识别（Scene Text Recognition）**
  - 许多场景图像中包含着丰富的文本信息，对理解图像信息有着重要作用，能够极大地帮助人们认知和理解场景图像的内容
  - **场景文字识别指识别自然场景图片中的文字信息**，即将图像信息转化为文字序列的过程



图像分析及应用

- **场景文字检测和识别的难点**
  - **自然场景中文本多样性和变异性**：文本的颜色、大小、字体、形状、方向、宽高比等属性变化较多
  - **背景的复杂性和干扰**：背景存在与文本相似的形状的物体或存在遮挡问题
  - **不完善的成像条件**：低分辨率、失真、模糊、低/高亮度、阴影等

图像分析及应用

- **场景文字识别的应用**
  - 图像文字检测和识别技术有着广泛的应用场景。已被互联网公司落地的相关应用涉及了**识别名片、识别菜单、识别快递单、识别身份证、识别营业证、识别银行卡、识别车牌、识别路牌、识别商品包装袋、识别会议白板、识别广告主干词、识别试卷、识别单据**等
  - 很多服务商提供图像文字检测和识别服务，包括了腾讯、百度、阿里、微软、亚马逊、谷歌等大型云服务企业。这些企业既可以使用提前训练好的模型直接提供场景图文识别、卡证识别、扫描文档识别等云服务，也可以使用客户提供的数据集训练定制化模型，以及提供定制化AI服务系统集成

图像分析及应用

- **图像生成（Image Generation）**
  - 图像生成是指根据输入向量，生成目标图像。这里的输入向量可以是随机的噪声或用户指定的条件向量。具体的应用场景有：手写体生成、人脸合成、风格迁移、图像修复、超分重建等

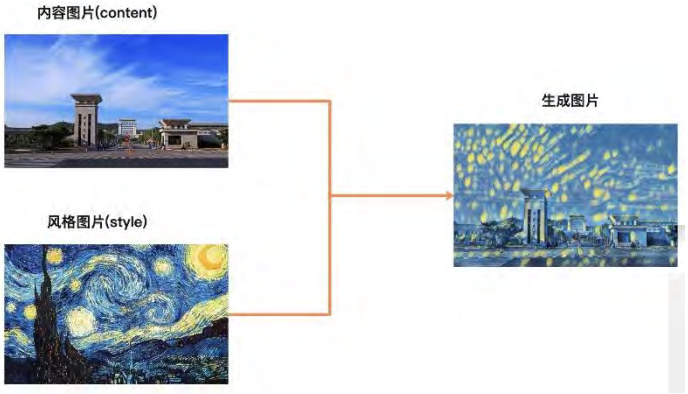
    - 生成手写字等图像数据集，扩充实验数据集

    - 用一张输入图像，合成不同姿势的人脸，作为人脸识别数据

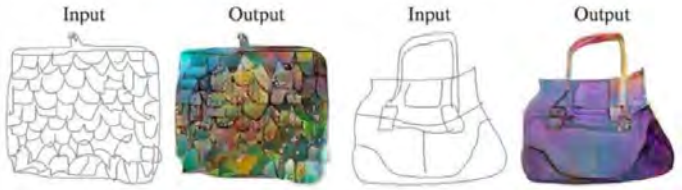
图像分析及应用

- **图像生成（Image Generation）**
  - 图像转换

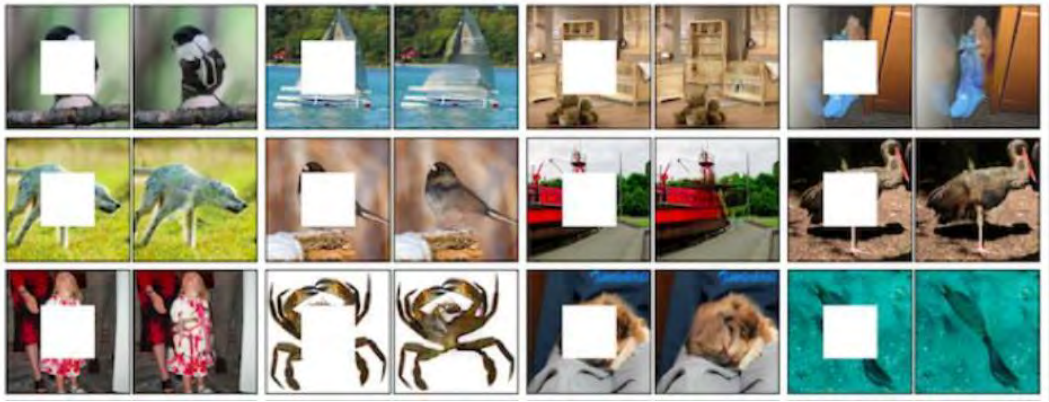
图像风格迁移



图像着色

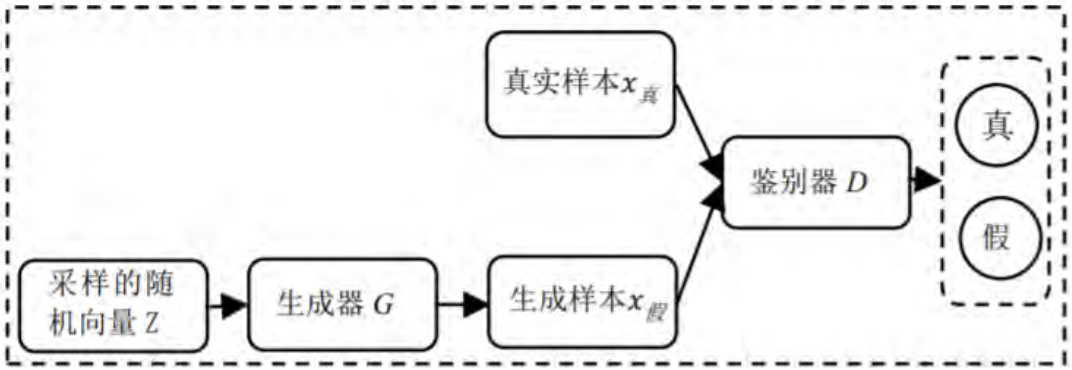


  - 图像修复



图像分析及应用

- 图像生成应用是基于**生成对抗网络**（Generative Adversarial Networks, GAN），GAN是蒙特利尔大学的Ian J. Goodfellow提出
- **GAN网络由两部分组成：生成器和识别器。**生成器的输入是随机噪声或条件向量，输出是目标图像。识别器是一个分类器，输入是一张图像，输出是该图像是否是真实的图像。在训练过程中，生成器和识别器通过不断的相互博弈提升自己的能力
- 基于GAN网络的图像生成是当前热门前沿的方向，在CVPR 2020上GAN的论文超110+篇之多



Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. arXiv preprint arXiv:1406.2661, 2014.

