南京大学
NANJING UNIVERSITY

人工智能导论

Applications

郭兰哲
南京大学 智能科学与技术学院
Homepage: www.lamda.nju.edu.cn/guolz
Email: guolz@nju.edu.cn

答辩顺序

1 陈仲石

2 朱睿

3 陆金仪

4 吴屿伦

5 袁宇枫

6 花鹏宇

7 徐云哲

8 蔡嘉豪

9 庞景豪

10 陈薛含

11 谭睿睿

12 何志烨

13 陈博闻

14 温慧林

15 周子涵

16 梁昊东

17 陈子昂

18 郎丛宁

19 马新凯

20 赵哲

21 倪建宇

22 王萌

23 钱子涵

24 苏珉乐

25 徐乐怡

26 杨锐然

27 顾浩译

28 钱禹昂

29 王未

30 柳凯

31 康靖

32 杨可

33 黄旭鸿

大纲

- 计算机视觉
- 自然语言处理
- 多模态AI
- 机器人学

计算机视觉(Computer Vision)

什么是计算机视觉

计算机视觉有哪些任务

人工智能技术在计算机视觉中的应用

计算机视觉(Computer Vision)

- 计算机视觉是一门研究如何对图像或者视频进行高层理解的交叉学科

人工智能视角：赋予机器“看”的智能，即，用机器自动实现人类视觉系统的功能，图像或视频的获取、处理、分析和理解等诸多任务

计算机视觉(Computer Vision)

计算机视觉(Computer Vision)

计算机视觉(Computer Vision)

图像分类(Image Classification)

- 图像分类：将图片划分到某个特定的类别中

目标检测(Object Detection)

- 目标检测：检测图像中某个类别的物体的任务，一般来说，就是在图像中寻找我们感兴趣的目标，并给出其位置

语义分割(Semantic Segmentation)

- 语义分割：关注如何将图像分割成不同类别的区域，与目标检测不同，语义分割可以识别并理解图像中每个像素的内容，其语义区域的标注和预测是像素级的

实例分割(Instance Segmentation)

- 实例分割：检测和划定出图像中出现的每个不同的感兴趣的对象，研究如何识别图像中各个目标实例的像素级区域。与语义分割不同，实例分割不仅需要区分语义，还要区分不同的目标实例

语义分割

实例分割

视频分类(Video Classification)

- 视频分类是给定一个视频，为其划分到指定的类别中

目标跟踪(Object Tracking)

- 目标跟踪：根据视频序列第一帧中目标的初始状态（中心位置和比例），自动获得该物体在后续视频帧中的状态

图像描述(Image Caption)

图像生成(Image Generation)

风格迁移(Style Transfer)

计算机视觉概述

- 20世纪50年代：研究生物视觉工作原理

D.H.Hubel

T.N.Wiesel

- Hubel和Wiesel的一些研究成果。他们从生理学的角度来分析猫（据说猫和人类的大脑比较相近）的视觉皮层系统，发现视觉通路中的信息分层处理机制，并提出了感受野的概念，也因此获得了诺贝尔生理学和医学奖

计算机视觉概述

- 20世纪60年代：计算机视觉开始发展

- 第一位计算机视觉博士，Larry Roberts，在1963发表的论文"Machine perception of three-dimensional solids”中，将物体简化为几何形状（立方体、棱柱体等）来加以识别

(a) Original picture (b) Differentiated picture (c) Feature points selected

20世纪60年代早期，在Marvin Minsky领导下MIT的人工智能实验室成立，同一时期John McCarthy在斯坦福也建立人工智能实验室。1966年MIT人工智能实验室试图解决计算机视觉领域的理论问题，虽然实际中结果没有达到预期，但从此计算机视觉领域开始蓬勃发展。

计算机视觉概述

- 20世纪70-80年代

- MIT的心理学教授David Marr在计算机视觉理论方面做出了大量贡献，融合了心理学、神经生理学、数学等多门学科，提出了有别于前人的计算机视觉分析理论，影响了计算机视觉领域的未来二十年的发展。

计算机视觉概述

- 20世纪70-80年代

- MIT的心理学教授David Marr在计算机视觉理论方面做出了大量贡献，融合了心理学、神经生理学、数学等多门学科，提出了有别于前人的计算机视觉分析理论，影响了计算机视觉领域的未来二十年的发展。

