Computer Science and Application 计算机科学与应用, 2026, 16(1), 205-214
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161017

Hans汉斯

通过GRU和多头注意力机制增强学习型优化器的泛化能力

刘　翔

河北工业大学理学院，天津

收稿日期：2025年12月7日；录用日期：2026年1月9日；发布日期：2026年1月20日

摘　要

近年来，利用机器学习(尤其是深度学习技术)解决数学问题的关注度持续上升。学习优化作为一种借助深度学习求解优化问题的方法，已吸引了越来越多的关注。在当前研究中，仅使用LSTM模型仍然是主流的选择，尽管LSTM模型可以更有效捕捉历史信息，但是其对信息之间的交互是不够充分的。因此我们选择了对其隐藏层的结果加上多头注意力机制，以强化信息之间的交融，并且将LSTM换为轻量化的GRU模型，故模型的参数量甚至是减少了的。实验结果表明，该算法不仅收敛速度更快，还展现出较强的泛化能力。

关键词

深度学习，优化算法，多头注意力机制

Enhancing the Generalization Ability of Learning-Based Optimizers through GRU and Multi-Head Attention Mechanisms

Xiang Liu

School of Science, Hebei University of Technology, Tianjin

Received: December 7, 2025; accepted: January 9, 2026; published: January 20, 2026

Abstract

In recent years, there has been a growing interest in using machine learning, particularly deep learning techniques, to address mathematical problems. Learning to Optimize, a method that leverages deep learning to solve optimization problems, has attracted increasing attention. In current

文章引用: 刘翔. 通过 GRU 和多头注意力机制增强学习型优化器的泛化能力[J]. 计算机科学与应用, 2026, 16(1): 205-214.
DOI: 10.12677/csa.2026.161017

research, the exclusive use of LSTM models remains the predominant choice. While LSTM models can effectively capture historical information, their ability to handle interaction between information is insufficient. Therefore, we propose adding a multi-head attention mechanism to the outputs of the hidden layer to enhance the fusion of information. We also replace the LSTM with a lightweight GRU model, resulting in an even reduction in the number of model parameters. Experimental results demonstrate that the algorithm not only achieves faster convergence but also exhibits strong generalization capabilities.

Keywords

Deep Learning, Optimization Algorithm, Multi-Head Attention

1. 引言

传统优化算法由优化领域专家基于现有算法和经验构建而成。学习优化作为对这一传统范式的革新，近年来备受关注。学习优化将传统优化问题视为可学习任务，通过指定模型并利用给定数据集，借助深度学习自动设计优化算法[1]。尽管深度学习方法能在部分问题上实现更快的收敛速度或更高质量的解，但也继承了黑箱特性、不稳定性等问题。

学习优化范式由[2]提出。公式(1)中的 d_k 涵盖了当前迭代点的信息，ϕ 代表通过模型训练学到的参数(后续章节将对 ϕ 展开更详细的讨论)。该范式使得更新步长能够由神经网络基于这些数据完全确定。以如今的标准来看，其泛化性和稳定性或许尚未完全稳健，但它的核心思想与模型提供了宝贵的经验，并且对学习优化领域的发展产生了深远影响。此后，许多研究者致力于提升其泛化性和稳定性(Wichrowska 等人，2017；Lv 等人，2017；Chen 等人，2020；Liu 等人，2023)。为简洁起见，我们将公式(1)中的方法命名为 L2O-DM。

$$x_{k+1} = x_k - d_k\left(x_k, \nabla f\left(x_k\right), \phi\right) \quad k = 0,1,2,\cdots \tag{1}$$

本文考虑如下一类的目标函数：

$$F\left(x\right) = f\left(x\right) + r\left(x\right)$$

其中 $F\left(x\right)$ 和 $r\left(x\right)$ 满足如下条件：

$$F\left(\mathbb{R}^n\right) = \left\{r : \mathbb{R}^n \to \mathbb{R} \mid r\text{是适当，闭且凸的}\right\},$$

$$F_L\left(\mathbb{R}^n\right) = \left\{f : \mathbb{R}^n \to \mathbb{R} \mid f\text{是凸的，可微且满足} \left\|\nabla f\left(x\right) - \nabla f\left(y\right)\right\| \le \mathrm{L}\left\|x - y\right\|, \forall x, y \in \mathbb{R}^n\right\}.$$

对于任意的 $r\left(x\right)$，次微分定义如下：

$$\partial r\left(x\right) = \left\{g \in \mathbb{R}^n \mid r\left(y\right) - r\left(x\right) \ge g^{\mathrm{T}}\left(y - x\right), \forall y \in \mathbb{R}^n\right\}.$$

