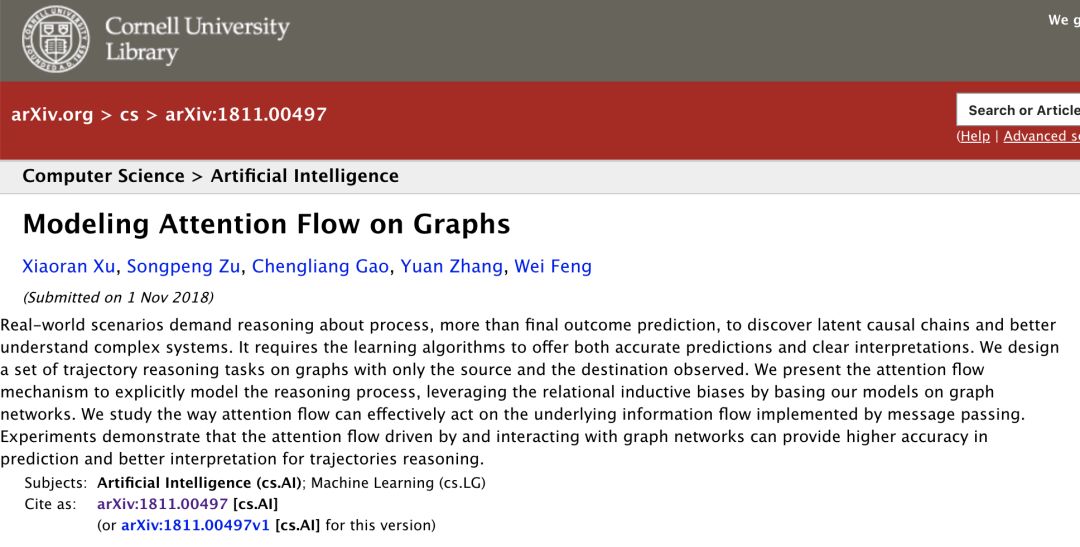
图网络上的注意力流

原创： 辛茹月 [集智俱乐部](javascript:void(0);) 2018-11-20



**1.因果推理召唤图网络**

现实世界中，很多实际应用都需要推断潜在的因果链或者构建对观察、预测的解释。

例如在物理世界中，我们想要在给定很少观察现象的情况下推理出移动物体的轨迹; 在视频流系统中，我们希望推荐模型能够跟踪不断变化的用户兴趣，以提供将用户观看与推荐视频相关联的个性化推荐原因。

在此，我们介绍一篇近期发表在 arxiv.org 上的论文，**研究者运用图网络（Graph Network）模型，专注基于图的场景，旨在推断出基于给定图的节点和连边导致观察到的结果的隐含的关系链。**

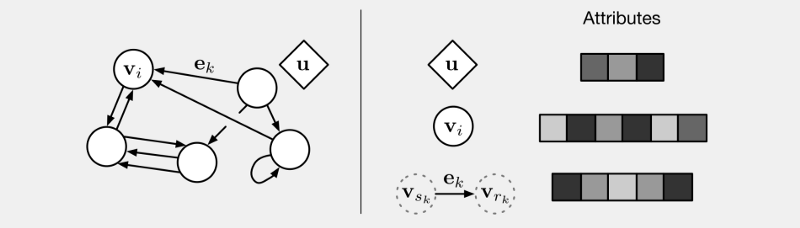
论文题目：Modeling Attention Flow on Graphs

论文地址：https://arxiv.org/abs/1811.00497

图网络模型是一系列神经网络，它们在图上运行并且带有很强的关系归纳能力。因此，图网络具有处理图结构数据的强大能力。然而，当建模可解释的推理过程时，图网络模型的黑盒性质使得它相比其他可微分的逻辑推理模型更不具竞争力。因此，这项工作在图上开发了一种新的注意力机制，称为注意力流，用于对推理过程进行建模，以便可解释预测的最终结果。

文中使用图网络中的消息传递算法来得到随时间步长演变的转换矩阵，以驱动注意力流；另外还让注意力集中在传递的消息上，称为信息流。为了评估模型，本文设计了一组轨迹推理任务，其中仅观察轨迹的源和目标，学习中间轨迹。

