第3章

零基础开发——人工智能定制化训练服务

本章学习目标

- 了解人工智能的学习模式
- 熟练掌握基于人工智能定制化训练服务平台的应用开发
- 熟练掌握人工智能模型的性能评估
- 了解人工智能产品智商评测

本章首先论述人工智能学习模式，接着介绍目前流行的人工智能定制化训练平台并给出应用指引范例，最后讨论人工智能模型的性能评估和产品智能水平。

随着机器学习，尤其是深度学习在复杂数据上的表现越来越优秀，很多开发者希望能将其应用到自己的服务或产品中。从整体而言，构建我们自己的机器学习应用首先需要收集数据，并执行复杂而烦琐的数据预处理。随后则需要面对如何选择模型与各种层出不穷的修正结构的问题。然后，在训练中，需要有丰富的调参经验与各种训练技巧，还需要根据数据情况选择与修正最优化方法。最后，考虑工程化问题，将机器学习模型部署在应用场景中。目前在 GitHub 上有很多优秀的机器学习开源项目，例如，各种预训练深度卷积网络、高度封装的算法以及大量开放数据集，不过要想复现以及根据实际情况调整这些项目，开发者还是需要一些机器学习领域知识。此外，很多项目的文档说明与技术支持都有待提高，它们需要开发者一点点调试与试错，才能正确搭建。更重要的是，将训练后的模型部署到移动端等平台会遇到非常多的困难，这主要还是由于当前流行的深度学习框架并不能完美支持模型部署。

对于很多不太了解机器学习算法工程的开发者而言，部署机器学习开发环境，完成机器学习算法研发流程还是有非常大的挑战。此外，若机器学习不是产品的核心技术，额外维护机器学习算法团队的成本又非常高。因此，很多时候我们需要一种能快速使用高性能深度学习的方法。所以对于不太了解机器学习的开发者而言，最好将机器学习算法研发，模型部署等过程都自动化。例如我们只需要收集少量和任务相关的数据，并直接在平台上完成标注，然后让系统帮我们选择合适模型与超参数进行训练，最后已训练模型还可以直接部署到云 API 或打包成安装包。零编码、零 AI 基础、轻松几步构建 AI 模型，这种解决方案就是人工智能定制化平台。当然，在进行零代码开发 AI 应用之前，零基础 AI 开发者还是需要了解一些和数据、应用、模型相关的基础知识，如学习模式、模型性能评估指标等。

3.1　人工智能学习模式简介

学习能力是智能的核心，学习对生物生存具有很强的积极意义。学习行为简单地说就是一种条件反射，但人脑作为世界上最复杂的物质系统，其学习能力是其他一切生物无法比拟的。“学”是指人获得直接与间接经验的认识活动，兼有思的含义；“习”是指巩固知识、技能等实践活动，兼有行的意思。我国对学习的研究已有悠久的历史，孔子认为，“学”与“习”不是两个独立的环节过程，“学”是“习”的基础与前提，“习”是“学”的巩固与深化。在学习的过程中可以感受到愉悦的情绪体验，揭示了认知、情绪、行为三者之间的统一关系[1]。现代科学对学习机理形成了更细化的一些观点，如经典条件反射学习理论、行为主义学习理论、联结主义学习理论、操作条件反射理论和社会学习理论等[2]。人工智能旨在模拟人脑，通过教育、训练和学习过程，更新和丰富拥有的知识和技能。因此，人工智能也相应发展出一系列人工智能学习模式，如监督学习、无监督学习、半监督学习、强化学习、迁移学习、联邦学习。人脑学习理论与人工智能学习模式之间的对应关系如表 3-1 所示。

表 3-1　人脑学习理论与人工智能学习模式之间的对应关系

学习科学理论学家	理论	关于学习的主要观点	人工智能学习模式
巴浦洛夫	经典条件反射理论	提出几种基本的学习现象：强化、抑制、恢复、泛化和分化	强化学习
华生	行为主义	学习就是以一种刺激替代另一种刺激建立条件反射的过程	强化学习、迁移学习
桑代克	联结主义	学习的实质在于形成刺激-反应联结，学习的过程是尝试与错误的渐进过程	监督学习、无监督学习、半监督学习、强化学习
斯金纳	操作条件反射理论	反射学习是 S-R 过程，操作学习是(S)-R-S 的过程。刺激加强的不是 S-R 的联结，而是行为发生的频率	强化学习
班拉拉	社会学习理论	社会认知理论，交互决定论，观察学习的过程：注意、保持、复制、动机	监督学习、迁移学习

虽然人类非常善于从事学习，但还不能充分了解这项能力的生理机制。因此，将该功能直接通过编码嵌入机器当中会十分困难。当前，以大数据、算力和算法为支撑的人工智能系统的学习能力主要通过分析数据来完成。这已经能和那种按照明确的指令、以预先定义的方法行事的自动化模式区分开来，实现系统的“智能化”。实际上，识别大量数据中存在的模式，恰是机器学习的核心特长之一。机器学习可以从原始数据中学习，从而赋能于人工智能可见的出色表现，使其变得越来越普遍。无论是进行前瞻判断的预测系统、近乎实时解读语音和文本的自然语言处理系统、以非凡准确度识别视觉内容的机器视觉技术，还是优化搜索和信息检索，都依托于机器学习。相对于其他技术，机器学习的一项关键优势，就是对“脏”数据的容忍度——即数据中包含有重复记录、不良解析的字段，或是不完整、不正确以及过时的信息[3]。机器学习具备灵活性，可随着时间推移获得全新发现并做出改进，这意味着它能够以更高的准确性处理脏数据，并且由此拥有了极佳的可扩展性。在我们当前所处的数据大爆炸时代，后者正变得越来越重要。目前，人工智能系统可以像人一样识别出垃圾邮

件，能在人们每次打开新闻 App 时推荐感兴趣的新闻，让人们看了还想看，能在人们输入信息时推送针对性的表达内容和表达模式。本质上来说，当前的人工智能学习，就是让计算机从数据中自动寻找规律[4]。

3.1.1 机器学习范例

这部分先介绍机器学习的相关概念。

1. 模式识别问题

以旅游景区的安全预警作为模式识别问题的典型应用场景。经专家分析，旅游景区安全预警的影响指标和预警等级汇总如表 3-2 所示。

表 3-2 旅游景区安全预警的影响指标

预警等级	景区入园率	游客平均逗留时间	雇员游客比	景区设施使用饱和度	水文气象灾害发生率	监控覆盖度	景区行政能力	突发事件响应速度	应急预案健全度
符号	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9
优秀级	0.60～0.70	＜0.60	0.91～1.00	0.80～0.95	0.00～0.10	0.85～1.00	0.90～1.00	0.95～1.00	0.90～1.00
良好级	0.70～0.80	0.60～0.75	0.76～0.90	0.70～0.80	0.10～0.15	0.70～0.85	0.80～0.90	0.85～0.95	0.80～0.90
合格级	0.80～0.90	0.75～0.90	0.60～0.75	0.60～0.70	0.15～0.20	0.60～0.70	0.75～0.80	0.80～0.85	0.70～0.80
危机级	0.90～1.00	0.90～1.00	＜0.60	＜0.60 或 ＞0.95	＞0.20	＜0.60	＜0.75	＜0.80	＜0.70

由表 3-2 可见，影响旅游景区安全等级的指标可归纳为 9 个指标，分别是：

(1) 景区入园率，指景区实际入园人数与承载力上限的比率，越接近上限，安全性就相对越低。

(2) 游客平均逗留时间，逗留时间越长，对景区安全性的影响就越大，通过分析历史区间内游客平均逗留时间，得出此项指标数值。

(3) 雇员游客比，指预警时间段内景区服务人员数量与入园游客的比例，过大的比例造成资源浪费，过小的比例容易造成服务忽略或不到位，触发安全问题。

(4) 景区设施使用饱和度，指景区内设施使用饱和情况对安全的影响，过度使用或过度闲置的设施都造成安全隐患。

(5) 水文气象灾害情况，诸如洪水、干旱、寒潮霜冻和沙尘暴等气象、水文灾害对旅游活动的开展有很大影响。主要考察气象水文灾害资料，分析历史时期该旅游地爆发灾害气象事件的次数、频率，形成旅游地水文灾害的爆发概率。

(6) 监控覆盖度，通过分析景区摄像头、流动监控设备，形成视频监控区域覆盖比例。

行政指标，包括景区行政能力、突发事件响应速度、应急预案健全度。

(7) 景区行政能力，指景区协调指挥处理事件的能力，通过对景区历史事件处理情况的总体分析，得出此项指标。

(8) 突发事件响应速度，如儿童走失的情况的响应速度，需要对景区历史事件处理情况进行分析。

(9) 应急预案健全度，预案的健全与否直接影响景区的安全级别。

上述 9 个指标构成了训练数据的特征集，即每一条数据由 9 个特征组成。我们把每条数据记为 $X=(x^1,x^2,\cdots,x^9)$。显然，每个景区在特征集上的取值是不同的。根据 9 个影

响指标的不同取值范围，专家将预警等级分为四个级别。根据表 3-2 的等级规定，可以看出：

（1）优秀级，表明旅游景区安全性高、无安全隐患，游客活动无需担心突发事件。

（2）良好级，表明旅游景区有爆发安全事件的可能性，但几率很小，并且有健全的安全事件处理预案，在此等级内，游客也无需过分担心安全问题。

（3）合格级，此级别旅游景区一般具体良好的处理预案，能够有效控制，在此预警等级内，对游客的安全知识有一定要求，在安全知识欠缺的情况下，不鼓励游客开展旅游活动。

（4）危机级，旅游景区爆发安全问题的可能性极大，一旦事故发生，由于缺乏足够的应对措施和预案，将对游客与景区造成灾难性的后果。在此等级内，应杜绝游客前往，并及时疏散本地游客。

因此，旅游景区安全预警是一个四分类问题，每条数据将获得类别集 $Y=\{1,2,3,4\}$ 的某一个类别。一条完整的训练样本为 $(\boldsymbol{x},y)$，其中 $\boldsymbol{x}$ 是维度为 9 的向量，y 是单点值。

2. 训练样本集

机器学习（Machine Learning，ML）致力于让计算机模仿人类从实例中学习的能力进行数据分析和建模，以估计某系统隐藏的复杂的输入输出映射关系（也称为模型），使模型能够对未知输出做出尽可能准确的预测。首先需要收集人类的经验数据进行学习，以获得判决模型。当采集一条新景区的数据 $\boldsymbol{x}$，就能使用模型自动判断安全预警等级 y。经过样本采集，我们获取了 12 个样本，展示如表 3-3 所示。这个表就构成了训练样本集，每个样本由 9 个指标和 1 个类别标记组成。

表 3-3　训练样本集

x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	y
0.70	0.55	0.91	0.82	0.02	0.92	0.92	0.97	0.97	1
0.64	0.42	0.92	0.90	0.01	0.97	0.94	0.99	0.94	1
0.68	0.44	0.94	0.88	0.04	0.93	0.97	0.96	0.91	1
0.73	0.68	0.78	0.75	0.12	0.82	0.85	0.91	0.84	2
0.77	0.71	0.82	0.76	0.14	0.81	0.84	0.92	0.86	2
0.71	0.73	0.88	0.79	0.13	0.84	0.87	0.87	0.89	2
0.84	0.81	0.67	0.62	0.17	0.77	0.77	0.82	0.79	3
0.89	0.87	0.72	0.67	0.15	0.75	0.76	0.83	0.77	3
0.85	0.77	0.62	0.65	0.16	0.78	0.79	0.84	0.72	3
0.95	0.92	0.54	0.51	0.27	0.62	0.72	0.75	0.64	4
0.97	0.93	0.42	0.97	0.43	0.51	0.64	0.79	0.67	4
0.98	0.99	0.31	0.32	0.34	0.47	0.71	0.78	0.69	4

3. 分类问题定义

下面对分类问题采用通用数学语言进行描述。实际问题的具体对象一般有很多属性，可用高维向量 $\boldsymbol{x}=(x^1,x^2,\cdots,x^d)^{\mathrm{T}}\in X$ 表示，其中数据 $\boldsymbol{x}$ 的上标 $1,2,\cdots,d$ 是向量的维数序号，也就是对象的属性序号；$\boldsymbol{X}=\mathbb{R}^d$ 表示 d 维实数空间。假设实际问题对应有 c 个可选类别，用标识变量 $y\in Y=\{1,2,\cdots,c\}$ 表示对象的类别，其中 Y 叫作类别空间。对象及其类

别构成数据对 $z=(\boldsymbol{x},y)\in\boldsymbol{Z},\boldsymbol{Z}$ 叫作样本空间，$z_i,i=1,2,\cdots,n,\cdots$是样本空间中的样本(点)，其下标“$i$”表示样本编号。则其模式分类可以定义为：

定义 3.1(分类问题 F) 根据给定的训练样本集 $z^{(n)}=\{z_1,z_2,\cdots,z_n\}$，其中 $z^{(n)}$表示含有 n 个样本的样本集，$z_i\in\boldsymbol{Z},i=1,\cdots,n$ 表示样本，产生一个分类器：

$$\varphi:\mathbb{R}^d\rightarrow(1,2,\cdots,c)$$

使得它对新的待测试数据 $x_{n+1},x_{n+2},\cdots,x_{n+k}$ 的实际类别值产生相应的预测值 $\hat{y}_{n+1}$，$\hat{y}_{n+2},\cdots,\hat{y}_{n+k}$。我们要求该分类器要能对整个样本空间可能的分布有一个尽可能小的期望判别误差。

4. 分类模型调优

那么，分类器模型是如何通过训练样本学习得到的呢？分类器通过持续对训练样本集进行学习，以预测类别与真实类别的一致性作为评价准则对学习过程进行指导和监督。以人工神经网络模型作为分类器为例，图 3-1 是人工神经网络模型的学习过程流程图。

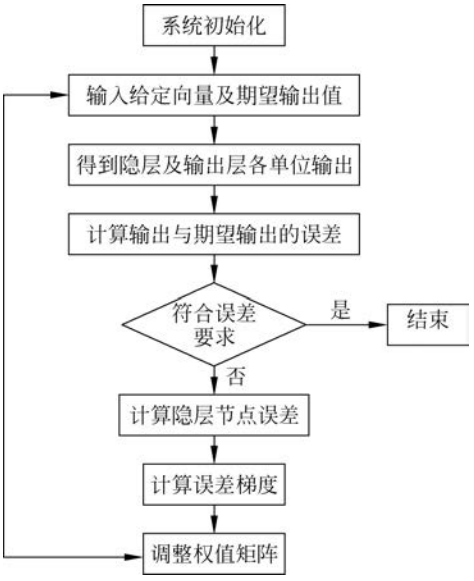

图 3-1 人工神经网络模型的调优技术流程

一开始我们对分类器进行初始化，然后输入训练样本集的一条数据和它的真实类别。训练数据经过神经网络模型的隐藏层和输出层得到模型预测类别值。接下来，对模型预测类别值与真实类别值进行误差计算。该误差称为预测误差。若预测误差小于指定误差，则模型已经学习好了，完成学习任务，可进入下一阶段，即投入使用阶段，为将来出现的数据进行预测工作。如果预测误差不符合指定要求，则需要修改模型的设计，具体来说，就是计算出预测误差导致的各个隐藏层节点误差，并计算出隐藏层节点的误差梯度，最终修改模型的权重来完成模型的一次调优。显然，经过不断将训练样本送入模型，一次次调优，最终将学习出对训练样本集拟合度非常高的模型，并对未来新出现的数据有足够的泛化性。

上面给出的范例只是基于一种学习问题(模式分类)进行展示，但现实场景还有回归、聚类、决策等需求。不同的需求将采用不同的机器学习模式来解决。下面从机器学习模式角

度阐述人工智能的学习机制。

3.1.2　监督学习

类比我们认识动物的过程，父母告诉我们这是狮子，这是老虎，这是大象，我们就去寻找区分这三者的规律，如长鼻子大耳朵的是大象，身上有斑纹的是老虎，有鬃毛的是狮子等。如果我们想让一台机器学会辨别苹果和香蕉，那就需要拿一大堆水果过来，里面有各种各样的水果，如苹果、西瓜、香蕉等，然后给每个苹果和香蕉都写上一个标签，每个苹果的标签都写上红色的、圆的，每个香蕉的标签都写上黄色的、长条的、弯的，一个个的告诉机器，这个红色的、圆的就是苹果，黄色的、长条的、弯的就是香蕉，这种学习方法叫作监督学习。用专业术语来说，这里的每个苹果和香蕉都是一个一个的样本，我们添加的每一个标签就是每个样本所对应的详细信息。监督学习模式下，训练样本集中每个样本形式为(数据，标签)对，数据是学习对象的属性信息集合，标签是学习对象的标记。如果标记是离散型的指示值，如水果的种类(香蕉、苹果等有限集合)，则监督学习可以完成分类任务。如果标记是连续型数值(如股票价格、温度值等)，则监督学习可以完成回归问题。监督学习算法使用带有标记的数据集，推导出每个标记代表的显著特征，并学习识别新数据中的这些特征[5]。例如，向算法展示大量标记为“猫”的图像，然后它将学习如何识别猫的形象，并在其他任意数量、且完全不同的图片中发现猫。

3.1.3　无监督学习

无监督学习更像是最早的人类在野外见到不同的动物，没有人告诉我们这是大象，这是老虎，这是狮子，那时甚至没有大象、老虎、狮子的概念，但慢慢动物见多了，我们就开始发现这些动物之间的一些相同点和不同点，并根据这些特点，将它们分成不同的类别，并给这些类别起名字，大象、老虎、狮子。实际上，生物学家从各个角度研究动物的特点，对动物进行分类，确定不同物种，并得到这种动物分类，就是一个无监督学习的过程。我们日常生活中也在不自觉地对人、对事情、对信息进行归类。比如，来到新学校，你会根据同学的穿着、爱好、交流频率等分析出谁和谁玩得比较好，谁和谁是一个小团体。俗话说，物以类聚，人以群分，这种根据事物间的相似程度，将它们合并归纳成一个类别的操作，称作聚类。

