第1章　人工智能与人工智能应用

学习目标

学生通过本章的学习，系统掌握人工智能核心知识体系（从基础概念到机器学习、深度学习技术架构），培养算法应用与模型构建的实践能力，同时树立科技伦理意识与社会责任感，最终实现技术认知、实操技能与人文素养的协同发展，为推动人工智能领域创新与社会价值落地奠定基础。

知识目标

1．理解人工智能的基本概念、发展历程及重要里程碑。

2．掌握机器学习的主要类型、基本原理及系统构成。

3．了解深度学习的核心概念、模型类型及应用领域。

技能目标：

1．学生应能够理解不同类型的机器学习算法（如有监督学习、无监督学习等）的基本原理和应用场景。

2．学生能够根据具体任务选择合适的机器学习算法，并进行模型训练、优化和评估。

3．学生应能够了解深度学习模型（如卷积神经网络、循环神经网络等）的基本原理和结构。

素养目标：

1．培养学生的科技伦理意识，使其在研究和应用人工智能技术时，能够遵守道德和法律规范。

2．引导学生关注人工智能技术的社会影响，思考如何将其用于促进社会公平、公正和可持续发展。

3．培养学生的创新意识和团队协作能力，鼓励他们在人工智能领域不断探索和创新，同时与他人合作，共同推动科技进步。

1.1 认识人工智能

在21世纪的科技浪潮中，人工智能（artificial intelligence, AI）无疑是最具颠覆性和前瞻性的技术之一。它不仅是计算机科学领域的一颗璀璨明珠，更是推动社会进步、经济发展和文化繁荣的重要力量。从智能家居到自动驾驶，从医疗诊断到金融风控，人工智能的应用已经渗透我们生活的方方面面，深刻改变着人类的生产方式和生活模式。

1.1.1 人工智能的概念

人工智能是计算机科学或智能科学中涉及研究、设计和应用智能机器的一个分支。它旨在使计算机系统能够模拟和执行人脑的某些智力功能，包括感知环境、逻辑推理、语言理解和自主决策等。人工智能的近期主要目标在于研究用机器来模仿和执行人脑的某些智力功能，而远期目标则是用自动机模仿人类的思维活动和智力功能。

人工智能作为一个学科领域，其涵盖范围极为广泛，涉及多个学科和技术的交叉融合。它不仅包括计算机科学中的算法设计、数据结构、编程语言等基础知识，还涉及自动控制、电子技术、数学、心理学、语言学、哲学等多个学科。人工智能的研究领域也十分广泛，包括计算机视觉、自然语言处理、机器学习、语音处理等。

人工智能的发展历史可以追溯到3000多年前，但现代意义上的人工智能研究则始于20世纪。1956年夏季，在美国达特茅斯学院举办的一次研讨会上，首次使用了“人工智能”这一术语，标志着国际人工智能学科的诞生。此后，人工智能经历了多次起伏和发展，逐渐形成了今天繁荣的研究和应用局面。

1.1.2 机器学习与深度学习

1．机器学习

机器学习（machine learning）是人工智能领域的一个重要分支，它专注于使用算法和统计模型来使计算机系统能够自动地从数据中学习与改进，并且无须进行明确的编程。机器学习的核心思想是让机器（计算机）通过观察大量的数据和训练，发现事物规律，获得某种分析问题及解决问题的能力。

机器学习可以分为多种类型，根据预期输出和输入类型的不同，可以分为有监督学习、无监督学习、半监督学习和强化学习等。有监督学习需要一组已知的输入—输出对，算法通过学习这些对来预测未知的输入。无监督学习则不需要已知的输入—输出对，算法通过学习数据中的模式来进行分类或聚类。半监督学习是介于有监督学习和无监督学习之间的一种混合学习方法，它使用有限的监督数据和大量的无监督数据来训练算法。强化学习则是通过奖励和惩罚机制来学习，在与环境的交互中优化决策。

机器学习的基本原理可以概括为：使用算法从大量数据中提取特征，建立模型，然后应用这些模型来对新数据进行预测或分类。机器学习系统通常由数据、算法、模型以及评估方法四部分组成。数据是机器学习的基础，包括结构化数据和非结构化数据；算法是机器学习中的核心部分，它决定了如何从数据中提取有用信息；模型是算法在数据上训练得到的结果，它代表了数据的内在规律和模式；评估方法则用于衡量模型的性能。

2．深度学习

深度学习（deep learning）是机器学习的一个子领域，它通过构建多层神经网络模型来模拟人脑神经元的工作方式，从而使计算机能够自主学习并提取数据中的高级特征。深度学习的“深度”通常指的是神经网络的层数，一般超过 8 层的神经网络被称为深度学习模型。

深度学习模型通过组合低层特征形成更加抽象的高层特征表示，从而发现数据的分布式特征。这些模型在训练过程中，通过反向传播算法不断调整网络中的权重和偏置，以最小化预测误差。深度学习模型根据其结构和应用领域的不同，可以分为多种类型，如卷积神经网络（CNN）、循环神经网络（RNN）和生成对抗网络（GAN）等。

卷积神经网络主要用于处理图像数据，通过卷积层和池化层提取图像中的特征。循环神经网络则擅长处理序列数据，如语音、文本等，能够记忆序列中的上下文信息。生成对抗网络由生成器和判别器组成，生成器负责生成逼真的数据，判别器则用于判断生成的数据是否真实。

深度学习具有强大的特征提取能力、非线性建模能力和泛化能力。它能够自动从原始数据中提取高级特征，无须人工设计特征提取器。通过多层非线性变换，深度学习模型能够捕捉数据中的复杂关系。同时，在训练数据上学习到的知识可以较好地应用于未见过的数据。

机器学习与深度学习的关系如图 1-1 所示。

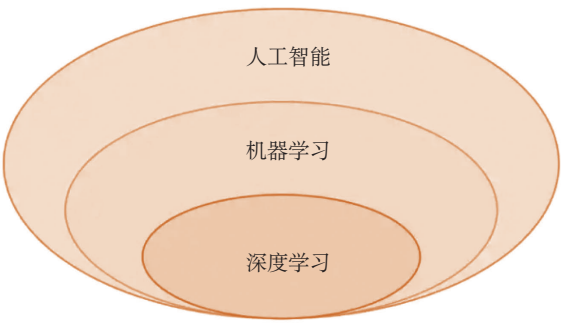


图 1-1

1.1.3 中国人工智能发展史

中国的人工智能研究起步较晚，且发展道路曲折坎坷。在20世纪五六十年代，由于受到苏联批判人工智能和控制论的影响，中国在人工智能领域几乎没有研究。直到20世纪七八十年代，随着改革开放的推进和思想解放的深入，中国的人工智能研究才逐渐活跃起来。

1．艰难起步

20世纪70年代末至80年代，知识工程和专家系统在欧美发达国家得到迅速发展，并取得重大的经济效益。当时中国相关研究处于艰难起步阶段，一些基础性的工作得以开展。例如，自1980年起，中国大批派遣留学生赴西方发达国家研究现代科技，学习科技新成果，其中包括人工智能和模式识别等学科领域。这些人工智能“海归”专家已成为中国人工智能研究与开发应用的学术带头人和中坚力量。

1981年9月，中国人工智能学会（CAAI）在长沙成立，秦元勋当选第一任理事长。这标志着中国人工智能研究进入了一个新的阶段。此后，一些人工智能相关项目也被纳入国家科研计划。例如，在1978年召开的中国自动化学会年会上，报告了光学文字识别系统、手写体数字识别、生物控制论和模糊集合等研究成果，表明中国人工智能在生物控制和模式识别等方向的研究已开始起步。

