基于强化学习的生成式人工智能综述

尹义鹏[1],施水才[1,2],黄自力[1]

(1. 北京信息科技大学 计算机学院,北京 100101;2. 拓尔思信息技术股份有限公司,北京 100096)

摘 要: 生成式人工智能一直是机器学习领域的热点话题,在文本生成、计算机视觉等多个领域都有广泛应用。最大似然估计方法确立了生成模型的训练目标,但在满足用户个性化需求方面尚有不足之处。近年来,强化学习通过引入人为定制的评价机制等新的训练信号,在构建高性能模型方面展示出一定潜力,为生成式人工智能模型的设计与应用带来突破性进展。对生成式人工智能领域的最新进展进行全面系统的回顾,从跨领域的视角进行分类总结,梳理各类模型及其应用场景;重点关注快速发展的大模型技术,探讨当前存在的限制及未来的发展方向。通过剖析模型的弱点全面深入地了解生成人工智能的理论基础与实践应用现状,以期有助于指导行业实践,亦为未来研究提供参考。

关键词: 强化学习;生成式人工智能;最大似然估计

DOI: 10. 11907/rjdk. 232227 开放科学(资源服务)标识码(OSID):

中图分类号: TP391.1 **文献标识码:** A **文章编号:** 1672-7800(2025)001-0183-10

Reinforcement Learning for Generative AI: A Review

YIN Yipeng[1], SHI Shuicai[1,2], HUANG Zili[1]

(1. *School of Computer, Beijing Information Science and Technology University, Beijing* 100101, *China*;
2. *TRS Information Technology Co., Ltd, Beijing* 100096, *China*)

Abstract: Generative artificial intelligence has always been a hot topic in the field of machine learning, widely used in various fields such as text generation and computer vision. The maximum likelihood estimation method establishes the training objectives for generating models, but there are still shortcomings in meeting the personalized needs of users. In recent years, reinforcement learning has shown certain potential in building high-performance models by introducing new training signals such as artificially customized evaluation mechanisms, bringing breakthrough progress to the design and application of generative artificial intelligence models. Conduct a comprehensive and systematic review of the latest developments in the field of generative artificial intelligence, classify and summarize various models and their application scenarios from a cross disciplinary perspective; Focus on the rapidly developing large-scale modeling technology, explore the current limitations and future development directions. By analyzing the weaknesses of the model, we aim to gain a comprehensive and in-depth understanding of the theoretical foundation and practical application status of generative artificial intelligence, in order to guide industry practice and provide reference for future research.

Key Words: reinforcement learning; generative AI; maximum likelihood estimation

0 引言

随着计算机技术的快速发展,生成式人工智能成为当前研究的前沿领域,其可通过模拟人类的创造性思维,自动生成文字、图片、音乐等多种类型的内容。从变分自编码器(Variational Auto Encoder, VAE)、对抗性生成网络(Generative Adversarial Network, GAN)、基于能量的模型(Energy-Based Model, EBM)到扩散模型,生成式技术正逐步推动多个领域的技术突破[1]。尤以自然语言处理、图像生成、药物研发等最为惹眼,例如通过机器学习分析大型数据集以预测药物的潜在不良反应,或通过蛋白质折叠和

收稿日期: 2023-12-02
扫描二维码阅读全文:

作者简介: 尹义鹏(2001-),男,北京信息科技大学计算机学院硕士研究生,研究方向为自然语言处理;施水才(1966-),男,硕士,北京信息科技大学计算机学院教授、硕士生导师,研究方向为信息检索、大数据分析和挖掘、人工智能;黄自力(2000-),男,北京信息科技大学计算机学院硕士研究生,研究方向为自然语言处理。

分子设计加速药物研发进程。AlphaFold的成功便证明了生成式人工智能在蛋白质结构建模方面的独特优势[2-4]。最近,大型语言模型(Large Language Model,LLM)如ChatGPT[5]的出现,更是将机器学习系统开发带入了一个新纪元,极大地缩小了实现通用人工智能的距离。

尽管生成式人工智能在多个领域取得了显著进展,但其训练过程的核心目标——设计有效的目标函数指导学习,依然面临挑战。最大似然估计(Maximum Likelihood Estimation,MLE)作为主流的目标函数,在处理某些复杂任务时的局限性逐渐显现。强化学习作为一种能够从交互中学习并具有灵活目标奖励函数的训练范式,为生成式人工智能的训练提供了新的途径[6]。通过将生成问题重新定义为决策问题,强化学习增强了模型对多样任务需求的适应性。因此,强化学习不仅是技术上的突破,而且是推动整个人工智能领域进步的关键力量。本文旨在深入探讨生成式人工智能的发展背景、当前应用以及面临的挑战,同时重点分析强化学习如何促进生成式人工智能发展,以期为未来研究提供新的视角和方向。

1 研究背景

1.1 生成式模型

生成式人工智能涵盖了多种模型,包括基于概率分布的VAE、通过对抗训练实现生成与判别的GAN,以及基于能量函数描述概率分布的EBM。生成式模型旨在学习数据的潜在概率分布以生成新的似然样本,可以将其表述为:对于一个随机变量x或变量序列x_1、x_2...x_n(其中n为序列长度),模型需要估计它们在数据空间X上的概率分布$p(x)$或$p(x_1, x_2, ..., x_n)$。表1为本文使用到的符号及其解释。

Table 1 Symbols used in this article and their explanations

表1 本文使用到的符号及其解释

符号	解释
x	输入变量,在某些应用中,目标变量即为输入变量
$x_1, .., x_i, .., x_n$	变量序列,有时指x的维度
z	潜变量
t	序列中的时间步长,通常用作索引
s_t	马尔可夫决策过程中时间步t的状态
a_t	马尔可夫决策过程中时间步t的动作
r_t	马尔可夫决策过程中时间步t环境给予的奖励
τ	轨迹,又名一系列状态、动作和奖励($s0, a0, s1, r1, ...sn, rn$)
R_t	在一条轨迹上的累计折现收益
Π	强化学习代理的策略
γ	以指数速度降低未来奖励影响的折扣因子
$V_\pi(S_t)$	给定时间步t状态的智能体的价值函数
$Q_\pi(S_t, a_t)$	以状态和动作作为输入的代理的价值函数
$p(\cdot), q(\cdot)$	给定变量的概率分布
$E_{x\sim p_x}[\cdot]$	某个变量对x分布的期望
$D_{KL}(q\|p)$	同一变量的两个分布p和q之间的Kullback-Leibler(KL)散度
$D(\cdot)$	GAN中的判别器
$G(\cdot)$	GAN中的生成器