1982年

视觉任务的挑战

- 数字图像由一个个像素(pixel)组成，每个像素的亮度、颜色等属性在计算机内表示为一个或多个数字
- 如果是黑白图像(灰度图像)，每个像素由一个字节表示，0-255的数值表示亮度，如果是彩色图像，通常用RGB三个字节表示

视觉任务的挑战

Illumination

Object pose

Clutter

Occlusions

Intra-class appearance

Viewpoint

视觉任务的挑战：intra-class variation

视觉任务的挑战

- 数据维度高
  - 一张高清图像包含数百万像素
- 数据量庞大
  - 几十亿可获得的图像
  - 每天产生的视频时长
- 涉及类别多
- …

计算机视觉的关键技术

- 马尔视觉理论的核心是从图像或视频中重建物体的“几何”形状模型，围绕此目标，研究人员在此后数十年开展了大量工作

- 到2000年前后，多视几何(Multiple View Geometry)、摄像机内外参数标定、分层三维重建等方法和技术得到长足发展，形成了计算机视觉的“几何”分支，并延续至今，近年来在虚拟现实、增强现实等领域得到了广泛应用

- 另一分支便是基于机器学习，通过将计算机视觉中的问题，定义为从输入图像/视频直接到求解标签的函数拟合问题，采用数据驱动的方法来求解待拟合的函数

机器学习解决视觉任务

X

$\rightarrow$ $f_\theta(X)$ $\rightarrow$ y

- 类别标签(如图像分类、物体识别等)
- 连续变量/向量/图像（如目标检测、语义分割）

基于浅层模型的方法

- 步骤1：图像预处理：用于实现目标对齐、几何归一化、亮度或颜色校正等处理
- 步骤2：特征设计与提取：从预处理的图像中提取描述图像内容的特征
- 步骤3：特征汇聚或特征变换：对提取的特征进行统计或降维处理，从而得到更利于后续分类或回归任务的特征
- 步骤4：分类器或回归器的设计与训练：采用机器学习算法训练学习器，决策树、SVM、LR等

特征设计与提取方法

- 人工设计特征本质上是一种专家知识驱动的方法，即研究者通过自身对所研究问题或目标的理解，设计某种流程来提取觉得“好”的特征

- 例如：人类识别早期，研究人员普遍认为应用面部关键特征点的相对距离、角度或器官面积等就可以区分人脸，但后来证明这些特征并不好

- 目前，多数人工设计的特征有两类，即全局特征和局部特征
  - 全局特征：建模图像中全部像素或多个不同区域像素蕴含的信息
  - 局部特征：从一个局部区域内的少量像素中提取信息

局部二值模式(LBP特征)

- LBP算法是一种简单有效地编码图像局部区域内变化模式的局部描述算子，只关注中心像素与其邻域像素的明暗关系

LBP用于人脸识别

- 给定一副人脸图像，假设是128*160像素，可以将其划分为20个32*32的子图像
- 每个子图像，有900个像素可以作为3*3邻域的像素中心计算LBP模式值
- 统计每个模式出现的次数，得到一个多维的直方图
- 将所有的直方图串接得到整个人脸图像的描述特征

各种算子

Roberts算子

$$S_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, S_y = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Prewitt算子

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Sobel算子

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Canny算子

Laplacian算子

…

各种算子

原始图片

Roberts算子

Prewitt算子

Sobel算子

Canny算子

Laplacian算子

基于深度学习的视觉方法

2012年, Hinton 组参加ImageNet 竞赛, 使用 CNN 模型以超过第二名10个百分点的成绩夺得当年竞赛的冠军

基于深度学习的视觉方法

基于深度卷积神经网络的图像分类与识别

基于深度卷积神经网络的图像分类与识别

卷积、池化操作实现了特征提取的功能，这种特征提取不是人为设定的，而是自动学习得到的

Image

Convolved Feature

卷积层

Convolved feature

Pooled feature

池化层

基于深度卷积神经网络的图像分类与识别

基于深度学习的视觉方法

从“特征工程” 到“特征学习”或“表示学习”

- 特征工程由人类专家根据现实任务来设计，特征提取与识别是分开的两个阶段

手工设计特征 → 分类识别

- 特征学习通过深度学习自动产生有益于分类的特征，是一个端到端的学习框架.

