浅谈生成式 AI 及其数据风险

黎有琦
liyouqi@bit.edu.cn

计算机学院
北京理工大学

June 29, 2023

Outline

Outline

智能问题 & 智能任务

我们希望计算机能完成，

- 判别：异常判别
- 识别：人脸识别
- 监测：恶意流量监测
- 生成：图像生成、文本生成
- 推理：因果推断
- 控制：自动驾驶、机器人
- …

你还能想到其他任务吗？

人工智能的定义

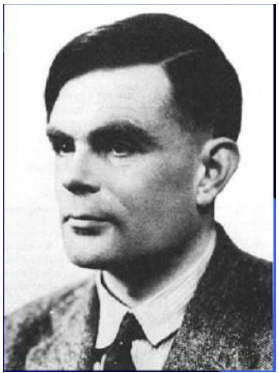
什么是智能？具有解决以下问题的能力

- 是什么（What）：判别、识别、监测
- 为什么（Why）：推理
- 怎么做（How）：生成、控制

人类具有解决以上问题的能力，计算机可以吗？
我们希望计算机可以模拟人类的智能，人工智能应运而生。

起源

- 1946 年：第一台电子计算机 ENIAC 诞生
- 1950 年：阿兰 · 图灵，《计算机器和智能》，图灵测试：如何知道一个系统是否具有智能？
- 1956 年：达特茅斯夏季人工智能研究会议
  - 克劳德 · 香农、约翰 · 麦卡锡、马文 · 明斯基、纳撒尼儿 · 罗切斯特...
  - 指明人工智能研究的未来方向、诞生了一些新名词：机器学习



(a)

(b)

Outline

传统人工智能

思路：

- 先了解人类是如何产生智能的，然后让计算机去模拟人思考
- 机器要像人一样思考才能获得智能
- 本质上，就是通过规则来定义智能

缺点：人类的智能过程还未完全研究透，无杀手级应用

- 根据图灵测试，实现智能的方式不唯一，不必像人类一样思考智能，只有智能效果一样即可。
- 例子：飞机并没有模仿鸟飞才会飞，飞机能飞，靠的是空气动力学

Outline

Outline

数据的定义

- 侠义：数字，或必须由数字构成的。
- 广义：刻画某个领域的信息组合
  - 互联网上的任何内容：文字、图片和视频
  - 医疗数据：病例、医学影像
  - 公司和工厂的各种设计图纸
  - 出土文物上的文字、图示，尺寸、材料
  - 关于宇宙的数据，宇宙基本粒子数量
- 数据、信息和知识：
  数据与信息的联系：数据是信息的表现形式，信息是数据有意义的表示。数据是信息的表达、载体，信息是数据的内涵，是形与质的关系。

大数据：

- 体量大
- 维度多，信息量丰富
- 覆盖面广：每个维度的样本都要有

Outline

集合

- 概念：由一个或多个确定的元素所构成的整体
- 大白话：一个容器
- 举例：全体自然数集合 $\mathbb{N}$，全体实数集 $\mathbb{R}$，各种自定义的集合
- 元素和集合的关系：属于/不属于，$a \in \mathcal{A}$，$b \notin \mathcal{A}$，
- 集合与集合之间的关系：子集/交并补/映射（函数）

集合上的操作

- 添加：从集合 $\mathcal{A}$ 添加一个元素 a
- 删除：从集合 $\mathcal{A}$ 删除一个元素 a
- 变换（映射、函数）：从一个集合变 $\mathcal{A}$ 换到另外一个集合 $\mathcal{B}$，$f : \mathcal{A} \to \mathcal{B}$
  - 排序
  - 求最值
  - 将某个数的集合 $\mathcal{A}$ 同时扩大两倍 $2\mathcal{A}$
  - ...

数据的多样性：维度

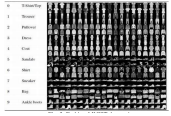
- 给定集合 $\mathcal{A} = \{a_1, a_2, a_3\}$ 和集合 $\mathcal{B} = \{b_1, b_2\}$，可以组合一个新的集合 $\mathcal{C} = \mathcal{A} \times \mathcal{B} = \{(a_1, b_1), (a_2, b_1), (a_3, b_1), (a_1, b_2), (a_2, b_2), (a_3, b_2)\}$，该操作称为笛卡尔积.
- 对于数据集 $\mathcal{C}$ 来说，有两个维度的信息，第一维度的数据来自集合 $\mathcal{A}$，第二维度的数据来自集合 $\mathcal{B}$。

用集合来定义数据集

数据集是一个集合，集合的每一个元素是一个数据样本，$\mathcal{D} = \{d_1, d_2, ..., d_n\}$

- 图像数据集：数字手写数据（MNIST）、时尚产品数据（Fasion-MNIST）、CAFAR-10
- 文本数据集：新闻分类、命名实体识别、序列标注、机器翻译
- 视频数据集：异常监测
- 声音数据集：语音数字数据（Audio-MNIST）
- 传感器数据集：步数监测数据
- 医疗数据：肿瘤监测
- 营销数据: 产品推荐
- ...

