第 39 卷 第 3 期
2024 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 39 No. 3
Jun. 2024

DOI: 10.13364/j.issn.1672-6510.20230145
数字出版日期：2024–04–18；数字出版网址：http://link.cnki.net/urlid/12.1355.n.20240416.1916.010

基于探针稀疏注意力机制的门控 Transformer 模型

赵婷婷，丁翘楚，马　冲，陈亚瑞，王　媛
（天津科技大学人工智能学院，天津 300457）

摘　要：在强化学习中，智能体对状态序列进行编码，根据历史信息指导动作的选择，通常将其建模为递归型神经网络，但其存在梯度消失和梯度爆炸的问题，难以处理长序列。以自注意力机制为核心的 Transformer 是一种能够有效整合长时间范围内信息的机制，将传统 Transformer 直接应用于强化学习中存在训练不稳定和计算复杂度高的问题。门控 Transformer-XL（GTrXL）解决了 Transformer 在强化学习中训练不稳定的问题，但仍具有很高的计算复杂度。针对此问题，本研究提出了一种具有探针稀疏注意力机制的门控 Transformer（PS-GTr），其在 GTrXL 中的恒等映射重排和门控机制的基础上引入了探针稀疏注意力机制，降低了时间复杂度和空间复杂度，进一步提高了训练效率。通过实验验证，PS-GTr 在强化学习任务中的性能与 GTrXL 相当，而且训练时间更短，内存占用更少。

关键词：深度强化学习；自注意力机制；探针稀疏注意力机制
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1672-6510（2024）03-0056-08

Gated Transformer Based on Prob-Sparse Attention

ZHAO Tingting，DING Qiaochu，MA Chong，CHEN Yarui，WANG Yuan
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：In reinforcement learning，the agent encodes state sequence and influences action selection by historical information，typically employing recurrent neural network. Such traditional methods encounter gradient issues such as gradient disappearance and gradient explosion，and are also challenged by long sequences. Transformer leverages self-attention to assimilate long-range information. However，traditional Transformer exhibits instability and complexity in reinforcement learning. Gated Transformer-XL（GTrXL）ameliorates Transformer training stability，but remains complex. To solve these problems，in this article we propose a prob-sparse attention gated Transformer（PS-GTr）model，which introduces prob-sparse attention mechanism on the basis of identity mapping rearrangement and gating mechanism in GTrXL，reducing time and space complexity，and further improving training efficiency. Experimental verification showed that PS-GTr had comparable performance compared to GTrXL in reinforcement learning tasks，but had lower training time and memory usage.

Key words：deep reinforcement learning；self-attention；prob-sparse attention

引文格式：
赵婷婷，丁翘楚，马冲，等. 基于探针稀疏注意力机制的门控 Transformer 模型[J]. 天津科技大学学报，2024，39（3）：56–63.
ZHAO T T，DING Q C，MA C，et al. Gated Transformer based on prob-sparse attention[J]. Journal of Tianjin university of science & technology，2024，39（3）：56–63.

强化学习是机器学习领域中的一个重要分支，智能体通过与环境交互学习，使奖励最大化[1]。在智能体与环境交互的过程中，智能体通过环境返回的状态给出动作。强化学习的核心是学习产生动作的策

收稿日期：2023–07–24；**修回日期**：2023–10–16
基金项目：国家自然科学基金项目（61976156）；天津市企业科技特派员项目（20YDTPJC00560）
作者简介：赵婷婷（1986—），女，内蒙古赤峰人，副教授，tingting@tust.edu.cn

略模型，智能体的动作不仅与当前的状态有关，还与之前的状态有关。

在对强化学习的智能体进行训练的过程中，智能体通常使用循环神经网络（recurrent neural network，RNN）和长短期记忆（long short-term memory，LSTM）模型作为记忆结构[2-4]。虽然 RNN 和 LSTM 都能够对状态序列建模，但是 RNN 容易出现梯度消失和梯度爆炸的问题，难以处理长序列。LSTM 在 RNN 的基础上引入了门控机制，在一定程度上解决了这些问题，但处理的序列长度仍然受限，限制了智能体能够记忆的状态长度。

自注意力（self-attention）机制是一种能够在长时间范围内有效整合信息的机制，它不需要引入循环神经网络结构，而是利用对输入的数据进行自注意力计算得到序列之间的关系[5]。自注意力机制摒弃了 RNN 和 LSTM 中的递归结构，直接连接了输入和输出，有效解决了梯度爆炸和梯度消失的问题，从而在处理长距离依赖中具有更显著的效果。相关研究[5-11]也表明自注意力机制相比于传统的循环神经网络和长短期记忆模型的性能有显著提升，基于自注意力机制的 Transformer 结构在自然语言处理（NLP）领域的性能也取得了突破性提升。