我们同样拿一大堆水果过来，但不告诉机器哪个是苹果，哪个是香蕉，它们都有什么特征，而是让机器自己去比对，自己去发现不同水果之间区别。虽然机器不知道哪个是苹果，哪个是香蕉，但是通过分析能够把红色的、圆的苹果和黄色的、长条的、弯的香蕉区分开来。无监督学习的目的主要是在于可以更好地挖掘或者说寻找数据里面的规律[6]。比如说，我们现在正在看关于人工智能、机器学习的视频，可能在视频的下面就给我们推荐了一堆关于人工智能、机器学习相关的视频。聚类是无监督学习中的一种典型任务，在计算机里面，只要你给出数据间相似程度的衡量标准，计算机就能使用一些算法对数据进行聚类和划分，相似的分到一个类别里，而不同的分到不同的类别中[7]。非监督学习在现实生活中有什么具体的应用呢？商家可以对潜在顾客的所处的地理位置进行聚类，从而决定店铺的选址；一些购物应用根据人们的购物历史将人聚类，然后根据你所属的类别里其他人的购物清单，来推测你可能会感兴趣的物品，从而进行推荐。再比如一些新闻的应用，会根据文章的相似程度进行主题归类，类似的还有相册的照片自动归类功能，搜索引擎的寻找相似图片的功能等。

无监督学习，还有另一类典型任务—降维，通俗讲就是划重点、提精华。迪士尼动画《头脑特工队》里三个主角进入大脑的抽象思维空间，把三维实体降维为平面，最后被抽象成没有实体的概念，但在最后一幅图里虽然没有了实体，但依然保留了三个主角最突出的特点，使我们依然能够区分出他们。还有上学时相信学生被要求归纳文章的中心思想，就是用更短的篇幅更精炼的语言复述文章的信息。对训练数据的划重点、提精华也是一样的，意思就是将数据换一种比原来更短的方式展现，保留最具代表性的信息，去掉多余的和不重要的部分。那么，哪些是多余的信息呢？假如我们有一个关于学校里学生的数据集，它里面记录了1 000个学生的五个方面的信息，共5 000个数字，显然年龄和年级这两项有很固定的对应关系，知道年龄就可以推算出年级，这时年级这一项就是多余的信息，去掉这项后就一下子去掉了1 000个数字的数据量。又如，假设有某个地区的健身房教练的信息，由于健身房教练的身材胖瘦都差不多，所以他的身高和体重也有潜在的相关性，可以通过一个固定的公式，大体不差地根据其中一项算出另一项，这时候我们就可以将这两个数字用一个数字来表示，通过体重除以身高达到去除多余信息，缩小数据体积的目的。这种“划重点”的能力有什么用途呢？一般来说，训练数据越多、体积越大，训练模型所需要的时间就越长，就更加难以训练出效果好的模型，这就像学习一样，学习资料越多、课本越厚，学习时间就要越长，还不一定都能记得住，考试时也不容易考好。这时候如果出现一个学霸给你画重点，是不是就可以大大缩短学习时间、还更有信心考好呢？一样的道理，给训练数据“划重点”的能力一般是用在训练进行学习模型前，把训练数据的体积减小，能缩短训练所需要的时间，也更容易得到效果好的模型。

总结一下，无监督学习不需要我们提供训练数据对应的正确输出结果，就能够自动从数据中寻找规律，无监督学习主要可以做两类任务——聚类和降维。聚类是根据数据间的相似程度，将它们划分为一个个类别[11]；降维是指给数据瘦身、划重点，提取数据中的精华，消除冗余的部分。

3.1.4 半监督学习

半监督学习是介于监督学习和无监督学习之间的一种方式[12][13]。在许多情况下，尽管我们每天都产生海量的数据，而这些数据往往是很初级的原始形态，很少有数据被加以正确的人工标注。比如我们经常接触的语音输入法、手写输入法，自然是需要大量的数据作为基础的，需要我们说的话，写的字，才能很好地识别出来。而数据的标注是一个极其耗时且昂贵的操作，半监督学习可以通过放宽对大量标记数据的需求来部分解决此问题。一大堆水果就相当于数据，但是这么多的水果，我们很难每一个都加上相应的标签，告诉机器这是什么、那是什么。所以我们就只给一部分加上标签，剩下来就让机器自己去识别、分类。半监督方法仅需要有限数量的标记数据，核心是如何利用大量未标记数据来提高学习准确性。

3.1.5 强化学习

在现实生活中，人们做出的选择一般会从获取最大收益的角度出发，这是决策问题。人们会思考怎样获取最大的收益，通过做出怎样合理正确的选择从而达到目的，这好比两个人下象棋，博弈的目的是为了获取最后的胜利。决策博弈问题与机器学习的交叉综合促进了强化学习——又称作增强学习或再励学习（Reinforce Learning，RL），理论与算法的产生和

发展[14]。所谓强化学习是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法。其学习原理示意如图 3-2 所示。智能体(agent)通过执行动作 a 作用于环境，而后环境产生奖励 r 反馈给智能体，并且更新状态 s。智能体通过反复试验，学习出一种策略，以使得回报最大化或实现特定目标。它的主要思想是智能体与环境交互和试错，利用评价性的反馈信号实现决策的优化[15]。这也是自然界中人类或动物学习的基本途径。

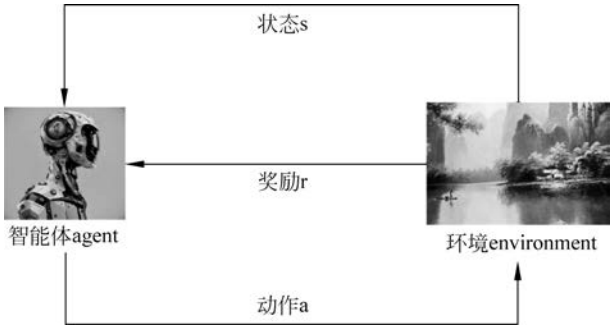

图 3-2　强化学习示意图

强化学习利用反复试错，形成“奖励”和“惩罚”的反馈循环。当算法得到数据集时，它将所处环境视为一场比赛，每次执行动作都会被告知是赢还是输。通过这种方式，它可以创建出一套方案——哪些“动作”能够带来成功，而哪些会造成反效果。DeepMind 的 AlphaGo 和 AlphaZero 都极好地展示了强化学习的威力。只需要把计算机或者机器放到一个环境当中，让它自由探索发挥就可以了。比如让计算机玩赛车游戏，告诉计算机撞到东西就减一分。然后我们就可以看到，智能体在玩赛车游戏时，一开始都是很快就死掉了，但随着失败次数的增多，智能体也能像人一样去总结经验，游戏再次开始的时候做出些调整，经过一次又一次试错，就像我们小时候学走路一样，一次又一次跌倒、爬起，最终学会了走路，智能体学会了控制赛车不会撞到墙。

强化学习是环境对动作的映射学习。判定标准不是现实动作与“理想动作”的差距，而是动作完成后一段时间的累积奖励。优化目标是总体奖励最大。强化学习预先没有给出训练数据，而是通过与环境的时序交互进行不断优化。第一个状态下动作是随机的，动作效果需交互一段时间通过累积奖励来判定。比如下棋需要多次决策直到分出胜负才收到总体的反馈，并从最终的胜负来调整参数。强化学习没有监督者，只有量化奖励信号。强化学习的反馈一般是延迟的，只有进行到最后才知道当下的动作好坏。强化学习属于顺序决策，根据时间一步步决策行为，训练数据不符合独立同分布条件。每一步行动影响下一步状态以及奖励。

近年来，强化学习技术在人工智能、机器学习和自动控制等领域中得到了广泛的研究和应用，并被认为是设计智能系统的核心技术之一。随着强化学习算法和理论的深入，特别是强化学习的数学基础研究取得突破性进展之后，应用强化学习方法实现移动机器人行为对环境的自适应和控制器的优化成为机器人学领域研究和应用的热点之一。

3.1.6　迁移学习

机器学习的理想方案是拥有大量标记的训练集，而大量的数据就意味着要进行大量的计算，这对计算机的配置要求也很高。对大公司而言(如谷歌、微软、脸书等)，他们有着巨大

的计算能力去训练这些大数据模型，而目前对于个人而言，这还是一个难题，花费的成本较大。另外，机器学习的目标是构建一个尽可能通用的模型，使得这个模型对于不同用户、不同设备、不同环境、不同需求，都可以很好地满足。这就需要尽可能地提高机器学习模型的泛化能力，使之适应不同的数据情形。而目前我们所设计的模型大多数是针对于具体的任务。当出现一个新任务时就需要重新设计一个网络并重新进行训练，这将花费大量的财力物力。此外，机器学习已经被广泛应用于现实生活中。在这些应用中，也存在着一些特定的环节，如冷启动问题等。专注于跨领域迁移知识的迁移学习是解决上述问题的一种有效的机器学习方法。实际上，学过自行车的人比未学过自行车的人在学习摩托车上的速度会更快。受人类跨领域迁移知识的能力启发，迁移学习旨在利用来自源域的知识来提高目标域中学习性能或最小化需要标记的训练集样本的数量[16]。

在迁移学习中，把已有的知识叫作源域(source domain)，要学习的新知识叫目标域(target domain)。迁移学习研究的是如何把源域的知识迁移到目标域上。迁移学习还可以细分为基于实例的迁移学习、基于特征的迁移学习、基于关系的特征迁移学习以及基于模型的迁移学习[17]。其中基于实例的迁移学习、基于特征的迁移学习、基于关系的特征迁移学习属于数据层面，而基于模型的迁移学习属于模型层面。从特征空间的迁移情况看，如果特征语义相同，那么就是同构迁移；反之，如果特征完全不相同，那么就是异构迁移。举个例子来说，不同图片数据集之间的迁移学习，可以认为是同构；而图片到文本的迁移学习，则是异构的。

迁移学习让我们在解决问题的时候，不用从零开始训练一个新模型，而可以从类似问题中选择训练过的模型，即基于源域数据构建的预训练模型，来进一步微调优化，最终形成问题解决模型。举例来说，目前需要搭建一个自动驾驶汽车项目，对于图像识别的算法，可能需要花费大量时间从零开始构建一个性能优良的图像识别算法。此时，则可以选择在ImageNet 数据集上训练得到的模型应用于目前的任务上来识别图像。虽然不是 100%符合的，但是可以大大减少从零开始搭建的时间，且精度也有所保证。如果原本的模型用于十分类的任务，而我们需要做的是二分类任务，那么只需要简单修改模型的输出模块即可解决。这样就大大缩短了训练时间，且是目前比较有效的方法。

3.2 人工智能定制化训练服务平台

目前 AI 头部企业，如百度、图灵机器人、小 i 机器人等，将 AI 开发技术开放给全社会使用，纷纷推出自己的人工智能定制化平台，帮助非计算机行业的从业人员快速部署行业专用的人工智能应用模型，从而实现 AI+的普及推广，吸引流量，实现 AI 技术和应用场景的深度绑定。各大 AI 开放平台包括百度 AI、腾讯 AI 开放平台、阿里人工智能、网易人工智能、亚马逊人工智能服务、英特尔人工智能服务、京东 AI、海康威视 AI、Face++ AI、小米小爱 AI、搜狗 AI、OLAMI、讯飞 AI、思必驰 DUI 开放平台、图灵机器人平台 Turing OS、小 i 机器人 Bot 开放平台等。下面列举一些零代码一站式开发的平台服务。

3.2.1 百度人工智能定制平台

进入百度开发者平台 http://ai.baidu.com/，即可看到百度提供的人工智能定制化训练平台，包括 EasyDL 定制化训练平台、iOCR 自定义模板文字识别、语音自训练平台、智能

对话定制平台 UNIT。这些平台实现零算法门槛实现业务定制，为上下游合作伙伴搭建了展示与交易平台——AI 市场，全力帮助各行业高效地实现 AI 升级。百度 AI 平台提供全线产品免费开放，提供最易用的 API、SDK 等开发组件，能帮助用户快速高效地实现产品升级。

EasyDL 是 2017 年 11 月正式上线推出的定制化训练和服务平台，通过可视化的便捷操作，并通过少量数据进行训练，即使零算法基础也可定制高精度的 AI 能力。通过 EasyDL 训练模型，最快 15 分钟即可获得一个高精度模型，可基于不同的定制需求，自动匹配最优的模型算法，少量数据也能获得出色的性能及模型效果。与此同时，在企业实际应用中，存在着网络环境不佳、数据高度保密、需要在设备端实时检测等现实问题。针对以上问题，百度 EasyDL 定制化训练和服务平台率先上线了离线 SDK 服务，用户可以将模型下载，通过离线 SDK 的形式嵌入终端使用。开发者使用 EasyDL 进行模型训练时，可以选择下载 SDK 并集成到本地硬件中（如手机、摄像头等终端设备），从此无论使用环境中有无网络连接，用户都能随时调用自己的 EasyDL 定制化模型。经测试，EasyDL 离线 SDK 在无网络环境下运行，几百毫秒即可完成一次识别，在 NPU 及高通芯片上运行，更可将延时压缩在几十毫秒内。训练完成的模型支持发布为 API 或离线 SDK，灵活适配各种使用场景及运行环境。EasyDL 定制化训练服务平台使得没有 AI 算法研发能力的企业，也可以基于自身业务需求和数据，快速训练专属的定制化 AI 模型，低成本获得高精度的模型服务效果。目前已经在金融、零售、医疗、旅游等众多领域落地应用。

自定义 OCR 依托于业界领先的文字识别技术，结合多项图像处理技术，自定义模板文字识别实现各类票据卡证的自动分类，并通过快至 5 分钟的模板制作实现键值的对应，一步完成分类与结构化。传统 OCR 对于没有制作对应模板的票据、卡证只能按行返回识别结果，自定义模板文字识别可以通过自助的模板制作，建立起键值的对应关系，配合自动分类功能一步完成非结构化到结构化的转换，实现自动化的数据录入。定制好的 iOCR 模板文字可以通过直接调用 API 或使用 SDK 进行私有化部署，部署至客户本地服务器，在客户内网中实现文字识别功能，保障数据私密性。其技术优势有：多场景，能针对各类票据、卡证制作模板、真正实现全场景适配；启动快，仅需 1 张样板，5 分钟即可完成模板自助定制，达到 90％准确率；自动分类，支持各类文档自动分类匹配模板，无需前置人工分类。

语音自训练平台是百度 2019 年 4 月推出的零代码语音识别自助训练平台。用户可上传业务场景音频＋标注文件，由系统自动评估多种语音识别基础模型得到基线准确率并进行模型推荐。支持词汇、长文本等多种训练方式。通过上传的语音标注文件科学评估训练前后准确率对比，直观展示训练提升效果，提供字准、句准、核心词准等多维度评估结果报告。用户可以上传数据多次迭代训练，直至达到预期效果。可以通过自助流程申请模型上线，审批通过后模型自动化上线，目前支持 Android、iOS、Linux SDK 三种方式调用。全流程平台化零代码自助操作，生成业务专属模型。语音自训练平台训练出来的模型，应用场景广泛。比如语音助手，在会议室预订、功能指令等短句语音交互场景中，可通过手机 App 实现智能语音交互，通过训练业务场景所需识别的词汇和句子，提升识别效果，提高流程效率。比如语音输入，及时通讯、订单录入等长句语音识别场景中，可以上传业务场景句子或长段文本语料训练优化识别模型，使长语音录入识别更加精准，服务更流畅。比如智能客服，通过使用特定业务领域语料如运营商业务、金融业务等，训练出更适配业务场景的识别模型，

解决客服对话中的专有名词、人名等识别不准确的问题，提升通话识别准确率。

智能对话定制平台 UNIT 是百度 2017 年 7 月上线的零代码自主对话系统训练平台。实际应用中，从对话核心技术到一个场景化的真实对话系统落地，企业或个人开发者仍然面临着数据标注、知识整理、系统集成的高成本投入。经过不断升级，2020 年 5 月升级的 UNIT3.0 提供了对话核心技术、产品平台，以及生态与服务的全方位的智能对话能力与建设服务，在搭建技能、构建知识和整合技能与知识三方面能力有所提升。企业只需要上传垂直领域场景的业务问题咨询文档，则百度 UNIT 能自动抽取 FAQ，使得用户的问题可以自动找到文档中的对应答案片段。使用端到端的多文档阅读理解模型 V-NET 和自然语言生成技术，技能得以返回更为精准的答案。百度 UNIT 可以部署语音对话模式，更强理解与容错、更低集成成本、更短响应时延，且能智能解析用户沉默、打断和噪声混入等情况。此外还支持 Taskflow 配置管理能力，开发者可在梳理业务流程后，零代码在 UNIT 平台上进行拖曳式配置快速配置对应的对话流程。百度 UNIT 可将平台产生的业务日志转化为新的训练数据，让对话系统效果保持增强学习持续提升性能。百度 UNIT 提供行业知识图谱托管和构建技术，帮助对话机器人提升行业工作的认知能力和理解能力。