(c) MNIST.



(d) FashionMNIST.

(e) CIFAR10.

数据集查找网站: `https://www.kaggle.com`

Outline

数据集有什么作用

- 检索信息，即查找一个元素（数据样本）是否在数据集中
- 存储信息，即加一个元素（数据样本）到数据集中
- 预测信息，
  - 通过现有数据集，泛化到未知数据，即人工智能任务，可简称为智能任务
  - 举一反三的能力

数据驱动智能的思路

核心：变智能问题为数据问题，步骤如下：

- 通过数据来（建立）定义统计数学模型
- 训练：优化模型参数，得到最优的模型
- 用训练好的模型进行预测未来数据表现

优点：随着数据量越大，模型性能就变好，接近甚至优于人类的智能
缺点：数据量少的时候，模型性能差
可以通过量变实现质变

Table: 从阿拉伯语到英语的翻译

公司	准确度（BLUE 值）	智能方法
谷歌	51.31%	数据驱动
南加州大学	46.57%	数据驱动
IBM 沃森实验室	46.46%	数据驱动
马里兰大学	44.97%	数据驱动
SYSTRAN 公司	10.79%	传统的

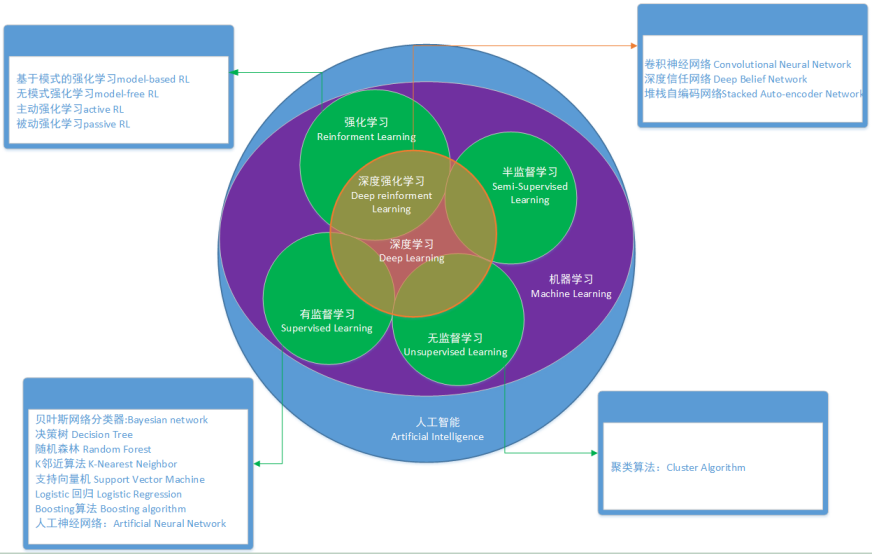
谷歌用的是老方法，但数据量是其他的几千倍到上万倍的数据

例子

- 语音识别: Siri、小度
- 机器翻译：谷歌翻译、ChatGPT
- 文本的自动摘要或写作: ChatGPT
- 战胜人类的国际象棋/围棋冠军：AlphaGo
- 自动回答问题：ChatGPT
- 路况信息实时监测：谷歌地图、百度地图

得益于互联网加速了数据的诞生和传播

人工智能的范畴


基于模式的强化学习model-based RL
无模式强化学习model-free RL
主动强化学习active RL
被动强化学习passive RL
卷积神经网络 Convolutional Neural Network
深度信任网络 Deep Belief Network
堆栈自编码网络Stacked Auto-encoder Network
强化学习
Reinforment Learning
深度强化学习
Deep reinforment
Learning
半监督学习
Semi-Supervised
Learning
深度学习
Deep Learning
机器学习
Machine Learning
有监督学习
Supervised Learning
无监督学习
Unsupervised Learning
人工智能
Artificial Intelligence
贝叶斯网络分类器:Bayesian network
决策树 Decision Tree
随机森林 Random Forest
K邻近算法 K-Nearest Neighbor
支持向量机 Support Vector Machine
Logistic 回归 Logistic Regression
Boosting算法 Boosting algorithm
人工神经网络：Artificial Neural Network
聚类算法：Cluster Algorithm