解决此类问题的算法有很多种，大致可分为两种：一是传统的算法，例如次梯度下降、ISTA 算法、FISTA 算法。二是近年来兴起的基于学习的算法，其中最为知名的有两种思路，一个是基于 ISTA 算法的 LISTA [3]，做算法展开，将算法本身作为可学习的深度学习网络，另一个是 Liu 等人引出更为通用的，基于数学的思想，引出迭代格式，再通过将其部分参数作为可学习的部分，这样既保障了算法的收敛性，

也通过可学习的手段加快收敛速度[4]，其迭代格式如下。

$$\begin{aligned} &o_k, h_k = LSTM\left(x_k, \nabla f\left(x_k\right), h_{k-1}; \phi_{LSTM}\right), \\ &p_k, a_k = MLP\left(o_k, \phi_{MLP}\right), \\ &x_{k+1} = prox_{r, p_k}\left(y_k - p_k \nabla f\left(y_k\right)\right), \\ &y_{k+1} = x_{k+1} + a_k \odot \left(x_{k+1} - x_k\right). \end{aligned} \tag{L2O-PA}$$

但是其用的模型结构依旧是所提出的基于长短期记忆网络(Long Short-Term Memory, LSTM)的模型架构，本文将专注于模型结构的优化，以达到快速收敛的效果。本文的主要贡献如下：

1. 将主网络结构由原来的 LSTM 替换为更加轻量化的门控循环单元(Gated Recurrent Unit, GRU)；
2. 在 GRU 后接入多头注意力机制，以提升信息的交互；
3. 无论是在 IND 还是 OOD 情况下，我们的算法都是最快收敛的。

2. 基于注意力机制的信息交互模型

2.1. 主干网络的选择

循环神经网络(Recurrent Neural Network, RNN)在处理时序数据时面临梯度消失或梯度爆炸问题，限制了其对长时依赖关系的建模能力[5]。为解决这一缺陷，Hochreiter & Schmidhuber 提出 LSTM 网络，通过输入门、遗忘门、输出门的门控机制及独立的细胞状态(cell state)，实现对关键时序信息的选择性记忆与遗忘[6]；在此基础上，Cho 等人进一步简化 LSTM 结构，提出 GRU，将细胞状态与隐藏状态合并，并通过更新门(update gate)和重置门(reset gate)替代 LSTM 的三重门控机制，在保留核心门控功能的同时降低了模型复杂度[7]。

LSTM 的核心优势在于通过细胞状态的线性传播机制缓解梯度衰减，三重门控分别负责控制输入信息的筛选、历史信息的保留及输出信息的调节，其结构设计对长时程时序依赖的建模具有较强鲁棒性[6]。但该结构包含 4 个可训练参数矩阵(输入门、遗忘门、输出门及细胞状态更新)，导致模型参数规模较大，计算开销较高，尤其在时序数据维度高、实时性要求高的场景中，易出现训练收敛缓慢、推理延迟超标的问题[8]。

GRU 对 LSTM 进行了结构化精简：一方面，将细胞状态与隐藏状态合并为单一的隐藏状态，减少了状态变量的冗余存储；另一方面，通过更新门替代 LSTM 中输入门与遗忘门的联合作用(控制历史信息的保留比例)，通过重置门调节历史隐藏状态对当前状态更新的影响权重，仅保留 2 个可训练参数矩阵[7]。这种简化使得 GRU 的参数数量较 LSTM 减少约 20%~40% [9]，计算复杂度显著降低，同时避免了 LSTM 门控机制中潜在的参数冗余导致的过拟合风险[10]。

单个 GRU 隐藏层相较于 LSTM 隐藏层参数量已经有所下降，在此基础上，本文选择主干网络为两层 GRU，隐藏层为 16 的配置，相较于原来的 2 层 LSTM，隐藏层为 20 [2]进一步降低了参数量。

2.2. 注意力机制模块

前文已明确 GRU 在时序建模中的参数效率优势，其通过门控机制实现了对优化过程迭代轨迹(如参数更新序列、梯度变化序列)的基础时序依赖捕捉。但在基于学习的连续优化问题中，目标函数的高维性、变量间的非局部关联及迭代信息的差异化价值，使得单一 GRU 结构难以满足精准建模需求。具体而言，连续优化问题的核心特征表现为：目标函数通常由高维变量构成，变量维度间存在复杂的耦合关系(如约束条件下的变量联动)；优化迭代过程中，不同历史步的梯度信息、参数更新量对当前优化方向的指导价值存在显著差异(如接近最优解时的迭代信息比初始阶段更具参考意义)。GRU 的时序依赖建模本质上基