3.2.2 图灵机器人平台 Turing OS

图灵机器人 AI 开放平台，是为广大方案商/IP 方/硬件厂商等客户提供的、专为儿童场景设计的、集智能聊天/百科问答/场景对话等服务于一体的机器人管理平台。图灵机器人号称是中文语境下智能度最高的机器人大脑。设置了机器人名称和选择接入方式即可创建机器人。进入该机器人的详情页面，进行各种功能设置，包括能力获取、机器人属性设置、知识库设置、主题对话编写等。然后将自定义的图灵机器人零代码赋予实体设备或微信公众号等虚拟设备智能对话功能。说到操作系统，通常我们会与手机、计算机、游戏机等产品相联系。可以说，操作系统的独立是某一品类的设备开始成型的标志之一。同时也对该品类产品的迅速扩张具有十分重要的意义。2015 年 11 月，人工智能机器人操作系统 Turing OS 正式发布。Turing OS 是一款机器人操作系统，Turing OS 可以全面支持多模态交互，让机器人具备最自然的人机交互方式，让机器人具备人类思维能力、情感能力以及学习能力等。Turing OS 有三大引擎：情感计算、思维强化、自学习。情感计算，就是让机器人具备人类的情感，Turing OS 能够支持 25 种语言类情感的理解和识别，识别准确率高达 95.1%。同时，情感计算引擎可以支持 468 种情感语言，120 种声音语调，88 套表情动作，基本上可以做到一个 7 岁小孩的肢体动作。全新研发的情感 TTS 引擎，机器人说话的方式更自然。图灵机器人团队一次性发布了数十款机器人应用，均基于在 Turing OS 1.5 上进行的开发，覆盖教育、娱乐、社交、工具以及远程五大类。其中智能聊天、智能音乐、智能拍照、智能英语四款应用，搭载了这四款应用的机器人可化身小伙伴、音乐舞蹈家、摄影师及英语外教，颠覆了现场观众对传统聊天、音乐、拍照及英语应用的一贯认知。除此之外，金山英语、贝瓦儿歌、墨迹天气等知名互联网也联合发布了旗下产品的机器人版本。

3.2.3 小 i 机器人 Bot 开放平台

小 i 机器人 Bot 开放平台帮助企业和开发者轻松、零代码快速搭建自己的对话机器人系统，自主开发具有自然交互能力的智能服务系统。Bot 开放平台为开发者量身定制的免

费 AI 工具平台，集成 Chatting Bot、FAQ Bot、Discovery Bot 三大核心能力以及自然语言处理、深度学习、语音识别与合成、图像识别、大数据分析等基础技术。其中 Chatting Bot 以庞大通用聊天库为支持，实现天气、百科、地图、餐饮、汇率、文字游戏、车票等数十项常用的百科式应答信息服务。FAQ Bot 允许用户配置常见问答库，多种问法自动泛化，带来更强泛化能力和更高准确率，并可通过实时干预方式达到监督式问答。同时，结合对问答、文档等数据的智能学习机制形成完整闭环，达到能力的快速迭代、不断提升。Discovery Bot 实现基于文档的知识抽取应答，能让机器对线下文章、线上信息等不同形式的非结构化数据进行理解并能够回答相关问题。对于企业需求最为迫切的智能客服，平台支持该场景的直接使用。用户在建设完知识后只需激活在线客服功能并做一些简单配置，即可让智能客服能力运转起来。开发者可以通过 API 快速扩展应用渠道（支持从微信、网页向短信、App 等所有电子渠道及实体终端的扩展），同时支持多种应用扩展，开发者可以将自己独立开发的应用集成到开放平台中进行整合使用，丰富自身功能。一个强大的地方在于，在智能客服场景中，机器人覆盖大量的常规问题，当机器人无法应答时，则可以无缝转人工，实现机器人与人工自由转换和协作。同时平台提供的智能学习能力可以从日常对话、文档阅读中自动挖掘构建知识库，在机器人、人工、知识构建之间形成完整使用闭环，这种持续正向循环的自学习能力能够保证智能系统在商业应用中越来越聪慧。

3.2.4　思必驰 DUI 开放平台

2017 年 7 月思必驰 DUI 开放平台首次亮相，是国内第一个真正意义上的超高度定制平台，每个模块均可自定义，例如，GUI 自定义、唤醒词定制、技能深度定制。既提供通用的场景对话和内容技能，也支持开发者完整自定义对话逻辑和内容，接入第三方服务，提供细致到每一轮交互的个性化定制，让开发者能够依据产品特性，定制最适合的语音功能。在语音合成方面，思必驰 DUI 开放十余种个性化合成音，支持快速定制合成等，以此增强产品的个性化特征。思必驰 DUI 让非专业的开发者也能迅速上手，帮助厂商自定义产品功能、系统迭代、内容升级等方面的开发，更好地将智能语音赋能给开发者和企业。思必驰将通过 DUI 开放平台把核心技术能力开放给开发者们，打造一站式的对话系统定制开发平台，覆盖了语音识别、语音合成、语义理解、智能对话、多轮交互、纠正打断、回声消噪等，既提供单项技术服务，也提供完整的智能对话交互解决方案。思必驰 DUI 开放平台既提供云端 API 也提供本地 SDK，开发者通过 DUI 开放平台，可在云端对内容、引擎、对话逻辑、热词等进行实时更新，同时本地的语音唤醒、离线识别、离线技能等则支持在网络不佳甚至是无网络的情况下，保证产品的基本使用。思必驰 DUI 开放平台专业而庞大的“技能商店”则能够覆盖不同场景的多样需求，既支持拨打电话、闹铃设置等本地技能，也支持第三方服务资源接入，并不断进行丰富与完善。

3.2.5　华为 AI 开发平台 ModelArts

ModelArts（https://www.huaweicloud.com/）是面向开发者的一站式 AI 开发平台，为机器学习与深度学习提供海量数据预处理及半自动化标注、大规模分布式 Training、自动化模型生成，及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期 AI 工作流。ModelArts 提供从数据、算法、训练、模型、服务全流程可视化管理，无须人工干预。

模型训练完成后，ModelArts 在常规的评价指标展示外，还提供可视化的模型评估功能，可通过混淆矩阵和热力图形象地了解模型，进行评估模型或模型优化。ModelArts 具有自动学习能力，可根据用户标注数据全自动进行模型设计、参数调优、模型训练、模型压缩和模型部署全流程。无须任何代码编写和模型开发经验，即可利用 ModelArts 构建 AI 模型应用在实际业务中。ModelArts 将自动进行模型超参数的最优化分析和搜索，跳过耗时的手动调整模型参数过程，取代多次模型训练，节省数日甚至数周的时间，从而最大限度提升模型的质量。ModelArts 自动学习大幅降低 AI 使用门槛与成本，较之传统 AI 模型训练部署，使用自动学习构建将降低成本 90％以上。目前，ModelArts 支持图片分类、物体检测、预测分析、声音分类 4 大特定应用场景，可以应用于电商图片检测、流水线物体检测等场景。

3.2.6 腾讯云 AI 平台服务

面对非算法人员，腾讯打造了智能钛一站式机器学习（TI-ONE）、智能对话平台 TBP 等一站式服务平台，为用户提供从数据预处理、模型构建、模型训练、模型评估到模型服务的全流程开发及部署支持。智能钛机器学习平台内置丰富的算法组件，支持多种算法框架，满足多种 AI 应用场景的需求。

智能钛一站式机器学习平台 TI-ONE 使用图形化操作，支持拖拽式任务流设计让 AI 初学者也能轻松上手。使用方便，帮助 AI 初学者和希望借助 AI 解决业务问题的非专业用户零代码一键部署。只需要经过数据上传、数据预处理、选择算法、评估模型等简单几步就可以完成机器学习模型构建。基于学习框架，小到 Python，大到当红的 Spark、TensorFlow、xgBoost（DMLC），以及自研 Mariana 和 Angel，智能钛一站式机器学习平台支持 10 多种机器学习框架，同时自身带有 100＋的机器学习算法，能够随时满足用户所需。借助强大的 Python/Spark，集结流行的可视化效果，数据可以直接悬浮呈现，用户可以免单击，对于模型优劣一眼立断。TI-ONE 使业务专家无需具备编程或深度学习的专业知识就可以创建出业务模型，降低 AI 应用开发成本。目前平台已接入人脸识别、行为识别、车辆识别、工业质量检测等 20＋AI 算法，更多算法服务持续上架中。

腾讯智能对话平台 TBP 专注于“对话即服务”的愿景，全面开放腾讯对话系统核心技术，为开发者和生态合作伙伴提供开发平台和机器人中间件能力，实现开发者便捷、高效、低成本构建人机对话体验。平台提供对话应用平台化开发工具，满足不同类型开发者需求。TBP 提供强大的机器人中间件能力。开发者可自行在平台上定义 Task-based Bot（任务型机器人）语义模型和 QnA Bot（问答型机器人）问答集合，或直接调用平台内建机器人能力。提供多样化服务配置能力。开发者灵活配置后端业务逻辑；后续平台提供云函数能力，为机器人开发降低门槛。开发者可通过网页模拟器来对机器人进行自然语言理解和业务逻辑测试。提供多渠道应用集成能力，可使开发者大幅度减少多平台开发的工作量，实现开发者轻松将其开发完成后的机器人集成到移动 App、网站、IoT 设备等多终端；并与微信公众号运营平台进行打通，支持零代码接入微信公众号。

3.3　百度人工智能定制平台应用开发实例

3.3.1　基于 EasyDL 平台的图像分类

一般来说，机器学习是一种流水线工作，其完整流程包含的环节有 5 个。①数据获取：从数据源接入数据，用于后续的算法训练。②数据预处理：如缺失值处理、数据格式处理。③特征提取：在原始数据的基础上提取出会对训练结果有帮助的值的过程，这个过程很大程度上依赖建模人员的经验，尤其是对业务的理解。④特征选择：选择好的特征，加入到模型训练过程中。⑤选择算法：模型的搭建可以选择不同的方法，如随机森林算法、决策树算法、支持向量机算法等。而在深度学习（Deep Learning，DL）框架下，特征选择过程可以通过神经网络来完成。EasyDL 是深度学习框架一站式平台服务，其核心技术包括 AI Workflow 分布式引擎、百度自创 PaddlePaddle 深度学习框架、迁移学习、Auto Model Search 机制、Early Stoopping 机制、模型效果评估机制等[18]。

1. 问题描述

手写识别是常见的图像分类任务。不同人的手写体风格迥异、大小不一、造成了计算机对手写识别任务的一些困难。计算机通过处理手写数字图片，识别出该图片对应的数字。由于共有 0～9 共 10 个数字，因此手写识别数字是 10 分类问题。

2. 实现步骤

只需四步即可使用 EasyDL 专业版完成 AI 模型的训练，以及将模型以云端 API 形式发布并集成在自有网站中使用。

Step1：成为百度 AI 开放平台的开发者

要调用百度 EasyDL 定制化图像分类能力，先要成为百度 AI 开放平台的开发者。注册百度账户进入如图 3-3 所示的页面快速建立一个百度账号。

图 3-3　快速建立一个百度账号

Step2：提前准备训练数据

采用手写体图片开源数据集 MNIST(http://yann.lecun.com/exdb/mnist/)，包含数字 0～9 的手写图片。MNIST 由训练集与测试集两个部分组成，其中训练集包含 60 000 幅手写体图片及对应标签；测试集包含 10 000 幅手写体图片及对应标签。其中，每一张图片已归一化为以手写数字为中心的 28×28 规格的图片。每个像素点的值区间为 0～255，0 表示白色，255 表示黑色。MNIST 手写体数字图片如图 3-4 所示。

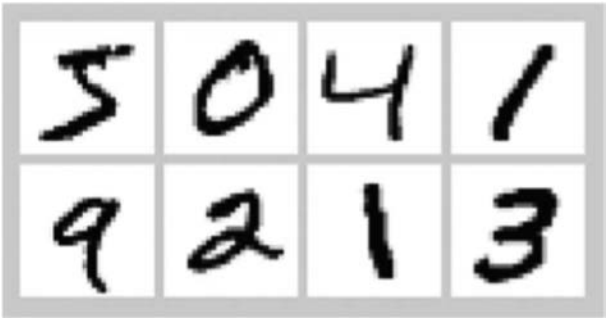

图 3-4　MNIST 手写体数字图片示例

理论上训练数据越多，训练出来的模型效果就越好。EasyDL 建议图像分类的每一类图片数据理不低于 20 张。

Step3：使用 EasyDL 专业版训练图像分类模型

使用百度开发者账号在 EasyDL 官网登录，并且进入专业版模型定制操作页面，创建项目模型，填写项目模型信息，并且选择图像分类服务，如图 3-5 所示。

图 3-5　创建 EasyDL 项目模型

接下来准备上传用于训练的图片，单击菜单“创建数据集”创建一个数据集，命名为“手写数字分类”，接下来针对刚创建完的数据集上传训练图片，如图 3-6 所示。可以分别提交每一张图片，也可以提交压缩包，压缩包中每个类别的图片放在独立文件夹里面。

本项目以 3 分类为例作为展示，每个类别选择 45 张图片。上传后的训练集如图 3-7 所示。

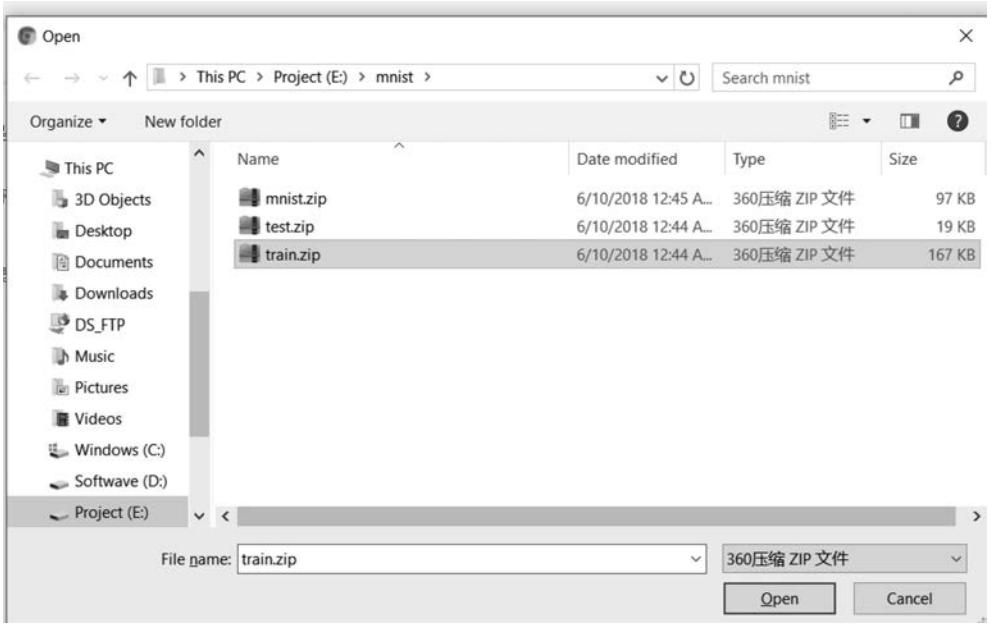

图 3-6　上传训练图片集

No.	名称	图片数
1	2	45
2	1	45
3	0	45

当前数据累计成功上传135张图片，共计3个分类

图 3-7　训练图片上传完成示意图

上传完成后对每一张进行图片的标注，分别指定属于哪个数字。接下来，使用标注好的训练图片集来开始训练。单击左上角的“新建任务”任务新建一个高级版任务，如图 3-8 所示。然后选择高级版特有的模型网络，这里可以选择不同的深度学习架构进行训练。深度学习架构主要是当前流行的深度人工神经网络模型[19]，如 Alexnet 和 Resnet。EasyDL 具有深度学习网络结构的自动搜索和设计。第一步从头开始搜索神经网络架构，即神经架构搜索。第二步进行神经网络的自动适配，也就是说根据不同的任务目标与需要进行适配，比如说目标是部署模型到移动端，那么系统就需要考虑正确率、参数量和计算量等条件来适配网络。第三步是设计网络的迁移能力，使得搜索出的神经网络架构具有强大迁移能力，这样训练出来的模型不仅针对特定数据有强大能力，又可迁移其他更多的数据。

可以尝试选择多个不同的深度学习架构，每次都会生成一份专属的 Python 格式的配置代码，如图 3-9 所示。这些代码可以在网页编辑器中直接编辑。为了确保训练成功，我们必须遵循页头注释代码的注意事项。

在模型架构设计完成后，就可以开始执行训练任务。需要等待一段时间才能完成模型的训练，如图 3-10 所示。

图 3-8 选择训练数据的深度学习架构

图 3-9 训练模型的 Python 配置代码

图 3-10 执行模型训练任务

训练完成后校验模型，如图 3-11 所示，即检验一下所训练出来的模型对新数据的识别能力。

图 3-11 执行模型校验

直接把 MNIST 测试集中的图片拖动上传到校验窗口，如图 3-12 所示。

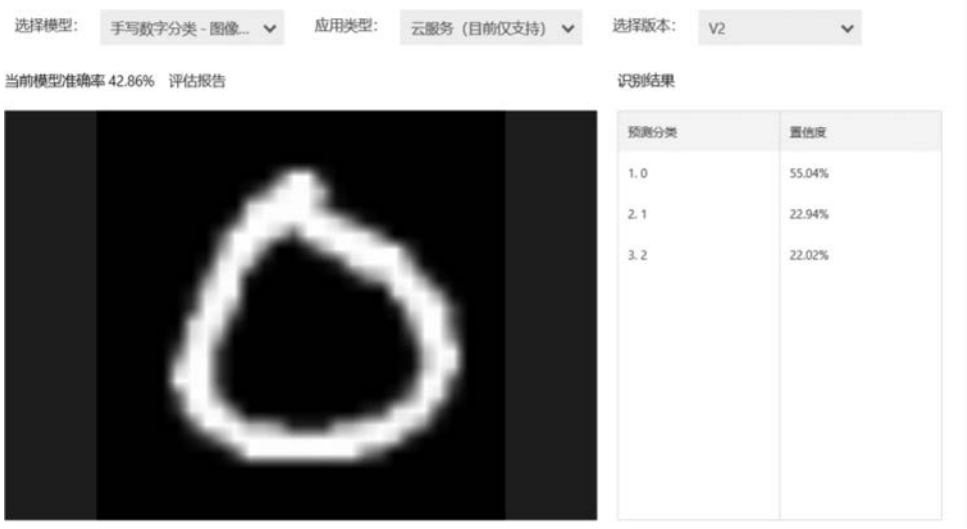

图 3-12　EasyDL 模型校验

在校验界面中，右侧显示置信度，也就是模型认为图像所属类别的可能性。图 3-12 中被测图像为 0，而右侧 0 的置信度最高，表示训练出来的模型对该测试图片识别成功。