Outline

现代智能：机器学习

三要素：数据、数学模型和硬件基础

- 数据：语音、图像、文本、无线信号（光、电磁波、核磁共振成像）、传感器数据（加速度计、陀螺仪）
- 数学模型：支持向量机、K-means、人工神经网络
- 硬件基础：CPU 和 GPU

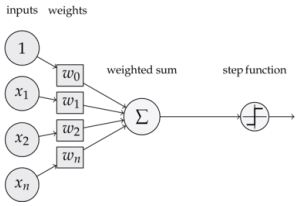
机器学习：让计算机从大量的数据中自己学习（优化）得到相应的模型参数
学习的过程：训练和测试
机器学习的效果取决于：

- 不断学习的深度：迭代次数，消耗计算资源
- 学习时使用的数据量：类比考试前，做题的数目
- 数据的质量：无噪声。类比考前，做题的质量。

Outline

概念

- 同义词：深度学习、连接主义和人工神经网络
- 与人的脑神经没有关系
- 由一个个 ‘‘神经元’’（感知器）按照一个规则连接
- 感知器本质上就是带权求和，加上激活函数



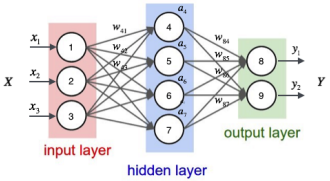
输入：$\boldsymbol{x} \in \mathbb{R}^n$，模型参数：$\boldsymbol{w} \in \mathbb{R}^n$，偏置参数：b
激活函数（有很多种、如 Sigmoid、step function）：

$$f(z) = \begin{cases} 1, & \text{if } z > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

模型计算：$y = f(\boldsymbol{w}^\intercal \boldsymbol{x} + b)$

最基础的版本：全连接神经网络

- 输入层、隐含层（至少 1 层）和输出层
- 可以拟合任何函数，包括非线性的（万有逼近原理）



- 给定输入 $\boldsymbol{x} \in \mathbb{R}^n$，模型的参数可以表示成 Θ，输出为 $\boldsymbol{y} \in S$，模型计算可以表示成 $\boldsymbol{y} = f_\Theta(\boldsymbol{x})$
- label 集合 S，为离散集合时候，神经网络接近分类任务；否则为回归任务
- 训练网络：找到最优的模型参数 Θ^*
- 定义训练目标，即损失函数：网络的输出和真实数据集的输出差距
- 训练方法（不断迭代）：反向传播得到梯度，然后进行梯度下降法，$\Theta_{t+1} = \Theta_t - \eta \frac{\partial f_\Theta}{\partial \Theta}$

变体

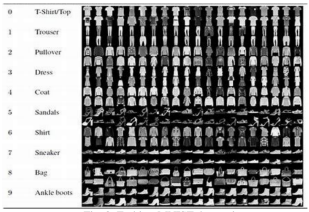
- 图像识别：卷积神经网络（CNN）、残差网络（ResNet）
- 自然语言处理：循环神经网络（RNN）、长短时记忆网络（LSTM）、Transformer
- 图数据任务：图神经网络（GNN）
- 生成类的网络：语言（生成）模型、对抗生成网络（GAN）、扩散模型（DDPM）

为什么可以学习

- 拟合了原始数据的概率分布，即对于输入数据 x_1，经过优化后的神经网络计算，输出真实 label 的概率最大，即 $\Pr(y^*|\boldsymbol{x})$
- 数据量大，潜在的概率分布规则就越明朗
- 硬件的进步：CPU 和 GPU
- 优点：1）强大的函数近似能力；2）很强的泛化能力
- 缺点：1）缺乏健壮性（robustness）；2）缺乏可解释性

现有最好结果（State-of-the-arts）

(f) MNIST.



(g) FashionMNIST.

(h) CIFAR10.

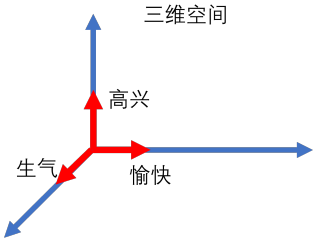
任务	模型	精度
MNIST	CNN	98.2%
Fashion-MNIST	ResNet	96.7%
CIFAR10	ResNet	94.7%