2．迎来曙光

进入20世纪80年代后期和90年代，中国的人工智能研究迎来了快速发展的时期。1984年召开了全国智能计算机及其系统学术讨论会，1985年又召开了全国首届第五代计算机学术研讨会。这些会议推动了中国人工智能研究的深入发展。

1986年起，智能计算机系统、智能机器人和智能信息处理等重大项目被列入国家高技术研究发展计划（“863计划”）。这为中国人工智能研究提供了有力的政策支持和资金保障。此后，中国的人工智能研究在多个领域取得了显著成果。例如，在机器翻译、语音识别、图像识别等方面都取得了重要进展。

3．蓬勃发展

进入21世纪后，中国的人工智能研究迎来了蓬勃发展的时期。随着大数据、云计算、物联网等技术的快速发展，人工智能在各个领域的应用越来越广泛。中国在人工智能领域的创新能力也不断增强。例如，在2023年，中国科学院自动化研究所推出了跨模态通用人工智能模型——紫东太初。这是全球首个图文音（视觉—文本—语音）三模态预训练模型，同时具备跨模态理解与跨模态生成能力。这一成果标志着中国人工智能的发展已经从狭义人工智能（ANI）跨入了通用人工智能（AGI）的阶段。

此外，中国在人工智能产业方面也取得了显著成就。例如，在工业机器人领域，中国

已经成为世界最大的工业机器人市场之一。多家中国企业在工业机器人研发、生产和销售方面取得了重要进展。在自动驾驶领域，百度等中国企业也推出了无人驾驶出行服务，并在多个城市进行了试点运营。

4．政策支持与战略规划

进入 21 世纪第 2 个十年，中国政府高度重视人工智能技术的发展，将其视为推动经济转型升级及提升国家竞争力的关键力量。2015 年，国务院发布了《中国制造 2025》，首次将人工智能纳入国家重大战略，明确提出要加快人工智能技术在制造业中的应用，推动智能制造发展。随后，一系列政策文件相继出台，为人工智能产业的发展提供了强有力的政策支持和资金保障。

2017 年，国务院印发《新一代人工智能发展规划》，这是我国首个面向 2030 年的人工智能发展国家战略，明确了我国新一代人工智能发展的战略目标、重点任务和保障措施。规划提出，到 2020 年，人工智能总体技术和应用与世界先进水平同步；到 2025 年，人工智能基础理论实现重大突破，部分技术与应用达到世界领先水平；到 2030 年，人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心。

5．科研突破与国际合作

在政策的引导下，中国在人工智能领域取得了显著突破。在基础理论研究方面，中国学者在机器学习、深度学习、自然语言处理等领域发表了大量高水平论文，部分研究成果在国际上处于领先地位。在应用技术研究方面，中国在计算机视觉、语音识别、智能机器人等领域取得了重要进展，涌现出一批具有国际竞争力的创新企业和产品。

同时，中国积极参与国际人工智能合作与交流，与多个国家和国际组织建立了合作关系。通过参与国际标准制定、举办国际会议、开展合作项目等方式，中国在国际人工智能领域的影响力不断提升。例如，中国科学家在多个国际人工智能顶级会议上担任重要职务，推动了中国与国际人工智能界的交流与合作。

6．产业应用与经济社会影响

随着人工智能技术的不断成熟和普及，其在经济社会各领域的应用日益广泛。在智能制造领域，人工智能技术被广泛应用于生产流程优化、质量控制、设备维护等方面，提高了生产效率和产品质量。在智慧城市领域，人工智能技术助力城市治理现代化，提升了公共服务水平和城市运行效率。在医疗健康领域，人工智能技术辅助医生进行疾病诊断和治疗方案制定，提高了医疗服务的精准性和效率。

人工智能的发展还催生了新的产业形态和商业模式。例如，基于人工智能技术的智能客服、智能推荐、智能金融等服务模式不断涌现，为消费者提供了更加便捷、个性化的服务体验。同时，人工智能技术的发展也带动了相关产业链的发展，形成了包括硬件制造、软件开发、数据服务、应用集成等在内的完整产业生态。

7．人才培养与教育创新

人工智能的发展离不开高素质的人才支撑。为了培养更多适应人工智能发展需求的专业人才，中国高校和科研机构纷纷开设人工智能相关专业和课程，加强人工智能领域的人才培养。同时，国家也出台了一系列政策措施，鼓励和支持高校、科研机构与企业开展产学研合作，共同培养人工智能领域的创新型人才。

在教育创新方面，中国积极探索人工智能与教育教学的深度融合。例如，利用人工智能技术开发智能教学系统、个性化学习平台等，为学生提供更加个性化、精准化的学习服务。同时，人工智能技术也被应用于教育评价、教育管理等方面，提高了教育管理的科学性和效率。

8．面临的挑战与未来展望

尽管中国人工智能发展取得了显著成就，但仍面临一些挑战和问题。例如，人工智能基础理论研究仍需加强，关键核心技术有待突破；人工智能伦理、法律和社会影响等问题亟待解决；人工智能产业发展还存在区域不平衡、人才短缺等问题。

展望未来，中国人工智能发展将呈现以下趋势：一是人工智能技术将不断创新突破，推动经济社会各领域智能化水平不断提升；二是人工智能产业将加速发展，形成更加完善的产业生态和竞争格局；三是人工智能将与其他技术深度融合，催生更多新的产业形态和商业模式；四是人工智能伦理、法律和社会影响等问题将得到更加关注和重视，推动人工智能健康、可持续发展。

1.2 认识 AIGC

生成式人工智能（AI generated content，AIGC）是一种人工智能技术，能够根据输入的数据生成新的内容，如文本、图像、音频或视频。生成式人工智能通过学习训练数据中的模式和特征，创建与训练数据相似但并不完全相同的内容。其核心是生成式模型，这些模型通过对数据的理解来生成新数据，而不是仅仅进行分类或预测。

常见的生成式模型包括生成对抗网络（GAN）、变分自编码器（VAE）和自回归模型（如 GPT 系列）。这些模型被广泛应用于多个领域，如自动文本生成（如 ChatGPT）、图像生成与修复（如 Deepfake）、音乐创作、游戏设计，以及增强现实和虚拟现实等。

1.2.1 AIGC的原理

生成式人工智能技术发展按时间顺序先后经历了图像生成、文本生成、跨模态生成三个主要阶段。其中，图像生成包括自动编码器生成、生成对抗网络生成、扩散模型生成三个主要阶段；文本生成包括 *N*-gram 模型生成、LSTM（long short-term memory，长短

期记忆网络）模型生成、基于 Transformer 模型生成 3 个主要阶段；跨模态生成包括图生文、文生图、文生视频 3 个主要领域。

1．图像生成

图像生成算法是一种使用算法模型从零开始创建新的图像或基于已有图像实现风格迁移的计算机视觉技术，相比传统的真实场景拍照或者人工创作等图像生成方式，基于人工智能的图像生成具有更显著的创造性和多样性，在艺术创作、视觉特效、虚拟现实、产品设计等领域得到了广泛的应用。典型的图像生成算法包括图像去噪自动编码器（denoising auto encoder，DAE）、变分自编码器（variational auto encoder，VAE）、去噪扩散概率模型（denoising diffusion probabilistic models, DDPM）。