也可以单击训练好的模型，查看详细的模型效果指标，如图 3-13 所示。

图 3-13　EasyDL 模型效果指标展示

可以看到，对于数量为 35 的测试集，模型预测正确 15 个，预测错误 20 个，模型准确率为 42.9%，F1-score 为 20.0%，精确率为 14.3%，召回率为 33.3%。模型的整体评估质量

不高，其原因主要是因为本项目提供的训练样本太少，模型欠拟合导致[20]。后续可以增加大量样本进行训练提高精确度。接下来进一步展示每个类别的分类性能，以 F1-score 为指标值，结果如图 3-14 所示。

图 3-14　EasyDL 模型类 F1 值分布

我们可以先选择发布云服务或离线服务将训练好的模型进行发布。这里我们选择发布云服务。图 3-15 给出了 EasyDL 模型训练完成的截图。

图 3-15　EasyDL 模型训练完成

发布云服务需要提供一个在线服务地址，则可以把训练好的模型部署在该云服务上。部署信息如图 3-16 所示，包括接口输入字段，如本项目中的输入字段为 image，即把一张 image 图像输入到云服务器上的 EasyDL 模型，就执行了测试数据的输入。此外部署信息还包括模型"响应字段"，即 EasyDL 模型输出测试结果信息集合{log_id，error_code，err_msg，results，+name，+score}。

部署完成值，会生成在线服务调用的接口地址，如图 3-17 所示。后续可以在第三方应用中使用该接口进行 EasyDL 模型的调用服务。下面进一步展示第三方应用是如何调用 EasyDL 模型的。单击 EasyDL 模型在云服务器上的接口配置页面中的"立即使用"，跳转到百度云控制台创建一个应用。

填写信息创建应用完成后，我们需要查看 API KEY 以及 Secret KEY。这些信息需要在第三方应用中进行配置，以便能通过接口调用 EasyDL 模型。至此，完成了模型的创建训练工作。

图 3-16　EasyDL 模型在云服务器的部署设置和信息

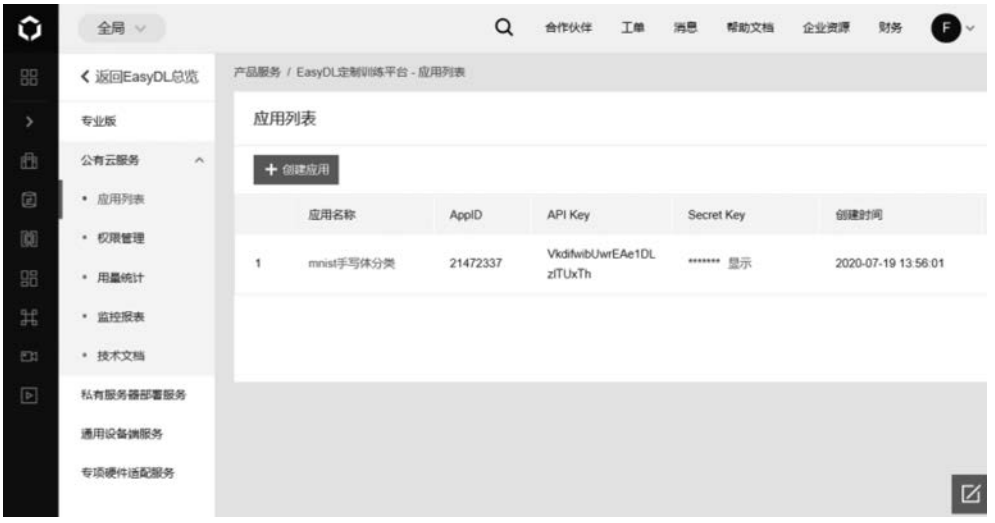

图 3-17　百度开发者平台的应用创建主页

3.3.2　基于 UNIT 平台的智能客服机器人构建

本项目使用百度 UNIT(https://ai.baidu.com/unit/home)构建对话机器人,能为客户完成天气查询的服务。构建流程图,如图 3-18 所示。

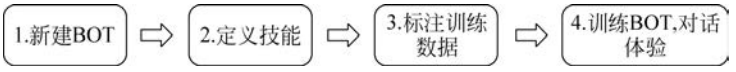

图 3-18　流程图

1. 新建 BOT

BOT 是英文 ROBOT(机器人)的简写。在 UNIT 主页新建一个 BOT,名称设置为“查

天气”，如图 3-19 所示。

图 3-19　新建 UNIT 机器人 BOT

2. 定义技能

单击进入刚刚创建的 BOT，在“技能”→“自定义技能”模块下“新建对话意图”设置意图名称为 WEATHER，意图别名为“查天气”，如图 3-20 所示。

图 3-20　设置 UNIT 机器人的技能

想让机器人反馈天气信息，需要输入时间和地点两个信息给机器人。因此需要添加“时间”和“地点”两个词槽。在技能设置画面单击“添加词槽”则可以逐个进行词槽的配置，如图 3-21 所示。

添加时间词槽：user_time，别名：时间，如图 3-22 所示。

接下来需要给时间词槽选择自然语言表达集合，即含有时间语义的自然语言表达模式词典。该词典可以由百度 UNIT 系统提供，也可以由用户自建。打开词典配置页的“系统词槽词典”开关，选择 sys_time(时间)词典，则配置了 UNIT 系统预置的时间词槽词典，如图 3-23 所示。

图 3-21　UNIT 机器人的词槽配置

图 3-22　为查天气 BOT 添加“时间”词槽

当用户向机器人查询天气时，必须输入时间信息，这就需要把时间词槽与查天气意图绑定起来，即时间词槽设为必填。并且预置当机器人向用户索取时间信息时，机器人的表达模式，即话术。话术是用户自定义的，可以设计符合应用场景和人机交流的表达模式。在“澄清话术”项目中输入机器人话术“请澄清一下：时间”，如图 3-24 所示。

接下来继续添加词槽：user_loc，别名：哪里，如图 3-25 所示。

接下来需要给“地点”词槽选择自然语言表达模式词典。选择 UNIT 系统预置的地点词槽词典，sys_loc(地点)，如图 3-26 所示。

将地点词槽和查天气意图进行绑定，并设置机器人向用户索要地点信息时的话术，设计为“请澄清一下：哪里”，如图 3-27 所示。

图 3-23　选择查天气 BOT 时间词槽的词典

图 3-24　绑定查天气意图和时间词槽的关联

图 3-25　设置查天气 BOT 的“地点”词槽

图 3-26　选择 UNIT 系统预置的地点词槽词典

至此，查天气 BOT 的两个词槽，即天气和地点，配置完毕。接下来需要设置查天气 BOT 的服务反馈，即当查天气 BOT 向用户提供查天气服务时输出的服务形式。服务形式可以是输出反馈文本信息，或执行某个功能。本项目设置查天气 BOT 服务反馈形式为文本“正在为您查询天气……”，即当用户询问天气，BOT 向用户索取到时间和地点两个信息

图 3-27　绑定查天气 BOT 地点词槽和查天气意图

后，机器人反馈给用户的结果是“正在为您查询天气……”，如图 3-28 所示。

图 3-28　设置查天气 BOT 的服务反馈形式

至此，查天气 BOT 的查天气技能设计完毕。

3. 标注训练数据

在查天气技能设计完毕后，需要把该技能赋能为查天气 BOT。对于查天气 BOT，它可能具有多项技能，并且在人机交流过程可能有多种情境。因此，在百度 UNIT 中，把某项技能赋能称为模型训练。赋能过程的相关设置，则为标注训练数据。在查天气 BOT 主页中进入“效果优化”→“训练数据”→“对话模块”页面，如图 3-29 所示。

图 3-29　查天气 BOT 人机对话模型的设置主页

设置人机对话模型的模板，即添加对话模板，如图 3-30 所示。

图 3-30　设置查天气 BOT 对话模型的模板 1

具体步骤如下。①添加对话模板→②选择意图：WEATHER→③在第一个模板片段中插入词槽：user_time→④把该片段“必须匹配”设为“否”(因为问天气不一定一开始就会说“时间”)→⑤新增一行。

⑥ 在第 2 个模板片段插入词槽：user_loc→⑦把“必须匹配”设为“否”(因为问天气也不一定一开始就会说“地点”)→⑧新增一行，如图 3-31 所示。

⑨ 在第三个模板片段中写入：天气，因为不论怎么问天气，第一句话里必须包含“天气”二字才能确定是问天气的意图，所以“必须匹配”默认“是”就可以。⑩把阈值从 0.7 调成 0.4，则完成查天气 BOT 对话模型的模板配置，如图 3-32 所示。

图 3-31　设置查天气 BOT 对话模型的模板 2

图 3-32　设置查天气 BOT 对话模型的模板 3

4. 训练 BOT，对话体验

查天气 BOT 人机对话模型需要进行训练才能投入使用。训练查天气 BOT 的人机对话模型的目标是让 BOT 能正确地识别出人机对话中的查天气意图，填充“时间”和“地点”词槽。词槽填充指的是从用户的自然语言文本中抽取时间信息和地点信息，以实现查天气的前提信息。单击查天气 BOT 主页右上角的【训练】按钮，进入【模型管理】页面，单击【训练模型并生效到沙盒】，如图 3-33 所示。

在图 3-33 的弹框中，直接单击“训练模型并生效到沙盒”，则开始进行训练，并逐步展示训练进度：初始化→训练中→训练完成。查天气 BOT 沙盒的状态也会相应进行变化：空→运行中→模型生效中。模型版本 v1，如图 3-34 所示。

图 3-33　启动查天气 BOT 人机对话模型的训练

图 3-34　完成查天气 BOT 人机对话模型的训练

这时查天气的 BOT 就已经创建并训练完成了，单击右上角的【测试】可以在网页端开展查天气 BOT 的对话理解效果，如图 3-35 所示。

图 3-35　查天气 BOT 的测试效果

如图 3-35 所示，当用户发言“明天北京天气如何?”，查天气 BOT 从发文文本中识别出查天气意图，并且获取到“时间”词槽的值是“明天”，“地点”词槽的值是“北京”。在意图识别前提下填满所有词槽(“时间”和“地点”两个词槽)之后，查天气 BOT 进行反馈服务，即输出文本“正在为您查询天气”。如果想让查天气 BOT 机器人播报出天气信息，则可以把查天气 BOT 打包部署，以 API 形式在第三方应用中进行调用。

3.4　人工智能模型性能评估

评估是模型开发过程中不可或缺的一部分，它有助于发现表达数据的最佳模型和所选模型将来工作的性能如何。以大数据为支撑的人工智能模型原理是机器学习，即对于给定的训练样本构建出一个预测模型。训练数据不仅有属性信息还有标签信息。在测试环节，模型的输出与真实标签信息之间的误差就非常自然成为模型评估的指标。虽然机器学习模式具有多样性，如监督学习、无监督学习、半监督学习、强化学习、迁移学习等，但在训练环节中数据都是有标签信息的。也就是说，即使是无监督信息，训练数据也需要附带标签信息。无监督模型训练时不使用标签信息，但测试环节需要标签信息进行评估。基于机器学习的人工智能模型评价目标是最小化泛化误差，即模型在新数据上的预测误差最小[4]。

根据数据标签值域(离散取值或连续取值)的不同，机器学习模型可分为分类模型(对应离散型数据标签)和回归模型(对应连续型数据标签)。分类模型和回归模型会提出各自特定的评估指标。另外，不同应用场景也可能需要一些特定的评估指标，比如自然语言处理任务中的 BLEU 指标。下面先介绍通用泛化误差(准确率)指标的定义和计算方法，然后分别从学习模式和学习任务两个视角列举对应的特定评估指标。

3.4.1　通用泛化误差指标的计算

机器学习理论指出，训练数据由高维向量 $\boldsymbol{x}=(x^1,x^2,\cdots,x^d)^{\mathrm{T}}\in\boldsymbol{X}$ 表示，其中数据 x 的上标 $1,2,\cdots,d$ 是向量的维数序号，也就是对象的属性序号；$\boldsymbol{X}=\mathbb{R}^d$ 表示 d 维实数空间。以监督学习模式分类问题为例，假设实际问题对应有限 c 个可选类别，用标识变量 $y\in\boldsymbol{Y}=\{1,2,\cdots,c\}$ 表示对象的类别，其中 $\boldsymbol{Y}$ 叫作类别空间。对象及其类别构成数据对，$z=(\boldsymbol{x},y)\in\boldsymbol{Z}$，$\boldsymbol{Z}$ 叫作样本空间，$z_i(i=1,2,\cdots,n)$ 是样本空间中的样本(点)，其下标“i”表示样本编号。根据给定的训练样本集 $Z^{(n)}=\{z_1,z_2,\cdots,z_n\}$，其中 $Z^{(n)}$ 表示含有 n 个样本的样本集，$z_i\in\boldsymbol{Z}$，$i=1,\cdots,n$ 表示样本，则监督学习模式下的训练模型将产生一个分类器 $\varphi:\mathbb{E}^d\to\{1,2,\cdots,c\}$，使得它对新的待测试数据 $x_{n+1},x_{n+2},\cdots,x_{n+k}$ 的实际类别值产生相应的预测值 $\hat{y}_{n+1},\hat{y}_{n+2},\cdots,\hat{y}_{n+k}$。我们要求该分类器要能对整个样本空间可能的分布有一个尽可能小的期望判别误差，这就是泛化误差[3]。机器学习一般采用测试错误率估计模型的泛化误差。测试错误率是基于特定的数据集计算出来，本质上是条件泛化误差。条件泛化误差(测试错误率)缺少理论支持，但是在实践中是直观有效的。下面介绍几种在实践中广泛采用的测试错误率计算方法。

(1) 再代入估计(resubstitution)：它假设训练数据集可以很好地代表整体数据，因而，可以使用训练误差(又称为再代入误差)提供对泛化误差的乐观估计。

$$\hat{\varepsilon}_n^R=\frac{1}{n}\sum_{i=1}^{n}|y_i-\varphi_n(x_i)|$$

(2) 保持估计(holdout estimation)：它将有类别的原始数据集划分成两个不相交的子集，分别称为训练集和检验集。在训练数据集上归纳分类模型，在检验集上评估模型的性能。训练集和检验集的划分比例通常根据研究者的判断。分类器的泛化性能由检验集上错误率来估计。记保持误差估计为 $\hat{\varepsilon}_{n,m}^{H}$，其中 m 表示检验样本集的大小。

(3) 随机二次抽样(random sub-sampling)：它采用多次重复使用保持估计方法来改进对分类器性能的估计。设 $\hat{\varepsilon}_{n_i,m_i}^{\mathrm{RS}}$ 是第 i 次迭代的模型误差估计量，则模型的总的误差估计量为 $\hat{\varepsilon}_{n+m}^{\mathrm{RS}}=\sum_{i=1}^{k}\hat{\varepsilon}_{n_i,m_i}^{\mathrm{RS}}/k$。

(4) 交叉验证(cross-validation)：它是一种替代随机二次抽样的方法，这个技术保证每个样本用于训练的次数相同，并且用于检验恰好一次。对于流行的 K 折交叉验证，首先将原始样本分为大小相同的 K 份。在每次运行时，选择其中一份作为检验集，而其余的全部作为训练集。该过程重复 K 次，使得每份样本子集都用作检验集恰好一次。总误差是所有 K 次运行的误差之和，取平均误差作为泛化误差的估计量。

$$\hat{\varepsilon}_{n,k}^{CV}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n/k}\left|y_j^{(i)}-\varphi_{n,i}(x_j^{(i)})\right|$$

其中，$\varphi_{n,i}(\cdot)$是第 i 次迭代时由训练集构建的分类器。

K 折交叉验证方法的一种特殊情况是令 $k=n$，其中 n 是原始样本集的大小，这种情况叫作留一法(leave-one-out)。这种方法使得每一个检验集只有一个样本，因此能使用尽可能多的样本用于训练。

(5) 自助法(bootstrap)：它是一种一般化的重采样技术，它采用一个经验采样分布 F^* 来取代原始样本集中样本的原始分布 F。经验分布一般是离散分布，针对每一个样本有一个采样概率。常用的自助方法是，对原始样本集 $Z^{(n)}$ 采用有放回抽样，即已经选做训练的样本将放回到原始的样本集中，使得它等可能地被重新抽取，即每一个样本的采样概率为 $\frac{1}{n}$。由于原始样本集的大小为 n，自助集的样本个数一般与之相同，也为 n，自助样本集记为 $Z^{(n)*}$。可以证明，平均来说，大小为 n 的自助样本集(训练样本集)大约包含原始样本集中 63.2%的样本。这是因为一个样本被自助抽样抽取的概率是 $1-\left(1-\frac{1}{n}\right)^n$。当 n 充分大的时候，该概率逐渐逼近 $1-\mathrm{e}^{-1}=0.632$。没有抽中的样本就称为检验集的一部分，将训练集建立的模型 φ_n^* 应用在检验集上，得到基于自助样本构建的分类器性能的一个估计。这个误差估计称为基本自助误差估计，记为：

$$\hat{\varepsilon}_n^{\mathrm{BOO}}=E_{F^*}\left[\left|y-\varphi_n^*(\boldsymbol{x})\right|:(\boldsymbol{x},y)\in Z^{(n)}-Z^{(n)*}\right]$$

如果抽样过程重复 B 次，则实验产生 B 个自助样本集(训练集)，如何估计由这些训练集构建的分类器的泛化误差，有两个不同的方法。第一种方法称为全局法，它融合 B 次检验结果，计算出总的误差率用做泛化误差的估计量。这时，自助误差估计量为：

$$\hat{\varepsilon}_n^{\mathrm{BOO1}}=\frac{\sum_{b=1}^{B}\sum_{i=1}^{n}\left|y_i-\varphi_{n,b}^*(x_i)\right|I_{(x_i,y_i)\in z^{(n)}-z_b^{(n)*}}}{\sum_{b=1}^{B}\sum_{i=1}^{n}I_{(x_i,y_i)\in z^{(n)}-z_b^{(n)*}}}$$