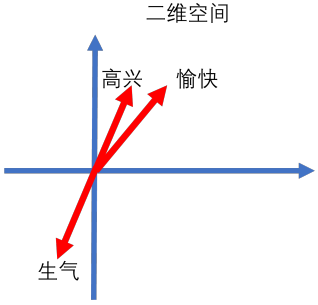
Outline

词向量：将文本数值化

- 把单词映射（embedding）到向量空间中的一个向量，又叫词嵌入（word embedding）
- 保持相似性（保测度）
- 例子：高兴、愉快、生气
- 独热编码，$[1,0,0]$、$[0,1,0]$ 和 $[0,0,1]$
- 低维度词向量，$[0.5,0.9]$、$[0.6,0.9]$ 和 $[-0.4,-0.9]$

确定词向量的方法：1）利用语料库自己训练，word2vec；2）用现有已预训练好的，Chinese-Word-Vectors，FastText



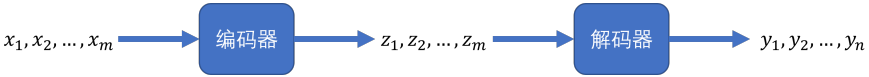


语言生成模型简介：序列到序列的学习

- 语言模型是定义在单词序列上的概率模型，用来计算一个给定的单词序列的概率 $\Pr(w_1, w_2, ..., w_T) = \prod_{t=1}^{T} \Pr(w_t | w_1, w_2, ..., w_{t-1})$
- 序列到序列的学习：给定单词序列 $x_1, x_2, ..., x_m$，输出单词序列 $y_1, y_2, ..., y_n$。语言模型计算以下概率，

$$\Pr(y_1, y_2, ..., y_n | x_1, x_2, ..., x_m) = \prod_{i=1}^{n} \Pr(y_i | y_1, y_2, ..., y_{i-1}, x_1, x_2, ..., x_m) \quad (2)$$

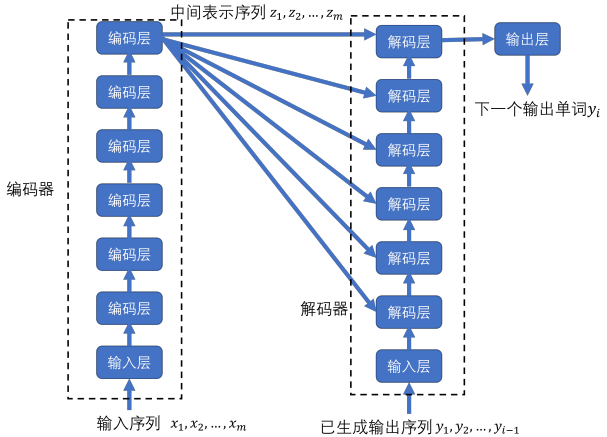
- 一般由编码器网络和解码器网络组成，编码器将输入 $x_1, x_2, ..., x_m$ 转化为中间表示 $z_1, z_2, ..., z_m$，解码器网络将中间表示 $z_1, z_2, ..., z_m$ 转化为输出序列 $y_1, y_2, ..., y_n$



基于注意力机制的序列到序列学习模型：Transformer

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 使用注意力机制来实现编码、解码以及编码器和解码器之间的信息传递
- 注意力机制是一种特殊的数学操作、类似的还有卷积神经网络的卷积、下采样
- 结构：编码器（1 个输出层、6 个编码层），解码器（1 个输出层、6 个解码层、1 个输出层）



预训练语言模型

- 事先使用大规模预料学习基于 Transformer 等的语言模型，之后微调，用于各种下游任务进行学习和预测
- 代表性模型：GPT（generative pre-training）和 BERT（bidirectional encoder representation from Transformers）
- GPT 的训练过程主要包括两个阶段。
  - 第一个阶段是利用一个大的文本语料库来学习一个高容量的语言模型
  - 第二个阶段进行微调，也就是利用标注数据将模型适配到一个下游任务

GPT 的模型架构

《机器学习方法》，李航

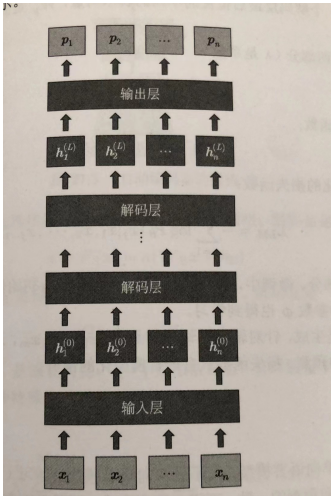
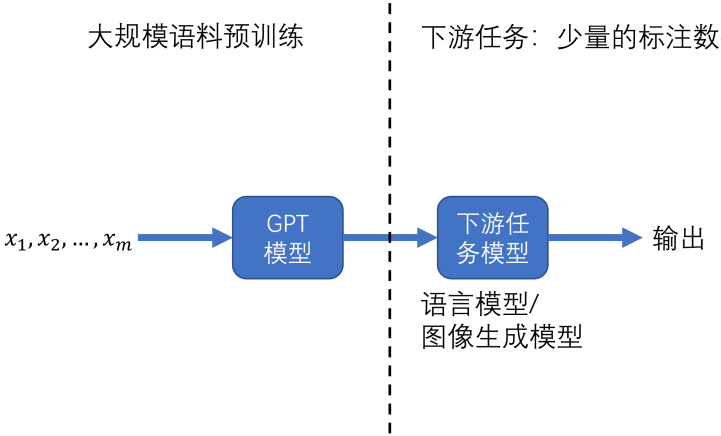