鉴于 Transformer 对于长序列问题的突出表现，研究者提出将其用于强化学习中作为记忆模块的预期性能提升会非常有效。然而，经典的 Transformer 结构在强化学习训练中存在难以收敛的问题，甚至在一些简单的任务中与随机策略表现相近[12]。

针对上述问题，Parisotto 等[13]提出 GTrXL（gated Transformer-XL）模型，实现了在强化学习任务中用 Transformer 编码状态序列并能够稳定训练智能体。GTrXL 模型在原有的 Transformer 结构中加入了恒等映射重排（identity map reordering）和门控层，使其在强化学习训练中能够更好地控制信息的传递。因此，GTrXL 模型能够在强化学习训练中很好地收敛，并且在性能上取得了比循环神经网络更好的效果。

GTrXL 模型使在强化学习中利用 Transformer 编码状态序列得以实现，然而它的时间复杂度和空间复杂度仍然是关于序列长度 L 的平方。因此，当智能体训练时，L 较大会导致训练时间较长和内存占用严重，使智能体能够编码的状态序列长度受限，进而限制了强化学习中自注意力机制的训练和部署[14]。

为了提升 Transformer 在训练时的效率，本研究将探针稀疏注意力（prob-sparse attention）机制[15]与恒等映射重排以及门控机制结合，提出了一种具有探针稀疏注意力机制的门控 Transformer（prob-sparse attetion gated Transformer，PS-GTr）模型。PS-GTr 模型引入探针稀疏注意力机制，通过筛选对注意力分数有主要贡献的查询，有效降低计算的复杂度并提高效率。通过强化学习领域中经典的 Atari-Breakout 任务，验证了本研究 PS-GTr 模型在与 GTrXL 性能相同的前提下，具有更短的训练时间和占用更少的内存，提升了训练效率。

1　相关工作

Transformer 是一种能够有效集成长序列信息的神经网络结构，它通过自注意力机制编码序列信息之间的相关性，在自然语言处理和计算机视觉领域取得了巨大的成功。基于自注意力机制这一特点，将 Transformer 用于处理强化学习领域中的状态序列十分值得研究。然而，将 Transformer 编码器的结构直接用于强化学习的智能体中效果较差，甚至与随机策略相当[12]。

Parisotto 等[13]的 GTrXL 模型在 Transformer 的基础上提出了恒等映射重排[16]并加入了门控层，取得了比使用循环神经网络更好的结果。恒等映射重排通过调整层标准化的位置，使模型在训练智能体时能更好地学习前期的反应型行为。门控层使用门控循环单元[17]替换残差连接以控制信息的流动，从而更好地控制了输入残差连接的未经编码的信息以及经过注意力机制和前馈神经网络编码的信息。通过恒等映射重排以及加入门控机制，GTrXL 模型能够在强化学习中稳定训练，取得了比 LSTM 更好的平均累积回报。

GTrXL 在训练时输入的状态序列的长度 L（即上下文长度）是固定的，这需要在训练之前设定适当的序列长度参数。Kumar 等[18]在 GTrXL 的基础上引入自适应注意力广度，使模型在训练时能够学习最佳的上下文长度。自适应注意力广度机制[19]通过网络学习每个注意力层的最佳上下文长度，使每一个注意力层具有不同上下文长度的输入。在通常情况下，Transformer 的低层级自注意力层的上下文长度比高层级的小，即每层需要的上下文长度不同。通过自适应注意力广度机制，每层都能达到最佳的上下文长度，从而减少了设定固定长度而导致的内存与性能的

浪费。

CoBERL(contrastive BERT for reinforcement learning)将对比学习的方法和 BERT 中使用的掩码语言模型(masked language model)方法用在了 GTrXL 上[8,20]。CoBERL 首先利用掩码语言模型对 Transformer 的输入进行掩码操作，之后将部分掩码的输入通过 Transformer 编码，得到新的编码序列。由于强化学习任务中没有明确的标签，因此 CoBERL 使用对比学习的方法计算损失函数，将经过 Transformer 还原的向量与掩码之前的该向量作为正样本，其余作为负样本，计算对比损失函数。最后，将对比损失函数与强化学习的目标函数加权平均得到最终的目标函数。此外，CoBERL 在 GTrXL 模型输出之后加入了 LSTM 模型，Banino 等[20]认为 LSTM 模型更擅长处理短期的记忆，而 GTrXL 处理长距离依赖非常有效，因此结合了两者的优点。CoBERL 将 BERT 方法融入 GTrXL 并与 LSTM 模型一同使用，增加了模型对整个状态序列的关注，从而提升了 Transformer 在强化学习中的性能。

虽然 GTrXL 及其相关研究解决了 Transformer 在强化学习中难以训练的问题，但是与循环神经网络相比，该方法在强化学习中训练 Transformer 需要大量的时间和内存。由于注意力机制计算的复杂度是序列长度 L 的平方，因此随着训练过程中设置的序列长度 L 的增大，计算所需的时间及内存都快速增长，极大地限制了其在强化学习中的使用。