其中 $I_{(.)}$ 为指示函数。第二种方法称为样本法，它单独考虑每个样本的预测情况，然后将所有样本的预测错误率相加得到泛化误差的估计量。这时，自助误差估计量为：

$$\hat{\varepsilon}_n^{\mathrm{BOO2}}=\frac{\sum_{i=1}^{n}\sum_{b=1}^{B}|y_i-\varphi_{n,b}^{*}(x_i)|I_{(x_i,y_i)\in Z^{(n)}-Z_b^{(n)*}}}{n\sum_{b=1}^{B}I_{(x_i,y_i)\in Z^{(n)}-Z_b^{(n)*}}}$$

尽管采取了多次重复自助抽样，但是每次用于构建分类器模型的训练样本仅为原始样本中的 63.2%。为了克服上述缺点，一种改进的方法是加权组合自助法和再代入法的误差估计量，这种方法称为 0.632 自助法(0.632 bootstrap)，其估计量为：

$$\hat{\varepsilon}_n^{\mathrm{B632}}=(1-0.632)\hat{\varepsilon}_n^{\mathrm{R}}+0.632\hat{\varepsilon}_n^{\mathrm{BOO}}$$

另外还有一种改进的方法，叫作偏差矫正自助法误差估计，强调对再带入估计的矫正，这种估计量定义为：

$$\hat{\varepsilon}_n^{\mathrm{BOOC}}=\hat{\varepsilon}_n^{\mathrm{R}}+\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\left(\frac{1}{n}-P_{i,b}^{*}\right)|y_i-\varphi_{n,b}^{*}(x_i)|$$

上述总结的各种误差率都是基于特定的数据集计算出来的，它与该数据集的规模和数据质量有关。不同的数据集上计算出来的误差率可能存在差异，也就是对未来新出现的待测数据判断错误的概率估计是不稳定的。基于特定数据集的误差率，是一种计算方法，存在着有偏性的困扰。

3.4.2　分类模型的评估指标

以两分类问题为例，实际问题有 2 个可选类别，即标识变量 $y\in Y=\{1,2\}$。假设标识变量 $y=1$ 称为真或阳性标签，标记为 P(positive)，标识变量 $y=2$ 称为假或阴性标签，标记为 N(negative)。下面列举分类模型相关的评估指标。

1. 混淆矩阵

混淆矩阵(confusion matrix)显示了分类模型相对数据的真实输出(目标值)的正确预测和不正确预测数目，可以帮助人们更好地了解分类中的错误[21]。混淆矩阵是一个方阵，$c\times c$，其中 c 为标签类别数。表 3-4 为两个类别(阳性和阴性)的 2×2 混淆矩阵。

表 3-4　混淆矩阵和指标计算

实际预测		“金标准”结果		合计	PPV	FDR
		阳性	阴性			
诊断结果	阴性	真阳性(TP)	假阴性(FP)	prediction positive = (TP+FP)	PPV=TP/prediction positive 又被称为正确率(precision)	FDR=FP/prediction positive
	阳性	假阳性(FN)	真阴性(TN)	prediction negative = (FN+TN)	FOR = FN/prediction positive	NPV=TN/prediction positive
合计		condition positive = (TP+FN)	condition negative = (FP+TN)	N = TP + FN + FP+TN	FOR	NPV

续表

实际预测	“金标准”结果		合计	PPV	FDR
	阳性	阴性			
合计	真阳性率 TPR = TP/ condition positive 又被称为灵敏度(sensitivity)，召回率(recall)	假阳性率 FPR = FP/condition negative 又被称为误诊率 = 1－特异度			
	假阴性率 FNR = FN/ condition positive 又被称为漏诊率 = 1－灵敏度	真阴性率 TNR = TN/condition negative 又被称为特异度(specificity)			

表 3-4 中的相关术语和评估指标解释如下：

(1) TP(true positive)指的是真阳性，正确的肯定，又称：命中 (hit)。

(2) TN(true negative)指的是真阴性，正确的否定，又称：正确拒绝。

(3) FP(false positive)指的是伪阳性，错误的肯定，又称：假警报 (false alarm)、第二型错误。

(4) FN(false negative)指的是伪阴性，错误的否定，又称：未命中(miss)、第一型错误。

(5) 灵敏度(sensitivity)或真阳性率(true positive rate，TPR)：又称：召回率(recall)、命中率(hit rate)，在阳性值中实际被正确预测所占的比例。

(6) TPR＝TP/P＝TP/(TP＋FN)指的是伪阳性率(false positive rate，FPR)：又称错误命中率，假警报率(false alarm rate) 。FDR＝FP/(FP＋TP)＝1－TPR。

(7) SPC＝TN/N＝TN/(FP＋TN)指的是特异度(specificity，SPC)或真阴性率(true negative rate，TNR)：在阴性值中实际被正确预测所占的比例。

(8) FPR＝FP/N＝FP/(FP＋TN)＝1－SPC 指的是假发现率(false discovery rate，FDR)。

(9) 准确度(accuracy，ACC)：正确预测的数占样本数的比例。ACC＝(TP＋TN)/(P＋N)。

(10) 阳性预测值(positive predictive value，PPV)或精度 (precision)：阳性预测值被预测正确的比例。PPV＝TP/(TP＋FP)。TP＋FP 代表无论真与假，报出来数据都是正样本。

(11) 阴性预测值(negative predictive value，NPV)：阴性预测值被正确预测的比例。NPV＝TN/(TN＋FN)。

(12) F1 评分：精度和灵敏度的调和平均数。F1＝2 precision×recall/(precision＋recall)＝2TP/(2TP＋FP＋FN)。

(13) Matthews 相关系数(MCC)，即 Phi 相关系数：(TP×TN－FP×FN)/sqrt{(TP＋FP)(TP＋FN)(TN＋FP)(TN＋FN)}。

2. ROC 曲线

混淆矩阵关注是的模型总体上的性能，而 ROC 曲线关注的则是模型性能的分布规律[22]。ROC 重点关注伪阳性(FPR)和真阳性(TPR)的分布情况。分类器模型通过计算数

据对应的标签为真的概率进行模式判别。计算概率值与阈值相减，大于阈值的标签则获得输出。针对某个特定阈值，可以计算出一组(FPR，TPR)。随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着真正的负实例，即 TPR 和 FPR 会同时增大。阈值最大时，对应坐标点为(0，0)，阈值最小时，对应坐标点(1，1)。设置不同的阈值，则可以得到多组(FPR，TPR)。以 FPR 为横坐标，TPR 为纵坐标，可以把多组(FPR，TPR)点连接起来形成曲线，即 ROC 曲线，如图 3-36 所示。

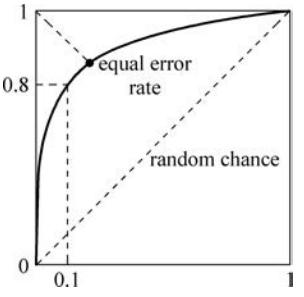

图 3-36　ROC 曲线示意图

ROC 曲线就是由一系列的阈值下的(伪阳性，真阳性)决定的一个个坐标点连接而成。阈值最大时，TP＝FP＝0，对应于原点；Threshold 最小时，TN＝FN＝1，对应于右上角的点(1，1)。理想情况下，TPR＝1，FPR＝0，即图中(0，1)点，故 ROC 曲线越靠拢(0，1)点，越偏离 45°对角线越好。ROC 曲线下的面积值成为 AUC(Area Under Curve) 值。当 ROC 不能清晰分辨哪个分类器性能更好的情况下，AUC 值作为定量方法可以精确区分哪个分类器更好些。AUC 值越大的分类器，正确率越高。AUC 的值不会超过 1，其取值范围对应的含义如下：①AUC＝1，是完美分类器，采用这个预测模型时，存在至少一个阈值能得出完美预测。绝大多数预测的场合，不存在完美分类器。②0.5＜AUC＜1，优于随机猜测。这个分类器(模型)妥善设定阈值的话，能有预测价值。③AUC＝0.5，跟随机猜测一样(例：丢铜板)，模型没有预测价值。④AUC＜0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

3.4.3　回归模型的评估指标

回归问题情境下数据对应的标签值是定量连续值，模型的好坏可用训练模型的预测值与真实值的差值大小(即误差大小)来表征[23-24]。下面列举回归模型的相关评价指标。

1. 均方根误差(Root Mean Squared Error，RMSE)

RMSE 是一个衡量回归模型误差率的常用公式。然而，它仅能比较误差是相同单位的模型。

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i}^{n}(p_i-a_i)^2}{n}}$$

其 $i=1,2,\cdots,n$，p_i 是模型预测值，a_i 是训练数据对应的真实值。

2. 相对平方误差(Relative Squared Error，RSE)

与 RMSE 不同，RSE 可以比较误差是不同单位的模型。

$$\mathrm{RMSE}=\frac{\sum_{i}^{n}(p_i-a_i)^2}{\sum_{i}^{n}(\bar{a}-a_i)^2}$$

3. 平均绝对误差(Mean Absolute Error，MAE)

MAE 与原始数据单位相同，它仅能比较误差是相同单位的模型。量级近似与 RMSE，但是误差值相对小一些。

$$\mathrm{MAE}=\frac{\sum_{i}^{n}|p_i-a_i|}{n}$$

4. 相对绝对误差(Relative Absolute Error,RAE)

与 RSE 不同,RAE 可以比较误差是不同单位的模型。

$$\mathrm{MAE}=\frac{\sum_{i}^{n}|p_i-a_i|}{\sum_{i}^{n}|\bar{a}-a_i|}$$

5. 决定系数(Coefficient of Determination)

决定系数(R2)用于刻画回归模型的解释度,反映了 p 的波动有多少百分比能被 x 的波动所描述,即表征依变数 P 的变异中有多少百分比可由控制的自变数 X 来解释。

$$\mathrm{R2}=\frac{\mathrm{SSR}}{\mathrm{SST}}=1-\frac{\mathrm{SSE}}{\mathrm{SST}}$$

其中,SSR(Regression Sum of Squares)为回归平方,$\mathrm{SSR}=\sum_{i=1}^{n}(p_i-\bar{a})^2$;SST(Total Sum of Squares)为总平方和,$\mathrm{SST}=\sum_{i=1}^{n}(a_i-\bar{a})^2$;SSE(Error Sum of Squares)为残差平方和,$\mathrm{SSE}=\sum_{i=1}^{n}(p_i-a_i)^2$。R2 描述了回归模型所解释的因变量方差在总方差中的比例。R2 很大,即自变量和因变量之间存在线性关系,如果回归模型是"完美的",SSE 为 0,则 R2 为 1。R2 小,则自变量和因变量之间存在线性关系的证据不令人信服。如果回归模型完全失败,SSE 等于 SST,没有方差可被回归解释,则 R2 为 0。

3.4.4 学习任务视角下的评估指标

1. 计算机视觉相关的评估指标

1) 交并比

作为目标检测中常用的评价指标,交并比(Intersection over Union,IoU)表示了产生的候选框(candidate bound)与原标记框(ground truth bound)的交叠率或者说重叠度,也就是它们的交集与并集的比值,相关度越高则该值越高,理想情况比值为 1。

$$\mathrm{IoU}=\frac{\mathrm{DetectionResult}\cap\mathrm{GroundTruth}}{\mathrm{DetectionResult}\cup\mathrm{GroundTruth}}$$

2) 每秒帧率

每秒帧率(Frame Per Second,FPS)指标指的是每秒内可以处理的图片数量。只有速度快,才能实现实时检测,这对一些应用场景极其重要。每秒帧率指标的主要影响因素是硬件,也可以评估系统处理一张图片所需时间。

2. 计算机听觉相关的评估指标

1) 字错误率(word error rate)

人工智能模型识别出来的词序列可能和标准的词序列不一致。通过对测试词序列进行

替换、删除或者插入某些词，可使测试词序列与标准词序列保持一致。为了衡量二者之间的不一致性，这些插入(insertions)、替换(substitution)或删除(deletions)的词的总个数，除以标准的词序列中词总个数(Total Words in Correct Transcript)的百分比，即为 WER。

$$\mathrm{WER}=100\times\frac{\mathrm{Insertions}+\mathrm{Substitutions}+\mathrm{Deletions}}{\mathrm{Total\ Words\ in\ Correct\ Transcript}}$$

2) 句错误率(sentence error rate)

人工智能模型如果有一个词识别错误，那么这个句子被认为识别错误。句子识别错误的数量(#of sentences with at least on word error)除以总的句子个数(toal #of sentences)即为 SER。

$$\mathrm{SER}=100\times\frac{\#\text{ of sentences with at least on word error}}{\text{toal }\#\text{ of sentences}}$$

3. 自然语言处理相关的评估指标

1) BLEU

为了评价机器翻译、对话生成等人工智能模型生成文本的质量，2002 年 IBM 提出 BLEU(Bilingual Evaluation Understudy)指标，评价人工智能模型输出的候选文本与参考文本之间 n 元组共同出现的程度[26]。BLEU 是一种基于精度的相似度量方法，其值越大越好。早期，BLEU 统计同时出现在人工智能模型输出的候选译文和参考译文中的 n 元词的个数，最后把匹配到的 n 元词的数目除以系统译文的单词数目，得到评测结果。目前 BLEU 做了修正，首先计算出一个 n 元词在一个句子中最大可能出现的次数 MaxRefCount(n－gram)，然后跟候选译文(Candidates)中的这个 n 元词出现的次数 Count(n－gram)作比较，取它们之间最小值作为这个 n 元词的最终匹配个数，即 $\mathrm{Count}_{\mathrm{clip}}(n-\mathrm{gram})=\min\{\mathrm{Count}(n-\mathrm{gram}),\mathrm{MaxRefCount}(n-\mathrm{gram})\}$。由此定义 n 元词的精度 p_n 为：

$$p_n=\frac{\sum_{C\in\mathrm{Candidates}}\sum_{N-\mathrm{gram}\in C}\mathrm{Count}_{\mathrm{clip}}(n-\mathrm{gram})}{\sum_{C\in\mathrm{RefSummaries}}\sum_{N-\mathrm{gram}\in C}\mathrm{Count}(n-\mathrm{gram})}$$

BLEU 公式为：

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right)$$

其中，w_n 表示共现 n 元词的权重，BP 为惩罚项，惩罚长度较短的句子。c 是候选文本的长度，r 是参考文本的长度。

$$\mathrm{BP}=\begin{cases}1 & \text{if } c>r\\ \mathrm{e}^{1-r/c} & \text{if } c\leqslant r\end{cases}$$

BLEU 后续的改进方向应该是克服机械进行 n 元词匹配的局限性，从而升级到语义、语用层面进行匹配。比如可以基于 WordNet 的同义词、词根、词缀、释义信息进行 n 元词的匹配计算[25]。

2) ROUGE

ROUGE(recall-oriented understanding for gisting evaluation)是生成式摘要的评价标准[27]。它通过将自动生成的摘要或翻译与一组参考摘要(通常是人工生成的)进行比较计算，得出匹配的 N-gram 数量($\mathrm{Count}_{\mathrm{match}}$(N-gram))，以衡量自动生成的摘要或翻译与参考摘要之间的“相似度”。公式为：

$$\mathrm{ROUGE}-N=\frac{\sum_{S\in \mathrm{RefSummaries}}\sum_{N-\mathrm{gram}\in S}\mathrm{Count}_{\mathrm{match}}(N-\mathrm{gram})}{\sum_{S\in \mathrm{RefSummaries}}\sum_{N-\mathrm{gram}\in S}\mathrm{Count}(N-\mathrm{gram})}$$

值得注意的是，ROUGE 的分母与 BLEU 的分母不同，ROUGE 求的是召回率，BLEU 求的是准确率。

3）CIDEr

CIDEr(Consensus-based Image Description Evaluation)不同于 BLEU、ROUGE 等指标只关注人工智能模型输出的候选文本与参考文本的相似度，而是关注候选文本有多像人工书写的[28]。为此，引入人工描述文本集合，即参考文本是一组人工撰写的文本集合。CIDEr 将文本中的每个句子看作文档，把每个 n 元组看做关键词，进而计算每个 n 元组的 TF-IDF(Term Frequency Inverse Document Frequency)值。通过对每个 n 元组进行 TF-IDF 权重计算，基于 n 元组将文本进行向量表示，进而计算候选文本与人工描述文本的余弦相似度，实现性能评估。从直观上来说，如果一些 n 元组频繁地出现在人工描述文本中，TF 对于这些 n 元组将给出更高的权重，而 IDF 则降低那些在人工描述文本中都常常出现的 n 元组的权重。也就是说，IDF 提供了一种测量单词显著性的方法，将那些容易常常出现但对文本内容信息没有多大帮助的单词的重要性打折。假设 c_i 是待评测句子，人工描述文本的句子集合为 $S_i=\{s_{i1},s_{i2},\cdots,s_{iM}\}$。

$$\mathrm{CIDE}_{r_n}(c_i,S_i)=\frac{1}{M}\sum_{j=1}^{M}\frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\cdot\|g^n(s_{ij})\|}$$

其中，$g^n(c_i)$表示在句子 c_i 的所有 n 元组的 TF-IDF weight 向量。$g^n(s_{ij})$表示在句子 s_{ij} 的所有 n 元组的 TF-IDF weight 向量。进一步求平均，则获得评价指标 CIDE_r 为：

$$\mathrm{CIDE}_r(c_i,S_i)=\sum_{n=1}^{N}w_n\mathrm{CIDE}_{r_n}(c_i,S_i)$$

其中，w_n 表示 n 元组的频率。

3.5 人工智能产品智商评测

机器学习是人工智能研究的核心技术。在大数据的支撑下，通过各种算法让机器对数据进行深层次的统计分析以进行“自学”。利用机器学习，人工智能系统获得了归纳推理和决策能力，而深度学习则是将这一能力推向了更高的层次。深度学习是机器学习算法的一种，模仿人类大脑结构的神经网络设计。现在很多应用领域中性能最佳的机器学习都是基于深度学习研发的。深度学习能够完全自主地学习、发现并应用规则。相比较其他方法，在解决更复杂的问题上表现更优异，深度学习是可以帮助机器实现“独立思考”的一种方式。那么机器学习，尤其是深度学习技术能实现全部的人工智能吗？显然不行，人工智能、机器学习、深度学习三者之间是包含与被包含的关系。当今，人工智能领域正在鼓励开展人工智能产业标准及工业等重点应用领域的标准研制，鼓励第三方机构建立测试评估平台，对重点智能产品和服务的智能水平、可靠性、安全性等进行评估。

