人工智能常用概念

杜晓宇

解决问题的基本流程

- 问题抽象
- 模型设计/选择
- 模型训练
- 模型推理
- 模型评价
- 产品化

问题抽象

- 了解用户需求，确定问题的**输入**和**输出**
- 测粮食体积的需求为例
  - 目前输入是不确定的，可以由我们来设计
  - 输出是体积

问题抽象

- 由于输入不确定，又没有现成的方案参考，所以是比较困难的
  - 目前猜想，可以在上方挂两个RGBD摄像头，采集两张图像
  - 两张图像做3D复原，形成立体模型
  - 根据立体模型计算体积
- 于是该问题由两个子问题构成：
  - 1、**输入**两张RGBD图像，**输出**立体模型
  - 2、**输入**立体模型，**输出**体积

模型设计/选择

- 模型的概念：
  - 它其实是一个基本的架子，数据由输入，经过这个架子，变成了输出
  - 架子的中间内容是可调的，也就是由模型**参数**来控制
- 类比
  - 机场安检通道，是预先定好的
  - 根据时间不同、人流量不同，会调整通道的顺序，调整开放的门
  - 最终旅客从哪个门出，是由安检通道当时的设定（**参数**）来决定的。

模型训练

- 现在人工智能都使用**机器学习**一套理论来实现
  - 定义一套框架，确定输入输出分别为y = f(x)，f函数中有一些参数是不确定，可以调整的。
  - 收集数据集，遇到什么输出，就给出什么输出，称之为标注数据
  - 让y=f(x)在所有标注数据上都能够有效
- f(x)一开始肯定是不能够全部为y的，那么需要根据数据集学习，调整里面的参数，这个过程就叫**模型训练**。
- 理想状况下，f(x)都能够生成准确的y，但是很多数据是没有在数据集中出现过的，学习过程中很难精准的把握，所以模型就有好有坏。

模型推理

- 训练好的f(x)，我们拿它到实际场景中去计算，就叫**推理**
- 例如，我们已经有一个f(羊群照片)的函数了，那么它的输出，就是我们希望得到的羊的数量了。
- 训练过程也会做推理，会根据推理准确性，去调整f函数中的参数
- 推理函数f的运行效率一般是不低的，资源消耗也不会特别夸张
- 推理过程是可以并行的，一般来说不是系统的效率瓶颈。

模型验证

- 一般会通过一些指标来验证模型好坏
  - 准确率：常用于分类任务。比如给一张图，这图里是猫还是狗，分对了为1，分错了为0，那么准确率就是为1的数量占总数的多少
  - IoU：有些输出是区域，那么模型输出的区域与理想区域的覆盖程度就是IoU。
- 这些指标可用于合同中体现我们算法的优越性。例如：
  - 通过人工筛选24小时，准确率到多少
  - 我们保证算法的准确率在人工的准确率的基础上提升10%。
- 验证指标特别多，不止上面两个，具体问题具体分析。

产品化

- 一般说来，模型训练好了以后，都是y = f(x)的形式
  - x可以是二维矩阵（黑白图、音频）、三维张量（彩色图）、四维张量（视频）等各种复杂的东西，也可以是上述内容的组合
  - y也同样可以有各种输出，类别编号、选择框（x,y,w,h)、文本、图像、视频等等，取决于任务是什么
- 产品化一般先做一套接口，输入x，返回y，然后使用web或者app远程调用接口即可。

人工智能的难点

- 问题的抽象。
  - 实际问题一般是多种问题的组合，如何划分子问题是难点。问题评价也是难点。
- 模型的选择与训练
  - 人工智能的模型都是在炼丹，效果好不好取决于模型选择和训练
  - 模型选得好，训练得不好，不行。（我学生就是训练牛，所以常拿冠军）
  - 模型选得不好，训练得再好也没戏。（这个也是看经验）
  - 训练一次，普通模型一天能出结果，大模型一周起步
  - 训练一次，效果不好，就得重新找模型或再训练一次找新的可能，时间主要就花费在这上面了
  - 更多的卡，在一次训练过程中，可以同时尝试更多的可能，有机会更快找到更好的参数。
- 推理的效率问题。粗略估计图像类算法在3090上一秒10张图。提供服务的并发限制很明显。

落地的代价

- 有现成的模型
  - 我们问题的输入输出都与现有模型完全一致，那大概率能快速落地
- 有相似的模型
  - 现在有数脸的模型，但没有数羊的模型，那么可以用数脸的模型在数羊的数据上训练一下
  - 我那个优秀学生一般一周能搞定；发表过一篇论文的学生大概要花1个月达到一个比较好的水平；新手得半年往上
- 没有相似的模型
  - 输入输出没啥类似的，需要自己设计。我那个优秀学生三个月到半年。别的学生可能1年往上
- 人力成本。前两种情况可以招本科生或硕士（有点科研经验的），第三种只能招博士（50万起步）
- 服务器成本。第一种，一张卡就够用了。第二种，1-10张卡。第三种，越多越好。（如果不做大模型，并不需要A100，3090或4090都可以的，成本会降很多）