稀疏注意力机制是一类特殊的注意力机制，能够减少自注意力机制中复杂度过高的问题。这些方法通过只关注输入序列中的一部分计算而不对所有的输入序列进行计算，从而有效地降低运算的复杂度。Li 等[21]注意到了注意力的周期性，LogSparse Transformer 缩小了注意力的关注窗口。Beltagy 等[22]提出的 Longformer 结合了局部窗口自注意力和任务驱动全局自注意力，降低了自注意力机制的运算复杂度。这些稀疏注意力机制方法在自然语言处理、时序预测、图像处理等领域取得了成功。稀疏注意力机制在强化学习中的应用具有重要研究价值。

综上所述，传统 Transformer 在强化学习中作为状态序列建模的方法都存在计算复杂度高的问题。本研究以门控机制和恒等映射重排为基础，利用注意力的稀疏性减少注意力计算时的复杂度，提出了基于探针稀疏注意力机制的门控 Transformer，将稀疏注意力机制的 Transformer 用于强化学习的状态序列建模中，进一步提升其在强化学习中的训练效率。

2　研究背景

2.1　点积注意力机制

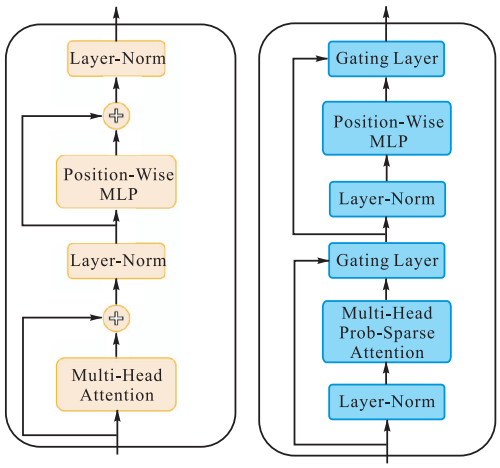
点积注意力机制是 Transformer 的核心机制，设输入向量经过线性变换后得到查询矩阵 $\boldsymbol{Q}\in\mathbb{R}^{L_Q\times d}$，键矩阵 $\boldsymbol{K}\in\mathbb{R}^{L_K\times d}$ 和值矩阵 $\boldsymbol{V}\in\mathbb{R}^{L_R\times d}$，其中 L 是序列长度，d 是嵌入维度。注意力分数计算为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

在 Transformer 中，实际使用的是多头自注意力机制，它用线性层将查询、键和值映射到多个独立的注意力运算中，最后将结果连接后再输出。多头注意力可使模型具有学习多种不同类型相关性的能力。

2.2　Transformer 结构

Transformer 由多个层堆叠而成，其中每个层都由多头自注意力层和多层感知机两个模块组成，具体结构如图 1(a)所示。记 $\boldsymbol{E}^l\in\mathbb{R}^{L\times D}$ 为第 l 层的输入，其中 L 是输入序列的长度，D 是隐藏层的维度。设 E^0 是维度为[L, D]的任意输入的嵌入，在语言建模中是单词的嵌入，而在强化学习中则是智能体与环境交互得到的观察(observation)的嵌入。



(a) Transformer　　(b) PS-GTr

图 1　Transformer 与 PS-GTr 的框架图

Fig. 1　Framework of Transformer and PS-GTr

记第 l 层自注意力层的输出为 $\boldsymbol{Y}^l$，且多头自注意力操作为 MHA，层标准化操作为 LN。则多头自注意力子模块的计算为

$$\boldsymbol{Y}^l=\text{LN}\left(\boldsymbol{E}^{l-1}+\text{MHA}\left(\boldsymbol{E}^{l-1}\right)\right) \tag{2}$$

多层感知机子模块实际上是对输入序列的每一

步应用一个核大小和步长都为 1 的卷积网络，产生新的嵌入向量 $\boldsymbol{E}^l$，记 f^l 为多层感知机的计算，则多层感知机子模块的计算为

$$\boldsymbol{E}^l = \mathrm{LN}\left(\boldsymbol{Y}^l + f^l\left(\boldsymbol{Y}^l\right)\right) \tag{3}$$

3　本文模型

首先对 GTrXL 中引入的恒等映射重排和门控层进行分析，然后对探针稀疏注意力机制进行详细阐述，提出基于探针稀疏注意力机制的门控 Transformer 模型，最后介绍实现 PS-GTr 模型的强化学习算法。

3.1　恒等映射重排

经典的 Transformer 模型很难在强化学习中稳定训练，GTrXL 在Transformer 的基础上引入了恒等映射重排和门控单元，使模型在强化学习中能稳定收敛。

Transformer 中存在一系列的非线性变换。恒等映射重排[16]通过调整 Transformer 中层标准化的位置，从而实现从模型第一层的输入到最后一层输出的恒等映射。