1.1.1 VAE

VAE通过学习数据的潜在表示来实现数据的生成与重构,其核心思想为通过最大化数据的边缘对数似然来训练模型,同时最小化近似后验与先验的KL散度,以实现对潜在变量的合理编码,通过考虑潜在变量z重建输入x来学习有用的表示。在形式上,可以通过潜变量分解分布$p(x)$,表示为:

$$p(x) = \int p(x|z) \cdot p(z)\, dz \tag{1}$$

$$\ln p(x) \geq D_{KL}\big(q(z|x)|p(z)\big) + E_{q(z|x)}\big[\ln p(x|z)\big] \tag{2}$$

式(1)中:$p(x)$为观测数据的边缘分布,$p(x|z)$为生成数据的分布,$p(z)$为潜在变量的先验分布。该公式表明,VAE试图通过对潜在变量的先验分布$p(z)$进行采样,然后通过生成器网络$p(x|z)$生成与观测数据相似的数据,从而实现对观测数据的重构,这个过程也被称为解码过程。同时,VAE也试图最大化观测数据的边缘对数似然$p(x)$,以实现对观测数据的生成,这个过程被称为编码过程。然而,该公式中的期望在实践中计算困难,一种常用的解决方法是利用变分推断(Variational Inference)引入一个简单的近似后验分布$q(z|x)$,将证据下界(Evidence Lower Bound,ELBO)作为优化目标。表示为:

$$\text{ELBO} = E_{q(z|x)}\big[\log p(x,z)\big] - E_{q(z|x)}\big[\log q(z|x)\big] \tag{3}$$

式中:第一项为重构项,最大化该项可以提高重构质量;第二项为KL散度项,最小化该项可使近似后验$q(z|x)$尽可能接近真实后验$p(z|x)$。通过最大化ELBO,VAE模型可以同时优化重构质量和潜在空间的结构性,从而实现良好的生成与重构能力。

1.1.2 GAN

GAN由鉴别器和生成器组成[7],两个网络彼此博弈。生成器的目标是生成一个对象(比如人的照片),并使其看起来和真的一样;而鉴别器的目标是找到生成出的结果与真实图像之间的差异。鉴别器经过训练可以对样本来源、真实数据集和生成数据集进行分类;而生成器的目的是欺骗鉴别器。因此,两个网络形成零和博弈,从而推动生成器的输出分布逼近真实分布。在形式上,GAN的目标可以定义为:

$$\min_G \max_D E_{x\sim p_{dt}(x)}\big[D(x)\big] + E_{z\sim p_g(z)}\big[1 - D\big(G(z)\big)\big] \tag{4}$$

式(4)描述了GAN中的最小最大优化问题。其中,$D(x)$为判别器网络,$G(z)$为生成器网络,$Ex\sim p_{dt}(x)$为真实数据的分布,$p_g(z)$为生成器的输入噪声分布。最小最大优化过程为GAN中的核心训练过程,即通过生成器与判别器相互博弈来实现生成器网络生成逼真样本,判别器网络准确区分真假样本的训练结果。

1.1.3 EBM

EBM是一类通过能量函数表示概率分布的生成模型。能量函数是一个关于输入数据的标量函数,用于衡量变量的可能性。EBM的核心思想为通过能量函数描述样本的概率分布,能量函数越小,样本概率越大。EBM中概率分

布的计算公式为:

$$p(x)=\frac{e^{-E(x)}}{\int\limits_{x'\sim X} e^{-E(x')}} \tag{5}$$

式中:分母为空间x上的积分,$E(x)$为能量函数,$p(x)$为样本x的概率分布。分母是对所有可能样本的能量函数的指数项求和,用于归一化得到概率分布。如果要计算概率,则需要评估所有能量函数x。如果空间太大,则计算概率非常棘手。

自回归模型是一种深度神经网络生成模型,其可通过链式法则将变量x的联合概率分解为条件概率的乘积。表示为:

$$p(x)=p\left(x_1,x_2,\cdots,x_n\right)=\prod_i p\left(x_i|x_1,\cdots,x_{i-1}\right) \tag{6}$$

自回归模型可应用于序列生成任务,如文本生成和分子设计,而应用于图像生成则尚不成熟。该模型生成的顺序是固定的,在训练和推理过程中不能更改。其通过链式法则分解概率分布,生成过程是逐步进行的,这种顺序生成方式虽然能保证每个时刻均利用到历史信息,但也使得采样过程无法并行化,计算效率较低。

综合来看,以上模型在概率分布描述、训练方式和核心思想上有所不同,但均致力于实现对数据的生成与重构,为生成式人工智能的发展提供了重要的理论与方法支持。

1.2　强化学习

强化学习是一种基于环境互动的机器学习范式,代理通过采取行动并获得反馈来学习最优决策策略,旨在最大化长期累积奖励[8]。强化学习在处理序列决策问题时表现出一定优势,但也存在延迟奖励等挑战。以下首先介绍马尔可夫决策过程(Markov Decision Process,MDP),其是一种可广泛应用于强化学习问题的公式;然后介绍强化学习方法的分类以及这些方法的关键细节。

1.2.1　MDP

在强化学习中,代理通过与环境互动来获得经验,这个互动过程是由一系列离散的时间步t= 0, 1, 2, 3...组成的。在每个时间步t,代理观察环境发出的状态s_t,基于这个状态决定其行动a_t,并接收来自环境的即时奖励r_{t+1},环境随后转换至下一个状态s_{t+1}。在理想情况下,状态s_t提供代理需要的所有信息并决定其最佳行动。代理的学习过程涉及收集新经验与优化其行动选择策略　。

MDP是强化学习问题中最常用的形式化框架,用于描述代理与环境的交互过程[8]。通常由以下5个要素构成:①S为环境状态的集合;②A为代理(学习模型和决策者)的一组动作;③T:$S\times A\rightarrow S$为转移概率分布$p(s_{t+1}|s_t,a_t)$;④$R(s,a,s')$为决定智能体目标的奖励函数;⑤$\gamma\varepsilon[0,1]$为累积奖励计算的折扣因子。

在有限情景的MDP中,状态空间和动作空间是有限的,序列长度也是有限的,这适用于多种生成任务。例如在图像生成中,常用方法是创建有限的RGB像素空间生成图像。由于图像尺寸是有限的,其状态空间和动作空间也因此受限。在情景MDP中,环境在经过T个步骤后会重置,这一系列步骤被称为一个剧集。智能体的目标为在一个剧集中最大化从环境中获得的累积奖励。在每个时间步t,代理根据策略$\pi(at|st)$,通过确定性或概率分布选择能够最大化累积奖励的动作。这个累积奖励即时间步t的回报,可以表示为:

$$R_t=\sum_{k=0}^{T-1}\gamma^k r_{k+t+1} \tag{7}$$

式中:γ为折扣因子,以指数速度降低未来奖励的影响。

代理的目标可被形式化为:

$$\pi^*=\arg\max_{\pi} E\left[R|\pi\right] \tag{8}$$

与无约束强化学习任务相比,MDP的一个关键性质为马尔可夫性,即给定当前状态,未来状态只依赖于当前状态和动作,与过去无关。其中:

$$p\left(s_{t+1},r_{t+1}|s_0,a_0,s_1,a_1,\cdots,s_t,a_t\right)=p\left(s_{t+1},r_{t+1}|s_t,a_t\right) \tag{9}$$

