第一章

肇始：什么是 AI 大模型

导语

在本章开始之前，我们先来思考一个问题：机器学习也好，深度学习也好，它们到底在做什么事情？它们的存在是要解决什么问题？

事实上，“学习”就是给计算机投喂训练数据，最后求解某种方式或方法的过程。这类似于我们学习英语。我们学习英语的过程就是不断接收单词、句型、语法等训练数据的过程，最后学会用另一门语言与人流畅交流，以及学会阅读另一种文字的书籍。

随着计算机技术及深度学习算法的不断进步，AI 大模型已成为当今人工智能领域的重要研究方向。AI 大模型及其相关应用也成为 IT 领域近年来最热门的话题。“AI”即 artificial intelligence（人工智能）的首字母缩写，那么，大模型又是什么呢？在这一章中，我们先来讲讲 AI 大模型中“大”的概念，再说明什么是“模型”。最后，再看看 AI 大模型究竟有什么特点。

第一节　大模型的“大”体现在何处

“大模型”通常指拥有大量参数和计算复杂度的神经网络模型，我们先来说“大”这个概念，后面会展开讲什么是神经网络。那么，到底有多少参数才能被称为“大”呢？它是否有一个确切的标准？

一、AI“大”模型的判定标准是什么

在特定场景下，一个参数量只有几千万的模型可能都可以被归为“大模型”。而在其他场景下，只有拥有数十亿，甚至近百亿个参数的模型才能被称为“大模型”。

那么，如何理解特定场景与其他场景呢？

比如，在金融、医疗和安全等行业，模型需要对数据的准确性和安全性拥有更高要求，不管是风险分析、疾病预测，还是安全检测，都需要进行高质量的预测。在这些特定场景下，即使参数量较小的模型也可以被归为“大模型”。这些模型的参数通常是高度定制的，针对特定的任务和数据集进行了优化，以获得更高的准确性和效率。

再如，ChatGPT 是一个生成式 AI 语言模型，而自然语言属于非常宽泛的领域，有千亿级参数及庞大的算力基础。随着 OpenAI 继续在商业上部署 ChatGPT 和该公司的生成式 GPT 模型，该语言模型可能需要超过 30 000 块显卡。这种显卡是价格相当昂贵的英伟达的 A100，它的最高显卡内存（简称为“显存”）为 80GB。而我们普通的游戏电脑显存超过 8GB 就已经算不错了。每块这种显卡的售价为 10 000～15 000 美元，这种价格绝对是普通大众难以负担得起的。随着 ChatGPT 的发布，各大厂纷纷跟随潮流，投身自研大模型的浪潮，因此该显卡已成为热门抢

手货。

那么，为什么机器学习训练模型都要用显卡的 GPU，而不是用计算机的 CPU 呢?

这是因为 GPU 在处理并行计算任务时，比 CPU 更加高效。首先，GPU 内部有数千个小处理器，这意味着 GPU 可以同时执行很多个计算命令，而 CPU 只能进行一小部分。其次，GPU 有更高的内存带宽，这意味着它们可以更快地将数据传输到处理器。

例如，在性能和能效方面，L4 GPU 的服务器能够提供比 CPU 解决方案高出 120 倍的人工智能视频性能，同时提供 2 倍的生成性人工智能性能，为深度学习和推理应用提供更高效的性能，如图 1-1 所示。

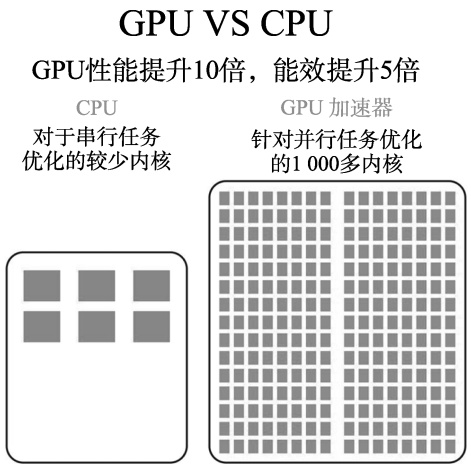

图 1-1　CPU 与 GPU 的对比

如图 1-2 所示：CPU 的工作流程就好像只有一个画家在画一朵花，这一个画家将绿叶红花的绘画工作一步一步地完成；GPU 就像一个个小画家，你画绿叶，我画花瓣，他画花蕊，明确分工，并行完成，速度上