Transformer 在每个子模块输出之后进行层标准化，恒等映射重排将层标准化的位置从每个子模块的输出位置调整到每个子模块的输入，为

$$\boldsymbol{Y}^l = \boldsymbol{E}^{l-1} + \mathrm{MHA}\left(\mathrm{LN}\left(\boldsymbol{E}^{l-1}\right)\right) \tag{4}$$

$$\boldsymbol{E}^l = \boldsymbol{Y}^l + f^l\left(\mathrm{LN}\left(\boldsymbol{Y}^l\right)\right) \tag{5}$$

通过恒等映射重排，能够提高 Transformer 在强化学习中训练的稳定性。这一现象产生一个可能的假设是：通过恒等映射重排，状态存在一条未经编码即可从输入端传输到输出端的通路。与之相比，原本的位置削弱了经过跳跃连接（skip connection）的信息，迫使模型依赖残差路径（residual path）。在许多强化学习任务中，需要智能体在记忆行为之前先学习反应行为，即智能体需要先学会如何走路，再学会如何记住自己在哪里。

3.2　门控层

门控循环单元（GRU）是一种循环神经网络的变体，比长短期记忆网络更简单[17]。通过加入重置门和更新门控制隐藏状态和输入的信息。GTrXL 将 GRU 引入 Transformer 中作为门控层，用来控制信息流。

门控层将 GRU 相对简单并能控制信息传递的特性用于 Transformer。Transformer 的残差连接操作是将未经编码的信息与经过编码的信息进行简单的加和。用 GRU 代替残差连接能够更好地控制信息的传递。在强化学习任务中，智能体能够更好地接受反应行为和记忆行为的信息并学习。

令 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别为经过跳跃连接的信息和经过注意机制或前馈网络编码的信息，则门控层的操作为

$$\boldsymbol{z} = \sigma\left(W_z^l \boldsymbol{y} + U_z^l \boldsymbol{x} - \boldsymbol{b}_g^l\right) \tag{6}$$

$$\boldsymbol{r} = \sigma\left(W_r^l \boldsymbol{y} + U_r^l \boldsymbol{x}\right) \tag{7}$$

$$\hat{\boldsymbol{h}} = \tanh\left(W_g^l \boldsymbol{y} + U_g^l\left(\boldsymbol{r} \odot \boldsymbol{x}\right)\right) \tag{8}$$

$$\boldsymbol{g}^l\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left(1 - \boldsymbol{z}\right) \odot \boldsymbol{x} + \boldsymbol{z} \odot \hat{\boldsymbol{h}} \tag{9}$$

其中：W 和 U 为线性层，$\boldsymbol{b}$ 是偏置，$\boldsymbol{z}$、$\boldsymbol{r}$、$\hat{\boldsymbol{h}}$ 为中间变量，σ 和 tanh 为激活函数，$\odot$ 表示矩阵的哈达玛积运算，$\boldsymbol{g}^l$ 为门控计算。用 GRU 替换原本残差连接的加和操作，智能体可以更好地获得信息并进行训练。

3.3　探针稀疏注意力机制和门控 Transformer 模型

探针稀疏注意力机制是一种稀疏机制，它从查询矩阵中筛选出一部分主要查询，从而降低了注意力运算的复杂度。

根据式（1）可知，在注意力计算时由于其中 $\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 导致了计算的复杂度 $O\left(L_Q L_K\right)$ 较大。为了降低复杂度，可以通过减少参与注意力计算的查询数目实现。

注意力分数通常呈现长尾分布[15]。在注意力分数矩阵中，通常得分高的仅仅占少数部分，多数得分通常很低，对最终的计算贡献很小，即只有少数的查询是主要的，称之为活跃（active）查询，而大部分查询的权重占比很小，称为惰性（lazy）查询。因此，只要找出活跃查询组成子查询 $\bar{\boldsymbol{Q}}$，就可以使复杂度降低。本研究通过探针稀疏注意力机制，利用 KL 散度计算查询的活跃度，以活跃查询组成查询矩阵计算注意力，从而降低注意力计算的复杂度。

记 $\boldsymbol{q}^i$、$\boldsymbol{k}^i$、$\boldsymbol{v}^i$ 分别为矩阵 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的第 i 行，则单个查询 $\boldsymbol{q}^i$ 的注意力分数为

$$A\left(\boldsymbol{q}_i, \boldsymbol{K}, \boldsymbol{V}\right) = \sum_j \frac{k\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right)}{\sum_l k\left(\boldsymbol{q}_i, \boldsymbol{k}_l\right)} \boldsymbol{v}_j = E_{p\left(\boldsymbol{k}_j \mid \boldsymbol{q}_i\right)}\left[\boldsymbol{v}_j\right] \tag{10}$$