图像分析及应用

- **基于人工智能的图像分析的前沿方向**
  - **算法方向**：
    - 自监督/半监督/弱监督/无监督学习
    - 少样本/单样本/零样本学习
    - 深度模型可解释性
    - 强化学习
    - 迁移学习
    - 模型压缩
  - **应用方向**：
    - 图像生成
    - 医学图像处理
    - 多模态融合

目录

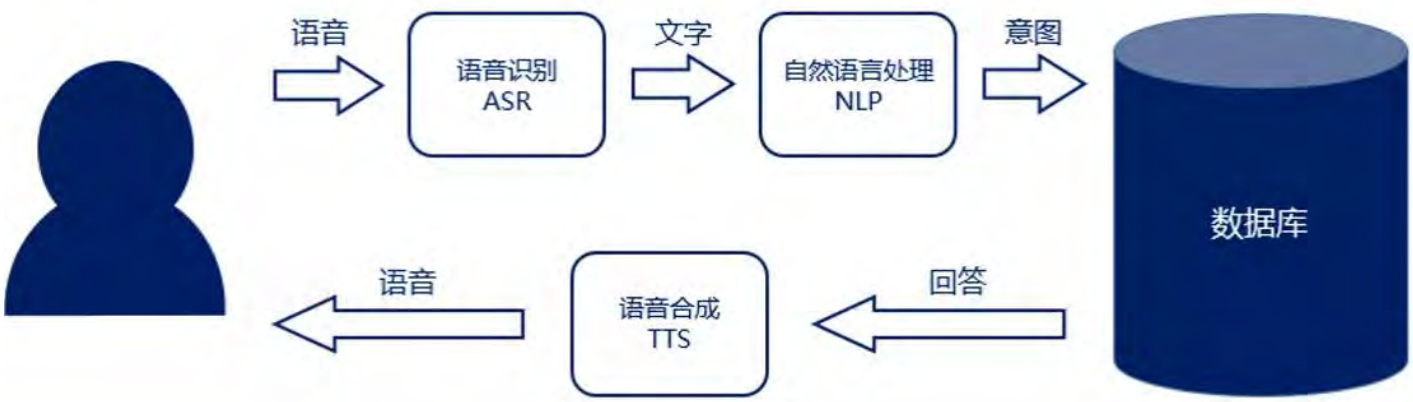
- 多媒体数据
- 图像分析及应用
- 语音分析及应用
- 视频分析及应用
- 多模态融合及应用
- 深度伪造

语音分析及应用

- **语音分析技术**
  - 语音是人类最自然的交互和传递信息的方式
  - 智能语音技术目的是让机器能听懂人类的，并实现人机交互，智能语音技术是人工智能应用中最重要的一部分

语音分析及应用

- **语音分析中的常见任务**
  - 语音识别
  - 说话人识别
  - 语音合成



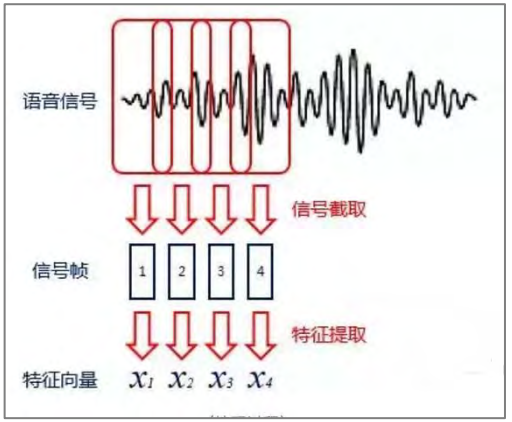
(语音交互流程图)

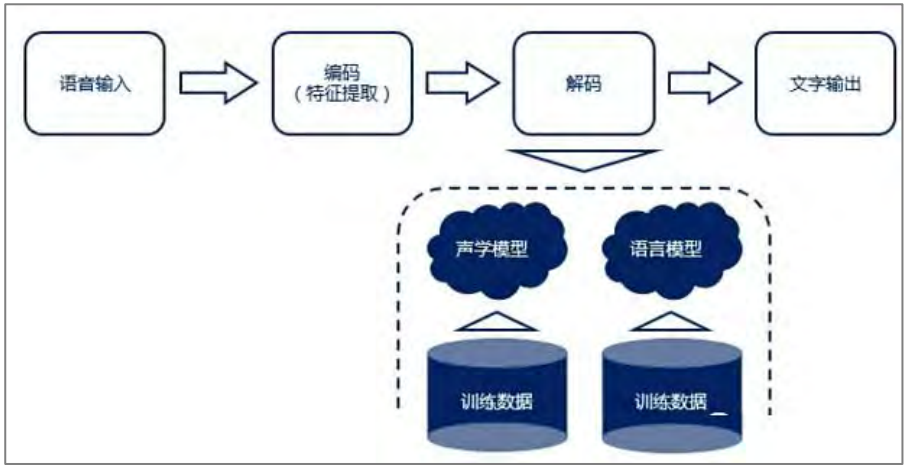
语音分析及应用

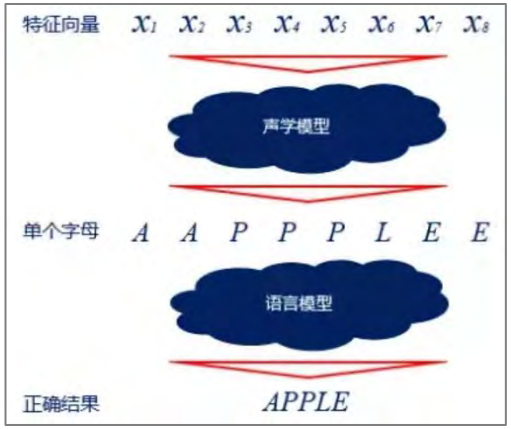
- **语音识别（Automatic Speech Recognition, ASR）**
  - 语音识别是指利用计算机**实现从语音到文字自动转换的任务**
  - 语音识别已经成为了一种很常见的技术，在日常生活中经常会用到：
    - 智能手机的语音助手
    - 微信里的一个功能是”语音转文字”，也利用了语音识别
    - 最近流行的智能音箱就是以语音识别为核心的产品
    - 汽车基本都有语音控制的功能，是语音识别