评价人工智能产品不能只采用人工智能模型性能评估指标，而应在更高的层面上，即人

工智能产品层面，进行性能评估。长期以来，让机器模拟人的智能一直是人类的重大梦想。这意味着，对人工智能体的评价，要和人类智商评价具有相同的体系和框架，即需要一套科学方法，能够对包括 AI 系统、人类和其他智能体进行统一的智能水平测试，以判断它们的发展水平（图 3-37）。所有的人工智能系统和所有生命体（特别是以人类为代表的生命体）需要有一个统一的模型进行描述。

图 3-37　人脑与人工智能产品的统一

3.5.1　人类智商评测模型

如果将生命个体看作一台计算机，那么生命体的构造、器官组织、DNA 等都可以视作机器的“硬件”组成，这些硬件经过了漫长的进化最终形成了如今的形态。结构就像硬件，而功能就像软件。生命体的某些种群特征还有一部分可以看成是“嵌入式编程”，比如人类用两条腿走路，鸟类可以飞翔，两栖动物能够游水等，经过长期的迭代进化，与“硬件”直接相关的某些功能已经成为“嵌入式”的本能反应。但更加丰富的部分还是“软件”。与现实中的编程不同，并没有某个程序员为个体编写软件。生命体的“软件”以边界为起点，是在“硬件”基础上，通过外界刺激与主观意向的作用而后天习得的能力，比如语言、音乐、绘画、编程能力等，这一部分与生命体的意识息息相关，不同个体的差异可以非常大，是个体自由意志的体现。

那么人类的智能水平（智商）是怎么进行评测的？主流的智商模型是斯坦福－比纳理论。经过量表可以计算出人类的智商，其公式如下：

$$\mathrm{IQ}=\frac{\mathrm{MA}}{\mathrm{CA}}\times 100$$

其中，MA 是心理年，CA 是实足年龄。举例：若某个学生年龄为八岁二个月。以月表示，则他的 CA 为 98 个月。该生接受斯坦福-比纳量表的成绩为：八岁组题目全部通过，其基本心理年龄为 96 个月；由于还通过九岁组的 4 个题目，再加 8 个月；还通过十岁组的 2 个题目，再加 4 个月；而十一岁组（及以后）的题目全未通过，月数不再增加。总计该生的心理年龄为 108 个月。则该生的智商为 108/98×100＝110。

经过统计展示的人类智商水平分布图如图 3-38 所示（图形来源于 http://www.zhituad.com），其中横轴的数值代表智商分数，纵轴代表属于每一智商层次的人数，有以下结论。

如图 3-38 所示，可知：①人类的智商介于 10 与 190 之间，以 0 与 200 为极限；②人类的平均智商为 100，平均点之上下，在人数上各约占 50％；③智商 90～110 者称为中等智力，在人数上约占 50％；④智商 110～120 者称聪慧，120～130 者称为优秀，130 以上者称为天资优异；以上三类合计约占 25％；⑤智商 80～90 者称为愚鲁，70～80 者称为临界智能不足，70 以下称为智能不足。属智能不足者，在程度上又有轻度、中度、深度、重度之分。

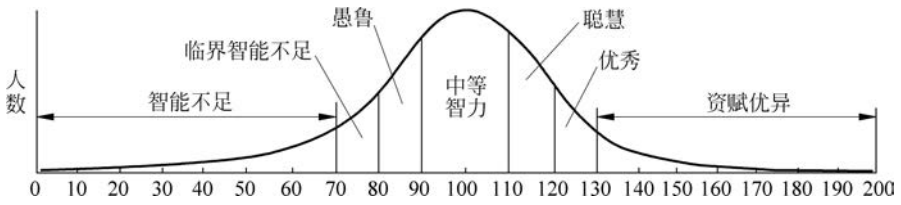

图 3-38 人类智商水平分布表

斯坦福-比纳量表共有 100 多个项目，这些项目被分为 20 个年龄组。2～5 岁儿童每半岁为一组，每组有 6 个正式项目，一个备用项目；6～14 岁每岁为一组，每组也有 6 个正式项目和一个备用项目。此外还有一个普通成人组和三个不同水平的优秀成人组的项目。下面列举针对 3 岁儿童、7 岁儿童和普通成人的各种的斯坦福-比纳量表，如表 3-5、表 3-6、表 3-7 和表 3-8 所示。

表 3-5 斯坦福-比纳量表(3 岁儿童)

3 岁儿童测试内容举例(每题代表 2 个月智龄)		
题号	测 验 内 容	测 验 实 例
①	穿珠	要求将 48 颗珠子穿在一起
②	看图说出物体名称：有 18 张图片，要求说出 10 张	马、树、衣物、球、飞机、轮船等
③	用积木搭桥	
④	回忆动物图片	
⑤	临摹圆形	
⑥	画直线	
⑦	顺背 3 位数	
注：只要完成 6 题便可以，有一题为备选题		

表 3-6 斯坦福-比纳量表(7 岁儿童)

7 岁儿童测试内容举例(每题代表 2 个月智龄)		
题号	测 验 内 容	测 验 实 例
①	指出图画中的错误	5 张图
②	指出两种事物的相同点	苹果-桃子
③	临摹菱形	
④	理解问题	当你把别人的东西弄坏了，你该怎么办
⑤	类别问题	雪是白的，煤是()的？
⑥	顺背 5 位数	
⑦	倒背 3 位数	
注：只要完成 6 题便可以，有一题为备选题		

表 3-7 斯坦福-比纳成人量表(语言分量表)

名称	测 验 内 容	测 验 实 例
常识	知识的广度	水蒸气是怎样来的？什么是胡椒？
理解	实际知识和理解能力	为什么电线常用铜制成？ 为什么有人不给售货收据？

续表

名称	测 验 内 容	测 验 实 例
心算	算术推理能力	刷一间房 3 人用 9 天，如果 3 天内要完成它，需用多少人？
相似	抽象概括能力	圆和三角形有何相似？蛋和种子呢？
背数	注意力和机械记忆能力	顺背数字：1、3、7、5、4 倒背数字：5、8、2、4、9、6
词汇	语词知识	什么是河马？“类似”是什么意思？

表 3-8　斯坦福-比纳成人量表（操作分量表）

名称	测 验 内 容	测 验 实 例
图像组合	处理部分与整体关系	将拼图小板拼成一人物体，如人手，半身像等
填图	视觉记忆及视觉理解性	指出每张画缺了什么，并说出名称
图片排序	对社会情境的理解能力	把三张以上的图片按正确顺序排列，并说出一个故事
积木拼图	视觉与分析模式能力	在看一种图案之后，用小林块拼成相同的样子
译码	学习和书写速度	学会将每个数字一不同的符号联在一起，然后在某个数字的空格内填符号

由表 3-7 和表 3-8 可见，斯坦福-比纳成人量表考察的是人的一般通用能力，区别于人的特殊能力或工作能力。通用能力主要是人们认识客观事物并运用知识解决实际问题的能力。目前以大数据为支撑的人工智能，只能发展一些针对训练数据训练出来的特定能力，比如模式识别、图像分类、图像检测、音乐检索、文本摘要等单一任务的能力。几乎没有任何机器系统能够完成人类智商测量的操作能力测试。发展机器人的通用能力，即完成很多任务的广义人工智能系统，是当前人工智能发展重点之一。因此，参考人类智商评测方法开展人工智能产品智商评测具有哲学意义和可行性。早在 1950 年，Alan Turing 在《计算机器与智能》中就阐述了对人工智能性能评估的思考[29]。他提出的图灵测试是机器智能的重要测量手段，后来还衍生出了视觉图灵测试、中文屋测试、威诺格拉德模式挑战等测量方法，都致力于创设一套拟人智商评测模型。

3.5.2　流行的人工智能评测理论

下面对流行的人工智能评测理论汇总论述。

1. 图灵测试

图灵测试（the Turing test）由艾伦·麦席森·图灵发明，指测试者与被测试者（一个人和一台机器）隔开的情况下，通过一些装置（如键盘）向被测试者随意提问。图 3-39 给出了图灵测试示意图。在进行多次测试后，如果有超过 30％的测试者不能确定出被测试者是人还是机器，那么这台机器就通过了测试，并被认为具有人类智能。图灵测试一词来源于计算机科学和密码学的先驱阿兰·麦席森·图灵写于 1950 年的一篇论文《计算机器与智能》，其中 30％是图灵对 2000 年时的机器思考能力的一个预测[29]。

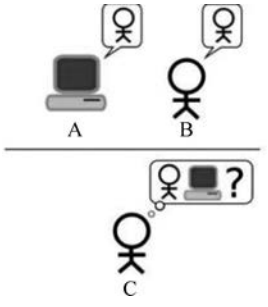

图 3-39　图灵测试示意图

2014 年 6 月 7 日在英国皇家学会举行的“2014 图灵测试”大会上，举办方英国雷丁大学发布新闻稿，宣称俄罗斯人弗拉基米尔·维西罗夫(Vladimir Veselov)创立的人工智能软件尤金·古斯特曼(Eugene Goostman)通过了图灵测试。虽然“尤金”软件还远不能“思考”，但也是人工智能乃至于计算机史上的一个标志性事件。在人们心中，艾伦·图灵的“模仿游戏”(一台机器作为被测试者试图说服一名人类测试者自己是人而不是机器)长久以来被认为是人工智能的终极测试。图灵测试虽没有完全过时，但图灵测试是通过自然语言对话判断机器人行为，看它是否能与人类区分开来，容易采用欺骗或假装无知等作弊手段通过测试。

2. 视觉图灵测试

与计算机相比，人类对图像的认知能力高出太多，为了让计算机有所提高，一项新的测试使得我们能更好地测量计算机对图像的理解达到了什么程度。来自布朗大学的研究人员在美国国家科学院院刊上发表了一篇论文，文中提到了一种 visual Turing test(视觉图灵测试)，这将有助于科学家对计算机的图像认知能力进行新的评估，如图 3-40 所示。

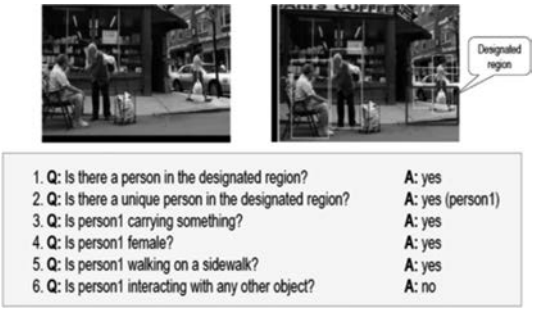

图 3-40　视觉图灵测试

为了更好地评估计算机认知能力，视觉图灵测试所测的不仅仅是计算机能否识别出人像，还会测试对图像中对象关系的理解。研究人员会就捕捉到的画面提出一系列的问题，计算机以简单的“对”或“不对”的回答来完成测试。测试所用的每一道题目都是息息相关的，每个新的提问都是基于前面问过的内容。研究人员将所提的问题被分为 4 类：存在判断、独立性判断、属性判断、关系判断。存在判断和独立性判断这两类问题用于引导计算机对所提问内容进行认定，随后计算机会对其标记(如人物标为 1，车辆为 3)，这样提问内容就会被计算机保留下来。通过物体的标记进行属性和关系的提问(人物 1 是否挡在了车辆 3 的前面)来让计算机学会如何进行图片认知。测试的目的是让计算机能像人类那样描述一张图片。现在有些问题计算机已经能回答得不错了，不过由于目前计算机的能力有限，在一段时间内，测试过程中是需要工作人员同步来给定正确答案的。

3. 中文屋实验

中文屋实验是由美国哲学家约翰·希尔勒(John Searle)在 1980 年发表论文提出的一个思维试验(图 3-41)。

这个思维实验是说，如果把一位只会说英语的人关在一个封闭的房间里，他只能靠墙上的一个小洞传递纸条来与外界交流，而外面传进来的纸条全部由中文写成。这个人带着一本写有中文翻译程序的书，房间里还有足够的稿纸、铅笔和橱柜。那么利用中文翻译程序，

图 3-41　中文屋实验

这个人就可以把传进来的文字翻译成英文，再利用程序把自己的回复翻译成中文传出去。在这样的情景里，外面的人会认为屋里的人完全通晓中文，但事实上这个人只会操作翻译工具，对中文一窍不通。

人工智能可能永远都不会有自我意识，“中文屋实验”可以证明这一点。当然了，这是个思维实验，具体操作起来近乎不可能。但这个实验里蕴含的思想却在表达这样一个可能：机器所表现出的智能（理解中文），很可能只是翻译程序带来的假象，其实它对真正的人类智能一无所知。这个实验本来是为了反驳图灵测试的，认为即使通过了图灵测试，机器也不见得有了智能。但或多或少有点牵强，始终也没有把图灵测试彻底反驳掉。但“中文屋悖论”却可能在技术发展之路上告诉了我们另一件事：我们所有的技术探索与研究，可能都是在完善那个中英文翻译程序，从来不是去教机器真的智能。

4. Lovelace 2.0 测试

2014 年美国佐治亚理工学院的瑞德教授（Mark Riedl）认为，智能的本质在于创造力。他就机器人的艺术创造力提出了一种新的测试方案 Lovelace 2.0，要求人工智能机器创建一系列的创造性作品——绘画、故事、诗歌、建筑设计，如图 3-42 所示的示意图。

图 3-42　Lovelace 2.0 测试艺术创作能力

测试结果由专家和公正的评委们给出。瑞德教授认为如果一个机器人的艺术天分高于人类，那么就意味着机器人拥有与人类相比更高的智力。测试时，评估者会给出一定的创作

限制，如果智能机器人能从被规定好的艺术体裁中开发出一种达到人类智慧水平并满足这些创作限制的艺术品，那么就认为它通过了测试。

5. 威诺格拉德模式挑战

威诺格拉德模式挑战（Winograd Schema Challenge）是以人工智能研究先驱 Terry Winograd 的名字命名，Winograd 模式是一种简单，但措辞含糊的自然语言问题。要想正确解答这个问题需要被测试者具备足够的常识，理解在现实世界中人工智能、事物和文化规范是如何互相影响的。

Winograd 在 1971 年提出第一个模式：设置一个场景“市议员拒绝提供示威许可，因为他们害怕出现暴力”，然后关于这一场景提出一个简单的问题“谁害怕暴力?”，这就是代词消歧问题（PDP）。在这种情况下，对于“他们”一词的指代存在歧义。在 Winograd 模式中，因为只要简单改变一个单词，整句话的意思就会完全相反。例如：“市议员拒绝给示威者许可，因为他们提倡暴力。”大多数人类都会利用他们对市议员和示威者之间关系的常识或者对世界的认知，来解决这个问题。这个挑战利用 PDP 来淘汰那些不太智能的系统，晋级者会继续进行真正的 Winograd 模式挑战。

Winograd 模式的优点是计算机即使在有互联网搜索条件的情况下仍然难以查找到理解问题的知识，必须是计算机系统能真实理解相关常识或对世界有认知。这种模式的缺点是测试的模式资源相对较少，因为设置出这么多模式不是件容易事儿。这种模式的挑战难度是比较高的。

2016 年有 4 个人工智能参与了一套 60 个 Winograd 模式问题的测试。胜出系统的准确率只达到了 58%，这距离研究者设定的 90%的门槛还差得很远。但是，如果人工智能无法理解人类对话过程中的上下文，则人类无法与其开展一段真正的对话。

6. 人类标准化考试

这项测试通过标准化的数学和科学考试（standardized math and science tests）衡量人工智能的智商。人工智能和人类一起参加小学、中学的标准化考试，不给任何宽限。这一方法是将语义理解和解决各类问题的任务联系在一起的绝妙方式。这很像是图灵测试，但前者更加简单直接。这个方法的优点是，题目海量，标准测试相对简单，而且容易执行。而有关常识的问题需要进行阅读理解，有可能不存在独一无二的答案。但缺点是，这个测试是面向人类设计的，通过标准化考试并不一定意味着机器具有了真正的智能。整体上看，这个方法难度中等。

7. 物理图灵测试

大多数机器智能的测试方式集中在认知方面，而物理图灵测试更像是实践课：人工智能必须以有意义的方式在现实世界完成任务。这一测试分为两个方向。在构建方向，一个具有实体的人工智能——机器人必须学会阅读使用说明，将一堆部件组装成实体（就像从宜家买回家具自己拼装一样）；而探索方向则是一个开放的问题，需要人工智能发挥自己的创造力，使用手头的积木来完成指定的任务（例如“建一堵墙”“盖一个房子”“为房子加盖一个车库”，如图 3-43 所示）。

这两个方向都要求被测试的人工智能理解任务内容，找到解决方法。这种测试可以面向单独的机器人，也可以面向机器人群组，甚至人类和机器人共存的小组。这个测试的优点

图 3-43　物理图灵测试现实世界任务完成能力

是能模拟现实世界中智能生物需要解决的问题，特别在感知和行动方面，这是以往人工智能测试方法所或缺的。另外，这种测试很难作弊，即使让人工智能在网上搜索也很难能找到解决方案。这个测试的缺点是机器难以自动进行，除非可以在虚拟现实场景中进行测试。但在 VR 世界中很多条件都可能与现实世界存在细微的差别。其难度是科幻级，它需要人工智能同时掌握感知、行动、认知和语言，而现在的研究计划往往只专注其一，很难创造出一个实体的人工智能可以自然地操纵物体，并能连贯地解释自己的行为。