其中：$\boldsymbol{q}_i$、$\boldsymbol{k}_i$、$\boldsymbol{v}_i$ 分别表示第 i 个查询、键和值。$p\left(\boldsymbol{k}_j \mid \boldsymbol{q}_i\right) = k\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right) / \sum_l k\left(\boldsymbol{q}_i, \boldsymbol{k}_l\right)$，$k$ 为非对称的指数操作，核函数 $k\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right) = \exp\left(\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}} / \sqrt{d}\right)$。由上式可知，第 i 个查询对所有键的注意力被定义为概率 $p\left(\boldsymbol{k}_j \mid \boldsymbol{q}_i\right)$ 与值的组合。由于活跃查询是那些能够找到与之关联最大的键的查询，因此主要的点积计算应该使相应的查询的概率分布远离均匀分布。当 $p\left(\boldsymbol{k}_j \mid \boldsymbol{q}_i\right)$ 接近均

匀分布 $q\left(\boldsymbol{k}_j \mid \boldsymbol{q}_i\right)=1/L_K$ 时，最终的注意力得分为各个值的微小的和，并且对残差连接是多余的。

基于此，可以用分布 p 和 q 的 KL 散度衡量查询的重要性，即

$$\mathrm{KL}(q\|p)=\ln\sum_{l=1}^{L_K}\mathrm{e}^{\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}-\ln L_K \tag{11}$$

其中 L_K 为键的长度，去掉其中常数项 $\ln L_K$，得稀疏度量公式，为

$$M\left(\boldsymbol{q}_i,\boldsymbol{K}\right)=\ln\sum_{l=1}^{L_K}\mathrm{e}^{\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}} \tag{12}$$

由于 LSE(log-sum-exp)计算存在潜在的数值稳定的问题，因此利用 max 操作经验近似，有

$$\bar{M}\left(\boldsymbol{q}_i,\boldsymbol{K}\right)=\max_j\left\{\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}}\right\}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}} \tag{13}$$

因此，计算自注意力时，首先随机选出 $U=L_k\ln L_q$ 组点积对，再代入式(6)选出 Top-u 查询，组成查询矩阵 $\bar{\boldsymbol{Q}}$，然后将 $\bar{\boldsymbol{Q}}$ 代入式(1)计算注意力得分。由此，稀疏注意力为

$$\mathrm{Attention}\left(\bar{\boldsymbol{Q}},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Softmax}\left(\frac{\bar{\boldsymbol{Q}}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{14}$$

通常在实际应用中，查询和键的长度在注意力计算时是相等的，即 $L_Q=L_K=L$。因此，探针稀疏注意力机制的时间复杂度和空间复杂度都是 $O(L\ln L)$。

通过将探针稀疏注意力机制与恒等映射重排和门控机制结合，最终得到 PS-GTr 的层块为

$$\bar{\boldsymbol{Y}}^l=\mathrm{ReLU}\left(\mathrm{SMHA}\left(\mathrm{LN}\left(\boldsymbol{E}^{l-1}\right)\right)\right) \tag{15}$$

$$\boldsymbol{Y}^l=g_{\mathrm{SMHA}}^l\left(\boldsymbol{E}^{l-1},\bar{\boldsymbol{Y}}^l\right) \tag{16}$$

$$\bar{\boldsymbol{E}}^l=\mathrm{ReLU}\left(f^l\left(\mathrm{LN}\left(\boldsymbol{Y}^l\right)\right)\right) \tag{17}$$

$$\boldsymbol{E}^l=g_{\mathrm{MLP}}^l\left(\boldsymbol{Y}^l,\bar{\boldsymbol{E}}^l\right) \tag{18}$$

其中：SMHA 为多头探针稀疏注意力操作，$\bar{\boldsymbol{Y}}$ 和 $\bar{\boldsymbol{E}}$ 为中间变量，g_{SMHA} 和 g_{MLP} 分别表示注意力子模块和前馈层子模块中的门控层。

3.4　基于 PS-GTr 的强化学习策略学习方法

使用 V-MPO 算法(on-policy maximum a posteriori policy optimization)作为强化学习的策略方法，用来训练 PS-GTr 模型。

V-MPO 采用了带有目标网络的 Actor-Critic 框架。V-MPO 首先从目标网络 $\pi_\theta^{\mathrm{old}}$ 中产生轨迹 τ；然后进行策略评估，从产生的经验数据中学习值函数 $V^{\mathrm{old}}(s)$ 并计算相应的优势函数 $A^{\pi^{\mathrm{old}}}(s,a)$；最后是策略改进，基于优势函数计算改进后的 π_θ，按照设置的步数 T_{target} 同步目标网络的参数。V-MPO 算法的总体损失函数 $L(\phi,\theta,\eta,\alpha)$ 可以分为策略评估损失 $L_{\mathrm{V}}(\phi)$ 和策略改进损失 $L_{\text{V-MPO}}(\theta,\eta,\alpha)$ 两部分，即

$$L(\phi,\theta,\eta,\alpha)=L_{\mathrm{V}}(\phi)+L_{\text{V-MPO}}(\theta,\eta,\alpha) \tag{19}$$