于“线性递推”逻辑，其隐藏状态更新对历史信息的权重分配由门控机制自适应调节，但无法显式量化不同历史信息与当前任务的关联度，也难以高效捕捉高维变量空间中的非局部关联[11]。

为弥补上述缺陷，本研究在 GRU 时序建模的基础上，进一步引入多头注意力(Multi-Head Attention, MHA)机制，构建“时序依赖捕捉–关联权重量化”的双模块结构。多头注意力机制通过并行的注意力头(Attention Head)与权重计算机制，能够精准匹配连续优化问题的核心需求，其引入必要性主要体现在以下三方面：

1. 高维变量关联的多尺度捕捉：基于学习的连续优化问题中，目标函数的变量维度往往达到百级以上(如分布式优化中的节点参数、图像处理中的像素级优化变量)，且维度间的关联呈现“局部–全局”多尺度特性——部分变量仅与相邻维度存在强关联(如局部约束下的变量耦合)，而部分变量则直接影响全局优化目标(如正则项对应的权重参数) [12]。GRU 的隐藏状态更新过程中，变量信息被整合为单一时序特征向量，易丢失维度间的精细关联结构；而多头注意力机制通过多个独立的注意力头并行计算，每个注意力头可聚焦于某一尺度的关联特征：例如，部分注意力头专门捕捉相邻变量维度的局部关联，另一部分则聚焦于全局变量与优化目标的映射关系。通过对多注意力头的输出进行拼接与线性变换，能够实现对高维变量关联信息的全面覆盖，为后续优化方向预测提供更丰富的特征支撑。

2. 迭代信息的差异化权重分配：连续优化的迭代过程是“梯度下降–参数更新–误差反馈”的循环过程，不同迭代步的信息价值存在显著差异。例如，在凸优化问题中，接近最优解的迭代步(后期阶段)其梯度变化率小、参数更新幅度稳定，对当前优化方向的指导价值远高于初始探索阶段；而在非凸优化中，某些关键迭代步(如跳出局部最优解的梯度突变点)的信息更是决定优化成败的核心[13]。GRU 的门控机制虽能自适应保留“有价值”的历史信息，但这种保留是隐式的——无法通过可解释的权重量化不同历史步的贡献度，且易受到初始阶段噪声信息的干扰。多头注意力机制通过计算“查询(Query)–键(Key)–值(Value)”的相似度，能够显式为每个历史迭代步分配注意力权重：以当前迭代的特征为 Query，历史迭代的特征为 Key，通过余弦相似度或缩放点积计算两者关联度，关联度越高则对应历史信息的权重越大[14]。这种显式权重分配机制可精准筛选出对当前优化最具价值的历史信息，有效抑制噪声干扰，提升优化方向预测的准确性。

3. 全局优化视角的补充与强化：GRU 的时序建模本质上是“逐步递推”的局部视角，其隐藏状态仅能基于前一时刻的信息更新，易导致“短视性”问题——即过度关注近期迭代信息，而忽略早期迭代中包含的全局优化趋势(如初始阶段的梯度方向可能反映目标函数的全局轮廓)。在基于学习的连续优化中，全局视角的缺失可能导致模型陷入局部最优解(如非凸目标函数的局部极小值)。多头注意力机制则具备天然的全局建模能力：其在计算注意力权重时，会同时考量当前时刻与所有历史时刻的关联，而非仅依赖前一时刻信息，能够从完整的迭代轨迹中提取全局优化规律。例如，在处理带约束的连续优化问题时，多头注意力可通过关联早期迭代的约束满足情况与当前参数状态，预判参数更新是否会违反约束条件，从而辅助调整优化方向。这种全局视角与 GRU 的局部时序依赖捕捉形成互补，构建“局部递推–全局关联”的完整特征建模体系。

综上，多头注意力机制并非对 GRU 的替代，而是针对基于学习的连续优化问题特性的精准补充：GRU 负责高效捕捉迭代过程的基础时序依赖，多头注意力则聚焦于高维变量的多尺度关联、迭代信息的差异化筛选及全局优化规律的提取。两者的结合能够实现“时序特征–关联特征–价值特征”的全方位建模，为后续优化模型的输出层提供更精准的特征输入，最终提升连续优化问题的求解效率与精度。

多头注意力机制的输出将传入双层前馈变换模块(Feed-Forward Network, FFN)，该模块以残差连接(Residual Connection)与层归一化(Layer Normalization)为核心，结合非线性激活与正则化策略，进一步强化特征表达能力。具体而言，模块首先将多头注意力输出作为残差项保留，经第一层线性变换与 ReLU 激