（1）图像去噪自动编码器（DAE）。自动编码器（auto encoder，AE）的工作过程通常是：首先使用编码器把高维信息编码为低维信息，其次使用解码器将低维信息解码回原来的高维信息，且要求解码尽可能保证能恢复出原来的样子。上述过程中的低维信息被叫作表征（representation）或嵌入表示（embedding）等。该类算法有以下显著特点。

① AE 是一个无监督的机器学习方法。无监督意味着 AE 无须人工标注数据集昂贵的时间和人力成本。通过筛选目标场景相关的图像数据进行无监督训练，AE 能将训练数据分布的冗余信息储存在模型的参数里。

② AE 可实现高效的压缩和表征。虽然降维几乎必然会损失信息，但是 AE 中的编码器部分和解码器相互配合，在实现高压缩率的同时，实现表征信息的语义有效性和泛化性，能够将损失的部分配合补齐，尽可能降低损失。

DAE 是一种深度学习模型。DAE 作为 AE 的一个常见变体，不同于 AE 直接还原输入图片，DAE 给训练数据加一个噪声（符合一些分布的随机数），然后让解码器还原出原始未加噪声的图片，即 DAE 在 AE 的基础上还要额外学会去噪。因此，通过训练 DAE，能够学习到从损坏数据到原始数据的映射关系，从而实现对图像的去噪处理，常用于从损坏数据中恢复原始未损坏数据。DAE 结构图如图 1-2 所示。

DAE 由编码器和解码器两大核心模块构成。在结构与功能层面，编码器是一个多层神经网络，它以包含信息损耗的图像（如添加高斯噪声、随机掩码部分区域的原始图像）作为输入，通过压缩编码机制，将图像数据转化为低维特征向量，实现对原始图像关键信息的提取与浓缩。

解码器同样采用多层神经网络架构，其核心任务是将编码器输出的低维特征向量还原为原始图像形态。在 DAE 的训练进程中，模型以原始图像作为参照标准，通过计算原始图像与解码器输出图像之间的差异，即重建误差，并利用反向传播算法最小化该误差，从而持续优化模型参数。在此过程中，DAE 逐步学习并掌握图像修复和去噪的能力，使得解码器能够在输入受损图像对应的特征向量时，输出尽可能接近原始图像的重建结果，有效恢复图像的原始细节与质量。

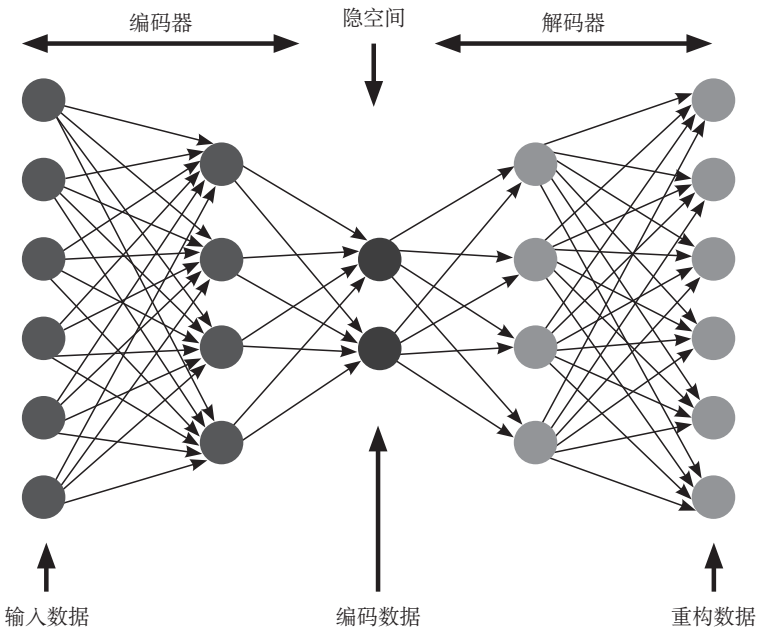
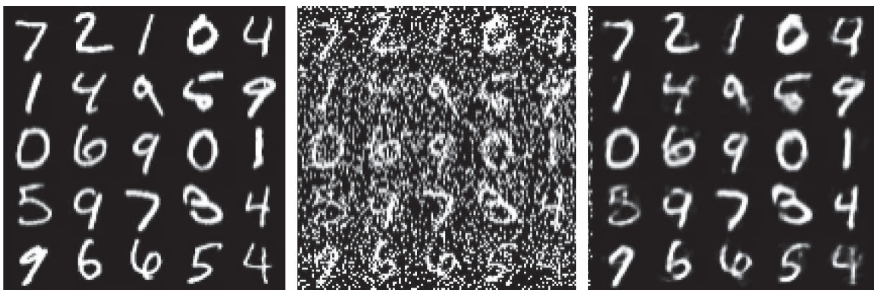


图　1-2

DAE 在图像处理领域具有广泛的应用，通过对 DAE 进行去噪训练，可以从带有噪声的图像中恢复出高质量的未损坏图像，从而提高图像的视觉效果和后续处理的准确性，类似能力也被应用到图像修复、超分辨率等场景。手写数字示意图如图 1-3 所示。



（a）原始图像　（b）有损图像　（c）重建图像

图　1-3

（2）变分自编码器（VAE）。训练好的传统自编码器或图像去噪自动编码器，理想情况下，将编码器生成的隐空间低维特征表示向量输入给解码器，解码器会输出一张与原始输入相似的图片。值得注意的是，上述过程是“重建”而非“生成”。以图片为例，将之前编码后的隐空间特征输入给解码器，解码器会输出和原始输入相似的图片。而如果我们尝试将与隐空间特征维度一致的随机噪声输入解码器，那么得到的将是无意义的噪声图片。因此，自编码器或图像去噪自动编码器的“生成”并不是真正意义上的“生成”，更准确地说应该是“重建”。自编码器图像生成示意图如图 1-4 所示，其中，z 表示潜在变量，长方形图表示无意义图像。

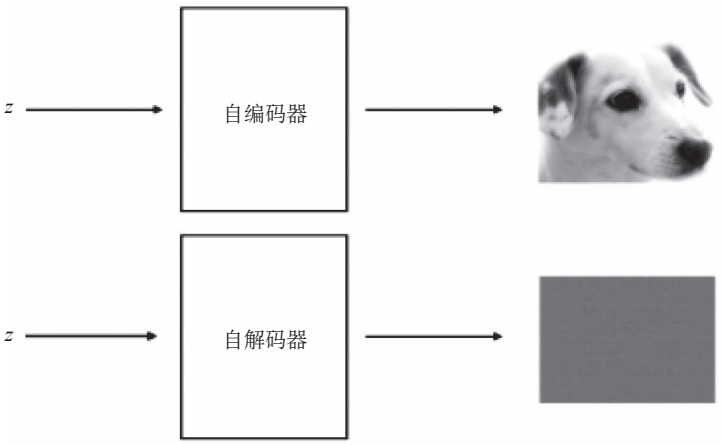


图 1-4

为了实现真正的“生成”，VAE 在 AE/DAE 的基础上进行了两个主要的更新：VAE 让编码器能够输出均值和方差，在推理阶段则从这样的正态分布里采样一个数据作为解码器的输入。与 AE 不同，给定一个输入样本 x，我们希望得到一个隐层分布，而不是一个固定的隐层表征，使原 AE 编码器的输出发生了一点随机扰动。

AE 的训练目标是，解码器输出尽可能与编码器的输入接近。VAE 在此基础上增加了一项训练目标：让编码器输出尽可能贴近标准正态分布，解码器则被强迫从标准正态分布重建，具体实现是在训练损失函数上添加 KL 散度项。通过学习数据的潜在分布，VAE 可以生成具有多样性和连续性的图像，它被广泛应用于图像压缩、图像去噪，甚至通过和传统分割模型结合，可以应用到图像分割领域。VAE 架构图如图 1-5 所示。