当前状态s_t和动作a_t完全决定了下一个状态s_{t+1}和奖励r_{t+1}。这一属性使得算法可以高效地工作,因为当前状态提供了足够信息来决定下一步的最优行动。

1.2.2　无模型方法

在深入探讨无模型方法之前,首先来看一个引人入胜的案例——神秘游戏。在这个游戏中,各种颜色和形状的色块组成一个网格界面,玩家对游戏规则一无所知,只能通过按键操作。玩家的目标是发现游戏规则,学会如何获得奖励并取得胜利。这个游戏的设计灵感来自于强化学习,玩家只知道可以采取的行动,而对环境信息一无所知。对于没有任何假设或直觉的计算机来说,这是一个巨大的挑战。在这种情况下,无模型强化学习成为解决问题的关键。

无模型强化学习主要包括基于价值函数和基于策略搜索两类方法。此外,还有一类行为批评方法,该类方法同时结合了价值函数与策略的优点。价值函数为式(7)中目标的变体,是对一系列奖励的期望,其在每个时间步骤中帮助优化策略达成目标。价值函数受主体和环境共同影响,但在无模型环境中,由于环境动态未知,价值函数主要由智能体决定。价值函数满足著名的Bellman方程,其递归性质使得可以通过简化模型来近似求解,从而使训练成为可能[9]。表2为经典的无模型强化学习算法。

1.2.3　基于模型的方法

基于模型的方法是一种强化学习技术,其依赖于环境

Table 2　Model-free reinforcement learning algorithm

表2　无模型强化学习算法

强化学习算法	相关算法
基于价值函数	Q-learning[8]、DQN[10]、Soft Q-learning[11]
基于策略搜索	Policy Gradient/REINFORCE[12]、TRPO[13]、PPO[14]
基于行为批判	Actor-Critic[8]、DPG[15]、DDPG[16]、SAC[17]、A3C[18]

的状态转移模型和奖励函数。该类模型可以帮助智能体更好地理解环境，并作出更明智的行动决策。基于模型的强化学习流程见图1。与无模型方法不同，基于模型的方法需要获取或学习环境的状态转移模型$P(s'|s,a)$；然后基于该模型通过规划或搜索等方式求解最优策略π或价值函数$V(s)$。根据表示形式，可将基于模型的方法分为以下几类：

（1）基于规划的方法。该类方法假设已知环境的精确转移模型P和奖励函数R，通过价值迭代或策略迭代等经典动态规划算法求解最优价值函数或策略。其中价值迭代通过不断应用Bellman方程更新价值函数，直至收敛得到V^*；策略迭代则通过不断评估并提升策略，直至收敛得到π^*。这两类算法计算代价高，主要用于小规模的MDP问题。

（2）采样规划方法。当环境模型维度过高时，上述经典算法将变得低效。采样规划通过蒙特卡罗方法对转移模型进行采样近似，提高了规划效率，代表性算法为蒙特卡罗树搜索（Monte Carlo Tree Search，MCTS），其在树的每个节点利用模型进行远程Look-ahead，通过反复模拟得到各个状态的值估计，从而逐步搜索出最优行动序列。

（3）模型辅助策略搜索方法。除了直接规划，转移模型还可以辅助策略搜索过程。例如，AlphaGo Zero中的MCTS策略改进便是利用一个通过自我对弈学习而来的模型用于Look-ahead模拟和引导搜索。另一种策略是首先利用模型辅助搜索出一个基准策略，然后将其作为导师策略进行策略迭代。

在基于模型的强化学习方法中，环境转移模型可来自于先验知识，也可通过从数据中学习获得。例如，对于下围棋的智能体来说，围棋的规则是固定的，可以通过编程得到完美的环境转移模型。当无法获得完美模型时，便需要从数据中学习模型的拟合。AlphaGo便属于这种情况，其构建了一个用于前向预测的MCTS模型[19]。一般来说，MCTS会创建一棵树，其节点代表强化学习中的状态，树的扩展过程正是在探索可能的动作和由此导致的新状态，决策过程基于各个节点的值估计。AlphaGo在原有MCTS的基础上引入策略网络和值网络两个深度神经网络模型，AlphaGo中使用策略网络来预测下一步最优策略，使用值网络来估计当前状态的价值。在推理时，AlphaGo沿着树结构逐步探索节点并获得值估计，策略网络和值网络的参数则在自我对弈的训练过程中分别通过策略梯度和监督学习的均方误差损失进行更新。最终，AlphaGo在选择具体动作时会综合考虑Q值估计、概率策略和节点遍历次数等因素。在AlphaGo的基础上，AlphaZero进一步简化了模型，专注于自我对弈的训练过程，从头开始学习策略与评估，不依赖于任何先验知识，极大提高了泛化能力[20]。

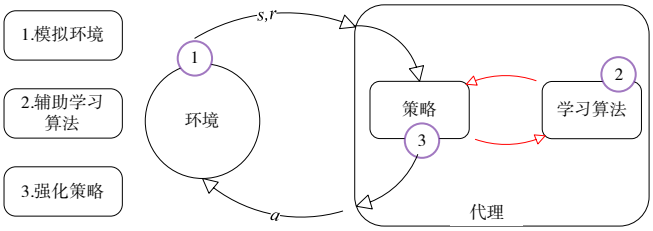

Fig. 1 Model-based reinforcement learning processes

图1 基于模型的强化学习流程

2 生成式人工智能的应用以及挑战

2.1 生成式人工智能的应用

2.1.1 自然语言处理

生成式人工智能在自然语言处理领域发挥着重要作用，涉及机器翻译、文本摘要、对话系统、文本生成等多个任务。传统机器翻译系统存在许多缺陷，如难以处理长句、需要大量人工标注数据等，而生成式神经网络模型（如Seq2Seq[21]、Transformer[22]）能够端到端地将源语言映射为目标语言，无需过多的人工特征工程，性能显著优于传统模型。近年来，预训练语言模型（Pre-trained Language Model，PLM）也被引入机器翻译任务中。例如，EDITOR模型融合了奖励建模与序列级别的强化训练，显著提升了机器翻译质量[23]。然而，即使采用PLM，长句翻译质量仍较低，为此Chen等[24]尝试将retrieval模型与生成模型相结合，以期获得更佳的长句翻译效果。

在文本生成任务中，大型生成式PLM，如GPT-3展现出强大的文本生成能力[25]。通过在大量文本数据上进行自监督训练，PLM能获得丰富的语言知识，微调这些模型即可生成高质量且多样化的文本。然而，PLM本身存在诸多缺陷，如缺乏长期记忆能力、生成内容的连贯性较差等。为此，Zhang等[26]尝试引入对抗性思路和增量式训练，以增强PLM的长期记忆和推理能力。