语音分析及应用

- **语音识别的流程**
  - 语音识别主要包括对输入语音的**编码特征提取**和**解码文本输出**两部分
  - 解码过程是将得到的向量变成文字的过程：
    - **声学模型:** 将特征向量转化成单个字母，即音素
    - **语言模型:** 将音素拼接起来成为单词或者汉字
    - 语言模型包括了N-gram、RNNLM等，声学模型包括HMM、DNN、RNN等模型





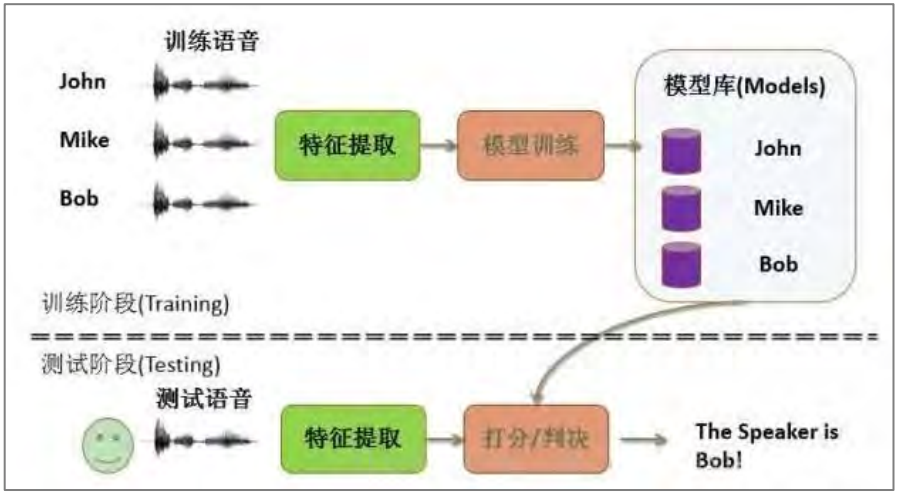


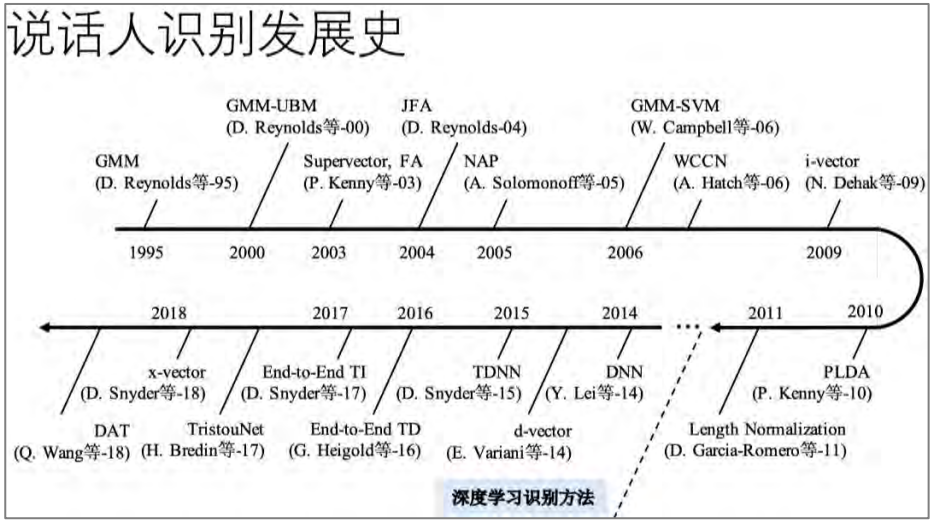
语音分析及应用

- **说话人识别** (Speaker Recognition)
  - 说话人识别或称声纹识别 (Voiceprint Recognition)，是根据语音中所包含的说话人个性信息，利用计算机以及现在的信息识别技术，自动鉴别说话人身份的一种生物特征识别技术。即**从语音中提取具有说话人表征性的特征，建立有效的模型和系统，实现自动精准的说话人鉴别**
  - 说话人识别可分为说话人辨认和说话人确认两种:
    - **说话人辨认**是判断待识别的人为模型用户集合中的哪一个
    - **说话人确认**是确定待识别者是否是所声称的参考者

语音分析及应用

- **说话人识别流程**
  - 一般的声纹识别过程是：
    - **训练阶段**：首先给定训练集用户语音，提取语音特征，再把特征投入模型中训练，得到训练好的声纹识别模型
    - **测试阶段**：给定识别用户的语音，利用模型预测，最后将分数最高或者最接近的用户作为识别结果





语音分析及应用

- **语音合成**(Speech Synthesis)
  - 语音合成也称为文语转换 (Text-to-Speech), 它是将任意的输入文本转换成自然流畅的语音输出
  - 语音合成将文本信息实时转化为近似的真人发声，为应用配上“说话”的能力，满足客户的定制化需求。常见使用场景为：
    - 语音导航
    - 有声读物
    - 智能教育
    - 人机交互

语音分析及应用

- **语音合成的流程**
  - 语音合成可以看作语音识别的逆过程
    - **语音识别**：通过语音波形提取得到声学特征向量，再变为文本特征向量，最后得到文本
    - **语音合成**：通过文本提取得到文本特征向量，再变为声学特征向量，最后反变换得到合成语音波形

目录

- 多媒体数据
- 图像分析及应用
- 语音分析及应用
- 视频分析及应用
- 多模态融合及应用
- 深度伪造

视频分析及应用

- 随着移动互联网、社交网络、安防行业的发展以及数字监控设备的广泛使用，海量的视频数据产生，并已经无法靠人力来完成管理挖掘
- 智能视频分析通过视频算法对视频内容进行挖掘分析，提取视频中有价值的关键信息，形成相应事件和告警的监控方式，并对视频进行有效的管理和检索等应用

监控系统

交通监控视频

安防监控视频

电影视频

新闻视频

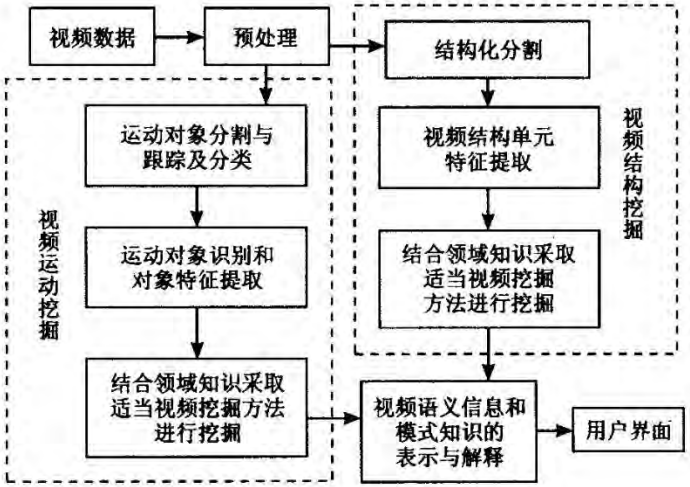
社交网络视频

视频分析及应用

- 视频不仅有图像和视频的**空间结构特征**, 还有**时间特征、视频对象特征、运动特征、音频特征**等内容，通过对视频表达的事物、事件及其特征的理解和总结, 还可以得出视频的语义信息和知识, 从而可以利用这些语义信息和知识辅助问题决策