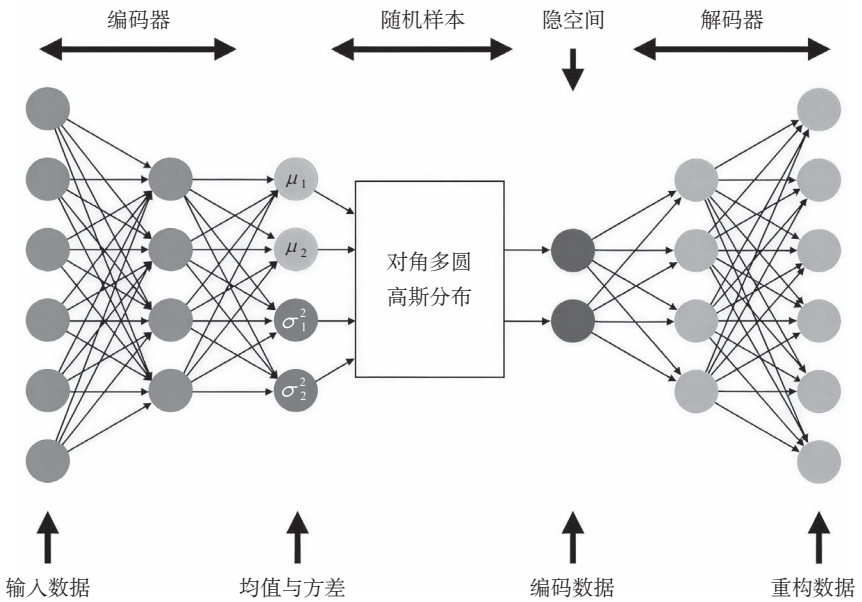


图 1-5

（3）去噪扩散概率模型（DDPM）。其生成过程如图 1-6 所示。

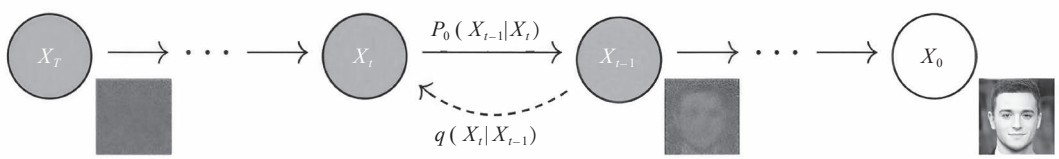


图　1-6

在扩散模型应用到深度生成模型之前，各类深度生成模型在多种数据模态上已经展示了高质量的生成效果，如生成对抗网络（GAN）、自回归模型、流模型和变分自编码器（VAE），生成的对象包括图像、音频等模态数据类型。DDPM 在 2020 年被提出，成功引领了扩散模型在生成领域的持续火热，后续绝大多数的扩散模型也是基于 DDPM 扩展，掌握 DDPM 技术原理有助于快速了解类似结构的扩散模型。

扩散原本是物理学中一个基于分子热运动的运输现象，是指分子通过布朗运动从高浓度区域向低浓度区域运输的过程，扩散模型的灵感来自非平衡热力学，其核心思想是通过向图片中加入高斯噪声模拟上述物理学过程，然后通过逆向过程从随机噪声中生成图片。

扩散模型跟 GAN 或者 VAE 的最大区别在于：扩散模型不是通过一个模型来进行生成的，而是基于马尔可夫链，通过学习噪声来生成数据。对比 GAN 网络模型，扩散模型训练过程中没有对抗，因为对抗训练过程互相博弈的两个模型，训练的难度大幅降低。另外，在训练效率方面，扩散模型还具有可扩展性和可并行性，在生成模态扩展及加速训练过程等方面也存在更广阔的空间。

DDPM 主要分为以下两个过程。

① 前向扩散加噪：输入真实图像，逐步对其添加高斯噪声（扩散过程是一种马尔科夫链），在进行了足够多次的加噪后，图像会被高斯噪声淹没，此时的图像可以认为满足各向同性的高斯噪声分布。与此同时，前向扩散加噪过程是构建训练样本标签至关重要的一步。

② 反向逆扩散去噪：输入是噪声图像，使用神经网络模型对其逐步去噪，最终复原出没有噪声的图像。和前向扩散不同的是，前向扩散里每一个条件概率的高斯分布的均值和方差是已经确定，而逆扩散过程里面的均值和方差是通过网络模型学习得到的。

2．文本生成

（1）*N*-gram 模型。*N*-gram 模型是一种常用的传统文本分析模型，*N* 表示 *N*-gram 的大小，意指基于 *N* 个连续词语或字符的序列模型。通过对大量文本数据进行统计分析，*N*-gram 模型可以学习到词语或字符之间的相关性，进而预测文本中下一个词或字符的概

率分布。它在文本生成、语言模型等场景得到了广泛的应用。通常 N 的取值为 1、2、3 等。

- Unigram（1-gram）：以一个单词或一个字符为一个单位，如 I、love、AI。
- Bigram（2-gram）：以两个相邻的单词或字符为一个单位，如 I love、love AI。
- Trigram（3-gram）：以三个相邻的单词或字符为一个单位，如 I love AI。

使用 N-gram 模型生成文本的方法：随机选择一个 N-gram 作为起始点，然后根据模型中的 N-gram 条件概率来选择接下来的 N-gram。以此类推，直到生成所需长度的文本。

（2）LSTM 模型。LSTM 模型是胡贝尔教授在 1997 年针对传统循环神经网络在长序列建模中存在的梯度消失问题而提出的一种新型网络架构。LSTM 模型示意图如图 1-7 所示。

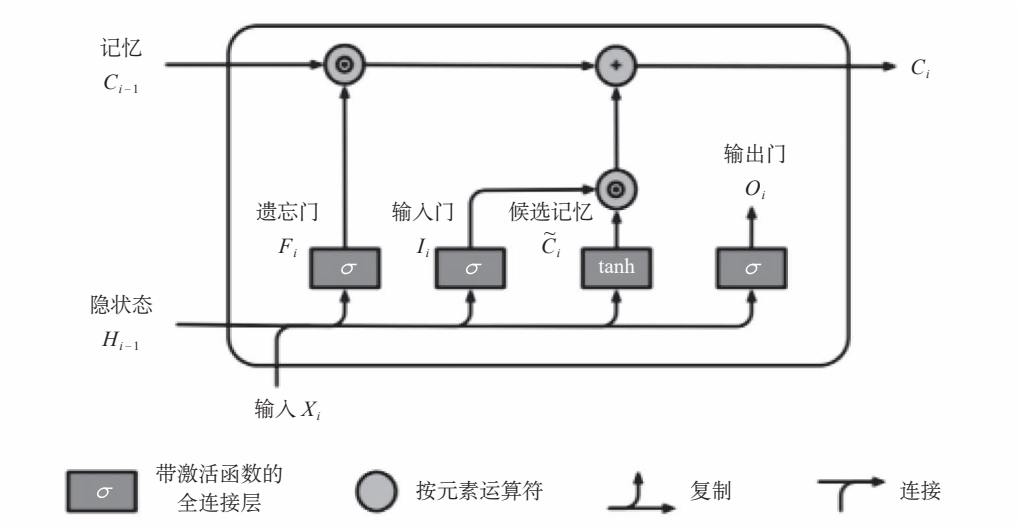


图 1-7

LSTM 模型包括以下两部分记忆。

- 长时记忆：记忆长期状态。
- 短时记忆：记忆当前状态。

LSTM 模型实现的关键是门控机制，一共包括以下三个门。

- 遗忘门：决定了上一时刻的单元状态有多少保留到当前时刻长时记忆中。
- 输入门：当前时刻网络的输入 X_t 有多少保存到长时记忆 C_t 中。
- 输出门：控制长时记忆 C_t 有多少输出是 LSTM 模型的当前时刻输出。