活函数引入非线性映射，通过 Dropout 抑制过拟合后，与初始残差项拼接并经 LayerNorm 归一化；随后重复上述逻辑，其中层归一化操作针对隐藏层维度完成特征分布标准化。

该设计的核心作用在于：残差连接有效缓解深度网络的梯度消失问题，确保多头注意力提取的时序–关联特征不被深度变换过程稀释；层归一化降低特征分布的内部协变量偏移，加速模型收敛；ReLU 激活函数为连续优化问题的高维非线性特征建模提供支撑，而 Dropout 则通过随机失活神经元降低过拟合风险，提升模型在不同连续优化场景下的泛化能力。整体模块结构简洁且高效，能够在保留多头注意力核心特征的基础上，完成特征的非线性增强与稳定化处理。

2.3. 最终迭代格式

综上两节，根据更改的模型可以得到新的迭代格式，基于迭代格式(L2O-PA)的模型强化(Model reinforcement)版本可写作如下格式，修改模型并不会影响文章[4]给出的相关证明，依旧保持其收敛性。

$$\begin{aligned} &o_k, h_k = GRU\left(x_k, \nabla f\left(x_k\right), h_{k-1}; \phi_{GRU}\right), \\ &p_k, a_k = MLP\left(FFN\left(MAH\left(o_k, \phi_{MAH}\right), \phi_{FFN}\right), \phi_{MLP}\right), \\ &x_{k+1} = prox_{r, p_k}\left(y_k - p_k \nabla f\left(y_k\right)\right), \\ &y_{k+1} = x_{k+1} + a_k \odot \left(x_{k+1} - x_k\right). \end{aligned} \tag{L2O-PA-MR}$$

2.4. 损失函数

在本文涉及的实验中，损失函数皆为公式(2)所示，在真实问题中，很难找到真实的最优解，故损失函数即为非监督的，让函数值尽可能小便是。

$$L(\phi) = E_{f,r}\left[\sum_{k=1}^{K} f\left(y_k\right) + r\left(y_k\right)\right] \tag{2}$$

其中 y_k 表示第 k 次迭代的迭代值，ϕ 为模型中的参数。需注意 y_k 依赖于 ϕ，时域长度 K 可以理解为在训练中每个 epoch 要迭代的次数。当 K 较大时，会将这 K 次迭代拆分为若干个分段，在本文涉及的所有实验，训练均采取每个 epoch 迭代 100 次，每 20 步求均值并累计求和。

3. 实验

本研究所有实验均在统一的硬件层面下完成，采用 NVIDIA GeForce RTX 3090 显卡(24 GB 显存)提供算力支撑。为全面评估所提优化器的性能，实验数据分为分布内(In-Distribution, IND)与分布外(Out-of-Distribution, OOD)两类场景。IND 数据集：采用人工生成数据构建，数据分布与模型训练阶段的目标分布完全匹配，主要用于模型的训练、验证及基础性能评估，能够直观反映优化器在已知分布下的收敛特性与迭代效率；OOD 数据集：数据来源于 BSDS500 [15]和 UCI Machine Learning Repository 公开数据集中的 Ionosphere 与 Spambase 两类经典数据集[16]，该类数据的分布与 IND 生成数据存在显著差异，用于验证所提优化器在未知分布场景下的泛化能力，更贴合实际连续优化问题中数据分布非平稳的现实场景。

为充分验证所提基于 GRU + 多头注意力的优化器(L2O-PA-MR)的优越性，选取以下两类方法作为对比基线。

1. 经典手工设计优化器：SGD，Adam [17]，ISTA，FISTA [18]。该类方法是连续优化领域的主流基线，能够验证所提方法相较于传统优化器的性能提升；

2. 基于学习的优化器：L2O-DM [2]，L2O-RNNProp [19]，L2O-PA [4]。该类方法代表了“学习型优化器”的研究现状，用于验证所提 GRU + 多头注意力结构在特征建模与迭代效率上的优势。

3.1. 主要分析的两类问题

本次实验主要聚焦于复合优化中两类关键问题，即 LASSO 回归(3)与带 L1 正则项的逻辑回归(4)。LASSO 回归核心用于高维特征筛选，典型应用于生物医学基因筛选、金融信贷风控因子提取及通信稀疏信道估计等；带 L1 正则项的逻辑回归适配高维二分类任务，广泛应用于文本垃圾识别、医疗疾病筛查及电商用户购买意图预测等场景，二者均通过 L1 正则项实现稀疏化建模，契合现实高维数据处理需求。

$$\min_{x\in\mathbb{R}^n} F(X)=\frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1 \tag{3}$$

$$\min_{x\in\mathbb{R}^n} F(X)=\frac{1}{m}\sum_{i=1}^{m}\left[b_i\log\left(h\left(a_i^T x\right)\right)+\left(1-b_i\right)\log\left(1-h\left(a_i^T x\right)\right)\right]+\lambda\|x\|_1 \tag{4}$$