其中：ϕ 为价值网络的参数，θ 为策略网络的参数，η 和 α 为拉格朗日乘子。在训练时策略网络和价值网络共享用于状态序列编码的 PS-GTr 模型的参数。

策略评估损失 $L_{\mathrm{V}}(\phi)$ 通过最小化平方损失函数更新，为

$$L_{\mathrm{V}}(\phi)=\frac{1}{2|D|}\sum_{s_t\sim D}\left(V_\phi^\pi(s_t)-G_t^n\right)^2 \tag{20}$$

其中：D 为收集到的轨迹数据，s_t 为轨迹中的状态，$V_\phi^\pi(\cdot)$ 为当前策略 $\pi_\theta^{\mathrm{old}}$ 的值函数，G_t^n 为标准的 n-step 目标。策略改进损失为

$$L_{\text{V-MPO}}(\theta,\eta,\alpha)=L_\pi(\theta)+L_\eta(\eta)+L_a(\theta,\alpha) \tag{21}$$

其中 $L_\pi(\theta)$ 为最大似然损失。

$$L_\pi(\theta)=-\sum_{s,a\sim\tilde{D}}^{n}\psi(s,a)\log\pi_\theta(a\mid s) \tag{22}$$

$$\psi(s,a)=\frac{\exp\left(\dfrac{A^{\mathrm{target}}(s,a)}{\eta}\right)}{\displaystyle\sum_{s,a\sim\tilde{D}}\exp\left(\frac{A^{\mathrm{target}}(s,a)}{\eta}\right)} \tag{23}$$

其中：$A^{\mathrm{target}}(s_t,a_t)=G_t^n-V_\phi^\pi(s_t)$ 为目标网络的优势函数，$\tilde{D}$ 表示采集样本 D 中优势函数的值更高的一半，s 和 a 分别为从中采样的状态和动作。η 的损失函数为

$$L_\eta(\eta)=\eta\varepsilon_\eta+\eta\left[\frac{1}{|\tilde{D}|}\sum_{s,a\sim\tilde{D}}\exp\left(\frac{A^{\mathrm{target}}(s,a)}{\eta}\right)\right] \tag{24}$$

最后是 KL 约束 $L_a(\theta,\alpha)$，即

$$L_\alpha(\theta,\alpha)=\frac{1}{|D|}\sum_{s\in D}\Big[\alpha\left(\varepsilon_\alpha-\mathrm{sg}[[D_{\mathrm{KL}}\left(\pi_{\theta_{\mathrm{target}}}\|\pi_\theta\right)]]\right)+\mathrm{sg}[[\alpha]]D_{\mathrm{KL}}\left(\pi_{\theta_{\mathrm{target}}}\|\pi_\theta\right)\Big] \tag{25}$$

其中：sg[[·]]为梯度停止，即其中的变量相对其他项为常数；D_{KL} 为 KL 散度计算；ε_α 为超参数。V-MPO 的算法具体流程见算法 1。

算法 1：基于 PS-GTr 的 V-MPO 算法

1. 初始化 Actor 和 Critic 的参数 θ 和 ϕ。

2. 用目标网络 $\pi_\theta^{\mathrm{old}}$ 生成轨迹 τ，生成轨迹时的状态经 PS-GTr 编码。

3. 策略估计，以编码后的 τ 计算 $L_{\mathrm{V}}(\phi)$。

4. 计算优势函数 $A^{\text{target}}(s,a)$。
5. 策略改进，分别计算 $L_\pi(\theta)$、$L_\eta(\eta)$、$L_a(\theta,\alpha)$。
6. 更新参数 θ 和 ϕ。
7. 每隔 T_{target} 更新目标网络的参数。

4　实验结果与分析

以如图 2 所示的 Breakout-v5 环境为模型验证任务，验证本文模型的有效性。环境包含砖块、球、反弹板，智能体的学习目标是控制反弹板左右移动弹回小球，使其尽可能多地击中砖块。环境返回的状态是一张 RGB 格式 210×160 三通道图片。动作空间是四维离散的，分别对应于反弹板的左右移动、开火以及原地不动。当智能体控制反弹板将小球弹回击中砖块时，环境返回奖励。

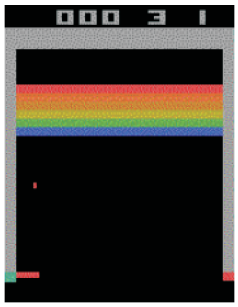


图 2　Breakout-v5 示意图

Fig. 2　Schematic diagram of Breakout-v5

本实验将通过 Breakout 环境对比 GTrXL 以及本文模型 PS-GTr 的各项训练指标。实验采用 V-MPO 作为智能体的训练算法，探索实验内容如下：

(1) 探针稀疏注意力机制的复杂度有效性检验：对比两个模型在不同的状态序列长度设置下的训练时间以及内存占用情况，以验证引入探针稀疏注意力机制的 Transformer 在强化学习中的作用。