LSTM 模型最终的输出是由输出门和长时记忆 C_t 共同决定的。

LSTM 模型提出后，相当长时间内 NLP 序列建模任务占据主导地位，但仍然存在以下两个主要瓶颈：

- LSTM 模型保留了循环神经网络的网络结构，使 LSTM 模型无法并行化处理数据，尤其是数据量较大时，效率问题更加明显。

- 尽管引入长时记忆在一定程度上克服了循环神经网络梯度消失爆炸问题，但在处理长序列数据时仍然会出现梯度消失或梯度爆炸的问题。

（3）Transformer 模型。原始的 Transformer 模型是标准的序列化文本生成模型，模型学习将输入序列 A 转化为输出序列 B。这是一个广泛应用于文本生成任务的框架，典型任务如机器翻译、摘要生成。基于 Transformer 模型的文本生成主要包括两种技术路线。

① 基于完整 Transformer 模型生成。典型应用如 BART 模型，其编码器端的输入是加了噪声的序列，解码器端的输入是右移的序列，解码器端的目标是输出原序列。模型设计的目的很明确，就是在利用编码器端的双向建模能力的同时，保留自回归的特性，以适用于生成任务。双向编码器与自回归解码器结构图如图 1-8 所示。

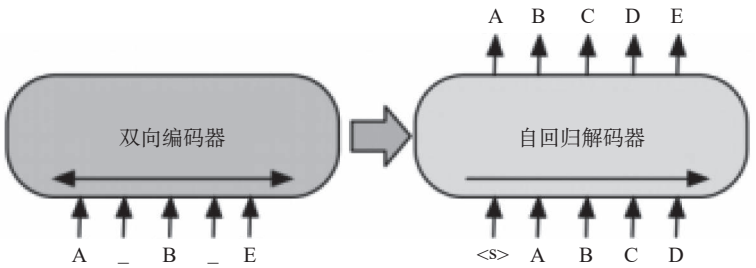


图 1-8

② 基于 Transformer 解码器的自回归模型。典型应用如 GPT 模型。GPT 模型的训练分为监督预训练和有监督的下游任务微调训练。预训练阶段，核心思想仍然是通过大规模的无监督预训练学习，将丰富的语言知识融入模型中，从而提高对各种自然语言处理任务的泛化能力。具体而言，预训练采用自回归的方法训练，给定 k 个连续的词，下一个词出现的概率，在注意力权重的计算过程中，通过添加注意力掩码（attention mask），迫使模型无法学习双向交互，在模型生成过程中每个位置只能关注到其左边上文已生成的词元（token），以匹配序列生成任务推理阶段无法看到未来的 token 的情况。即 GPT 模型在训练阶段和推理阶段的自回归性完全保持一致。自回归解码器流程图如图 1-9 所示。

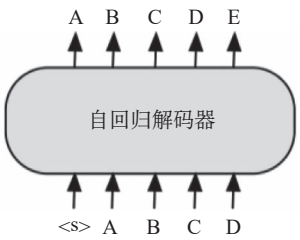


图 1-9

3．跨模态生成

因得益于 Transformer 模型的发展，计算机视觉图像生成和自然语言处理文本生成的技术架构得到统一，进而发展出更多形态的跨模态生成算法模型。

（1）图生文。图生文模型即输入为图像，输出为文本的跨模态模型，主要应用于看图说话、图文问答等领域。

UNITER 模型是微软公司 2020 年提出的一个典型多模态预训练模型，如图 1-10 所示。在该模型中可输入匹配的图像和文本对，目标是实现图像和文本之间的多模态理解与表示学习。通过将图像和文本信息结合起来，UNITER 模型具备能更好地理解和处理同时包含了图像和文本数据任务的能力，例如图像标注、视觉问答、文本图像检索等任务。

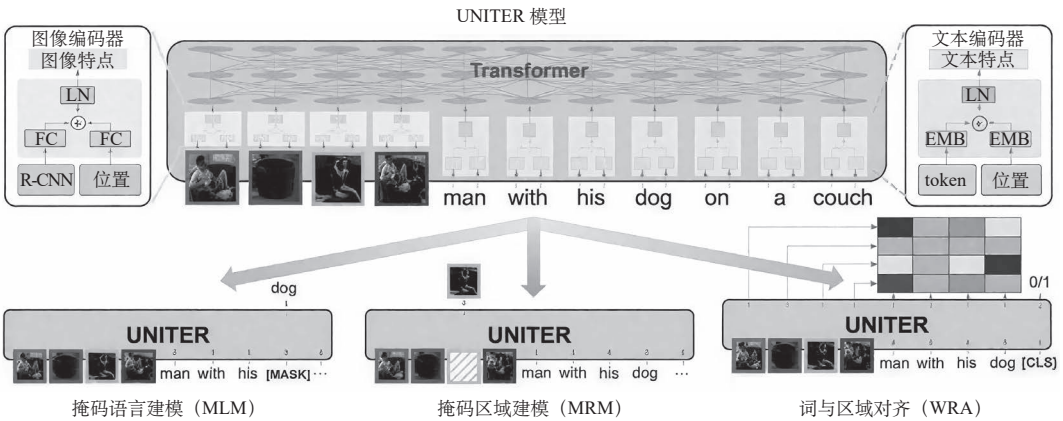


图 1-10

UNITER 模型包含以下三个主要模块。

① 图像编码器：使用 Faster R-CNN 模型抽取每个区域的区域特征，同时用一个七维的向量来编码每一个区域的位置特征，将区域特征和位置特征融合后，可以获得图像的表征信息。对应的预训练任务是掩码区域预测（masked region modeling，MRM），随机遮挡掉图片中提取出来的一些区域，然后让模型学习如何恢复这些遮挡的区域。

② 文本编码器：类似 Bert 模型，序列嵌入和位置嵌入融合后得到文本特征嵌入。对应的预训练任务是掩码语言建模（MLM），即随机遮挡掉一些词，然后训练模型让其学习恢复原始 token。

③ 多层 Transformer：将图像和文本两部分特征输入多层 Transformer，即是 UNITER 模型，对应的预训练任务是图文匹配（image-text-matching，ITM）。ITM 任务中抽取图文匹配的正样本对或者图文不匹配的负样本对，让模型去预测输入是正还是负。

UNITER 模型通过在大规模图像和文本数据集上进行预训练，以学习图像和文本的多模态表示。模型通过学习将图像和文本的表示对齐，使得在多模态任务中能够有效地跨越模态的统一表达，因此预训练 UNITER 模型可以迁移到各种不同的任务中。

UNITER 模型的推理过程如下。

① 推理输入：一幅图像和对应的文本描述 / 问题。

② 模型计算：模型提取输入图像—文本数据对的联合嵌入表征，然后使用该嵌入生成文本答案。

③ 推理输出：根据图像内容回答文本问题，输出文本描述答案。

（2）文生图。文生图模型即输入为文本提示词，输出为图像的跨模态模型，主要应用于艺术创作、动漫游戏设计等领域。Stable Diffusion 图像生成流程图如图 1-11 所示。