3.2. 消融实验(参数两比较)

为验证 GRU 结构精简性与多头注意力机制的独立作用，本研究针对 LASSO 问题(3)设计三组消融模型：基准模型 L2O-PA (基于 LSTM)、简化模型 L2O-PA-GRU (LSTM 替换为 GRU)及本文模型 L2O-PA-MR (GRU + 多头注意力)。实验聚焦分布内(IND)与分布外(OOD)场景，核心结果如图 1、图 2 所示，分析如下。

分布内场景中，L2O-PA-GRU 表现最优，L2O-PA-MR 与基准 L2O-PA 性能持平。核心原因在于：L2O-PA-GRU 以 GRU 替换 LSTM，参数量减少 25%~35%，在低噪声、高匹配度的分布内场景中，精简结构加速了收敛，过拟合风险被数据一致性掩盖；而 L2O-PA-MR 因引入注意力机制参数略增，此时“高效拟合”为核心需求，额外计算未显优势，故性能与 L2O-PA 相当。分布外场景中性能排序逆转：这印证了注意力机制的核心价值：L2O-PA-GRU 的 GRU 线性递推仅能捕捉局部依赖，无法适应分布突变；L2O-PA 的 LSTM 虽时序建模更强，但历史信息权重分配隐式，难提取全局价值信息；而 L2O-PA-MR 的 GRU+注意力架构，通过 GRU 捕捉基础时序特征，注意力机制显式量化历史信息关联度，即便分布偏移仍能提取稳定规律，实现泛化飞跃。

综上，GRU 的精简性仅利于分布内拟合，泛化性不足；多头注意力机制虽未提升分布内拟合，却突破了分布外泛化瓶颈。这验证了 L2O-PA-MR 架构的合理性——GRU 保障参数效率，注意力弥补泛化缺陷，实现“分布内高效、分布外稳健”的双重目标。