- **视频挖掘主要有两条技术路线**

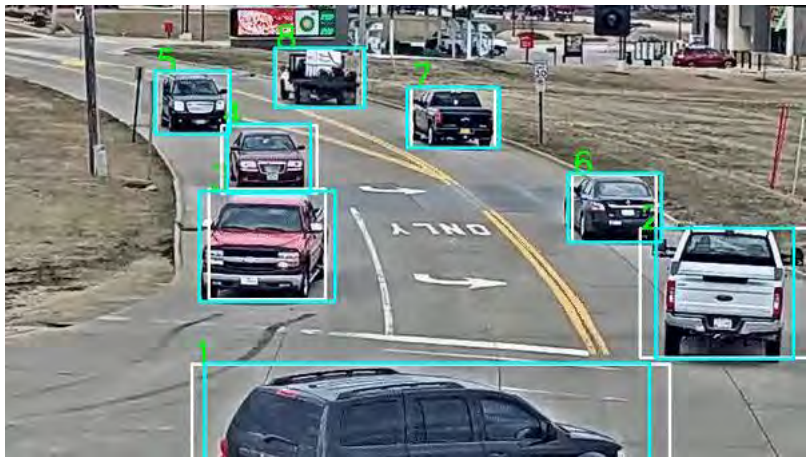
  - **视频结构挖掘**: 以一定的模型或规则将视频划分为视频帧、视频段、场景或镜头组等几个层次结构单元。可挖掘镜头内容随时间变化时, 特征差别体现出的事件变化模式



  - **视频运动挖掘：**从视频中分割并跟踪运动对象, 提取运动对象的运动特征, 得出运动对象特征的含义, 或者运动对象行为趋向和事件模式, 由此挖掘视频表达的高层语义信息

视频分析及应用

- **目标检测**：对视频中的对象进行**定位和识别**，如车辆识别、行人识别等
- **目标跟踪**：在连续的视频序列中，建立所要跟踪物体的位置关系，得到物体完整的运动轨迹。给定图像第一帧的目标坐标位置，计算在下一帧图像中目标的确切位置。
  - 在运动的过程中，目标可能会呈现一些图像上的变化，比如姿态或形状的变化、尺度的变化、背景遮挡或光线亮度的变化等



- 周界入侵检测
- 目标移动方向检测
- 目标消失和出现检测
- 人流量统计和车流量统计

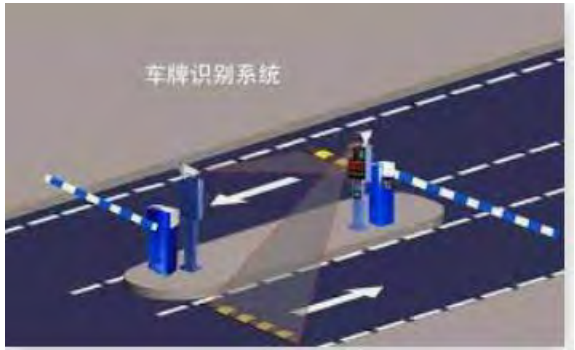
视频分析及应用

- **行为识别**：基于视频的行为识别包括两个问题，**即行为定位和行为识别**
  - **行为定位**：找到有行为的视频片段，与 2D 图像的目标定位任务相似
  - **行为识别**：对该视频片段的行为进行分类识别，与 2D 图像的分类任务相似

- 常见的面部动作(smile，laugh，chew，talk)
- 复杂的面部动作(smoke，eat，drink)
- 常见的肢体动作(climb，dive，jump)
- 复杂的肢体动作(brush hair，catch)
- 多人交互肢体动作(hug，shake hands)

- **视频人脸识别**

- **视频车牌识别**



视频分析及应用

- **视频内容理解：**学习理解视频的具体事件内容等信息
  - **视频语义表示学习**
  - **视频描述和摘要**：用文本描述视频的内容
  - **视频评论**：用文本对视频进行评论
  - **视觉问答**：给定视频和文本问题，输出文本答案

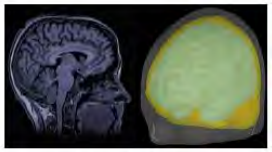
目录

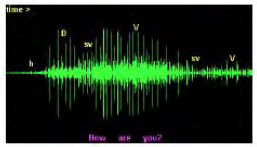
- 多媒体数据
- 图像分析及应用
- 语音分析及应用
- 视频分析及应用
- **多模态融合及应用**
- 深度伪造

多模态融合及应用

- **多模态融合:** 指机器从文本、图像、语音、视频等多个领域获取信息，实现信息转换和融合，从而提升模型性能的技术
- 多模态数据的融合可以为模型决策提供更多的信息，实现多种异质信息的互补，从而提高了决策总体结果的准确率，目的是建立能够处理和关联来自多种模态信息的模型，是典型的多学科交叉领域，并已成为当前的研究热点

Psychology


Medical


Speech


Vision

Language


Multimedia

Robotics


Learning

多模态融合及应用

http://www.cas.cn/syky/202107/t20210708_4797513.shtml?from=singlemessage

多模态融合及应用

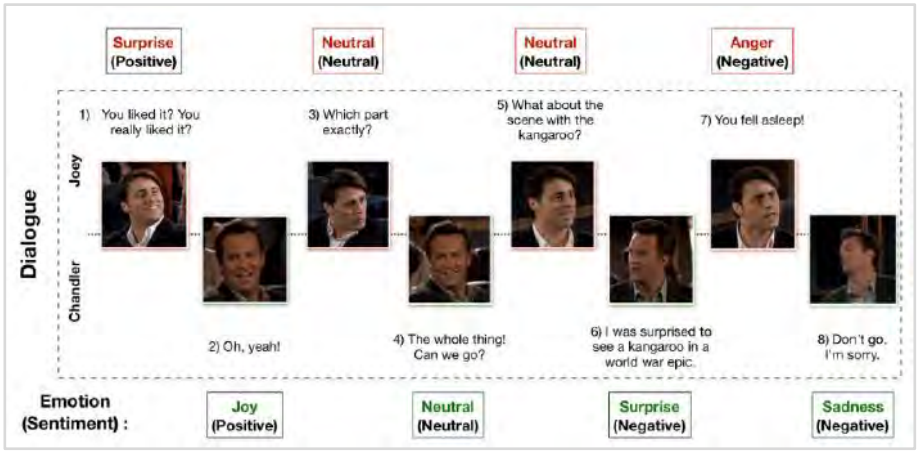
- 多模态融合的常见应用
  - 多模态情感分析
  - 跨媒体检索
  - 多媒体描述
  - 事件检测
  - 语音识别