自然要快很多。

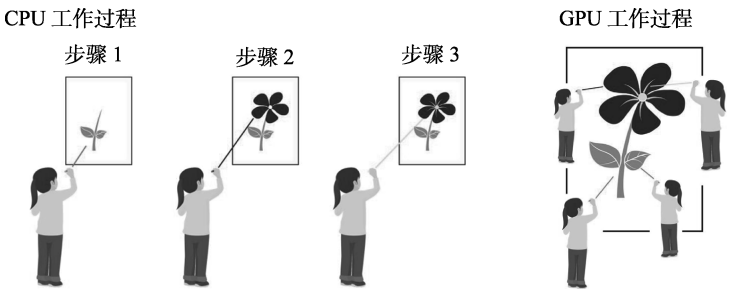

图 1-2　CPU 与 GPU 工作过程的类比

此外，一个模型能否被称为“大”，也可以和它要解决的问题有关。大模型解决“大事”，当处理比较复杂的语言、图像、视频、自然语言处理等领域的问题时，AI 大模型通常表现出比传统方法更好的鲁棒性、泛化能力和精度，使其在各种任务中较传统方法更具优势，并取得更好的效果。

二、AI 大模型的三个特性

下面，我们来对鲁棒性、精度及泛化能力这三个概念进行简要阐释。

1. 鲁棒性

鲁棒是英文单词“robust”的音译，有“健壮”“强壮”的意思。这个概念最早出现在 1979 年南开大学两位教授发表的文章“鲁棒（robust）调节器”中，后有学者认为，“robust”有“健壮”“强壮”之意，它被译为“鲁棒”是一个“音义兼顾”的绝好译法，便沿用至今。

鲁棒性指对噪声或异常值（如数据中的错误或干扰）有很好的抵抗

能力。大模型通常可以在处理含有噪声和错误的训练及测试数据时，保持更好的表现能力。在图像处理领域中，大模型通常能够更好地识别不同场景下的物体，即使背景是复杂的，物体角度是多样变化的，它也能够准确地进行识别。

我们以小明学习开车为例。

小明刚考完驾照，车技还不熟练，对外界各种因素的变化较为敏感，抗干扰能力很弱。复杂的路况、后视镜的角度，甚至是女友的唠叨都可能成为他出错的原因。这时我们说小明的驾驶能力还不够“鲁棒”。但随着不断学习，开车经验的增长，他具有了较强的鲁棒性，不但能在开车时跟身边的女友谈笑风生，还能从容掌控各种外界状况。当车外面突然刮大风时，他可以稳定操控方向盘；当前车急刹车时，他可以立即反应过来，踩下刹车；当一个行人突然横穿马路时，他也可以镇定地避让。这时候，我们就可以说小明的驾驶能力已经越来越“鲁棒”了。

对小明来说，他的驾驶技能的鲁棒性体现在对各种复杂路况和突发事件的适应能力。只有不断训练和积累经验，才能达到比较“鲁棒”的性能。

我们可以把这个例子类比机器学习和人工智能领域，理解算法和模型的鲁棒性。在机器学习领域，一个好的算法会具有较强的鲁棒性，它可以处理异常值，对数据的变化和噪声都能够做出较好的表现。

这样的算法能力对于实际应用非常重要，因为它可以帮助我们在面对各种复杂的数据集时，仍然能够获得良好的实际效果。

在实际应用中，我们是难以保证输入和环境始终“干净”的，系统部署后，实际输入数据的分布可能与训练数据有差异，这需要模型自身对这种分布的变化有适应性。生活中存在太多随机事件和复杂情况，这

对模型应对各种突发事件的稳定性提出了很高的要求。因此，我们希望机器学习算法和 AI 系统都能具备较强的鲁棒性。

下面，我们继续用小明学习开车的例子解释精度和泛化能力这两个概念。

2. 精度

那么什么是精度呢？精度指模型在特定的数据或任务上的正确率或准确度。

小明在刚学开车的时候，如果遇到宽敞一点的停车位，他基本能将车停进去；但如果遇到空间较小、更复杂一点的侧方停车位，则很容易让车被剐蹭到。随着他停车的技术越来越熟练，他终于可以停进前后只有一个拳头距离的侧方位了。在这种情况下，我们就可以认为他操控车辆的精度提高了。