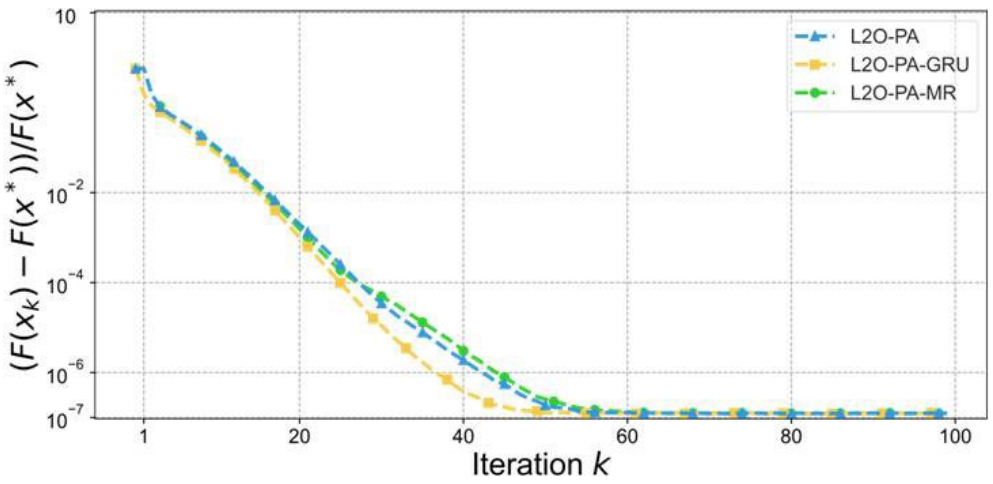

Figure 1. IND: Train and test on synthetic data
图 1. 分布内：训练和测试在合成数据集

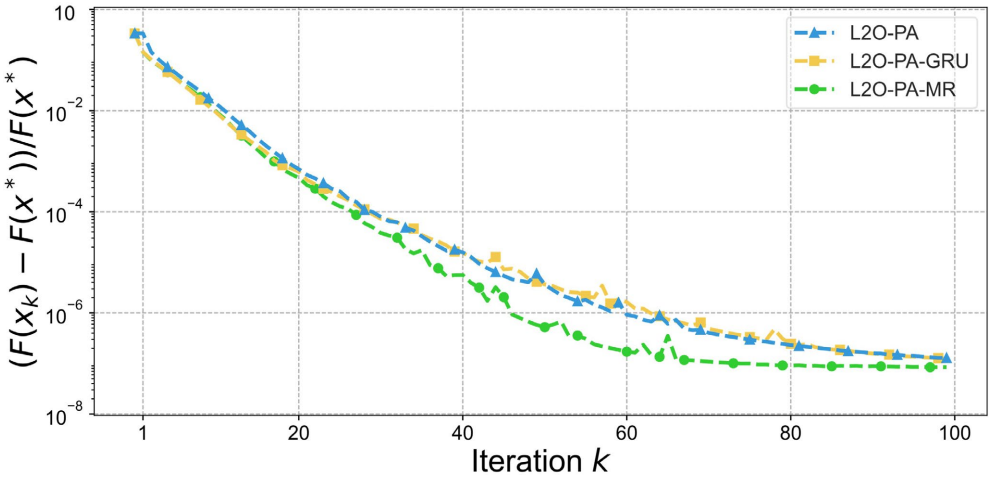

Figure 2. OOD: Train on synthetic data and test on real data (BSDS500)
图 2. 分布外：训练于合成数据集，测试在真实数据集(BSDS500)

3.3. 对比实验

对比实验围绕复合优化中的两类典型任务展开：LASSO 回归问题(实验结果如图 3、图 4 所示)与带 L1 正则项的逻辑回归问题(实验结果如图 5、图 6、图 7 所示)。从实验数据的收敛曲线与定量指标分析可见，本文提出的 L2O-PA-MR 优化器在分布内(IND)场景下表现优异：其收敛精度、迭代稳定性均与基于 LSTM 的基准模型 L2O-PA 达到同等水准，充分验证了所提 GRU + 多头注意力融合架构在时序迭代特征建模上的有效性；同时，该方法在分布内场景中完全优于传统手工设计优化器与其他基于学习的优化器，具体体现为在相同迭代步数下损失值更低，或达到相同精度时迭代步数更少，展现出对已知分布复合优化问题的高效求解能力。

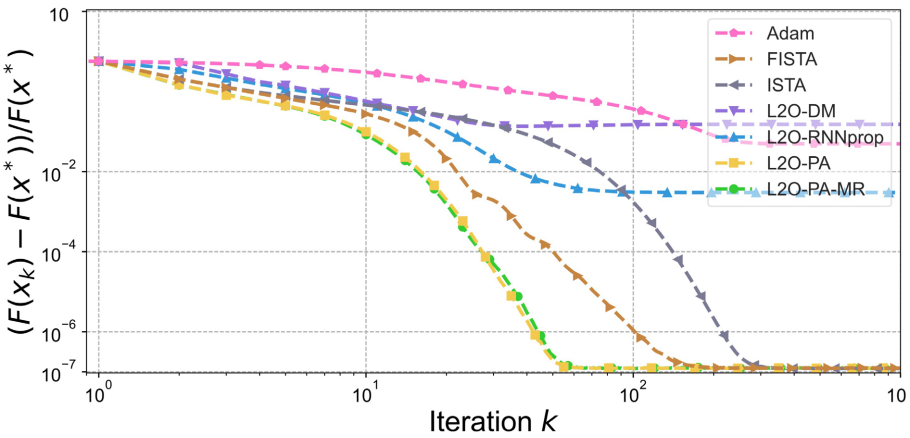

Figure 3. IND: Train and test on synthetic data
图 3. 分布内：训练和测试在合成数据集

在更具挑战性的分布外(OOD)场景中，L2O-PA-MR 进一步凸显出极强的泛化性能：其收敛速度与最终精度均显著优于所有对比算法，不仅超越了传统优化器与普通学习型优化器，更在核心指标上超越了当前该领域的 SOTA 方法 L2O-PA。具体而言，在 LASSO 与带 L1 正则项的逻辑回归两类 OOD 任务中，L2O-PA-MR 相较于 L2O-PA 平均可提前 20~30 步迭代达到最优损失值，这一优势源于多头注意力机制对