APPLICATIONS
Speech recognition and synthesis
Audio-visual speech recognition
(Visual) speech synthesis
Event detection
Action classification
Multimedia event detection
Emotion and affect
Recognition
Synthesis
Media description
Image description
Video description
Visual question-answering
Media summarization
Multimedia retrieval
Cross modal retrieval
Cross modal hashing

多模态融合及应用

- **多模态情感分析（Multimodal Sentiment Analysis）**
  - 随着社交网络的快速发展，人们在平台上的表达方式变得越来越丰富，如通过图文和视频表达自己的情绪和观点。
  - 多模态数据与单模态数据相比，包含了更多的信息，多个模态之间可以互相补充，帮助机器更好更准确地理解人的真实情感

◆ **文本+图像**

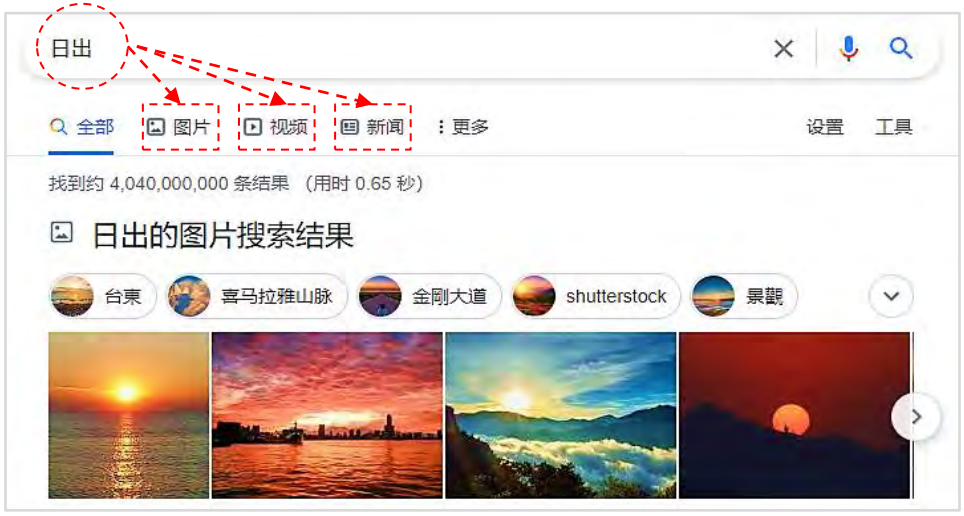


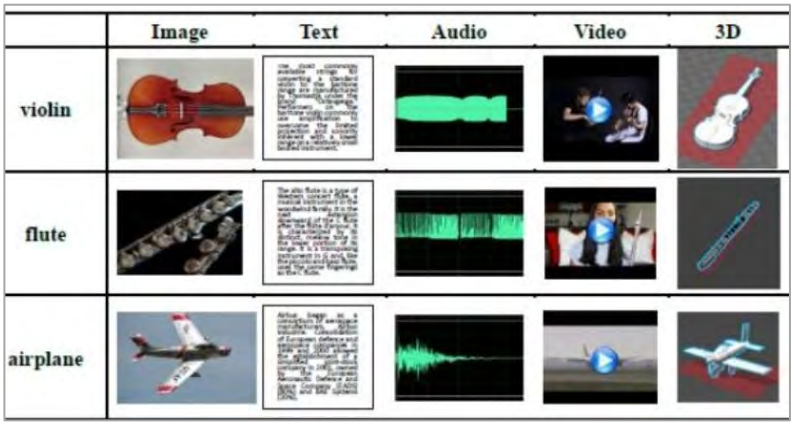
◆ **文本+图像+音频**

- 视频评论
- 新闻视频
- 对话视频

多模态融合及应用

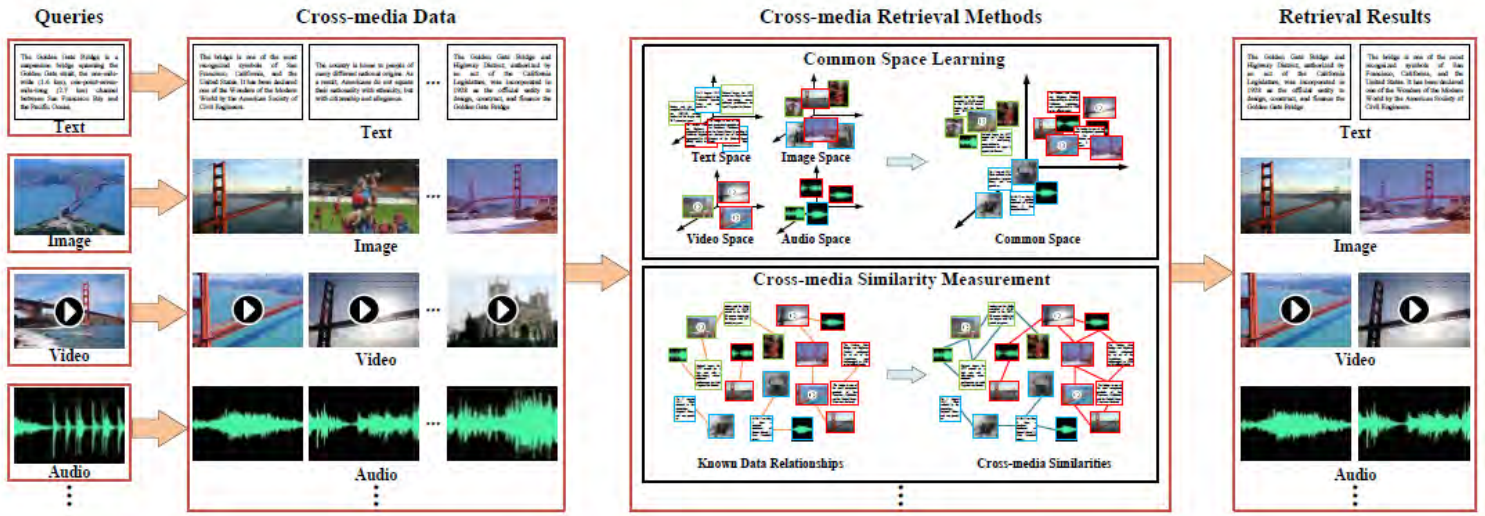
- **跨媒体检索（Cross-media Retrieval）**
  - 跨媒体检索的目标是计算不同媒体数据间的相似度，对于给定的查询样例，检索出与查询样例相关的不同媒体数据
  - 跨媒体检索中多媒体类型主要包括图像、文本、语音、视频。大多数跨媒体检索方法主要是限制在用图像和文本，还有少量语音和视频等