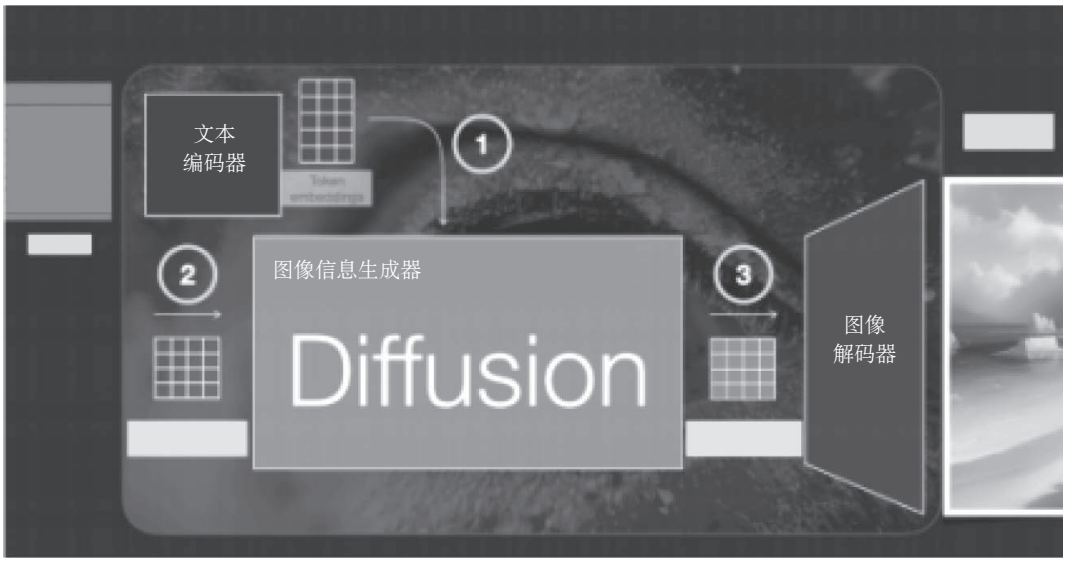


图　1-11

Stable Diffusion（稳定扩散）模型是 2022 年发布的典型文生图模型，技术原理层面，这是一种潜在扩散模型（latent diffusion model，LDM）。它包含三部分：文本编码器（text encoder）、变分自编码器（VAE）、UNet 网络。和大部分扩散模型原理类似，Stable Diffusion 模型也是通过在一个潜在表示空间中迭代去噪来生成图像表征，然后将表示结果解码为完整的图像。Stable Diffusion 模型的出现，让文生图模型能够在消费级 GPU 上，在 10s 内生成图片，大大降低了落地门槛，带动了整个文生图领域的热潮。

Stable Diffusion 模型推理过程如下。

① 推理输入：文本提示描述。

② 模型计算：主要包括三部分。

- 文本编码器：主要功能是编码文本提示词（prompt），获得文本嵌入。这一步非常重要，是有效理解提示词文本并最终能够生成图像的前提。谷歌的 Imagen 模型中提到，语言模型甚至比图像生成模型更关键。
- 图像信息创建器：整个 SD 模型的核心所在，由 UNet 神经网络和调度算法组成，主要功能是在获得了提示词等条件后，生成图像信息表征向量。SD 模型需要根据这些提示词抽象成的数字信息来对一张随机的噪声图进行扩散（diffusion），最终得到图像信息表征向量。

- 图像解码器：在得到图像信息创建器生成的图像信息表征向量后，利用图像解码器将图像信息表征向量转化为最终生成的图像。

③ 推理输出：文本提示中描述对应的图像。

（3）文生视频。文生视频模型即输入为文本提示词，输出为视频的跨模态模型，其主要应用于新闻媒体、动漫创作等领域，如图 1-12 所示。

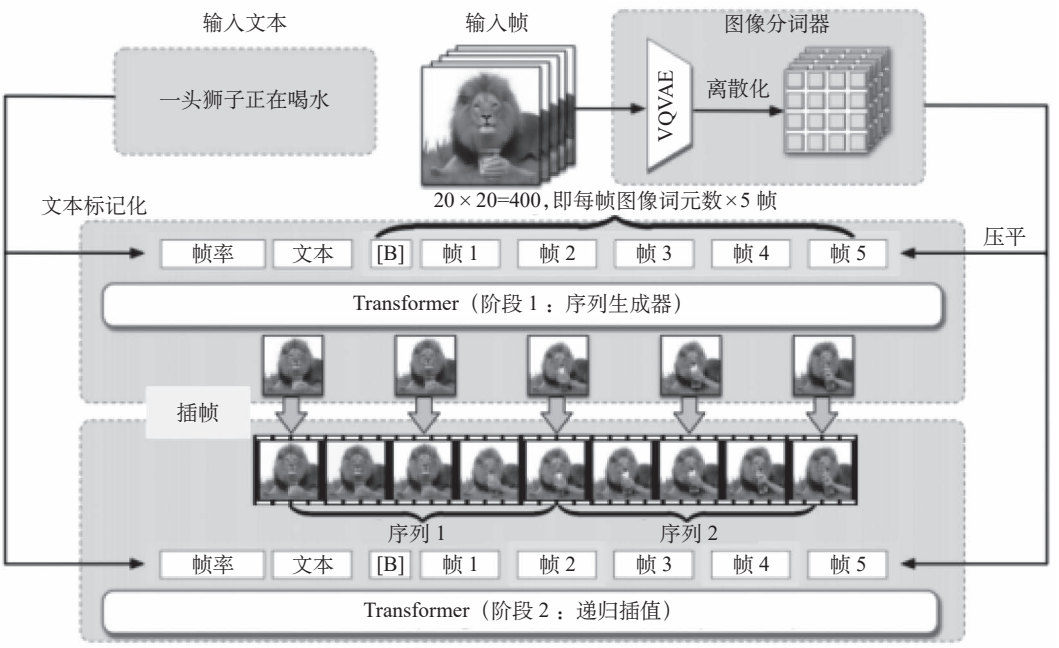


图 1-12

CogVideo 模型是智谱公司在 2022 年发布的典型文生视频模型，目标是构建一个统一的、多任务的视频理解系统，该系统综合利用 CNN、RNN、Transformer 多种网络结构实现了多模态信息的融合，能够对视频中的多个层次的信息进行建模，包括场景、物体、动作、人物、对话等，以便更好地理解视频内容。

CogVideo 模型采用多帧率分层训练策略，提出了一种基于递归插值的方法，即逐步生成与每个子描述相对应的视频片段，并将这些视频片段逐层插值得到最终的视频片段，赋予了 CogVideo 控制生成过程中变化强度的能力，有助于更好地对齐文本和视频语义，最终实现了从文本到视频的高效转换。

CogVideo 模型的主要特点如下。

- 多任务学习：通过联合训练多个任务，如行为识别、场景分类等，提高整个系统的泛化能力和性能。
- 多模态信息融合：利用视频、音频和文本数据之间的互补信息，提高视频理解的准确性。
- 可扩展性：框架设计灵活，易于扩展和集成新的数据集和任务。

- 开源和易于使用：项目采用 Python 实现，并提供了详细的文档和教程，方便研究者和开发者使用和定制。

CogVideo 模型的推理过程如下。

① 推理输入：文本提示描述。

② 模型计算：主要包括两部分。一是 CogView2，即通过文本生成几帧图像，这时候合成视频的帧率还很低；二是递归插值，循环插帧基于双向注意力模型理解前后帧的语义，对上一步生成的几帧图像进行插帧，进而生成帧率更高的完整视频。