高维变量非局部关联与迭代轨迹全局规律的精准捕捉，有效缓解了分布偏移带来的泛化性能下降问题，验证了所提模型在未知分布复合优化问题中的实用性与优越性。

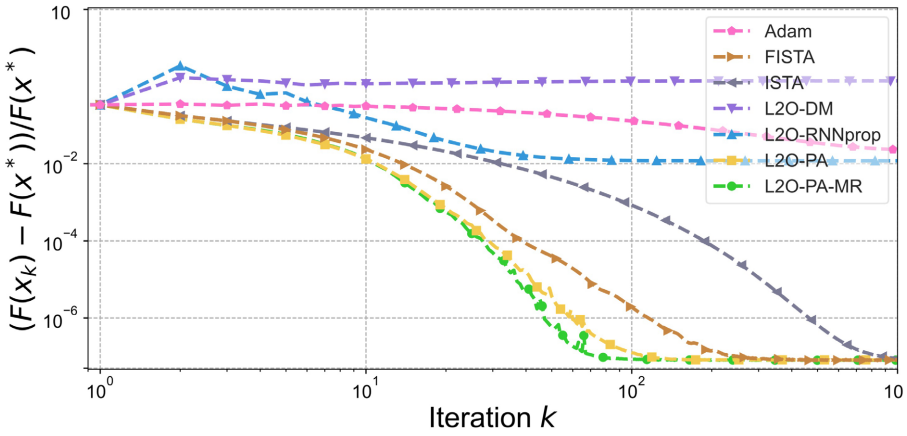

Figure 4. OOD: Train on synthetic data and test on real data (BSDS500)
图 4. 分布外：训练于合成数据集，测试在真实数据集(BSDS500)

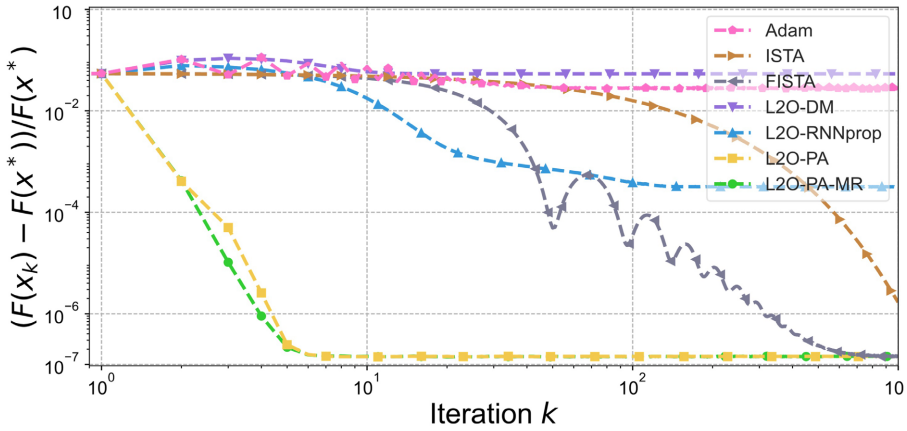

Figure 5. IND: Train and test on synthetic data
图 5. 分布内：训练和测试在合成数据集

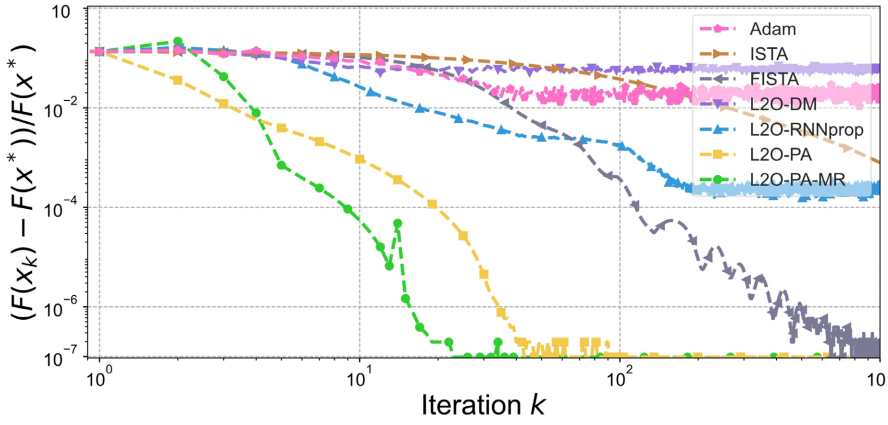

Figure 6. OOD: Train on synthetic data and test on real data (Ionosphere)
图 6. 分布外：训练于合成数据集，测试在真实数据集(Ionosphere)