精度是衡量一个模型在特定领域或任务上的表现程度的重要指标。通常，我们会在多个数据集和应用场景上评估一个模型精度，这有助于全面评价其性能。

3. 泛化能力

泛化能力体现了一个模型对未见过的数据或情况的适应性，它可以通过模型在训练数据集和测试数据集上的表现来评估。一个具有良好泛化能力的模型能够在看不见的数据上表现良好。我们依旧以小明学车的例子来说明。

小明刚学会开车，只在空旷平坦的道路上练习过如何开车，这时，如果带他去山区的山路上开车，他可能就开到山沟里面去了。这时，我

们可以认为他的泛化能力是较差的，因为他没见过，也无法处理那么复杂的路况。但经过不断练习，小明对各种路况都了如指掌，这时候，无论是城市街道，还是乡间小路，他凭借精良的驾驶技巧都能很好地适应路况，这证明他的泛化能力提高了。

总之，泛化能力就是模型对新数据、新环境、新情况的适应与推广能力。这个能力越强，模型的实用价值就越高。泛化能力是一个较为综合的性能维度。

以上，我们通过案例对三个概念进行了一番说明。然而，刚接触这些概念的读者可能对泛化能力和鲁棒性的区别还有些模糊。那么，如何给它们做一个区分呢？

泛化能力是一种更广义的概念，它关注的是模型对一般新的情况的适应力，与之意思接近的成语是“触类旁通”；鲁棒性则是一种更狭义的概念，专注模型对异常或极端情况的抵抗力，与之意思接近的成语是“坚韧不拔”。

泛化能力是模型的功能指标，鲁棒性更侧重模型的非功能属性。泛化能力决定模型的使用范围，鲁棒性决定模型的使用安全性。理想的模型应同时具备较强的泛化能力和鲁棒性。泛化能力保证模型在一般情况下的高性能，而鲁棒性使其即使在极端情况下，仍然能够保持稳定。

第二节　AI 大模型的优势

在第一节中，我们主要介绍了 AI 大模型的判定标准和主要特点。在本节中，我们将深入探讨 AI 大模型的优势，以帮助大家明晰 AI 大模型具体对我们有什么好处。

一、AI 大模型具备高性能的因素

视线回到我们的 AI 大模型。AI 大模型之所以比传统模型表现出更好的精度、鲁棒性和泛化能力，主要有以下几个原因。

（1）超大规模数据集。AI 大模型可以训练超大规模的数据集，这使其学到的数据更丰富，对各种情况的覆盖更广，从而提高精度和泛化能力。

（2）强大的计算能力。AI 大模型可以利用强大的 GPU 等计算资源进行大量计算，使其设计模型可以更复杂，网络层次可以更深，参数规模可以更大，从而捕获数据的内在模式和特征。

（3）复杂的算法。最新研发的机器学习算法，如深度学习等，可以自动学习模型的参数，提取数据的高级特征交互，这大大提高了模型的表达能力、精度和泛化能力。

（4）端到端学习。AI 大模型采用端到端的学习方式，使输入直接关联最终输出，避免手工提取特征等中间步骤，保留了更多原始信息，这也有助于提高泛化能力。

（5）自动调整能力。AI 大模型可以自动调整众多超参数，如学习率、网络层数、结点数等，找到模型性能的最优组合，这也是其精度和泛化能力较高的原因之一。

（6）噪声抵抗能力。大型深度模型的参数较多，对局部数据扰动的敏感度低，这使其表现出较强的鲁棒性，对异常值和噪声更有抵抗力。

例如，在图像识别领域，深度学习模型可以训练数十万张，甚至上百万张图像，这使其学到的视觉模式和特征丰富全面，可以识别的数据类别多，所以其精度和泛化能力远超传统方法。此外，深度模型可以自动学习图像的低层视觉特征（线条、形状）和高层语义特征（物体及场景），

这种端到端的学习方式保留了丰富信息，有利于提高其泛化能力。

在自然语言处理（natural language processing，NLP）领域，像 GPT-3、GPT-4 这样的大语言模型可以训练超过 1 000GB 的文本数据，所以其学习到的词汇、句法、语义知识都非常丰富，这使得它们生成的文本涵盖话题广、连贯性强。这证明了它们拥有很强的泛化能力。而 GPT 模型的大小达到百亿参数的量级，这使得它们对词语的敏感度较低，对文本噪声有较强抵抗力，表现出较好的鲁棒性。