③ 推理输出：文本提示中描述对应的视频。

1.2.2 Transformer模型简介

在自然语言处理的漫长演进过程中，我们见证了众多技术的兴起与更迭。从最早的基于规则的方法到后来的统计模型，再到深度学习时代的递归神经网络（RNNs）和长短期记忆网络（LSTMs），NLP 领域一直在寻求更高效、更准确的语言理解和生成方式。尽管 RNN 及其变种在处理序列数据时表现出色，但它们在处理长距离依赖关系时仍存在一定的局限性，并且在训练速度和并行化方面也面临挑战。

正当业界寻求突破之时，一种全新的架构应运而生。2017 年 6 月，Google 发布了一篇论文 *Attention is All You Need*，中文译名为《注意力就是你所需要的一切》，提出了 Transformer 模型。它摒弃了传统 RNN 中的时间步依赖性，转而采用一种完全基于注意力机制（attention mechanism）的编码—解码结构，从而实现了对序列数据的高效处理，并极大提升了模型训练的速度和效果。更重要的是，Transformer 模型不仅用于语言任务领域，它在图像处理等领域也展现出了强大的潜力。Transformer 模型是如何重新定义自然语言处理领域的？我们来看之前的 NLP 领域，它存在两个值得注意的问题。

（1）以 LSTM 长短时记忆网络为代表的循环神经网络虽然缓解了梯度消失的问题，但串行处理长序列需要进行文本建模，模型训练速度依然是瓶颈。

（2）word2vec 无监督训练获得向量表征的方法已经出现，但是该表征向量是静态的，因此出现了一些局限性，比如，无法适应不同的上下文，无法处理新词，缺乏全局信息等。

而 Transformer 模型将自注意力机制作为其核心计算单元，它将源于 RNN+Attention 模型的注意力机制提升为核心功能，解决了上面提到的 RNN 的顺序计算瓶颈和 CNN 的局部特征的限制，实现了真正的全局交互；其完全并行的架构极大提升了训练效率，为超大规模模型的实现铺平了道路；其强大的上下文表示能力为完成复杂 NLP 任务奠定了坚实基础。

1. Transformer 模型的 Seq2seq 架构

Seq2seq 架构是由序列生成序列的架构，是一种可以将序列 A 映射到序列 B 的神经

网络机器学习模型。Seq2seq 架构早期主要应用于机器翻译、语音识别等方面。但受限于网络表达能力有限，它在文本生成任务上表现不足，直到 Transformer 模型出现，文本生成局面才被彻底颠覆。

Transformer 模型的 Seq2Seq 架构的核心是“编码器—解码器 + 注意力机制”，它的本质是用 Transformer 的注意力机制（而非早期的 RNN）来实现序列生成序列转换目标的架构。它继承了 Seq2Seq 的经典任务框架，又通过 Transformer 的核心机制解决了早期 RNN 和 Seq2Seq 的不足。

与 Transformer 模型类似，早期的 Seq2seq 架构也包含编码器和解码器两部分，编码器负责将输入序列编码表征为一个特征向量，解码器基于这个特征向量迭代预测生成下一个字符。两者的区别在于：早期 Seq2seq 架构的编码器和解码器通常使用不同循环神经网络 RNN 模型，如果输入序列的长度很长，RNN 网络的编解码能力对上下文依赖关系的表征不足，因此任务表现会随着输入序列长度的增长而逐步降低；Transformer 模型通过堆叠注意力和前馈网络层来构造编码器和解码器，并引入了残差连接和层归一化，可以实现更深层网络的构建和训练，对长序列上下文依赖关系的捕捉能力也大幅增强，从而实现了序列生成能力的突破性提升。Transformer 的 Seq2seq 架构如图 1-13 所示。

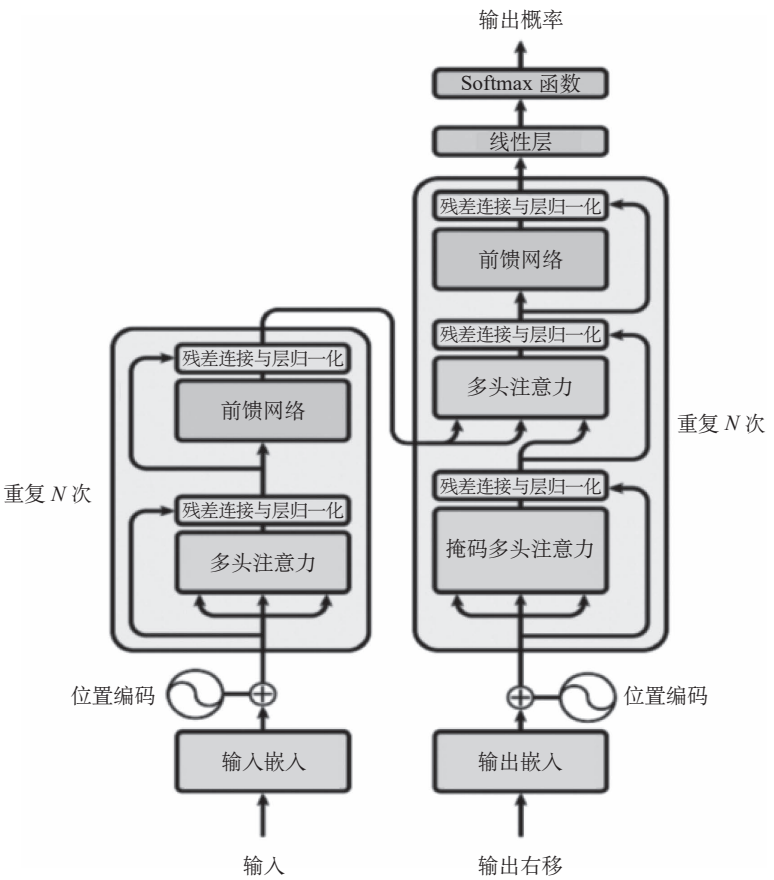


图 1-13

Seq2seq 架构的推理过程如下。

文本序列通过分词、嵌入等步骤输入编码器，获得每个中心词基于上下文的加权表征的输出为稠密向量。例如，BERT Base 中每个 token 字符表征的维度为 768 维，解码器基于编码器获得表征信息，然后生成目标文本序列，如图 1-14 所示。