除文本生成外，对话系统构建也是生成式人工智能应用广泛的领域。例如，Sheu等[27]利用强化学习，通过奖励建模来捕捉人类在对话中真实的行为偏好，构建了一个开放域对话系统，对话质量有了大幅提升。针对人机对话中的一些特殊问题，Stahl等[28]将合成参考对话作为奖励函数，用于指导生成更自然流畅的对话回复。

2.1.2 计算机视觉

生成式人工智能在计算机视觉领域也取得了突破性进展，主要集中在图像/视频生成、图像编辑、图像修复等任务方面。GAN被广泛应用于图像生成，通过生成器与判别器相互对抗训练获得逼真的图像[29]。然而，GAN存在训练不稳定、模式丢失等缺陷。近年来，扩散模型因其简洁有效的框架而受到极大关注，已成为图像生成领域的新热点[30]。其通过学习从纯噪声到图像的转换过程，能生成出质量较高的高分辨率图像。此外，自回归模型依靠强大

的语义理解能力，也展现出优秀的图像生成性能[31]。

当前的主流图像生成模型虽然能生成高分辨率、高质量的图像，但仍面临一些挑战：一是生成图像缺乏整体的语义一致性，物体摆放位置不合理；二是难以对生成内容进行精准控制[32]。为解决这些问题，一些研究将生成模型与强化学习等范式相结合，旨在引入更多先验知识以及人类偏好。例如，Chen 等[33]通过建模人类在图像中的注意力设计奖励机制，从而使生成图像能够满足人类的审美需求。

除了图像生成，生成式人工智能也被大量应用于图像编辑与修复等任务。例如，Jing 等[34]提出一种基于强化学习的语义驱动图像编辑框架，能够根据文字描述对图像进行自然、精细的编辑。该框架引入了显式记忆模块，更好地实现了编辑的逻辑一致性。Chen 等[35]利用 GAN 和半监督学习范式成功修复了存在部分缺失区域的图像。

2.1.3　艺术创作

艺术创作向来被视为人类智慧的高级体现，生成式人工智能的应用也为其带来了新的机遇和挑战。例如，Wang 等[36]提出的 StyleNeuralPainter 模型通过融合强化学习、风格嵌入、注意力机制等技术，实现了基于文字描述的绘画生成。在该模型中，风格嵌入模块用于捕获绘画的独特艺术风格，而强化学习则起到优化生成质量和风格表达一致性的作用。生成的绘画作品不仅在主体内容上符合文字描述，而且富有艺术特色；Li 等[37]提出的 PoeticGAN 模型将生成式人工智能应用于诗歌创作领域。该模型在 GAN 结构的基础上增加了词义注意力和写作规则约束等模块，使生成的诗歌在内容、意境和格式上都更加贴近人类写作风格。目前，生成式人工智能用于艺术创作仍面临诸多挑战，例如难以复现人类独特的创意思维，无法精准把控审美体验等。为此，Drigas 等[38]提出利用元认知框架来增强模型的创造力，但距离彻底解决这些问题仍有很长的路要走。

总体来说，生成式人工智能在自然语言处理、计算机视觉、艺术创作等领域均取得了长足进步，展现出广阔的应用前景。相较于传统机器学习模型，生成式人工智能模型无需大量人工标注，只需基于原始数据自主提取知识即可完成相应任务，具有更强的数据利用能力。

2.2　生成式人工智能面临的挑战

虽然强化学习开辟了生成式人工智能全新的研究领域，但也面临着一系列新的挑战，需要研究者们进一步探索。这些挑战包括利用—探索权衡、奖励赋权权衡和信用分配问题。

在生成式强化学习框架下，智能体需要在利用当前已掌握的策略获得较高奖励与主动探索新的可能性之间进行权衡。一方面，生成任务往往存在海量的可能输出空间，例如在文本生成中，词汇表通常包含数万个词，而句子长度也是不固定的。在这样高维、广阔的空间中，过度追求利用已知的生成模式而不探索新的可能性，会使模型极易过早收敛至次优解而错失更优输出；另一方面，如果过于偏重主动探索，不仅会使模型效率低下，还可能导致其生成内容质量失衡、连贯性不足等。因此，如何在这两种策略之间寻求合理的平衡并实现高效学习是生成式强化学习面临的一大挑战。为此，Zou 等[39]通过引入双重控制器（生成器与评分器）权衡探索与利用策略；Mikalsen 等[40]将人类反馈信号作为奖励来指导模型探索方向。但总体来说，在大规模生成空间中有效平衡二者的通用解决方案仍未被提出。

奖励函数反映了生成结果质量的主观评判标准。然而在实际应用中，不同目标之间往往存在一定的矛盾与冲突，如何量化和权衡这些目标并进行合理的奖励赋权是生成式强化学习面临的一大难题。以对话系统为例，Li 等[41]认为在设计奖励函数时需要平衡语义一致性目标与易理解性目标，过于追求语义一致性可能会导致对话内容缺乏多样性，而过分追求易理解性又可能牺牲语义连贯性。类似的，在文本生成任务中，Yang 等[42]研究发现需要在语法流畅性、生成多样性和反对复制等多重目标间作出取舍与权衡。生成式强化学习需要权衡利用与探索之间的矛盾，以及均衡多重奖励目标，这不仅需要对任务属性有深入理解，还需要具备一定的人工经验。如何在面临多重奖励目标时设计自动化奖励赋权机制，确保策略优化不会过度向某个单一目标倾斜，同时也不会陷入无所适从的困境，是一个极具挑战的难题，需要深入研究与探索。

此外，在生成系统中还存在着信用分配问题。智能体需要确定哪些先前的行为对当前的奖励有影响，然而随着时间步长的增加，先前动作数量增加变得越来越困难[43]。为此，分层强化学习策略被提出，其设置了管理器和工作者两个学习模块，管理器采用潜在表示并在低维空间中生成目标向量，工作者融合潜在向量和目标作出决定[44]。

管理器通过以下公式进行训练：

$$\nabla_\theta g_t = \left(R_t - V_M^t\right)\nabla_\theta d\left(s_{t+c}, g_t(\theta)\right) \tag{10}$$

式中：g_t 为目标，M 为管理器，$s_\cdot$ 为状态。

工作者通过以下公式进行训练：

$$\nabla_\theta \pi = \left(R_t + R_I^t - V_W^t\right)\nabla_\theta \log \pi_\theta\left(s_t\right) \tag{11}$$

式中：$R_I^t = \frac{1}{c}\sum_{i=1}^{c} I\left(d\left(s_t - s_{t-i}, g_{t-i}\right)\right)$，为内在奖励；$V_t^M$ 为工作者的值函数。

一些研究尝试通过分层强化学习策略解决长序列生成任务中的信用分配问题。例如，Guo 等[44]在 Montezuma's Revenge 游戏的迷宫环境中将代理分为管理器和工作者两部分，其中管理器负责生成低维目标向量，工作者根据目标向量作出决策；Vezhnevets 等[45]提出的 LeakGAN 模型在图像字幕任务中采用了类似的分层架构，不同之处在于引入了策略梯度算法，奖励来自 GAN 中的鉴别器。