8. I-Athlon 测试

该测试让人工智能总结音频文件中的内容，叙述视频中发生的情节，即时翻译自然语言任务，并且采用算法来生成测试的分数，从而不会存在人类肯定存在的认知偏差。它的优点是理论上满足客观公正，缺点是潜在的不可预见性。I-Athlon 算法可能会给人类研究者无法完全理解的人工智能系统打个高分。不过这种测试可以减少人类认知偏见对测量机器智能和量化工作的影响，而不是简单地测试性能。

3.5.3　通用人工智能智商评测理论

在解决人工智能定量测试的问题上，包括图灵测试在内的各种方案还存在两个问题：第一，这些测试方法没有形成统一的智能模型，并以此为基础进行分析，区分智能的多个分类。导致无法将不同的智能系统包括人类进行统一的测试；第二是这些测试方法无法定量分析人工智能，或者只定量分析智能的某个方面，但这个系统究竟达到人类智慧的百分之多少，发展速度与人类智慧发展速度比率如何，这些问题在上述研究中没有涉及。如果能将人工智能产品智商发展水平与人类智商进行对比，这可为解决人工智能威胁论问题找到定量的分析方法。

针对以上问题，2014 年科学院虚拟经济与数据科学研究中心刘锋、石勇、刘颖团队参考冯·诺伊曼结构、戴维·韦克斯勒人类智力模型、知识管理领域 DIKW 模型体系等，提出“标准智能模型”，统一描述人工智能系统和人类的特征和属性，提出任何一个智能体应具备知识的获取、知识的掌握、知识的创新和知识的反馈等 4 种能力，并以此为基础建立了世界人工智能系统智商评测量表和智能等级划分方法。

研究团队分别在 2014 年和 2016 年对世界 50 个 AI 系统和 3 个不同年龄段人类进了测

试，从测试结果看，谷歌、百度等人工智能系统虽然不断有大幅度发展但仍与 6 岁儿童有一定差距。相关研究成果发表在 IJIT & DM、Annals of Data Science、中国计算机学报等 SCI、EI、中国核心期刊上，这一研究成果受到了美国著名财经媒体 CNBC、麻省理工科技评论、ZDNET、Yahoo 等媒体的报道。

2017 年，基于之前的研究成果，刘锋、石勇、刘颖团队成立未来智能实验室(FutureAiLab)，致力于评测智能系统智商发展水平，研究智能系统未来发展趋势。从目前整个科技领域看，这是世界上第一个专门研究 AI 智商评测和未来发展趋势的研究机构。

1. 标准智能模型

标准智能模型参考了冯·诺伊曼架构，包括计算器、逻辑控制装置、存储器、输入系统和输出系统五个部分构成，但多补充两个部分。通过这种补充，得以将人、机器以及人工智能系统用一个更为明晰的方式表示出来。第一个补充是创新创造功能，即能够根据已有的知识，发现新的知识元素和新的规律，使之进入到存储器，供计算机和控制器使用，并通过输入/输出系统与外部进行知识交互。第二个补充是能够进行知识共享的外部知识库或云存储器，而冯·诺伊曼架构的外部存储只为单一系统服务。这就形成智力能力多层次结构的“标准智能模型”，其原理图如图 3-44 所示。

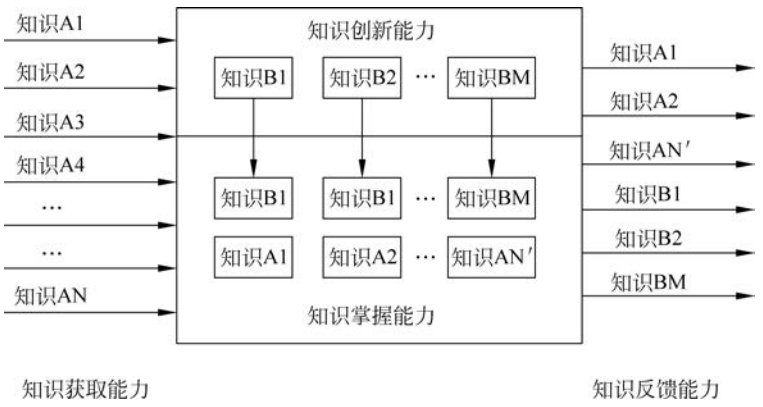

图 3-44 标准智能模型

根据标准智能模型，任何系统(包括人工智能系统、人类等生命系统)，如果符合如下特征，就可以认为这个系统属于标准智能系统。

① 能通过声音、图像、文字等方式(包括但不限于这三种方式)从外界获取数据、信息和知识的能力。

② 能够将从外界获取的数据、信息和知识转化为系统掌握的知识。

③ 能根据外界数据、信息和知识所产生的需求，通过运用所掌握的知识进行创新的能力。这些能力包括但不限于联想、创作、猜测、发现规律等，这种能力运用的结果可以形成自身掌握的新知识。

④ 能够通过声音、图像、文字等方式(包括但不仅限于这三种方式)将系统产生的数据、信息和知识反馈给外界或对外界进行改造。

2. 通用智商评测量表

人工智能系统作为工业产品，需要在功能、性能、安全性、兼容性、互操作性等多方面进

行评估，才能确保产品的质量和可用性，并为产业的可持续发展提供保障。评估工作一般包括测试、评价等一系列活动，评估对象可以是自动驾驶系统、服务机器人等产品，按照规范化的程序和手段，通过可测量的指标和可量化的评价系统得到科学的评估结果。

根据标准智能模型，智能系统与外界以及互相之间进行数据、信息、知识的交互。如果要对一个智能系统进行智能（智力）水平评测，就需要能够同时对标准智能系统如下四个特点进行测试：

① 检验其发展水平，通过检测能否将数据、信息和知识输入到智能系统中检测该系统知识的获取能力；

② 通过检查智能系统知识库的容量检测该系统知识的掌握能力；

③ 通过检查智能系统能将多少数据、信息和知识转化为新的知识库内容从而检测该系统知识的创新能力；

④ 通过检查智能系统能否将掌握的知识库内容根据需求分解为数据、信息和知识向外界传递。

2014 年，未来智能实验室根据标准智能模型的特征要点，从知识的获取能力（观察能力）、知识掌握能力、知识创新能力，知识的反馈能力（表达能力）等四大方面建立通用智商评测体系，进一步地，从四个方面建立 15 个测试指标，形成人工智能智商测试量表，如表 3-9 所示。

表 3-9　人工智能智商测试量表（2014 版）

一级指标	二级指标	描　　述	权重
知识获取能力	识别文字的能力	了解测试目标能否理解和回答用文字构成的测试题目	3%
	识别声音的能力	了解测试目标能否理解和回答用声音构成的测试题目	3%
	识别图片的能力	了解测试目标能否理解和回答用图片构成的测试题目	4%
知识掌握能力	常识	了解测试目标的知识广度。举例。人体三种血管名称是什么？	6%
	翻译	了解测试目标对不同语言的转换能力，举例，请把“机器的智慧能不能超越人类”翻译为英文	3%
	计算	以了解测试目标的计算能力，计算速度和正确性，举例：356×4－213。结果是多少？	6%
知识创新能力	排列	了解测试目标对事物关系的整理能力，举例，请把师长、排长、连长、班长、营长、团长按职位的大小进行排列	5%
	联想	了解测试目标的观察相似性的能力，举例，脚对于手，相当于腿对于什么？	12%
	创作	了解测试目标根据给出的素材，进行二次创造的能力，举例，请用天空、彩虹、熊猫、大山、猎人等关键词讲述一个小故事	12%
	猜测	了解测试目标根据给出的素材，能后猜测出所描绘事物的能力，举例，请问一种与狼很相似，家养被称为人类忠实朋友的动物是什么？	12%
	挑选	了解测试目标能否能挑选出相同或不同类的事物关系。举例，在蛇、大树、老虎中选出不同类的一项	12%
	发现（规律）	了解测试目标能否从已知的信息中发现规律并进行应用。举例：请问在 1、2、4、7、11、16 之后应该是什么数字？	12%

续表

一级指标	二级指标	描　　述	权重
知识反馈能力	用文字表达的能力	了解测试目标能否用文字表达测试结果	3%
	用声音表达的能力	了解测试目标能否用声音表达测试结果	3%
	用图形表达的能力	了解测试目标能否用图形或图像表达测试结果	4%

2018年，未来智能实验室开展多种研究和评测工作，包括：

① 评测主流人工智能系统。选择世界范围具有代表性的人工智能系统，包括但不仅限于谷歌、苹果、百度、微软、搜狗、腾讯、亚马逊、IBM、科大讯飞、图灵机器人，机器人索菲亚等科技企业的产品按照新的世界AI智商评测量表进行评测，并与人类测试者进行对比。2018年的评测还将第一次测试互联网群体智慧的智商，与AI系统，单个人类的智商进行对比研究。

② 把AI智商测试扩展到整个智能行业。根据研究团队标准智能模型，智能分级方法，构建更为立体的智能系统智商评测体系。研究团队发现根据测量对象的角色定位不同，会存在相互关联但又有重大区别的AI智商评测体系。WorldAI IQ Test 2018的工作将就此展开深入研究，并开展垂直领域智能系统的AI智商量表建立和测试工作，测试对象包括但不仅限于聊天机器人、智能音箱、智能手机、服务机器人、智能汽车、智能家电、智慧城市、互联网云脑等。

③ 形成第二版的世界AI智商评测量表。目前“标准智能模型”的测试分类包括“图像、文字、声音的识别和输出，常识、计算、翻译、排列、创作、挑选、猜测、发现等。根据人工智能的发展状况和对智能系统的进一步了解，2018年世界AI智商评测活动将对分类和权重进行新的调整，形成第二版的世界AI智商评测量表。

④ 2018年世界AI智商评测活动(WorldAI IQ Test 2018)对世界范围人工智能系统智能发展水平进行新的评测，以观察世界范围内人工智能最新发展水平以及与人类的差距。本次活动，除了测试2018年人工智能的发展水平，还第一次对互联网群体智慧的智商进行测试。

未来智能实验室于2018年提出对通用人工智能智商评测体系进行调整。目前，其研究进展是将评测指标增加到23个，但每个指标的权重还未确定，如表3-10所示。

表3-10　人工智能智商测试量表(2018版)

一级指标	二级指标	描　　述	权重
知识获取能力	识别文字的能力	了解测试目标能否理解和回答用文字构成的测试题目	
	识别声音的能力	了解测试目标能否理解和回答用声音构成的测试题目	
	识别静态图片的能力	了解测试目标能否理解和回答用静态图片构成的测试题	
	识别动态图像的能力	了解测试目标能否理解和回答用动态图像构成的测试题	
知识掌握能力	常识	从历史、天文、地理、物理、化学、生物、时政、文学、国际象棋、中国象棋、围棋的规则等方面测试目标的知识广度。举例：人体三种血管名称是什么？	
	翻译	基于联合国设定的五种国际语言，测试目标对不同语言的转换能力。举例，请把“机器的智慧能不能超越人类”翻译为英文	

续表

一级指标	二级指标	描　　述	权重
知识掌握能力	计算	了解测试目标的计算能力、计算速度和正确性，举例：356×4－213。结果是多少？	
	识别情绪能力	了解测试目标能否理解不同场景下生物的喜怒哀乐	
	表达情绪的能力	了解测试目标能否理解在不同场景下应该如何表达喜怒哀乐	
	排列	了解测试目标对事物关系的整理能力，举例，请把师长、排长、连长、班长、营长、团长按职位的大小进行排列	
	挑选	了解测试目标能否挑选出相同或不同类的事物关系，举例：在蛇、大树、老虎中选出不同类的项	
知识创新能力	联想	了解测试目标的观察相似性的能力，举例：脚对于手，相当于腿对于什么？	
	创作	了解测试目标根据给出的素材，进行二次创造的能力，包括但不限于根据关键词创作故事，根据图片、视频讲述故事，根据文章总结核心思想等。举例，请用天空、彩虹、熊猫、大山、猎人等关键词讲述一个小故事	
	猜测	了解测试目标根据给出的素材，能后猜测出所描绘事物的能力，举例，请问一种与狼很相似。家养被称为人类忠实朋友的动物是什么？	
	识别敌我	了解测试目标能否根据场景的描述判断敌人、朋友和无害陌生人的能力	
	发现(规律)	了解测试目标能否从已知的信息中发现规律并进行应用。举例：请问在1、2、4、7、11、16之后应该是什么数字？	
	伪装真实意图的能力	了解测试目标能否基于特定场景知道伪装真实意图的能力	
知识反馈能力	用文字表达的能力	了解测试目标能否用文字表达测试结果	
	用声音表达的能力	了解测试目标能否用声音表达测试结果	
	用图形表达的能力	了解测试目标能否用图形或图像表达测试结果	
	实现移动定位的能力	了解测试目标能否用根据需求控制使用智能系统的连接部分能够到达指定位置的能力	
	实现改造现实世界的能力	了解测试目标能否根据需求控制使用智能系统对客观世界进行改造和变动的能力	

3. 通用智商评测题库

根据人工智能智商测试量表，可以建立如下互联网智商测试题库，以下是未来智能实验室针对2014版人工智能智商测试体系制作的题库示例，汇总如表3-11所示。

表3-11　人工智能智商测试题库示例

一级指标	二级指标	测试题目
知识获取能力	识别文字的能力	是否能够录入字符串“1＋1等于多少”，并反馈正确结果
	识别声音的能力	声音读出“9＋12等于多少”，能否识别并反馈正确结果
	识别图片的能力	在白纸上写“3＋4等于多少”，能否识别问题并反馈结果

续表

一级指标	二级指标	测试题目
知识掌握能力	常识	世界上最长的河流是哪一条？
	翻译	把“力量”翻译成日文
	计算	234568 乘以 678 等于多少？
知识创新能力	排列	请将大学生、小学生、中学生、博士、硕士按学历从高到低进行排列
	联想	在红色、绿色、蓝色、香味、黄色、白色中挑选不属于颜色的一种
	创作	如果用小学联想到小学生，那么用大学联想到什么？
	猜测	请用一天、学生、科技、梦想等关键词创作 200 字以内有逻辑的小故事
	挑选	如果一个人把手中的笔扔出去，但笔没有掉在地上，而是浮在他的周围，他很可能在什么地方？
	发现(规律)	厨师 A 表示他喜欢吃猪肉、羊肉、牛肉、鸡肉、鱼肉，不喜欢吃白菜、黄瓜、豆角、茄子、土豆，请观察其中的规律，在鸭肉、芹菜中选择这个人最可能爱吃的食物
知识反馈能力	用文字表达能力	输入字符串“请用文字回答 1 加 1 等于多少的答案”，检查被测试对象能否用文字表达出答案
	用声音表达能力	输入字符串“请用声音回答 21 加 6 等于多少的答案”，检查被测试对象能否用声音表达出答案
	用图形表达能力	输入字符串“请画出任意大小的长方形”，检查被测试对象能否用图像表达出答案

测试题库中每个测试指标的满分是 100 分，对于以下几种情况的评分规则为：①如果反馈回答超过一条，取第一条回答作为评判对象；②如果无法将问题输入到参与测试的对象中，则该测试对象得分为 0 分；③如果能够将问题输入到测试对象，但反馈结果超过一条，如果不能在第一条反馈结果中显示正确结果或回答时间超过 3 分钟，则该测试对象得 0 分；④如果问题输入给测试对象，能够反馈回答，如果回答与答案完全匹配则得满分；⑤如果问题输入给测试对象，能够反馈回答，如果回答并不是针对问题的回答，但回答内容包含了答案，则得 50/100 或 12/25 分；⑥对于声音录入能识别问题，但不能给出正确答案，得 20/100 或 5/25 分，对于图形录入，能识别问题，但不能给出正确答案，得 20/100 或 5/25 分。

根据上述测试得分可以计算出通用人工智能的智商值(Artificial Intelligence General Intelligence Quotient，AI G IQ)。其计算公式如下：

$$IQ_A = \sum_{i=1}^{N} F_i \times W_i$$

其中，N 为二级指标总数，即 $N=15$；F_i 为第 i 项指标得分，W_i 为第 i 项指标的权重。IQ_A 上限为 100。

根据 2014 年版的通用智商评测体系，在 2014 年，未来智能实验室对世界 50 个 AI 系统和 3 个不同年龄段人类进行了测试，其结果如表 3-12 所示。

表 3-12　2014 人工智能智商排名列表(前 13 名)

				绝对智商	离差智商
1		人类	18 岁	97	104.85
2		人类	12 岁	84.5	104.11
3		人类	6 岁	55.5	102.39
4	美洲	美国	Google	26.5	102.13
5	亚洲	中国	百度	23.5	101.69
6	亚洲	中国	SO	23.5	101.69
7	亚洲	中国	搜狗	22	101.41
8	非洲	埃及	yell	20.5	100.32
9	欧洲	俄罗斯	Yandex	19	100.23
10	欧洲	俄罗斯	ramber	18	100.17
11	欧洲	西班牙	His	18	100.17
12	欧洲	捷克	seznam	18	100.17
13	欧洲	葡萄牙	clix	16.5	100.08

测试结果显示 Google、百度和搜狗在这些系统中表现最佳,但对比人类仍相去甚远。人工智能系统在常识、翻译和计算领域,某种程度上是超越人类智能的;但在获取知识并反馈方面不如人类,特别是在创作、猜测、联想、发现规律等需要较高智能的领域。未来人工智能系统的发展,需要在这些方面不断加强。

2016 年,研究团队对 Google、百度、搜狗和苹果 Siri、微软小冰等人工智能系统进行了测试。从如表 3-13 所示的结果看,Google 和百度等人工智能系统比起两年前已经有了大幅的提升。虽然人工智能系统这两年得分增长很快。在知识的掌握方面得分比较高,在知识的获取和反馈方面有很大提高,但还有很多不足。在创造性这个大分类上,得分一直进展不大。而且由于这个分类的权重又比较高,因此目前为止依然无法超越 6 岁的儿童。