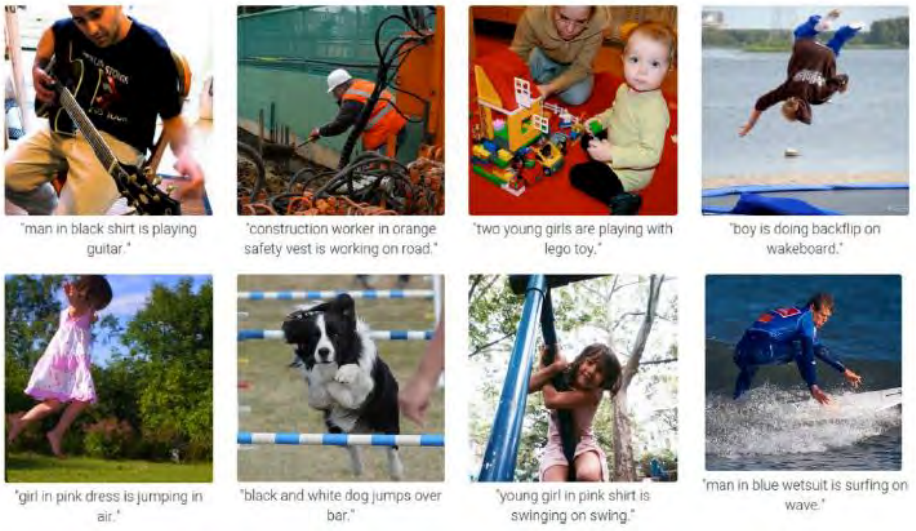
多模态融合及应用

- **跨媒体检索（Cross-media Retrieval）**
  - 跨媒体检索关键挑战在于**不同媒体的表示形式不一致，难以进行直接的相似性度量，即“媒体鸿沟”问题**
  - 两种主要的跨媒体检索方法：
    - **共同空间学习法**：为不同媒体数据学习一个统一的共同空间
    - **跨媒体相似性度量法**：不学习共同空间，而是直接计算跨媒体的相似性



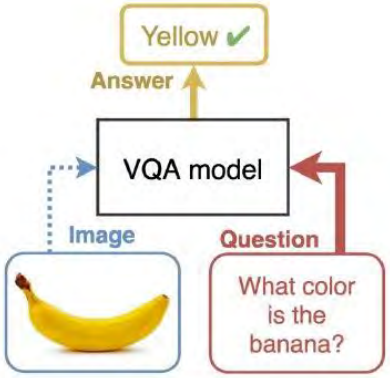
多模态融合及应用

- **多媒体描述（Multimedia Description）**
  - 给定某一个多媒体数据，比如图像、视频、音频等，根据多媒体信息为其生成一个简单的文本描述
  - 比如图像描述生成，即常说的“看图说话”，根据图片情景写出对应的描述语句，已经成功应用到图文搜索，视力受损人士的生活辅助及儿童教育等方面



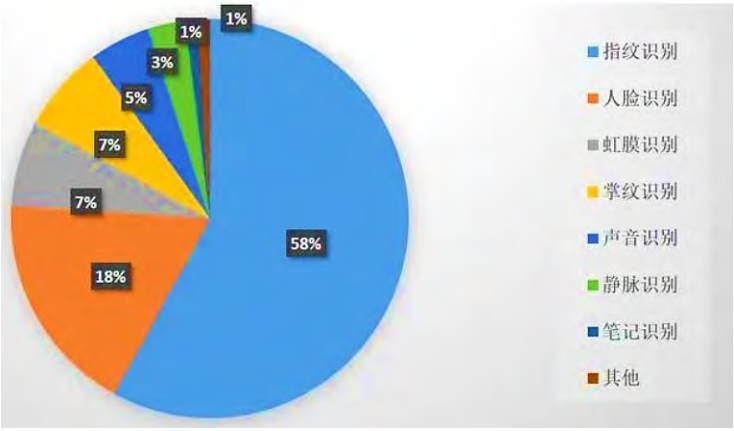
多模态融合及应用

- **视觉问答（Visual Question Answering）**
  - 视觉问答是一个比较新的领域，给定一张图片和一个关于图片的自然语言问题，以生成一条自然语言答案作为输出
  - 视觉问答可分为三个步骤：
    - 1) 从图像中提取特征；
    - 2) 从问题文本中提取特征；
    - 3) 结合图像和文本特征来生成答案，关键在如何把文本和图像结合起来



多模态融合及应用

- **多模态生物识别**
  - 指纹、人脸、虹膜、声纹等生物识别技术不断发站，但单模态的识别无论在识别性能还是安全性上存在瓶颈
  - 不同模态的生物特征具备不同的特性和分辨能力，每融合一个新模态的生物特征都能使得系统的识别能力和安全能力上一个等级