在游戏领域，像 DeepMind[①]的 AlphaGo 可以训练海量的人类棋谱，并且进行大量的自我博弈实践，这使 AlphaGo 学到的棋策丰富广博，可以很好地适应人类棋手的不同棋风。这证明了 AlphaGo 具有很强的泛化能力和鲁棒性。此外，AlphaGo 使用深度神经网络来学习棋局的内在规律，并和搜索算法相结合，这种混合模型设计有利于实现预测的高精度和广度。

对于无人驾驶车辆而言，它需要训练“星球级别”的数据量，学习各种天气、光照下的道路和环境，以适应复杂多变的行驶条件，这需要强大的计算资源和深度学习的能力才能实现。只有具备较强的泛化能力、鲁棒性和精度，无人驾驶车辆的安全性才能得到保证。

所以，通过这些示例，我们可以清晰地看到，AI 大模型之所以获得比较优异的机器学习性能，关键还是得益于海量数据、强大算力和深度学习算法等手段，这些手段使其学习的知识和模式更加全面深入，从而达到较高的泛化能力、鲁棒性与预测精度。

---

① DeepMind，位于英国伦敦，是由人工智能程序师兼神经科学家戴密斯·哈萨比斯（Demis Hassabis）等人联合创立的 Google 旗下的前沿人工智能企业。

二、强大的自学习能力

除了以上优势，在深度学习领域，AI 大模型还有一个神奇的本领，那就是自学习的能力。它们能够通过自动学习处理输入数据，不需要人工手动设计和提取特征。这种能力是基于深度学习和神经网络的构造及算法而实现的。

例如，在自然语言处理领域，AI 大模型可以对输入数据中的语法、语义和复杂关系等特征进行自学习，并逐渐优化模型的神经网络结构，以达到更高的精度和泛化能力。

深度学习模型通常由多个层次组成，每个层次都包含了一定数量的神经元。当看到新的数据时，模型的参数将根据最新数据的特征进行自动调整，同时优化自身结构，以便更好地表示数据的复杂结构。自学习的能力使得 AI 大模型可以更加适应不同类型的任务，并且在更广泛的应用场景中找到更好的解决方案。

我们可以用视觉系统对深度学习模型进行类比解释。视觉系统通过眼睛接收到大量的视觉数据，并通过大脑的神经网络进行处理和解释，最终产生对视觉输入模式的认知。

就像深度学习模型一样，视觉系统也是由多个分层级别的神经元组成的。其中，最底层处理的是像素级别的信息，中间层次处理的是更高阶的形状和特征，最高阶层次处理的则是更抽象的概念和概括。

在人的学习过程中，我们通过大量的视觉输入数据，不断锻炼和调整视觉神经网络的参数和结构，从而提高对事物的认知和理解。这里的“参数和结构”类比于深度学习模型中的“权重和层次结构”。

正如人类的学习和认知过程一样，深度学习模型通过大量的数据不断学习和优化，可以构建更复杂、更有效的神经网络，从而提高对数据

的理解和表现能力。

举个例子，假设有一辆我们之前从未见过的车，我们可以用“看到它—认识它—记住它”的过程来类比机器学习的过程，如图 1-3 所示。

图 1-3　如何记住一辆我们从未见过的车

（1）数据准备。首先，我们需要准备学习数据。那一辆之前从未见过的车的数据就是它的图像、车牌、车辆品牌、颜色等各种特征。

（2）数据输入和处理。其次，我们将这些数据输入 AI 大模型中进行处理。将这个步骤类比人脑，就是在大脑中处理这些数据，并进行分析和归纳的过程。

（3）训练和优化。再次，在机器学习中，我们通常需要将数据集分为训练集和测试集，并对模型进行训练和优化。将这个步骤类比人类的学习过程，就是我们记住了汽车的外观、品牌和车牌等特征，然后通过多次观察和回忆及想象，不断优化我们的记忆和理解的过程。

（4）预测和输出。最后，我们将经过训练和优化的 AI 大模型用于新的数据预测和输出。将这个步骤类比人类的学习过程，即当我们再次看见这辆车时，我们能否判断它就是之前我们见过的那一辆车的过程：我们可以根据它的外观、内饰，甚至是发动机的声音来做出判断。