学习生成特征 → 分类识别

基于深度模型的目标检测技术

R-CNN最早将卷积神经网络应用在目标检测中卷积，一般包括以下步骤：

- 候选区域生成： 一张图像生成1K~2K个候选区域
- 特征提取： 对每个候选区域，使用深度卷积网络提取特征
- 类别判断： 特征送入每一类的SVM 分类器，判别是否属于该类

实例：CIFAR-10图像分类

- https://www.kaggle.com/c/cifar-10

CIFAR-10 - Object Recognition in Images

Identify the subject of 60,000 labeled images

231 teams · 4 years ago

Overview Data Discussion Leaderboard Rules

Overview

Description

Evaluation

CIFAR-10 is an established computer-vision dataset used for object recognition. It is a subset of the 80 million tiny images dataset and consists of 60,000 32x32 color images containing one of 10 object classes, with 6000 images per class. It was collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

实例： ImageNet狗品种识别

- https://www.kaggle.com/c/dog-breed-identification

扩展阅读

- 《动手学深度学习》： https://zh.d2l.ai/
- Fei-Fei Li’s Lecture： http://cs231n.stanford.edu/

- 常用数据集：
- 图像分类：ImageNet、 CIFAR、MNIST
- 目标检测：COCO数据集
- 语义分割： Pascal VOC2012
- 视频数据集：Kinetics-700
- …

自然语言处理(Natural Language Processing)

什么是自然语言处理

自然语言处理有哪些任务

如何用人工智能技术解决自然语言处理任务

自然语言处理(Natural Language Processing)

- 自然语言：人类语言，区别于人工语言(比如程序语言)
- 自然语言处理：对词、句子、篇章进行分析，对内容进行理解，并在此基础上支持一系列技术，如翻译、问答等。

自然语言处理示例

这是一棵语法树
分词
这 是 一 棵 语法 树
词性标注
这 是 一 棵 语法 树
代词 动词 数词 量词 名词 动词
句法分析
这 是 一 棵 语法 树
代词 动词 数词 量词 名词 动词
语义分析
{ 这, 是, 语法树 }
知识库
应用
语义分析
机器翻译
自动问答
情感分析
… …

自然语言处理的难点：歧义性

- 组合歧义：
  - 两个/人/一起/过去、个人/问题
  - 从马/上/下来、马上/就来
  - 南京市长江大桥

- 句子级歧义：
  - 白天鹅在水里游泳
  - 该研究所获得的成果

自然语言处理的难点：歧义性

灵魂八问

配钥匙师傅：你配吗？

食堂阿姨：你要饭吗？

算命先生：你算什么东西？

快递小哥：你是什么东西？

上海垃圾分拣阿姨：你是什么垃圾？

滴滴司机：你搞清楚你自己的定位了么？

理发师傅：你自己照照镜子看看你自己，觉得还行么？

小区保安：你是谁？你从哪里来？要到哪里去？

自然语言处理的典型应用

- 智能问答系统
  - 通过中文分词、短文本相似度、命名实体识别等相关技术计算两个问题对的相似度，可解决问答、对话、语料挖掘、知识库构建等问题
- 文本分析
  - 通过关键词提取、文本聚类、主题挖掘等算法模型，挖掘突发事件、公众话题导向，进行话题发现、趋势发现等。多维度分析公众情绪、热点、趋势、传播途径等，及时全面的掌握话题动态
- 内容推荐
  - 通过关键词提取、短文本相似度等技术，提取关键语义信息，准确匹配出语义相似的内容，从而快速构建内容推荐场景
- 机器翻译
  - 通过文本语言分析，准确翻译语句内容，从而帮助用户跨语言沟通

自然语言处理的发展历程

数据驱动

规则驱动

2012年~

基于深度学习的方法

- 神经网络
- 分布式表示
- 非线性模型

1990年~2012年

基于统计学习的方法

- 离散表示
- 线性模型

1990年以前

基于规则的方法

基于规则的翻译

资源：规则库

1: If 源 = 我, **then** 译 = I

2: If 源 = 你, **then** 译 = you

3: If 源 = 感到 满意,
then 译 = be satisfied with

4: If 源 = 对... 动词 [表态度]
then 调序 [动词 + 对象]

5: If 译文主语是 I
then be 动词为 am/was

6: If 源语是主谓结构
then 译文为主谓结构

数据驱动的翻译

- 翻译模型从双语平行语料中学习翻译知识，得到各种单词的翻译及其概率
- 语言模型从单语语料中学习目标语言的词序列生成规律，来衡量目标语言译文的流畅性

基于神经网络的翻译

目标语言: I am satisfied with you
解码器（Decoder）
分布式
表示
.2 -1 6 5 .7 -2
编码器（Encoder）
源语言: 我 对 你 感到 满意
表示结果 （Representation）
输入 （Word）
输入 （Word）
编码器网络示例