(2) 性能对比实验：为了验证 PS-GTr 引入探针稀疏注意力机制后的性能，实验对比两个模型在训练时的平均累积回报的曲线图。

4.1　训练时间和内存占用

本节对比 GTrXL 与本文模型 PS-GTr 训练时所用的时间及内存。为了保证实验的公平性，两个模型的参数相同，设置层数 n_layer = 3 和注意力头数 n_head = 4，学习率 lr = 0.000 4，折扣因子 $\gamma=0.99$。此外，两个模型的训练在同一硬件平台上进行（Intel E5-2658，Nvidia RTX2080Ti）。为了验证本文方法的有效性，在实验时分别设置不同的序列长度，分别记录训练时不同序列长度的时间消耗以及内存占用情况并记录。

为了验证本文方法对计算复杂度的影响，实验设置序列长度分别为 50、100、150、200 时，智能体在环境中训练迭代 1 000 次所用时间，并绘制训练时间关于序列长度的变化图（图 3）。

从图 3 的实验结果可以看出，在序列长度大于 100 时，PS-GTr 的千轮训练时间小于 GTrXL，并且随着序列长度增大差距显著增大。图 3 的实验结果验证了加入探针稀疏注意力机制的门控 Transformer 的时间复杂度更低。

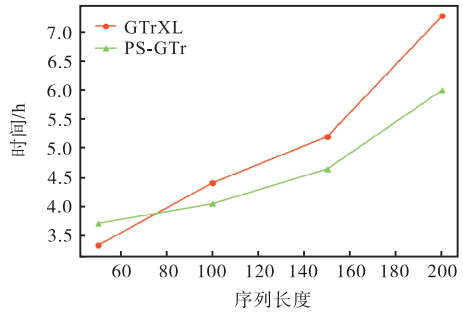


图 3　不同序列长度下的千轮训练时间对比

Fig. 3　Comparison of training time for a thousand episodes under different sequence lengths

测试内存占用的实验设置与测试训练时间的实验设置一致，分别测试不同序列长度下训练智能体时的内存占用情况，实验结果如图 4 所示。

图 4 结果表明，在相同的序列长度设置下训练智能体，PS-GTr 的内存占用小于 GTrXL。这表明基于探针稀疏注意力机制的门控 Transformer 能够降低空间复杂度，减少训练时的内存占用。

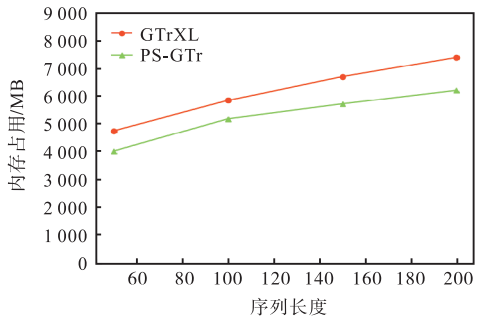


图 4　不同序列长度下的内存占用

Fig. 4　Memory usage under different sequence lengths

4.2　性能对比

对比 PS-GTr 与 GTrXL 在强化学习训练时的平均累积回报曲线变化，以验证引入了探针稀疏注意力机制后的模型性能几乎没有损失。为了保证对比实

验公平，设置两个模型的参数和策略参数都相同。每次实验进行 18 万次迭代，记录 10 次实验训练时的平均累积回报，实验结果如图 5 所示。

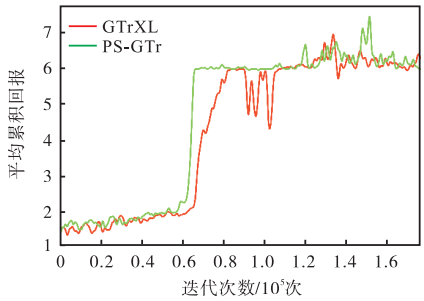


图 5　迭代过程中的平均累积回报

Fig. 5　Average cumulative return over the iteration

图 5 结果表明，引入了探针稀疏注意力机制的门控 Transformer 相对于 GTrXL 在智能体的训练中并无性能损失，两者收敛速度和最终性能一致。

综上所述，探针稀疏注意力机制的引入成功地在不损失性能的情况下降低了 GTrXL 的复杂度，提升了强化学习中以 Transformer 建模状态序列的效率，证明了本文模型在复杂度上优于 GTrXL，具有更高的效率。

5　结　语

GTrXL 的提出解决了 Transformer 在强化学习任务中难以训练的问题，本研究以减少 GTrXL 的计算复杂度、提升训练效率为研究目标，提出了基于探针稀疏注意力机制的门控 Transformer 模型。通过实验证明，基于探针稀疏注意力机制的门控 Transformer 模型比 GTrXL 具有更低的复杂度和更高的训练效率。在未来研究中，将研究如何进一步提升模型的性能，并在实验中验证模型的参数敏感性和有效性。

参考文献：

[1] SUTTON R S，BARTO A G. Reinforcement learning：an introduction[M]. 2nd ed. Cambridge：MIT Press，1998.