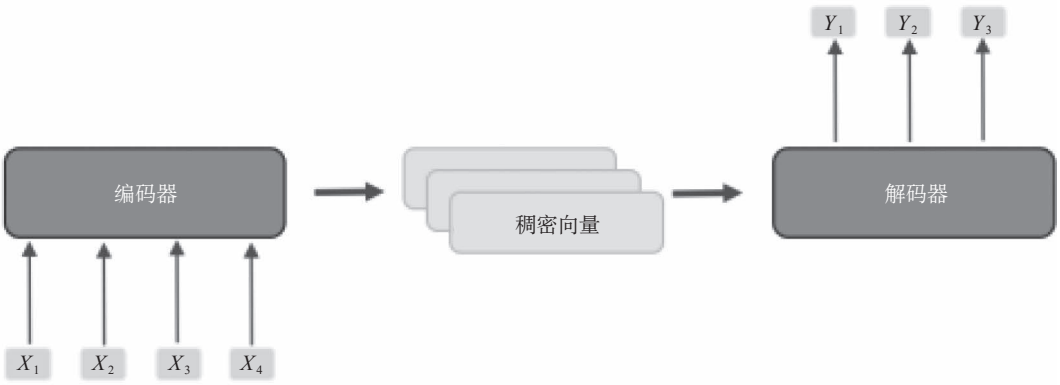


图　1-14

2．Transformer 模型的编码器—解码器架构

编码器—解码器架构是一种功能强大且常见的端到端机器学习架构，主要包括编码器和解码器两个部分。编码器核心功能是提取输入的特征表达，解码器则基于编码器特征生成内容。广义上的编码器—解码器架构可以根据不同的任务类型选择不同的编码器和解码器，因此不仅在自然语言处理领域广泛使用，对全部模态的数据类型都适用，即无论输入是文本、图像还是音频，都可以使用编码器对其进行特征抽取表征，再使用解码器生成符合训练标签预期的内容。例如，在计算机视觉领域，编码器—解码器架构是一种流行的深度学习图像分割模型，编码器采用卷积神经网络实现，负责提取图像的特征，而解码器使用反卷积层实现，将编码器特征解码为像素级别的分割结果，在大量生产实践中已实现准确的分割。Transformer 模型继承了 NLP 领域标准的编码器—解码器架构，与之前基于编码器—解码器架构的模型不同，Transformer 模型不再使用 CNN、RNN、LSTM、GRU 等网络结构构建编码器和解码器，而是使用堆叠以注意力机制为核心的标准网络层来构建编码器和解码器，从而获得更好的序列生成能力。

结合实际应用和后续网络发展，Transformer 模型的编码器—解码器架构主要包括以下三种应用类型。

（1）仅编码器：只使用 Transformer 模型的编码器部分，输入是文本序列，通过多层编码器层的表征，得到强大的上下文语义表征，如图 1-15（a）所示。

（2）仅解码器：只使用 Transformer 模型的解码器部分，输入是文本序列 A，通过自回归解码器模型的计算，迭代预测生成目标文本序列 B，如图 1-15（b）所示。

（3）编码器—解码器：完整使用 Transformer 的编码器和解码器，编码器输入为文本序列 A，解码器输入为编码器计算得到的文本序列 A 语义表征和已预测目标序列 B，并由解码器最终生成完整的目标序列 B，如图 1-15（c）所示。

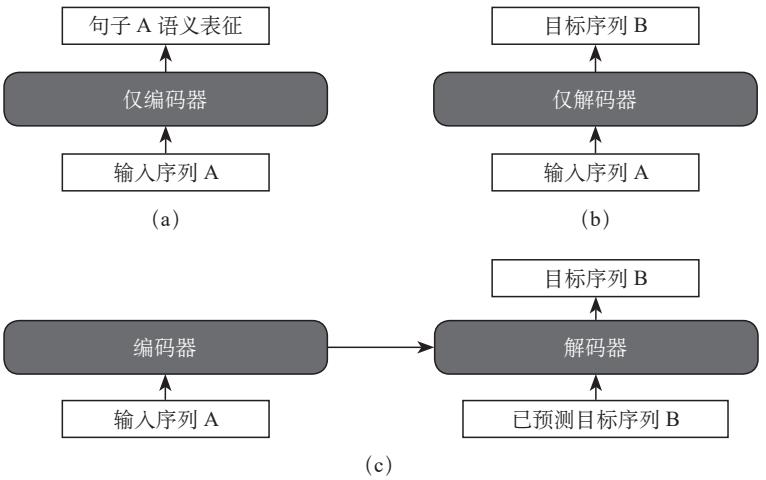


图 1-15

1.2.3 AIGC的发展

AIGC 发展阶段可以按内容生产方式发展阶段和技术发展阶段两个维度，从业务层和技术层分别进行分析和回顾。

1．生成内容方式的发展阶段

1）专业生成内容

（1）定义。专业生成内容（professional generated content，PGC）是指由专业的内容创作者或团队进行创作、编辑和发布的内容。PGC 创作方式起源于传统媒体时代，如报纸、杂志、电视和电影等，由传统的广电工作者按照几乎与电视节目相同的方式来进行制作。得益于创作者的专业性，且经过专业的编辑、制作和策划，这类内容通常具备较高的质量和可靠性，已经应用到各种领域，如网站、应用程序、短视频和音乐等。

（2）特点。PGC 的特点包括以下方面。

① 专业团队制作：PGC 由专业人士负责，通常是拥有专业知识、有内容相关领域资质、具有一定权威的团队或机构。

② 内容质量高：专业化内容、优质化内容、有价值内容。

③ 制作成本高：由于对制作团队专业性要求高，后期加工质量控制严格，PGC 制作的成本高昂。

2）用户生成内容

（1）定义。用户生成内容（user generated content, UGC）是一种用户使用互联网的

新方式，用户在网络上向他人展示自己的原创作品或向他人提供内容；泛指用户以任何形式在网络上发表创作的文字、图片、音频、视频、音乐、博客、评论等内容。这种创作方式是由 Web 2.0 时代随着社交网络和博客的出现而流行起来，典型的应用场景包括社交网络、在线论坛、博客、知识共享平台等。UGC 同类用户逐步聚集，逐步发展形成网络社区团体，代表性的社区或应用有小红书、抖音、百度贴吧等。

（2）特点。UGC 的特点包括以下方面。

① 用户自主生成：互联网侧重平台功能，内容主要由平台的用户自主生成，平台协调和维护秩序，充分利用流量优势提升用户参与度。

② 体现用户个性化：各大论坛、博客和微博客站点等内容均由用户自主创作并提交发布，可以充分发挥想象力和创新性，生成的内容也更加灵活。

③ 制作过程相对随意：由于创作内容用户的创作能力参差不齐，因此内容平均质量不高，且存在不可控风险，需要平台设计规则加以约束或遴选出优质内容。

3）AI 辅助生成内容

（1）定义。AI 辅助生成内容（AI-assisted generated content，AAGC）是介于 UGC 和 AIGC 之间的一种内容生成方式。在以 PGC 和 UGC 为主的创作框架内，平台通过开放 AI 工具协助用户创作，创作者发出提示词使 AI 半自动辅助生成内容，指示 AI 完成复杂的定制生成任务。但受限于技术现状，AI 暂不具备成为创作者进行自主创作的能力，仅是扮演辅助角色，创作者依然需要在关键环节创作内容或输入提示词。随着数据、算法等核心要素不断地升级迭代，AAGC 将逐步向 AIGC 方向发展，从而突破人工限制，提升到自主创作的水平，进而创作出更丰富多样的内容。

（2）特点。AAGC 的特点包括以下方面。

① AI 为辅且人工为主：AI 暂不具备自主创作的能力，因此仅是扮演辅助角色，用户依然需要在关键环节创作内容或输入提示词，对生成过程进行指导和控制。

② 个性化能力有限：由于 AI 为辅且人工为主，生成过程中用户在关键环节输入内容或提示词，生成结果通常会遵从用户的主观意图，因此一定程度上和 UGC 更接近，结果更加可靠和专业，但个性化能力不足。

③ 多阶段制作筛选：由于是以半自动的方式生成内容，生成过程伴随用户的关键节点介入而分为若干阶段，用户对阶段性的生成结果进行审核、评价、筛选，据此对下一阶段的生成提示词进行调整。

4）AI 生成内容

（1）定义。AI 生成内容（AI generated content, AIGC）是指人工智能通过学习海量现存数据，利用人工智能的理解力、想象力和创作力，让 AI 全程端到端完成内容创作，或者根据指定的需求和风格，创作出各种形态的内容，比如图片、视频、音乐、文字、3D 模型、代码等。得益于 Transformer 预训练模型和跨模态技术的快速发展，AIGC 的通用化