将会改变世界的对话式大型语言模型ChatGPT

- 2022年11月30日，OpenAI发布了AI对话模型，被认为是人工智能里程碑式的应用

- 仅用两个月时间，月活跃用户已达1亿，是史上用户增速最快的消费应用

像ChatGPT这样的AI聊天机器人将变得与个人电脑或互联网同样重要

OpenAI发布AI对话模型ChatGPT，开启生成式AI商业化新机遇 ...

2022年12月19日 — 当地时间11月30日，美国人工智能公司OpenAI发布全新产品ChatGPT，一款基于GPT-3.5的免费对话模型。公司CEOSamAltman透露上线五天该模型的全球用户数量 ...

AI新物种：ChatGpt 不会止于写代码调Bug - 巴比特

2022年12月11日 — GPT-3发布于2020年，作为一个自监督模型，几乎可以完成自然语言处理的绝大部分任务，在参数上，GPT-1包含了1.17亿个参数，GPT-2包含了15亿个参数，而GPT-3 ...

ChatGPT 通过了美国MBA、法律和医学考试 - Showmetech

2023年1月26日 — 这次的新颖之处在于，根据在美国进行的一项研究， OpenAI 会轻松通过即使是学生也难以通过的复杂测试。 重点是创建包含所有重要细节的法律文件和聊天GPT ...

微软100亿美元砸向OpenAI，ChatGPT要加入Office全家桶了？

2023年1月12日 — 如果100 亿美元的交易成真，OpenAI 将获得巨额资金，微软赢得广阔未来，双赢局面就此达成。 过去一段时间，对话式AI 模型ChatGPT 火遍了整个社区，它 ...

GPT-4：人工智能的新语言方法被定义为“强大” - Showmetech

2023年1月2日 — GPT-2024 计划于4 年发布，应该会为ChatGPT 带来更好的理解和文本创建。... 不浪费时间，该公司已经在准备下一代AI 语言方法，预计将于2023 年推出， ...

美国大学89%的学生居然用ChatGPT写作业 - 国际竞赛

1天前 — ChatGPT的崛起并在高等教育领域的突然普及，让众多美国高校感觉措手不及！ ... Nature早就很有先见之明地发文，担心ChatGPT会成为学生写论文的工具.

ChatGPT

- 用户的输入和模型的输出都是文字形式，一次用户输入和一次模型对应的输出，叫做一轮对话，我们可以把 ChatGPT 的模型抽象成如下流程：

- 多轮对话

ChatGPT

- ChatGPT 在接收到用户的提问输入后，输出的文字并不是一口气直接生成的，而是一个字、一个字生成的

ChatGPT 的神经网络结构 Transformer

Transformer 的核心是自注意力机制（Self-Attention），它可以帮助模型在处理输入的文字序列时，自动地关注到与当前位置字符相关的其他位置字符

Vaswani, et al., Attention is All you Need. NeurIPS 2017

GPT-1.0(GPT=Generative Pre-trained Transformer)

不完整的句子

“花谢花飞花满__”

GPT-1.0(GPT=Generative Pre-trained Transformer)

- GPT 初代所做的事：根据上文补全下文
  - 从大规模的文本语料中，将每一条文本随机地分成两部分，只保留上半部分，让模型学习下半部分学习到底该填写什么

GPT-2

- 基本想法：给模型出其它的语言任务，应该也会对模型训练有极大的帮助

- **句子打乱顺序再排序、选择题、判断题、问答题、寻找文中的错别字、把预测单字改成预测实体词汇**等等，都可以制定数据集添加在模型的预训练里

- GPT-2在GPT初代的基础上，又添加了多个任务，比如机器翻译、问答、文本摘要等等，扩增了数据集和模型参数，又训练了一番

GPT-2

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%

GPT-3

大数据+大参数量+大计算力

模型名	参数量	训练数据
Original GPT (GPT-1)	1.17 亿	4.5 GB文本
GPT-2	15 亿	40 GB文本
GPT-3	1750亿	570 GB文本

ChatGPT

- 有监督微调

ChatGPT

- 强化学习

大模型幻觉(Hallucination)