3 强化学习在生成式人工智能模型中的应用

上述种种挑战对生成式人工智能模型的发展造成了极大困扰，而强化学习范式恰恰能为这些挑战带来新的解决思路。强化学习的代理通过与环境交互获取反馈信号，持续优化生成策略，可有效应对利用—探索权衡，合理赋权不同奖励目标，甚至为信用分配问题提供新的解决方案。强化学习的一大优势在于其能力可传播梯度通过非微分模块提高了神经网络的训练能力，特别是在计算流程中包含离散模块时可对监督学习和无监督学习目标形成有益补充。

以下讨论强化学习在两类不可微学习问题中的应用。图2展示了强化学习代理如何在应用程序中充当生成器；图3展示了生成模型中可能出现的不可微情况。这些步骤可能会导致模型在训练时出现求导困难的不可微分条件。强化学习通常作为一个顺序决策者，使用代理生成序列 $x_1, x_2, x_3, \ldots, x_n$。此时，先前生成的结果会参与环境交互，同时任务特定的上下文应包括其他输入和奖励函数。例如，当应用程序为视觉问答时，不仅应将生成的序列视为输入，还应将包含必要信息的图像和问题视为输入。

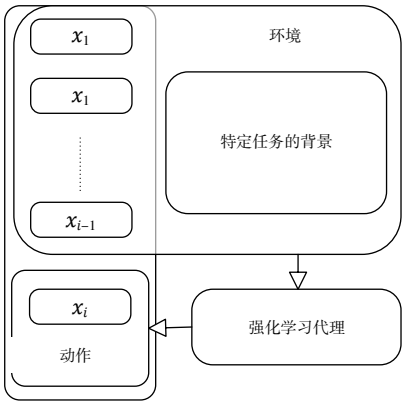

Fig. 2 Reinforcement learning agent as generator

图2 强化学习代理作为生成器

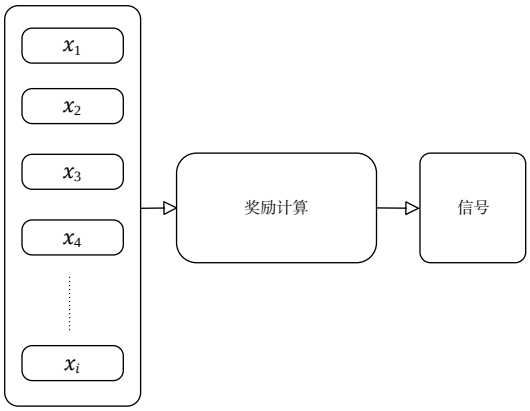

Fig. 3 Non-differentiable situation for generation models

图3 生成模型的不可微情况

3.1 解决不可微学习问题

强化学习通过非微分模块传播梯度，可用于扩展神经网络的能力，其允许在计算管道中存在离散模块时训练模型，这一特性是对监督学习和无监督学习目标的补充，因此它们都需要可微的训练管道。以下介绍两类不可微问题。

3.1.1 生成变量不可微

在计算机视觉、自然语言处理和分子设计等领域，离散值生成是一个常见问题。例如，在自然语言处理和分子设计中，文本和分子元素经常被转化为高维空间的嵌入，这些嵌入通常为离散值或独热向量。在计算机视觉中，典型的 RGB 图像格式包含3个颜色通道的离散值。虽然对于连续生成模型来说，对这些离散值进行归一化是有可能的，但是这种转换可能会导致模型鲁棒性下降等问题[46]。

强化学习是解决此类离散生成问题的有效工具。在强化学习中，策略梯度优化方法是一种常用且有效的算法，其通过直接对策略进行优化来寻找最优策略，但未对奖励与策略梯度之间施加明确约束。例如，文献[46-48]利用策略梯度方法和蒙特卡罗策略梯度估计优化模型，避免了监督学习的可微要求；Hjelm 等[46]提出的 BGAN 模型将策略梯度与 GAN 目标相结合，通过蒙特卡罗估计优化数据分布，解决了 GAN 在处理离散数据时的不可微限制；Fan 等[47]将 SeqGAN 模型[49]应用于视觉对话系统中，利用强化学习生成不可微分的单词序列，然而 GAN 的泛化能力可能受到影响。

3.1.2 训练目标不可微

除了生成变量外，训练目标本身也可能是不可微分的。策略梯度方法允许直接将这些不可微的目标纳入奖励，无需额外约束，机器翻译和文本摘要的广泛评估指标就是很好的例子[50-52]。例如，Ranzato 等[50]提出的 MIXER 方法使用 BLEU 和 ROUGE 测试指标直接优化模型，但无法解决序列生成中训练与测试阶段之间的暴露偏差；Cai 等[51]提出的 TAG 结合类型辅助指导可通过强化学习生成代码注释，并以 BLUE 和 ROUGE 作为奖励，与 MIXER 方法相比，该方法很好地解决了暴露偏差问题；Dognin 等[52]提出的 ReGen 使用强化学习进行不可微分评估，根据 BLEU、METEOR 和 chrF++指标指导文本生成。以上研究说明强化学习可以直接将不可微分的评估指标（如 BLEU、METEOR 等）作为奖励，用于指导和优化文本生成模型。

3.2 引入新的训练信号

强化学习方法具有一定的灵活性，可解决不可微问题，因此可以方便地设计有用的奖励函数从而引入新的训练信号，将各种目标纳入生成模型的训练过程中。以下分别介绍判别器奖励和神经架构搜索（Neural Architecture Search，NAS）两种方法。