[2] ESPEHOLT L，SOYER H，MUNOS R，Et al. IMPALA：scalable distributed deep-RL with importance weighted actor-learner architectures[C]//PMLR. International conference on machine learning. New York：PMLR，2018：1407–1416.

[3] KAPTUROWSKI S，OSTROVSKI G，QUAN J，et al. Recurrent experience replay in distributed reinforcement learning[C]//ICLR. International conference on learning representations. New Orleans：ICLR，2019：1–19.

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

[5] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York：ACM，2017：6000–6010.

[6] DAI Z，YANG Z，YANG Y，et al. Transformer-XL：attentive language models beyond a fixed-length context[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.1901.02860.

[7] RADFORD A，WU J，CHILD R，et al. Language models are unsupervised multitask learners[J]. OpenAI blog，2019，1(8)：9.

[8] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[EB/OL].[2023–05–01]. https://arxiv.org/pdf/1810.04805.pdf.

[9] YANG Z，DAI Z，YANG Y，et al. XLNet：generalized autoregressive pretraining for language understanding[EB/OL]. [2023–05–01]. http://doi.org/10.48550/arXiv.1906.08237.

[10] EDUNOV S，OTT M，AULI M，et al. Understanding back-translation at scale[EB/OL].[2023–05–01]. https://arxiv.org/pdf/1808.09381.

[11] DEHGHANI M，GOUWS S，VINYALS O，et al. Universal transformers[EB/OL]. [2023–05–01]. https://arxiv.org/pdf/1807.03819.pdf.

[12] MISHRA N，ROHANINEJAD M，CHEN X，et al. A simple neural attentive meta-learner[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.1707.03141.

[13] PARISOTTO E，SONG F，RAE J，et al. Stabilizing transformers for reinforcement learning[C]//PMLR. International Conference on Machine Learning. New York：PMLR，2020：7487–7498.

[14] LI W，LUO H，LIN Z，et al. A survey on transformers in reinforcement learning[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.2301.03044.

[15] ZHOU H，ZHANG S，PENG J，et al. Informer：beyond efficient transformer for long sequence time-series forecasting[C]//AAAI. Proceedings of the AAAI Conference

on Artificial Intelligence, 2021, 35(12): 11106–11115.

[16] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]//ECCV. Computer Vision-ECCV 2016: 14th European Conference. Berlin: Springer International Publishing, 2016: 630–645.

[17] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. [2023–05–01]. https://arxiv.org/pdf/1406.1078.pdf.

[18] KUMAR S, PARKER J, NADERIAN P. Adaptive transformers in RL[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.2004.03761.

[19] SUKHBAATAR S, GRAVE E, BOJANOWSKI P, et al. Adaptive attention span in transformers[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.1905.07799.

[20] BANINO A, BADIA A P, WALKER J, et al. CoBERL: contrastive BERT for reinforcement learning[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.2107.05431.

[21] LI S, JIN X, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[C]//ACM. Proceeding of the 33rd International Coference on Neural Information Processing Systems. New York: ACM, 2019: 5243–5253.

[22] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document transformer[EB/OL]. [2023–05–01]. https://doi.org/10.48550/arXiv.2004.05150.

责任编辑：郎婧

（上接第 14 页）

[19] ZHU R, ZHANG G, JING M, et al. Genetically encoded formaldehyde sensors inspired by a protein intra-helical crosslinking reaction[J]. Nature communications, 2021, 12(1): 581.

[20] 汪世华，张晓鹏，陈明，等. 有机溶剂和变性剂对枯草芽孢杆菌溶栓酶活性的影响[J]. 应用与环境生物学报，2008，14(6)：825–829.

[21] 林建城，吴建洪，林娟娟. 有机溶剂来源污染物对中国蛰 N–乙酰–*β*–*D*–氨基葡萄糖苷酶的影响[J]. 中国海洋大学学报(自然科学版)，2021，51(6)：42–49.

[22] OHS R, LEIPNITZ M, SCHÖPPING M, et al. Simultaneous identification of reaction and inactivation kinetics of an enzyme-catalyzed carboligation[J]. Biotechnology progress, 2018, 34(5): 1081–1092.

[23] PETROTCHENKO E V, SERPA J J, CABECINHA A N, et al. “Out-Gel” tryptic digestion procedure for chemical cross-linking studies with mass spectrometric detection[J]. Journal of proteome research, 2014, 13(2): 527–535.

责任编辑：郎婧