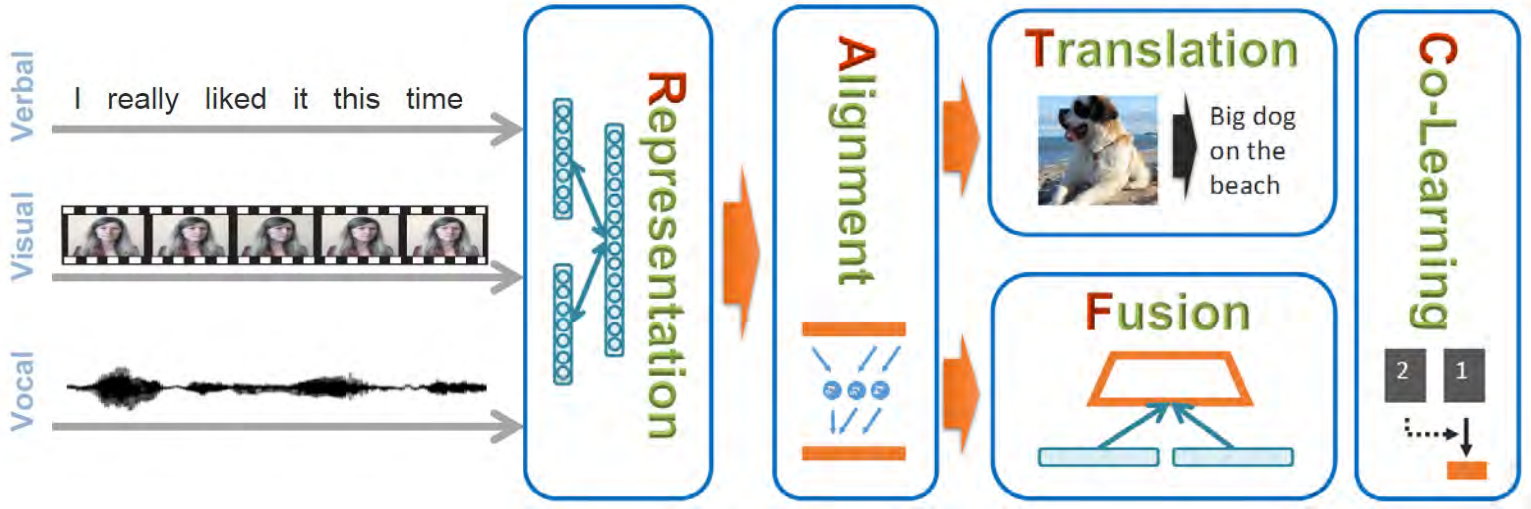


多模态融合及应用

- **人机交互-汽车智能助手**
  - 智能汽车正在从原本单一的车载语音识别，实现融合视觉、语音、车内外场景图像的多模态识别的转变
  - 在实际的语音交互中，车载智能助手不仅可以实现语音的识别，也可以通过摄像头识别人的表情神态、动作，比如识别疲劳驾驶、分心、发热等状况，以进行即时的语音提醒，语音交互也可以更加以人类的自然语言进行交互

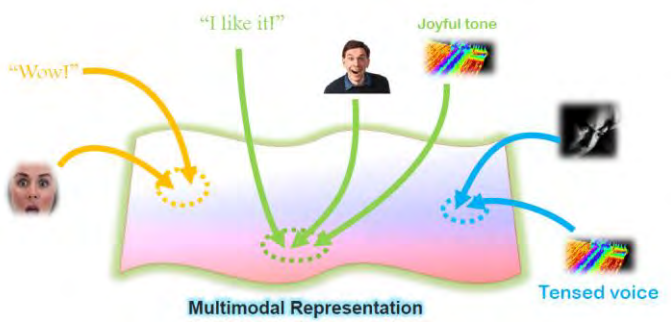
多模态融合及应用

- **多模态学习可以划分为以下四个研究方向：**
  - 多模态表示 (Multimodal Representation)
  - 模态转化 (Translation)
  - 对齐 (Alignment)
  - 多模态融合 (Multimodal Fusion)



多模态融合及应用

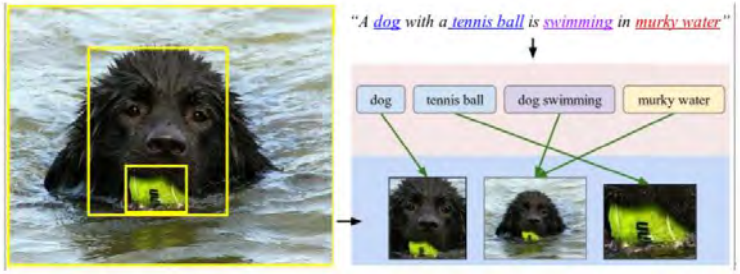
- **多模态表示 (Multimodal Representation)**
  - 多模态表示学习是指通过利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示



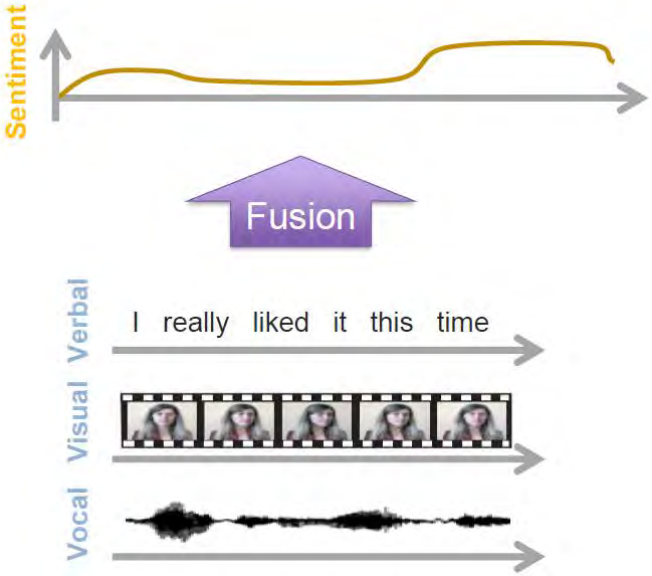
- **模态转化 (Translation)**
  - 将一个模态的信息转换为另一个模态的信息
  - 常见的应用包括：机器翻译、唇读、语音翻译、图片描述、视频描述、语音合成

多模态融合及应用

- **对齐 (Alignment)**
  - 多模态的对齐负责对来自同一个实例的不同模态信息的子分支/元素寻找对应关系。
  - 常见的应用包括：图像语义分割、文本图像对齐、电影的视频-音频-文本字幕对齐



- **多模态融合 (Multimodal Fusion)**
  - 多模态融合负责联合多个模态的信息，缩小模态间的异质性差异，同时保持各模态特定语义的完整性，获得更好的特征表示进行目标预测，是目前应用和研究最广的方向