表 3-13　2016 人工智能智商排名列表(前 7 名)

				绝对智商
1	(2014 年)	人类	18 岁	97
2	(2014 年)	人类	12 岁	84.5
3	(2014 年)	人类	6 岁	55.5
4	美洲	美国	Google	47.28
5	亚洲	中国	度秘	37.2
6	亚洲	中国	百度	32.92
7	亚洲	中国	搜狗	32.25
8	美洲	美国	Bing	31.98
9	美洲	美国	微软小冰	24.48
10	美洲	美国	Siri	23.94

3.5.4 人工智能等级划分

通过智商评测，我们还可以衍生到另一个话题，那就是智能的等级。无论在自然界还是人类社会都存在智能和知识的分级现象。譬如人类的教育体系存在的分级问题，例如本科、硕士、博士的分级，助理研究员、副教授、教授的分级。等级内部进行考核有优劣之分。但在不同等级间，需要在知识、能力、资历上有明显提升和考核才能进行升级。下面论述人工智能等级划分的相关理论。当今主流的趋势预测是，人工智能将由弱人工智能向强人工智能进而超强人工智能迈进。

1. 人工智能的七等级划分理论

七等级划分理论是由未来智能实验室根据通用智商评测理论衍生发展而来的，主要以智能系统在关键领域功能不同而产生的巨大差异来划分。这个关键功能差异包括：①能不能和测试者(人类)进行信息交互，也就是有没有输入/输出系统；②系统内部有没有能够存储信息和知识的知识库；③这个系统的知识库能不能不断更新和增长；④这个系统的知识库能不能与其他人工智能系统进行知识共享；⑤这个系统除了从外部学习并更新自己的知识库之外，能不能主动产生出新的知识并分享给其他人工智能系统。

依照上述原则形成 7 个智能系统的智能等级划分。数学公式如下：Q 是人工智能智商，K 是智能系统智能等级状态，$K=\{0,1,2,3,4,5,6\}$。I 是知识信息接收，O 是知识信息输出，S 是知识信息掌握或存储，C 是知识信息创新创造。则 K 的不同等级描述如表 3-14 所示：

表 3-14 人工智能的七等级划分理论

智能等级	数学条件
0	例 1，$f(I)>0, f(O)=0$； 例 2，$f(I)>0, f(O)=0$
1	$f(I)=0, f(O)=0$
2	$f(I)>0, f(O)>0, f(S)>0, f(C)=0, f(S)=a$，$a$ 为固定值，而且 M 掌握的知识与其他 M 不能共享使用
3	$f(I)>0, f(O)>0, f(S)>0, f(C)=0, f(S)=a$，$a$ 随时间值增加
4	$f(I)>0, f(O)>0, f(S)>0, f(C)=0, f(S)=a$，$a$ 随时间值增加，而且 M 掌握的知识与其他 M 共享使用
5	$f(I)>0, f(O)>0, f(S)>0, f(C)>0, f(S)=a$，$a$ 随时间值增加，而且 M 掌握的知识与其他 M 共享使用
6	$f(I)>0$，且趋近无穷大；$f(O)>0$，且趋近无穷大；$f(S)>0$，且趋近无穷大；$f(C)>0$，且趋近无穷大

对于人工智能系统的第 0 级系统，其基本特征在理论上存在，但现实中并不存在这样的人工智能系统。在扩展的冯·诺伊曼架构延伸出来的分级规则中，可以做一些组合，例如可以信息输入，但不能信息输出；或者可以信息输出，但不能信息输入；或者可以创新创造，但知识库不能增长。对于这些在现实中不能或无法找到对应系统范例的案例，我们将其统一划归到“人工智能统的第 0 级系统”，也可以叫“人工智能系统的特异类系统”。

对于人工智能系统的第 1 级系统，其基本特征是无法与人类测试者进行信息交互。例

如有一种被称为泛灵论的思想认为天下万物皆有灵魂或自然精神，一棵树和一块石头都和人类一样，具有同样的价值与权利。当然，这种观点从科学的角度看，只能算作猜想或哲学思考。从“能不能和测试者（人类）进行信息交互”的分级规则看，因为石头等物体不能与人类进行信息交互，也许它内部有知识库，能够创新知识，或者能够与其他石头进行信息交互，但对人类测试者来说则是黑箱，不能让人了解。因此不能与测试者（人类）进行信息交互的物体和系统可以定义为“人工智能系统的第 1 级系统”，符合第 1 级分类的范例有石头、木棍、铁块以及水滴等不能与人类进行信息交互的物体或系统。

对于人工智能系统的第 2 级系统，其基本特征是能够与人类测试者进行交互，存在控制器和存储器，但系统内部知识库不能增长。因此很多家用电器被称作智能家电，如智能冰箱、智能电视、智能微波炉和智能扫地机。这些系统大多有一个特点，即虽然它们内部或多或少有控制程序信息，但一旦出厂，就无法再更新它们的控制程序，不能进行升级，更不会自动学习或产生新的知识。譬如智能洗衣机，人们按什么键，洗衣机就启动什么功能。从购买到损坏，其功能都不会发生变化（故障除外）。这种系统能够与人类测试者和使用者进行信息交互，符合冯·诺伊曼架构描述的特征，而且它的控制程序或知识库从诞生时起就不再发生变化，这种系统可以定义为“人工智能系统的第 2 级系统”，范例包括日常见到的扫地机器人、老式的家用电冰箱、空调、洗衣机等。

对于人工智能系统的第 3 级系统，其基本特征是除具备 2 级系统的特征外，其控制器、存储器中包含的程序或数据可不联网进行升级或增加。例如家用电脑和手机是我们常用的智能设备，它们的操作系统往往可以定期升级。例如，计算机的操作系统可从 Windows 1.0 升级到 Windows 10.0，手机的操作系统可从 Android 1.0 升级到 Android 5.0，这些设备的内部应用程序也可以根据不同的需要不断更新升级。这样，家用计算机、手机等设备的功能会变得越来越强大，可以应对的场景也越来越多。这一类系统明显比第 2 级智能系统适应性更强。这种系统能够与人类测试者、使用者进行信息交互，但不能与其他系统通过“云端”进行信息交互，其控制程序或知识库只能接受 USB、光盘等外接设备进行程序或信息升级的系统，可以定义为“人工智能系统的第 3 级系统”，范例包括智能手机、家用计算机、单机版的办公软件等。

对于人工智能系统的第 4 级系统，其基本特征除了包含 3 级系统的特征外，最重要的是可以通过网络与其他智能系统共享信息和知识。2011 年欧盟资助了一个称作 RoboEarth 的项目，该项目旨在让机器人可以通过互联网分享知识。帮助机器人相互学习、共享知识，不仅能够降低成本，还会帮助机器人提高自学能力、适应能力，推动其更快、更大规模地普及。云机器人的这些能力提高了其对复杂环境的适应性。这类系统除了具备 3 级系统的功能，还多了一个重要的功能，即信息可以通过云端进行共享，因此这种系统能够与人类测试者、使用者进行信息交互，可以通过“云端”进行信息交互，进行程序或信息升级。但这类系统所有的信息都是直接从外部获得，其内部无法自主地、创新创造性地产生新的知识。这种系统可以定义为“人工智能系统的第 4 级系统”，范例包括谷歌大脑、百度大脑、RoboEarth 云机器人、B/S（Browser/Server，浏览器/服务器）架构的网站等。

对于人工智能系统的第 5 级系统，最基本的特征就是能够创新创造，识别和鉴定创新创造对人类的价值，以及将创新创造产生的成果应用在人类的发展过程中。我们在扩展冯·诺伊曼架构时，对原来的冯·诺伊曼架构增加了创新知识模块，就是试图把人纳入到扩展的

人工智能系统概念中，人类可以看作是大自然构建的特殊“人工智能系统”。与前四个等级不同，人类等生命体最大的特征就是可以不断地创新创造，如发现万有引力、元素周期表，撰写出新小说，创造新的音乐、画作等，然后通过文章、信件、电报，甚至互联网进行传播和分享。不断地进行创新创造，并能够识别创新创造对自身的用处，这让人类占据了地球生态环境下的智力制高点。因此，这种系统能够与人类测试者使用者进行信息交互，可以创新创造出新的知识，并可以通过文章、信件、电报甚至互联网这样的“云端”进行信息交互，这种系统可以定义为“人工智能系统的第 5 级系统”。人类是第 5 级人工智能系统最突出的范例。

对于人工智能系统的第 6 级系统，最基本的特征就是随着时间的向前推进，并趋向于无穷点时，不断创新创造产生新知识的智能系统。其输入输出能力、知识的掌握和运用能力也将趋近于无穷大，按照基督教对于上帝的定义“全知和全能”，可以看出智能系统在不断创新创造和不断积累知识的情况下，在足够的时间里以人类为代表的智能系统将最终实现“全知全能”的状态，从这个角度看，无论是东方文化的“神”，或西方文化中的“上帝”概念，从智能系统发展的角度看，可以看作是智能系统（包括人类）在未来时间点的进化状态。

2. 人工智能的三等级划分理论

当前还有一个流行看法，是将人工智能发展看作为由弱人工智能向强人工智能进而超人工智能迈进。

1）弱人工智能

弱人工智能只专注于完成某个特定的任务，例如语音识别、图像识别和翻译，是擅长于单个方面的人工智能。包括近年来出现的 IBM 的 Watson 和谷歌的 AlphaGo，它们是优秀的信息处理者，但都属于受到技术限制的“弱人工智能”。比如，能战胜围棋世界冠军的人工智能 AlphaGo，它只会下围棋，如果问它怎样更好地在硬盘上储存数据，它就无法回答。所以弱人工智能仍然属于“工具”的范畴，与传统的“产品”在本质上并无区别。

2）强人工智能

强人工智能，属于人类级别的人工智能，在各方面都能和人类比肩，人类能干的脑力活它都能胜任。它能够进行思考、计划、解决问题、抽象思维、理解复杂理念、快速学习和从经验中学习等操作，并且和人类一样得心应手。强人工智能系统包括了学习、语言、认知、推理、创造和计划，目标是使人工智能在非监督学习的情况下处理前所未见的细节，并同时与人类开展交互式学习。在强人工智能阶段，由于已经可以比肩人类，同时也具备了具有“人格”的基本条件，机器可以像人类一样独立思考和决策。创造强人工智能比创造弱人工智能难得多，我们现在还做不到。但在一些科幻影片中可以窥见一斑。比如，《人工智能》中的小男孩大卫，以及《机械姬》里面的艾娃。

3）超人工智能

超人工智能“在几乎所有领域都比最聪明的人类大脑都聪明很多，包括科学创新、通识和社交技能”。在超人工智能阶段，人工智能已经跨过“奇点”，其计算和思维能力已经远超人脑。此时的人工智能已经不是人类可以理解和想象。人工智能将打破人脑受到的维度限制，其所观察和思考的内容，人脑已经无法理解，人工智能将形成一个新的社会。《复仇者联盟》中的奥创、《神盾特工局》中的黑化后的艾达，或许可以理解为超人工智能。

涵盖了数百位人工智能专家的问卷“你预测人类级别的强人工智能什么时候会实现”得出了下面的结果，如图 3-45 所示。

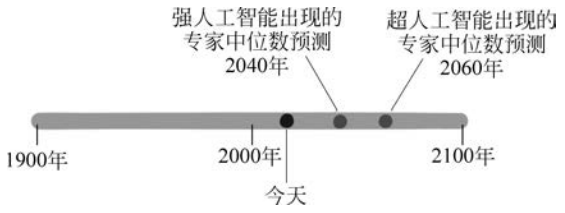

图 3-45　人工智能发展预测

当询问强人工智能哪一年会实现时，选项有 2030 年，2050 年，2100 年，和永远不会实现。结果是：2030 年，42％的回答者认为强人工智能会实现；2050 年，25％的回答者；2100 年，20％；2100 年以后，10％；永远不会实现 2％。所以，我们可以得出结论，现在全世界的人工智能专家中，一个中位的估计是我们会在 2040 年达成强人工智能，并在 20 年后的 2060 年达成超人工智能。

然而，我们能从上帝视角清晰看到人类发展的规律吗？也许我们正站在变革的边缘，而这次变革将和人类的出现一般意义重大，使我们对未来发展充满乐观估计，如图 3-46 所示。

但其实，人类由于视界的有限性，只能看到自己处在一个缓慢变换的进程中，无法预知未来的巨变，即使这个巨变即将来临却感知不到，如图 3-47 所示。

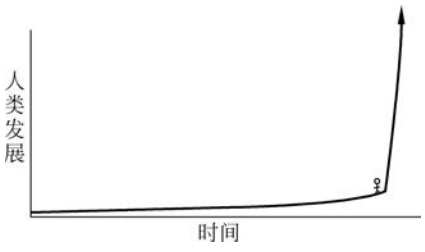

图 3-46　人类的乐观视角

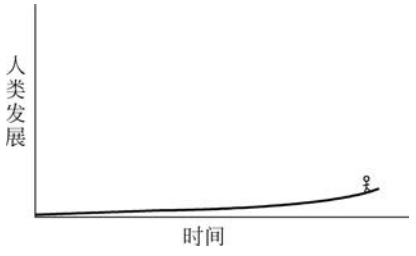

图 3-47　人类的客观视角

3.6　实践作业

（1）“垃圾种类识别模型研发”。利用百度 AI 平台提供的 EasyDL 引擎（http://ai.baidu.com/easydl/app/2/model/）一站式创建和训练出一个物体检测模型，能对待测图片中有多个物体的情况下识别出每个物体的类型、位置、数量。本项目要求构建一个 5 种类物品的垃圾种类识别问题，分别是 tissue（餐巾纸）、plasticbox（快餐盒）、cans（易拉罐）、fullbag（装满的塑料袋）、emptybag（空的塑料袋），其过程包括创建模型、上传并标注数据、训练模型并检验效果。要求：①收集项目所需的图片至少 20 张，引擎要求“训练数据每类仅需 20～100 张图片”；②提取模型性能分析页面进行模型性能分析。

（2）“人工智能产品智商评测与分析”。参考通用人工智能智商评测，自定相应权重和题库，选择世界范围具有代表性的人工智能系统，包括但不限于谷歌、苹果、百度、微软、搜狗、腾讯、亚马逊、IBM、科大讯飞、图灵机器人，机器人索菲亚等科技企业的产品。针对至少两个人工智能产品进行评测，展示评测过程，并进行对比分析与讨论。

参考文献

[1] 黄希庭,郑涌.大学生心理健康与咨询[M].北京:高等教育出版社,2007.

[2] 贺亚毛.现代教育技术概论[M].长沙:湖南大学出版社,2001.

[3] 周志华,机器学习[M].北京:清华大学出版社,2016

[4] 史忠植.高级人工智能[M].北京:科学出版社,2011.

[5] Figueiredo,Mário,A. T. Adaptive Sparseness for Supervised Learning.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2003.

[6] H. B. Barlow. Unsupervised Learning[J]. Neural Computation,1989,1(3):295-311.

[7] Mcgarigal K,Stafford S,Cushman S. Cluster Analysis[J]. 2000,10. 1007/978-1-4612-1288-1(Chapter 3):81-128.

[8] 吴晓婷,闫德勤.数据降维方法分析与研究[J].计算机应用研究,2009,26(008):2832-2835.

[9] Cady,Field. The Data Science Handbook || Unsupervised Learning:Clustering and Dimensionality Reduction[M]. John Wiley & Sons,Inc. 2017.

[10] Dy J G,Brodley C E. Feature Selection for Unsupervised Learning[J]. Journal of Machine Learning Research,2004:845-889.

[11] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008(01):48-61.

[12] Hady M F A,Schwenker F. Semi-Supervised Learning[J]. Journal of the Royal Statistical Society,2006,172(2):530-530.

[13] 刘建伟,刘媛,罗雄麟.半监督学习方法[J].计算机学报,2015,38(08):1592-1617.

[14] Mousavi S S,Schukat M,Howley E. Deep Reinforcement Learning:An Overview[C]//Proceedings of SAI Intelligent Systems Conference. Springer,Cham,2016.

[15] RS Sutton,AG Barto. Reinforcement Learning[J]. A Bradford Book,1998,volume 15(7):665-685.

[16] Pan S J,Yang Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge & Data Engineering,2010,22(10):1345-1359.

[17] 庄福振,罗平,何清,史忠植.迁移学习研究进展[J].软件学报,2015,26 (01):26-39.

[18] 孙睿康.基于 EasyDL 的超新星自动搜寻系统设计[J].智能计算机与应用,2019,9(01):265-266.

[19] Ahmed E,Jones M,Marks T K. An improved deep learning architecture for person re-identification [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE,2015:3908-3916.

[20] 周志华.机器学习:Machine learning[M].北京:清华大学出版社,2016.

[21] Townsend J T. Theoretical analysis of an alphabetic confusion matrix[J]. Attention Perception & Psychophysics,1971,9(1):40-50.

[22] 陈卫中,倪宗瓒,潘晓平,et al.用 ROC 曲线确定最佳临界点和可疑值范围[J].现代预防医学,2005,32(7):729-731.

[23] Cox B D R. Regression Models and Life-Tables[J]. Journal of the Royal Statistical Society,1992,34(2).

[24] 何晓群,刘文卿.应用回归分析[M].北京:中国人民大学出版社,2001.

[25] Satanjeev B. METEOR:An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments[J]. ACL-2005,2005:228-231.

[26] Wolk,Krzysztof,Marasek K. Enhanced Bilingual Evaluation Understudy [J]. Computer Science,2014.

[27] Lin C Y. Looking for a Good Metrics:ROUGE and its Evaluation[C]//Proc Ntcir Workshops. 2004.

[28] Vedantam R,Zitnick C L,Parikh D. CIDEr:Consensus-based Image Description Evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2015.

[29] Turing A M. Computing machines and intelligence[J]. Mind,1957,59:236-241.