3.2.1 判别器奖励

从训练信号的角度来看，GAN 架构中的鉴别器扮演着

类似于强化学习背景下奖励的角色。例如,Yu 等[49]提出的SeqGAN通过引入 GAN 来生成序列标记,该方法的奖励信号源自输出概率,作为真实样本与生成样本之间的区分度量。随后的研究通过利用精心设计的判别器[53]、开发新颖的奖励公式[54]、实施演员批评技术[55]、结合排名公式[56]等方法对GAN框架进行了扩展与改进,还有一些方法使用多个鉴别器[57]等策略实施了一系列修改。例如,Guimaraes 等[53]提出的目标强化GAN将SeqGAN扩展到使用具有特定领域目标的奖励函数,以引导生成的样本朝特定方向调整,该方法的优势在于能够更有针对性地引导生成模型;Wang 等[54]提出的 GRL(GAN-based Reinforcement Learning)模型采用对抗性奖励,取两个概率 $E_{x\sim p_{\text{data}}}D(x)-E_{x\sim\pi_\theta}D(x)$ 之间的差异,通过测量真实样本与生成样本之间的概率差异来构建奖励信号;Fedus 等[55]提出的MaskGAN使用批评者取代 REINFORCE 中的基线,形成一种用于文本生成的行动者批评者算法,与 SeqGAN 中的策略梯度不同,该方法利用判别器中真实单词的概率作为奖励,并通过填充任务来训练代理;Zhou 等[57]提出的SAL模型使用比较判别器改变奖励函数,鉴别器采用一对样本 (x_1,x_2) 作为输入,这些样本是从当前生成的样本与之前生成的样本中收集的,同时定义了3种类型的判别器 $D(>)$、$D(<)$ 和 $D(\approx)$ 来描述质量样本之间的关系,该方法的优势在于使用系数平衡勘探与利用之间的关系;Scialom 等[58]提出的SelfGAN模型将MCTS与协作解码结合起来,以改善训练离散顺序决策问题的不稳定性。协作解码意味着在解码阶段使用鉴别器将生成的序列重新排列为值函数。与 MCTS 一样,该模型根据 Q 值、策略概率和遍历计数来选择操作,因此鉴别器被集成到MCTS解码过程中以提高模型性能;Sarmad 等[59]提出的RL-GAN-Net模型使用强化学习代理为点云生成中的GAN模型生成潜在代码,其使用的奖励包括鉴别器损失、点云重建损失和基于切角距离的损失;Wu 等[60]提出的TextGAIL模型使用近端策略优化算法(Proximal Policy Optimization,PPO)评估序列的真实性。PPO是一种用于解决强化学习问题的优化算法,其通过在每一次迭代中根据大量采样数据不断优化策略,同时限制策略的变化范围,避免过大的策略更新。该算法的优势为通过优化策略提高智能体在环境中的性能,从而优化其决策与行为。

3.2.2　NAS

前文介绍了强化学习弥补了不可微分学习系统的差距,合并了新的训练信号,并充当了采样器。实际上,NAS本身可被视为一系列标记,也可称为主观强化生成器。而代理本身也是一个生成器,可应用于几乎所有使用神经网络作为学习器的任务。NAS可用于优化神经网络架构,其奖励通常为任务指标。例如,当代理优化分类器的架构时,分类器的准确性通常用作奖励。Zoph 等[61]提出使用强化学习指导神经架构设计,其采用循环神经网络网络通过REINFORCE、ENAS[62]生成架构描述,通过参数共享提高效率。该神经架构设计的优势在于其不是从头开始构建网络,而是在预定义的卷积单元上构建网络,这样可以减少搜索空间。类似的,Hsu 等[63]提出的MONAS(Multi-Objective Neural Architecture Search)模型删除了训练中的状态,只考虑动作和奖励,将功耗纳入奖励函数,优势为奖励计算中考虑了各种优化目标,例如混合、阈值和替代指标;Guo 等[64]提出的IRLAS(Inverse Reinforcement Learning for Architecture Search)模型将逆强化学习纳入 NAS 中,定义了将架构映射到代理轨迹的特征计数 $\mu=\sum_{t=1}^{T}\gamma^t\phi(s_t)$,其中 s_t 为架构信息,γ 为折扣因子,$\phi(\cdot)$ 为嵌入函数。其使用 μ 为镜像刺激函数创建一个线性模型,同时使用ResNet[65]等专家模型的拓扑作为指导。

NAS中的状态空间和动作空间均经过精心设计,使得训练易于处理。在NAS中,通常将层参数作为状态和动作空间的元素[66-67]。例如,Rijsdijk 等[66]在 NAS 上应用MetaQNN 进行旁道分析,优势为奖励函数的定义考虑了具有不同数量攻击痕迹的猜测熵;Zhong 等[67]提出的BlockQNN采用神经模型执行图像分类任务,其定义了网络结构代码,该代码的优势为量化了计算图中的架构信息,例如层索引、操作类型、内核大小以及其他相关节点。

在NAS任务中,提高模型采样效率也是一个重要目标。一次性学习和元学习便是两种可以提升采样效率的技术[68-70]。例如,Chuahan 等[68]提出的DQNAS架构将基于强化学习的NAS与一次性训练相结合,以使模型获得更好的性能;Chen 等[69]提出的CATCH策略采用元强化学习框架加速元测试任务的架构设计,其关键在于使用一次性训练从一些常用层转移权重,以快速建立训练任务;Pang 等[70]提出的RL-DARTS亦采用元学习方法,其使用的元优化器将梯度和控制超参数定义为状态,将控制超参数的转变定义为动作,将有效数据集上的性能定义为控制架构搜索器方向的奖励。

4　未来方向

将强化学习应用于生成式模型始于2015年,一些研究方向已经得到深入探讨,而新的方向仍在不断涌现。通过对现有研究进行分析,认为未来该领域的研究可能集中于以下几个方面。

4.1　奖励函数设计与多目标优化

在通过强化学习手动设计新信号引导模型训练方面,研究者们通常使用多个目标处理各种约束和引导建模,在理想情况下可以一次性实现最优值。在这种情况下,寻求平衡多个目标函数并找到最优解是一个挑战。未来研究可以关注如何在矛盾目标之间进行权衡,探索多目标优化和帕累托优化在生成式建模中的应用,以获得更强大的

模型[71-72]。

4.2 模型增强与控制

最新研究探索将强化学习应用于改进生成模型，例如辅助 EBM 的采样[73]或提高去噪扩散概率模型（Denoising Diffusion Probabilistic Model，DDPM）的效率[74]，这为利用强化学习增强生成模型性能开辟了新途径。不同于传统方法中引入新训练信号或架构建设的方式，强化学习在策略优化、方差减小等方面的新进展可能会进一步促进生成模型性能提升。

4.3 人类偏好建模与可解释性

近年来，通过人类反馈进行强化学习（Reinforcement Learning with Human Feedback，RLHF）越来越受关注，被应用于指导 LLM 的训练[75]。例如直接偏好优化（Direct Preference Optimization，DPO）直接利用基于偏好的奖励函数替代强化学习[76]。未来可进一步研究更好的人类偏好建模方法。此外，由于人类偏好是动态变化的，捕捉这种动态性并在此基础上改进生成模型也是一个有趣的方向，可用于提高模型的可解释性。

4.4 样本效率与泛化

目前，即使对于最好的 GPT-4 模型来说，推理逻辑任务的泛化也很困难[77]。强化学习算法可用于解决深度生成模型的泛化困难问题。今后应致力于设计一种能够更好泛化并在分布之外数据上取得更好结果的模型。重新训练[78]和因果机器学习[79]等方法可能有助于增强基于强化学习的生成器在学习和适应方面的能力。

4.5 新的强化学习方法引入

目前应用于生成式模型的多为经典的策略梯度等强化学习算法，且多采用离线设置。未来可以探索将离线强化学习的最新进展（如保守策略优化等）与经典模型相结合，可能有助于解决当前生成式模型所面临的困境，如 LLM 的幻觉问题等[80]。