图 27.2　GPT 模型的架构

GPT+ 下游任务


大规模语料预训练
下游任务：少量的标注数据进行微调
$x_1, x_2, \ldots, x_m$
GPT
模型
下游任
务模型
输出
语言模型/
图像生成模型

生成任务

- 文本生成文本：问答、翻译、摘要、对话
- 文本生成图像：根据文本描述画图，过程（输入一句话的词向量，神经网络计算，输出图像）
- 图像生成文本：根据图生成文本描述，过程（输入图像，神经网络计算，输出一句话的词向量或者词的独热编码）

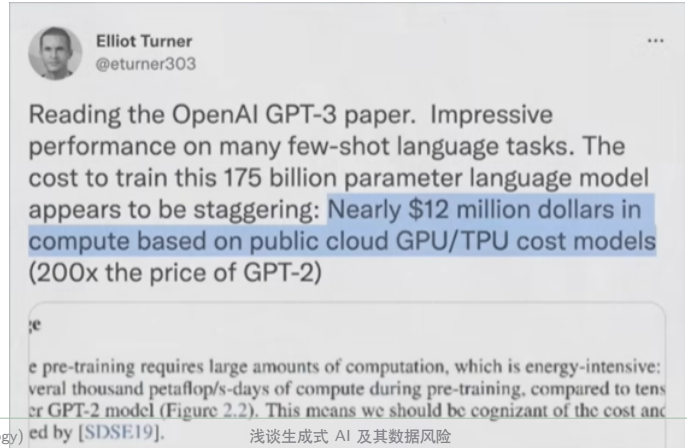
ChatGPT

- GPT1 和 GPT2 开源，往后的 GPT3、GPT3.5、InstructGPT、GPT4 闭源
- ChatGPT 是 GPT 技术应用到问答对话领域，核心技术与 InstructGPT 类似。采集互联网大规模进行预训练，再使用强化学习进一步优化
- 能回答什么问题
  - 简单问题：有明确答案，关于事实的问题，比如某某明星是哪儿人
  - 复杂问题：要有逻辑，需要整合各种信息；问过程的问题
- 擅长内容整合，不擅长原创内容创造，新内容。

模型	层数	头数	词向量长度	参数量	预训练数据量
GPT-1	12	12	768	1.17 亿	约 5GB
GPT-2	48	-	1600	15 亿	40GB
GPT-3	96	96	12888	1750 亿	45TB
GPT-4	-	-	-	100 万亿	-

训练费用

- 训练费：设备费 + 电费 + 人员费用 + 其他
- 训练 GPT-3 一次要花费 1200 万美金
- ChatGPT 训练一次要耗多少电？大概可能是 3000 辆特斯拉的电动汽车，每辆跑到 20 万英里，把它跑死，这么大的耗电量，才够训练一次，这个非常花钱的一件事。



Outline

数据风险

恶意平台：

- 训练阶段的风险：数据要上传到平台进行训练，造成泄露
- 部署阶段的风险：要上传输入数据到平台，进行计算推理，返回推理结果，输入数据也泄露给平台

恶意用户：

- 训练阶段的风险：投毒攻击、后门攻击
- 部署阶段的风险：逃逸攻击、模型窃取攻击

	诚实平台	恶意平台
诚实用户	无风险	有风险
恶意用户	有风险	有风险

防御手段

恶意平台：

- 训练阶段的风险：联邦学习
- 部署阶段的风险：密码学加解密、同态加密、密态计算

恶意用户：

- 训练阶段的风险：对抗训练
- 部署阶段的风险：差分隐私、恶意用户识别

Outline

总结

本次报告讲了，

- 数据驱动的人工智能
- 深度学习、语言生成模型
- 深度学习中的数据安全风险及其技术对策

参考文献

- 吴军，《智能时代》、《数学之美》
- 李航，《机器学习方法》
- GPT-4 Technical Report，https://cdn.openai.com/papers/gpt-4.pdf