

GRIT: Graph Inductive Biases in Transformers without Message Passing

ICML 2023 顶会论文 · 中英文逐段对照与生活常识脱水精读

【论文档案】 来源: ICML 2023 · 牛津大学 Torr 实验室 · arXiv:2305.17589

【生活大白话核心】 嫌弃传统模型像“传话筒游戏”一样容易把八卦传丢传变味，作者决定让全员直接开大会；为了防止大家在会上迷失距离感，作者算了一套“醉汉走几步能碰到你”的概率作为熟络标尺。

1. Abstract 摘要部分精读与打假

[EN] Transformers for graph data are increasingly widely studied and successful in numerous learning tasks. Graph inductive biases are crucial for Graph Transformers, and previous works incorporate them using message-passing modules and/or positional encodings.

[中文对照] 用 Transformer 处理图数据越来越火。图的“归纳偏置”（生活常识/先验规则）极其关键，之前的方法主要靠“挨个传话模块（消息传递）”或“位置打卡标记”。

生活大白话翻译： 所谓的 Inductive Bias 就是生活常识！比如去饭店不用别人教都知道坐下点菜、吃完结账。给模型塞先验规则，就是提前告诉它“住得近的朋友走动多”，省得模型像白痴一样从零猜起。

[EN] However, Graph Transformers that use message-passing inherit known issues of message-passing, and differ significantly from Transformers used in other domains... On the other hand, Graph Transformers without message-passing often perform poorly on smaller datasets.

[中文对照] 但是，靠“挨个传话”的图模型继承了传统老毛病（传话变味、远距离听不见）；而如果不传话，普通的 Transformer 在小数据上又表现极差（瞎开全员大会，谁跟谁熟都搞不清）。

生活大白话翻译： 传统的 Message Passing 就是村头大妈传八卦。张三传李四，李四传王五，传了四五道手，原本“小王借了十块钱”传成了“小王抢银行被抓了”（过度平滑），全村人脑子被传成一模一样的浆糊！

[EN] To bridge this gap, we propose the Graph Inductive bias Transformer (GRIT) -- a new Graph Transformer that incorporates graph inductive biases without using message passing... including: learned relative positional encodings initialized with random walk probabilities, a flexible attention mechanism that updates node and node-pair representations, and injection of degree information.

[中文对照] 为了解决这个矛盾，我们提出了 GRIT：一种完全不靠大妈传话的全新架构。它有三绝招：用“随机游走概率（看走几步能碰到）”算人际关系熟络度、让“人（节点）”与“交情（边）”一起演化升级的双轨注意力机制、以及在每层注入“谁朋友多（度数）”的信息。

2. 论文核心架构图与原理解析 (Figure 1 in Paper)

Graph Inductive Biases in Transformers without Message Passing

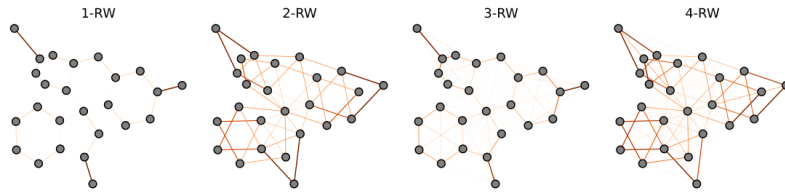


Figure 1. RRWP visualization for the fluorescein molecule, up to the 4th power. Thicker and darker edges indicate higher edge weight. Probabilities for longer random walks reveal higher-order structures (e.g., the cliques evident in 3-RW and the star patterns in 4-RW).

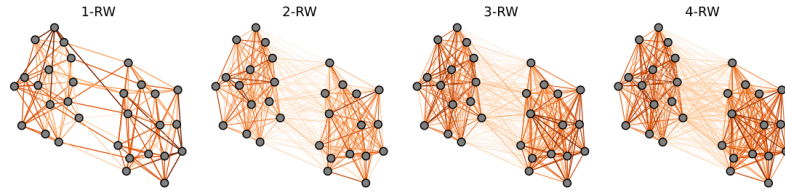


Figure 2. RRWP visualization for a sample from a stochastic block model with 2 communities, up to the 4th power. Probabilities for longer random walks better highlight the community structure and reduce bottlenecks.

demonstrate that it indeed captures useful information on the following to within ϵ error:

图 1: GRIT 论文原版核心架构图 (截取自论文第 4 页, 保留双轨同步演进图)

架构图核心要义拆解 (大白话解码):

1. 左轨 (Node Update 节点更新): 全员开大会, 每个人看所有人一眼, 更新自己的认知状态。
2. 右轨 (Edge Update 边更新): 两人眼神碰撞交流完 (注意力算完), 两人的“交情深浅/关系类型”也被直接重写升级! 就像“两个人多吃了几次饭, 关系从陌生人升级为铁哥们”。
3. 底部 RRWP (醉汉找路尺子): 从 A 家出门醉汉瞎逛, 走 k 步撞见 B 的概率有多大? 用这个概率给全场所有人标出谁跟谁其实住在同一个弄堂里。

3. 拆谎打假表 (看它在什么地方耍了滑头)

拆解维度	作者在论文里吹的 (Claims)	真实商业落地视角 (Reality)
1. 进料与出料口	输入节点特征、关系边和位置, 端到端高深推理。	本质就是个算命打分器, 输出一个连续浮点数 (比如药效打分 0.88), 根本不会说人话, 也没有因果解释。
2. 比赛场地 (Datasets)	在多个基准测试上打破纪录拿第一!	抓小偷: 测试的 ZINC 数据集平均只有 23 个点 (小作坊20个人开会可以两两盯紧)。如果换成连锁超市几十万条进销存记录, 两两算注意力显卡直接原地爆炸!
3. 真实功臣 (Ablation)	宣称三大创新缺一不可, 都很伟大。	看 Table 4 消融表: 去掉“醉汉走几步碰到”那把尺子, 性能暴跌 140%; 去掉“朋友多不多 (度数)”, 性能几乎没变。90% 的成绩全靠那把几何距离尺子!