4.6 基础模型与多模态生成

LLM 展现出跨任务的强大迁移能力，将其与视觉等其他模态的基础模型相结合是实现更强大的多模态生成模型的潜在途径[81]。该领域的快速发展反映出当前 LLM 的巨大潜力，同时对 RLHF 之类的方法提出了新挑战。

5 结语

生成式人工智能近年来受到广泛关注。强化学习作为一种无需标注样本便能进行训练的方法，通过判别器奖励函数与 NAS 等新的训练信号优化模型性能，为解决生成式模型中目标函数的不可导问题提供了新方案。尽管本文对各个应用领域进行了全面回顾，由于篇幅限制，某些细节未能详细展开。此外，强化学习在实际应用中面临信用分配、利用取舍均衡等新挑战，这些问题需要深入研究解决。未来可以探索如何优化强化学习在特定应用中的性能，探索满足更复杂生成模型需求的新训练信号，深入解决样本效率和泛化等问题。相信通过持续研究，基于强化学习的生成式人工智能将有更广阔的发展前景。

参考文献：

[1] HO J，JAIN A，ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems，2020，33：6840-6851.

[2] JUMPER J，EVANS R，PRITZEL A，et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature，2021，596(7873)：583-589.

[3] BAEK M，DIMAIO F，ANISHCHENKO I，et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science，2021，373(6557)：871-876.

[4] LIN Z，AKIN H，RAO R，et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science，2023，379(6637)：1123-1130.

[5] OpenAI. ChatGPT [EB/OL]. https://openai. com/chatgpt. Accessed：2023-08-01.

[6] SUTTION R S，BARTO A G. Reinforcement learning：an introduction[M]. Cambridge：MIT Press，2018.

[7] MOHEBBI M M，BOROOMAND B，JALALI M，et al. Games of GANs：game-theoretical models for generative adversarial networks[J]. Artificial Intelligence Review，2023，56(9)：9771-9807.

[8] ATANCE S R，DIEZ J V，ENGKVIST O，et al. De novo drug design using reinforcement learning with graph-based deep generative models[J]. Journal of Chemical Information and Modeling，2022，62(20)：4863-4872.

[9] BELLMAN R. The theory of dynamic programming[J]. Bulletin of the American Mathematical Society，1954，60(6)：503-515.

[10] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[11] HAARNOJA T，TANG H，ABBEEL P，et al. Reinforcement learning with deep energy-based policies[C]//International Conference on Machine Learning，2017：1352-1361.

[12] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning，1992，8：229-256.

[13] SCHULMAN J，LEVINE S，ABBEEL P，et al. Trust region policy optimization[C]//International Conference on Machine Learning，2015：1889-1897.

[14] SCHULMAN J，WOLSKI F，DHARIWAL P，et al. Proximal policy optimization algorithms[DB/OL]. https://arxiv. org/abs/1707. 06347.

[15] SILVER D，LEVER G，HEESS N，et al. Deterministic policy gradient algorithms[C]//International Conference on Machine Learning，2014：387-395.

[16] LILLICRAP T P，HUNT J J，PRITZEL A，et al. Continuous control with deep reinforcement learning [DB/OL]. https://arxiv. org/abs/1509. 02971.

[17] HAARNOJA T，ZHOU A，ABBEEL P，et al. Soft actor-critic：off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning，2018：1861-1870.

[18] MNIH V，BADIA A P，MIRZA M，et al. Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning，2016：1928-1937.

[19] SILVER D，HUANG A，MADDISON C J，et al. Mastering the game of

Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[20] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[21] SCOTTI V, SBATTELLA L, TEDESCO R. A primer on seq2seq models for generative chatbots[J]. ACM Computing Surveys, 2023, 56(3): 1-58.

[22] HAN D, PAN X, HAN Y, et al. Flatten transformer: vision transformer using focused linear attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 5961-5971.

[23] ZHANG H, LIU X, ZHANG J. Summit: iterative text summarization via ChatGPT[DB/OL]. https://arxiv.org/abs/2305.14835.

[24] CHEN G, CHEN Y, LI V O K. Lexically constrained neural machine translation with explicit alignment guidance [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 12630-12638.

[25] LEVINE D M, TUWANI R, KOMPA B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model[DB/OL]. https://europepmc.org/article/PPR/PPR610800.

[26] ZHANG Y F, YU W, WEN Q, et al. Debiasing large visual language models[DB/OL]. https://arxiv.org/pdf/2403.05262.pdf.

[27] SHEU J S, WU S R, WU W H. Performance improvement on traditional Chinese task-oriented dialogue systems with reinforcement learning and regularized dropout technique[J]. IEEE Access, 2023, 11: 19849-19862.

[28] STAHL B C, ANTONIOU J, BHALLA N, et al. A systematic review of artificial intelligence impact assessments[J]. Artificial Intelligence Review, 2023, 56(11): 12799-12831.

[29] CHEN Y, YANG X H, WEI Z, et al. Generative adversarial networks in medical image augmentation: a review[J]. Computers in Biology and Medicine, 2022, 144: 105382.

[30] LUGMAYR A, DANNELLJAN M, ROMERO A, et al. Repaint: inpainting using denoising diffusion probabilistic models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11461-11471.

[31] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[DB/OL]. https://arxiv.org/abs/2204.06125.

[32] RAHBAR M, MAHDAVINEJAD M, MARKAZI A H D, et al. Architectural layout design through deep learning and agent-based modeling: a hybrid approach[J]. Journal of Building Engineering, 2022, 47: 103822.

[33] CHEN M, LIU Y, YI J, et al. Evaluating text-to-image generative models: an empirical study on human image synthesis[DB/OL]. https://arxiv.org/abs/2403.05125.

[34] JING S, ZHANG H, ZENG P, et al. Memory-based augmentation network for video captioning[J]. IEEE Transactions on Multimedia, 2023, 26: 2367-2379.

[35] CHEN X, TAN J, WANG T, et al. Towards real-world blind face restoration with generative diffusion prior [DB/OL]. https://arxiv.org/pdf/2312.15736v2.pdf.

[36] WANG Q, GUO C, DAI H N, et al. Stroke-GAN painter: learning to paint artworks using stroke-style generative adversarial networks [J]. Computational Visual Media, 2023, 9(4): 787-806.

[37] LI J, TANG T, ZHAO W X, et al. Pretrained language models for text generation: a survey[DB/OL]. https://arxiv.org/abs/2201.05273.

[38] DRIGAS A, MITSEA E, SKIANIS C. Meta-learning: a nine-layer model based on metacognition and smart technologies[J]. Sustainability, 2023, 15(2): 1668.

[39] ZOU W, HUANG S, XIE J, et al. A reinforced generation of adversarial examples for neural machine translation[DB/OL]. https://arxiv.org/abs/1911.03677.

[40] MIKALSEN M, DINGSØYR T. Feedback as a process in a large semi-capstone software engineering course[C]//Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, 2023: 475-479.