目录

- 多媒体数据
- 图像分析及应用
- 语音分析及应用
- 视频分析及应用
- 多模态融合及应用
- 深度伪造

深度伪造

https://www.bilibili.com/video/av499530276/

深度伪造

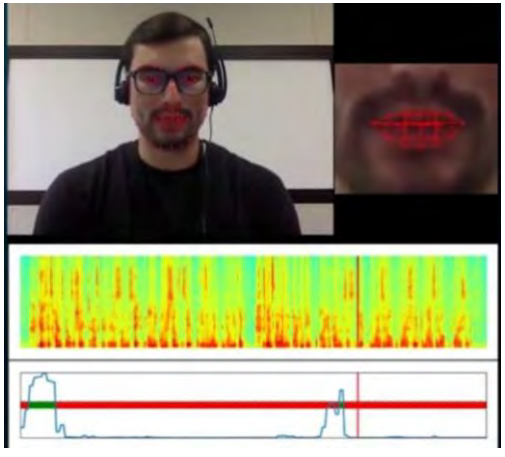
https://www.bilibili.com/video/BV1WK4y1Q7ew

深度伪造

- **深度伪造 (Deepfake)**
  - 深度伪造 (Deepfake) 是英文“deep learning”(深度学习) 和“fake”(伪造)的混合词，**利用深度学习算法来创建或合成图像、音视频、文本等视听觉内容**
  - 深度伪造技术可用于误导舆论、扰乱社会秩序, 甚至可能会威胁人脸识别安全系统、干预政府选举和颠覆国家政权, 已成为当前最先进的新型网络攻击形式
  - 深度伪造能“以假乱真”，能实现换脸、模仿真人语言和动作，创造出不存在的人物及活动,正在挑战人们“眼见为实”的传统认知

深度视觉伪造

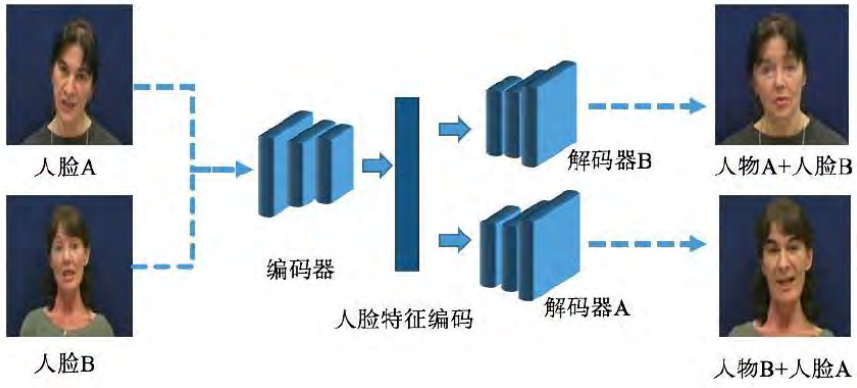
深度音频伪造



深度文本伪造

深度伪造

- **视觉伪造**：主要是针对图像或视频，是当前最主要的深度伪造技术，视觉深度伪造可划分为重现、替换、编辑、合成四类，其中重现和替换是最大的隐患，它们可以让攻击者控制身份和欺骗
  - **替换**：就是常见的AI换脸，指用另一个人脸来替换一张图片或视频中的一个人脸，是最常见的伪造技术
    - 视频换脸常用于恶搞抹黑政治人物及公众人物，造成了恶劣的负面影响，同时换脸还威胁着人脸识别等系统安全



深度伪造

- **编辑**：编辑是指添加、更改或删除目标身份的某种属性，比如更换发型、衣服、胡须、年龄、体重、颜值、眼镜和种族等属性

- **重现**：使用源身份Xs驱动目标身份Xt，使Xt的行为和Xs一样，包括表情模仿、嘴部、眼部、头部及人体姿势迁移。攻击者能假冒他人说或做，比如诽谤、散布错误信息、篡改证据、骗取信任诈骗等

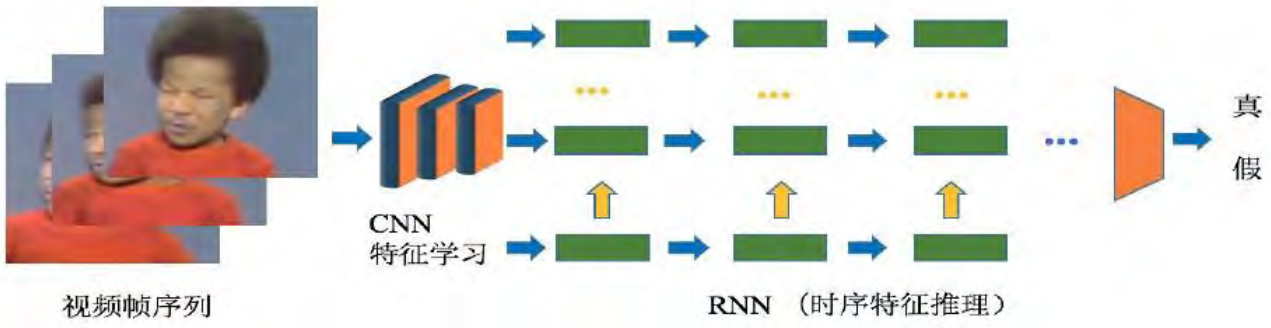
- **合成**：指在没有目标身份作为基础的情况下创建一个新的不存在的deepfake角色，类似直接用GAN或者其它生成模型生成人脸。攻击者可以在线创建虚假角色进行诈骗等违法活动

深度伪造

- **音频伪造**：即语音克隆技术，每个人的语速、语调等声道信息都可以表示成一个声音模型，通过模型模仿目标人的身份
  - 案例-语音诈骗：2019年9月英国一家能源公司发生了一个刑事案件。犯罪分子使用电话会议，YouTube，社交媒体甚至是TED演讲中获得的音频训练了模型，以复制公司老板的声音，并欺骗其下属将数十万美元汇入一个秘密帐户。
- **文本伪造**：通过伪造修该文件或图像视频中的文本内容来进行篡改信息、诈骗等活动
  - Facebook在2021年推出了TextStyleBrush研究项目，新AI工具可以复制和再现图像中的文本样式，实现图像文本内容的修改

深度伪造

- 深度伪造内容具有**辨别难度大、制造成本低、传播速度快和破坏能力强**等特点，对个人隐私数据，社会稳定甚至国家安全等造成严重的潜在威胁,所以亟需提出切实有效的深度伪造内容检测方法来应对深度伪造内容带来的严峻挑战
- 现有的深度伪造内容检测方法多依赖于深度学习模型，基于深度伪造内容数据集的训练，实现特征提取并**构建伪造检测分类器**
  - 深度伪造图像检测技术
  - 深度伪造视频检测技术
  - 深度伪造语音检测技术



Thank You!

Tel: 86-10-82546890 Fax: 86-10-82546891
E-mail: general@iie.ac.cn
地址：北京市海淀区闵庄路甲89号 100093