You

请介绍一下林黛玉倒拔垂杨柳的故事

ChatGPT

林黛玉倒拔垂杨柳是《红楼梦》中的一个描写，发生在小说的第五十六回。这一场景表现了林黛玉情感的纠葛和心境的复杂。

故事发生在贾府中，林黛玉因为家族事务和个人感情的纷扰而感到烦闷和心情沉重。在贾府的后花园有一片垂杨柳，林黛玉在心情沮丧的情况下，倒拔起这片垂杨柳。这个举动被认为是林黛玉情感的宣泄，同时也反映了她坚韧、叛逆的一面。

垂杨柳在中国文学中通常被赋予柔弱、婉约的象征意义，而林黛玉倒拔垂杨柳的举动则被解读为她个性的张扬和对命运的反抗。整个场景揭示了林黛玉内心的矛盾、痛苦与坚持，是《红楼梦》中一个富有象征意味的描写。这一情节也成为后来研究《红楼梦》时常被引用和评论的经典场景之一。

扩展阅读

- CS224d: Deep Learning for Natural Language Processing
- GPT: Improving Language Understanding by Generative Pre-Training
- GPT-2: Language Models are Unsupervised Multitask Learners
- GPT3: Language Models are Few-Shot Learners

多模态学习(Multi-Modal Learning)

多模态学习(Multi-Modal)

多模态(Multi-Modal)

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

mixing sparkling chemicals as mad scientists shopping for groceries working on new AI research

as kids' crayon art on the moon in the 1980s underwater with 1990s technology

→

DALL·E 2

多模态(Multi-Modal)

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

mixing sparkling chemicals as mad scientists shopping for groceries working on new AI research

as a 1990s Saturday morning cartoon as digital art in a steampunk style

→

DALL·E 2

多模态(Multi-Modal)

This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

机器人学(Robot)

- 1920年，捷克斯洛伐克作家Karel Capek发表了科幻剧本《罗萨姆的万能机器人》，将捷克语“Robota”写成了“Robot”，“Robota”是奴隶的意思，该剧预告了机器人的发展对人类社会的悲剧性影响

- 机器人的定义：机器人是一种能够通过编程和自动控制来执行诸如作业或移动等任务的机器

- 阿西莫夫的机器人三定律：
  - 机器人不得伤害人类，不得看到人类受到伤害而袖手旁观
  - 机器人必须服从人类给予的命令，除非这种命令与第一定律相冲突
  - 只要与第一或第二定律没有冲突，机器人就必须保护自己的生存

终极目标：通用机器人

人工智能在机器人中的应用

- 机器人学是集机械、电子、控制、计算机、传感器、人工智慧能等多学科和前沿技术于一体的高端设备

人工智能在机器人中的应用

- 人工智能在机器人“视觉”中的应用
  - 目标的识别、定位、分类等，需要借助机器学习模型

人工智能在机器人中的应用

- 人工智能在机器人“听觉”中的应用
  - 借助人工智能技术构建语音识别系统

人工智能在机器人中的应用

- 机器人智能导航和规划
  - 搜索、知识推理、机器学习等

具身智能(Embodied AI)

近来人工智能进展均集中在离身智能（Disembodied AI）

2012年，ImageNet引发的图像识别革命
旁观式地图像标签学习

2016年，AlphaGo战胜人类冠军棋手
虚拟世界计算与规划问题

2023年，ChatGPT通用对话
活在文字世界中，缺乏真实世界理解

离身智能(Embodied AI) 缺少身体体验，无法支持机器人智能行为

具身智能(Embodied AI)

离身智能
（当前主流AI学习方式）

人类告诉我长这样子是盒子
（“标签”学习）

具身智能

亲身体验盒子是什么
可以打开，可以装东西

具身智能：具有身体体验智能

具身智能(Embodied AI)

- 具身智能是指一种基于物理身体进行感知和行动的智能系统，其通过智能体与环境的交互获取信息、理解问题、做出决策并实现行动，从而产生智能行为和适应性

机器人本体

反馈进化

驱动载体

具身智能

基于身体，理解物理世界、并与物理世界互动达成人类需求的智能系统

自动驾驶

- 感知：

利用神经网络行人识别、车牌识别、信号灯识别等

- 决策：

借助强化学习做决策

- 推理：

利用环境知识、物理规则、交通法规

- …

Next

- See your application in next 2 weeks
- Conclude & Discuss & QA in the last week