测一测，你的大脑通过图 1-3 的“训练”，能在如图 1-4 所示的车辆局部信息中找到那一辆车吗？

图 1-4　4 辆汽车的局部信息

上述内容阐述了 AI 大模型的很多优点，然而，以现阶段的技术水平来看，AI 大模型也存在一些缺点。

首先，AI 大模型需要大量的计算资源来进行训练和推理，这对硬件平台的要求非常高。其次，AI 大模型对系统、开发语言、版本及存储空间等有一定的依赖，升级、维护和部署也需要大量的人力和资源投入。

尽管 AI 大模型存在这些缺点，但它在未来的应用前景非常广阔，有专家称其为“第四次工业革命前夜”，笔者认为这种说法并不为过。我们需要拥抱 AI 大模型时代的到来，并不断研究和推广新的技术和应用场景，为人类社会带来更多的创新和改变。

第三节　机器学习和深度学习

在本章的最后，我们再补充一点：关于机器学习和深度学习是怎么回事？它们之间有什么区别？

机器学习是一门研究如何使计算机系统自动学习和改进自身的技术学科。它的目标是设计和研发算法，通过训练数据使计算机具有学习能力。机器学习算法包括监督学习（回归、分类）、无监督学习（聚类）、强化学习等。它已经存在很久了，最早可以追溯到 1950 年左右，当人们收集到一些数据时，如果单凭自己去找规律，会比较麻烦，于是将这些数据扔给计算机。通过机器学习，计算机会告诉人类，这些数据可能存在什么样的规律，如温度变化或价格变化等，是呈线性分布，还是像某个波形函数那样分布，进而再对数据进行预测或者归纳分析。

而深度学习是机器学习的一个分支，兴起于 2006—2010 年之间，它尝试使用类似人脑神经网络的结构来学习数据的深层特征表示或分布。深度学习模型通常包含多隐藏层的神经网络，可以自动学习数据的高阶

特征。它常被用于大规模数据（如图像、语音、视频等）的特征学习和识别任务。我们所说的 AI 大模型就是基于深度学习产生的。人工智能与机器学习、深度学习的关系如图 1-5 所示。

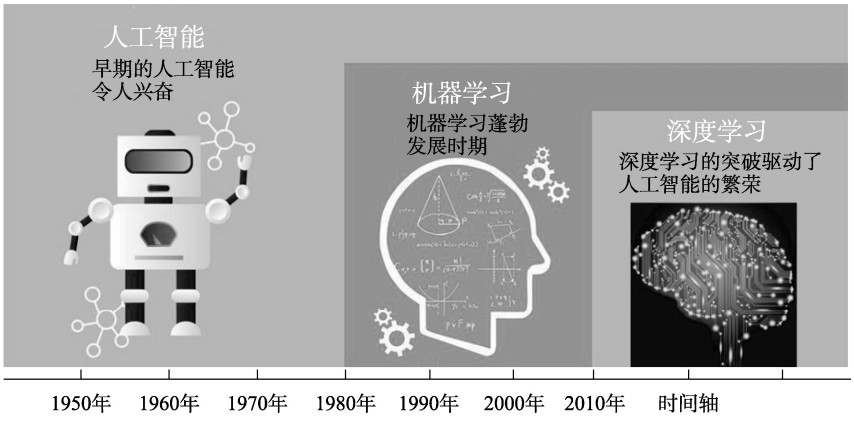

图 1-5　人工智能与机器学习、深度学习的关系图

如图 1-5 所示，深度学习是机器学习的子集，我们将在本书后面的章节对深度学习进行重点讲解。

小　结

AI 大模型正以其规模化的体量和深不可测的智能带领人类走向未来。它是海量数据与算力的产物，拥有复杂的神经网络结构，可以像人脑一样自动学习各种特征。超强的计算能力支撑起了它千亿级的参数量，让它在处理语音、图像、视频等领域展现出极高的精度、鲁棒性与泛化能力。

那么，AI 大模型的深层神经网络究竟是如何工作的？它又是如何实现数据的特征提取和表示的呢？我们只有继续掀开深度学习与神经网络的黑匣子，才能够对 AI 大模型的真容一窥究竟。