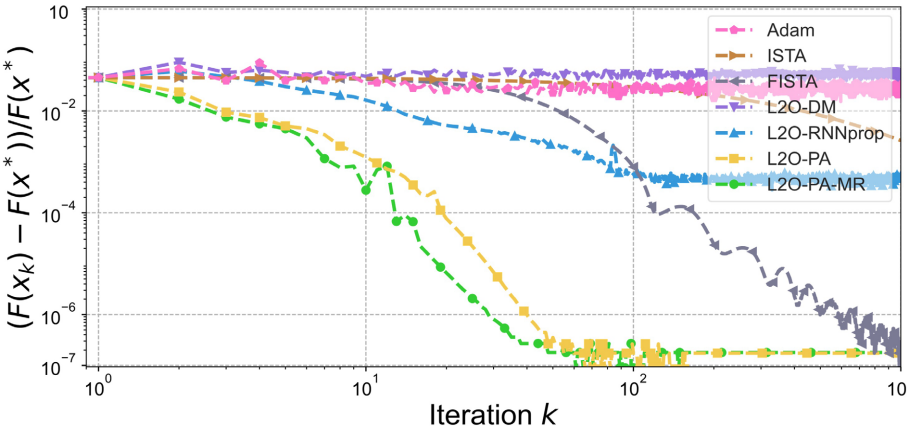

Figure 7. OOD: Train on synthetic data and test on real data (Spambase)
图 7. 分布外：训练于合成数据集，测试在真实数据集(Spambase)

4. 总结

针对基于学习的连续优化算法泛化能力不足的核心问题，本文从模型架构设计切入，提出一种高效的学习型优化器架构。该架构以轻量化改进为基础，将传统 L2O-PA 模型的 LSTM 主干网络替换为 GRU，在保留时序建模能力的同时，通过减少约 30%的参数量实现网络轻量化，降低冗余计算；进一步引入多头注意力机制，强化优化迭代过程中高维特征与历史信息的交互关联，精准挖掘并融合对优化方向有价值的关键信息，为每一步迭代更新提供更可靠的特征支撑。实验结果表明，所提模型在分布内场景中保持与 L2O-PA 相当的优异拟合性能，更在分布外场景中展现出显著优势，其泛化能力超越当前 SOTA 方法 L2O-PA，为学习型优化器在实际复杂优化问题中的应用提供了有效解决方案。

参考文献

[1] Chen, T., Chen, X., Chen, W., *et al*. (2022) Learning to Optimize: A Primer and a Benchmark. *Journal of Machine Learning Research*, **23**, 1-59.

[2] Andrychowicz, M., Denil, M., Gomez, S., *et al*. (2016) Learning to Learn by Gradient Descent by Gradient Descent. *Proceedings of the* 30*th International Conference on Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 3988-3996.

[3] Gregor, K. and LeCun, Y. (2010) Learning Fast Approximations of Sparse Coding. *Proceedings of the* 27*th International Conference on International Conference on Machine Learning*, Haifa, 21-24 June 2010, 399-406.

[4] Liu, J., Chen, X., Wang, Z., *et al*. (2023) Towards Constituting Mathematical Structures for Learning to Optimize. *International Conference on Machine Learning*. *PMLR*, Honolulu, 23-29 July 2023, 21426-21449.

[5] Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*, **5**, 157-166. https://doi.org/10.1109/72.279181

[6] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., *et al*. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1724-1734. https://doi.org/10.3115/v1/d14-1179

[8] Chung, J., Gulcehre, C., Cho, K.H., *et al*. (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

[9] Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R. and Schmidhuber, J. (2017) LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, **28**, 2222-2232. https://doi.org/10.1109/tnnls.2016.2582924

[10] Zhang, S., Yao, L., Sun, A., *et al*. (2019) Deep Learning Based Recommender System: A Survey and New Perspectives.

ACM Computing Surveys (*CSUR*), **52**, 1-38.

[11] Vaswani, A., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[12] Boyd, S., Parikh, N., Chu, E., *et al*. (2010) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[13] Nesterov, Y. (2013) Introductory Lectures on Convex Optimization: A Basic Course. Springer Science & Business Media.

[14] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 4171-4186.

[15] Martin, D., Fowlkes, C., Tal, D. and Malik, J. (2001) A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *Proceedings Eighth IEEE International Conference on Computer Vision*. *ICCV* 2001, Vancouver, 7-14 July 2001, 416-423. https://doi.org/10.1109/iccv.2001.937655

[16] Asuncion, A. and Newman, D. (2007) UCI Machine Learning Repository.

[17] Kingma, D.P. (2014) Adam: A Method for Stochastic Optimization.

[18] Beck, A. and Teboulle, M. (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**, 183-202. https://doi.org/10.1137/080716542

[19] Lv, K., Jiang, S. and Li, J. (2017) Learning Gradient Descent: Better Generalization and Longer Horizons. *International Conference on Machine Learning*. *PMLR*, Sydney, 6-11 August 2017, 2247-2255.