[41] LI J, MONROE W, RITTER A, et al. Deep reinforcement learning for dialogue generation[DB/OL]. https://arxiv.org/abs/1606.01541.

[42] YANG Z, HU J, SALAKHUTDINOV R, et al. Semi-supervised QA with generative domain-adaptive nets [DB/OL]. https://arxiv.org/abs/1702.02206.

[43] JIN W, BARZILAY R, JAAKKOLA T. Multi-objective molecule generation using interpretable substructures[C]//International Conference on Machine Learning, 2020: 4849-4859.

[44] GUO J, LU S, CAI H, et al. Long text generation via adversarial training with leaked information[DB/OL]. https://arxiv.org/abs/1709.08624.

[45] VEZHNEVETS A S, OSINDERO S, SCHAUL T, et al. Feudal networks for hierarchical reinforcement learning[C]//International Conference on Machine Learning, 2017: 3540-3549.

[46] HJELM R D, JACOB A P, CHE T, et al. Boundary-seeking generative adversarial networks[DB/OL]. https://arxiv.org/abs/1702.08431.

[47] FAN H, ZHU L, YANG Y, et al. Recurrent attention network with reinforced generator for visual dialog[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020, 16(3): 1-16.

[48] NGUYEN T M, GARG A, BARANIUK R G, et al. InfoCNF: an efficient conditional continuous normalizing flow with adaptive solvers[DB/OL]. https://arxiv.org/abs/1912.03978.

[49] YU L, ZHANG W, WANG J, et al. Seqgan: sequence generative adversarial nets with policy gradient[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017: 2852-2858.

[50] RANZATO M A, CHOPRA S, AULI M, et al. Sequence level training with recurrent neural networks [DB/OL]. https://arxiv.org/abs/1511.06732.

[51] CAI R, LIANG Z, XU B, et al. TAG: type auxiliary guiding for code comment generation[DB/OL]. https://arxiv.org/abs/2005.02835.

[52] DOGNIN P L, PADHI I, MELNYK I, et al. Regen: reinforcement learning for text and knowledge base generation using pretrained language models[DB/OL]. https://arxiv.org/abs/2108.12472.

[53] GUIMARAES G L, SANCHEZ-LENGELING B, OUTEIRAL C, et al. Objective-reinforced generative adversarial networks (organ) for sequence generation models[DB/OL]. https://arxiv.org/abs/1705.10843.

[54] WANG Q, JI Y, HAO Y, et al. GRL: knowledge graph completion with GAN-based reinforcement learning [J]. Knowledge-Based Systems, 2020, 209: 106421.

[55] FEDUS W, GOODFELLOW I, DAI A M. Maskgan: better text generation via filling in the_[DB/OL]. https://arxiv.org/abs/1801.07736.

[56] LIN K, LI D, HE X, et al. Adversarial ranking for language generation [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3158-3168.

[57] ZHOU W, GE T, XU K, et al. Self-adversarial learning with comparative discrimination for text generation [DB/OL]. https://arxiv.org/abs/

2001. 11691.

[58] SCIALOM T, DRAY P A, STAIANO J, et al. To beam or not to beam: that is a question of cooperation for language GANS[J]. Advances in Neural Information Processing Systems, 2021, 34: 26585-26597.

[59] SARMAD M, LEE H J, KIM Y M. RL-GAN-NET: a reinforcement learning agent controlled GAN network for real-time point cloud shape completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5898-5907.

[60] WU Q, LI L, YU Z. TextGAIL: generative adversarial imitation learning for text generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14067-14075.

[61] ZOPH B, LE Q V. Neural architecture search with reinforcement learning [DB/OL]. https://arxiv. org/abs/1611. 01578.

[62] PHAM H, GUAN M, ZOPH B, et al. Efficient neural architecture search via parameters sharing[C]//International Conference on Machine Learning, 2018: 4095-4104.

[63] HSU C H, CHANG S H, LIANG J H, et al. Monas: multi-objective neural architecture search using reinforcement learning[DB/OL]. https://arxiv. org/abs/1806. 10332.

[64] GUO M, ZHONG Z, WU W, et al. IRLAS: inverse reinforcement learning for architecture search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9021-9029.

[65] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[66] RIJSDIJK J, WU L, PERIN G, et al. Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis [J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2021, 71: 677-707.

[67] ZHONG Z, YAN J, WU W, et al. Practical block-wise neural network architecture generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2423-2432.

[68] CHAUHAN A, BHATTACHARYYA S, VADIVEL S. DQNAS: neural architecture search using reinforcement learning[DB/OL]. https://arxiv. org/abs/2301. 06687.

[69] CHEN X, DUAN Y, CHEN Z, et al. Catch: context-based meta reinforcement learning for transferrable architecture search[C]//Computer Vision-ECCV 2020: 16th European Conference, 2020: 185-202.

[70] PANG D, LE X, GUAN X. RL-DARTS: differentiable neural architecture search via reinforcement-learning-based meta-optimizer [J]. Knowledge-Based Systems, 2021, 234: 107585.

[71] BAKER B, GUPTA O, NAIK N, et al. Designing neural network architectures using reinforcement learning [DB/OL]. https://arxiv. org/abs/1611. 02167.

[72] PHAM N Q, NIEHUES J, WAIBEL A. Towards one-shot learning for rare-word translation with external experts[DB/OL]. https://arxiv. org/abs/1809. 03182.

[73] PARASHAKOVA T, ANDREOLI J M, DYMETMAN M. Distributional reinforcement learning for energy-based sequential models [DB/OL]. https://arxiv. org/abs/1912. 08517.

[74] FAN Y, LEE K. Optimizing DDPM sampling with shortcut fine-tuning [DB/OL]. https://arxiv. org/abs/2301. 13362.

[75] STIENON N, OUYANG L, WU J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020, 33: 3008-3021.

[76] RAMÉ A, VIEILLARD N, HUSSENOT L, et al. Warm: on the benefits of weight averaged reward models [DB/OL]. https://arxiv. org/abs/2401. 12187.

[77] LIU H, NING R, TENG Z, et al. Evaluating the logical reasoning ability of ChatGPT and GPT-4[DB/OL]. https://arxiv. org/abs/2304. 03439.

[78] TRIPP A, DAXBERGER E, HERNÁNDEZ-LOBATO J M. Sample-efficient optimization in the latent space of deep generative models via weighted retraining[J]. Advances in Neural Information Processing Systems, 2020, 33: 11259-11272.

[79] SCHÖLKOPF B, LOCATELLO F, BAUER S, et al. Toward causal representation learning [J]. Proceedings of the IEEE, 2021, 109(5): 612-634.

[80] PANG R Y, HE H. Text generation by learning from demonstrations[DB/OL]. https://arxiv. org/abs/2009. 07839.

[81] COSTA L, SAJID N, PARR T, et al. Reward maximization through discrete active inference[J]. Neural Computation, 2023, 35(5): 807-852.

(责任编辑:尹晨茹)