这篇文章首先**基于图网络建立了注意力流机制（attention flow mechanism）**，相当于引入了一种新的方法来提高应用图网络时的透明度。 其次，文章还展示了注意力流如何能够有效地影响在图网络中进行的消息传递，类似于强化学习**。**实验表明，具有显式和后向作用（explicit and backward acting）注意力流的图网络模型在预测准确性和可解释性方面都较优。

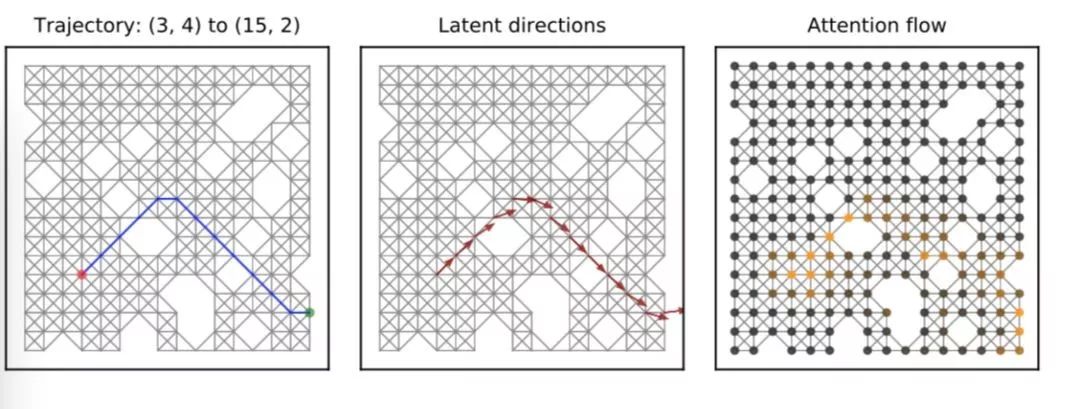


图网络的基本架构，图的特征以三种形式表示：节点（nodes），边的关系（relations）以及全局属性（global attributes）

**2.设计推理任务**

现实场景通常需要对过程进行推理，即通过列出一系列将开头与结果联系起来的因果关系来构建解释。因此，本文通过一个模拟系统来生成事件轨迹，由潜在因素控制其动力学，例如由外力控制的运动物体的轨迹，只允许观察源和目标，进而进行轨迹推理。

文中构建了一个不完整的N×N的网格，随机删除其中一小部分节点或边。每个节点和它的邻居最多有8种类型的有向边，例如东（E）和东北（NE）。任意选取其中一个节点作为源v0，然后绘制了一系列连续节点来构建轨迹（v0，v1，...，vT），vt为最终目标。除了源之外的轨迹上的每个节点都是从前一个节点V（x，y）的邻域中选择的，即根据由时间t和位置（x，y）决定的隐方向函数。



如果选择到不存在的边或达到最大步骤，轨迹就会终止。具体来说，生成了四种类型的轨迹：

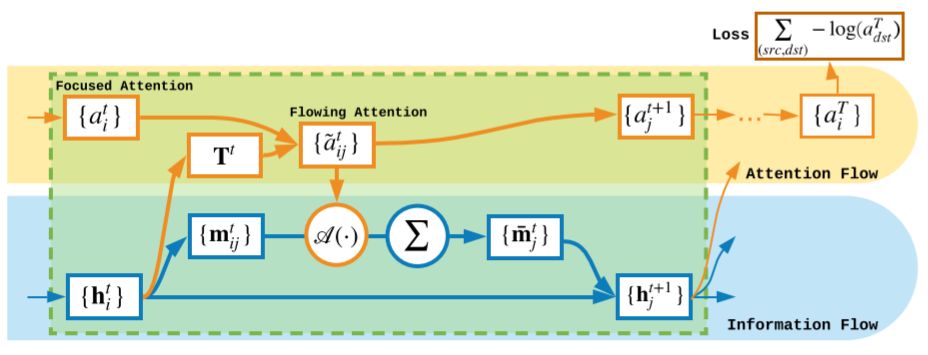
* 倾斜的直线
* 方向随时间变化的sin曲线
* 方向随当前位置变化
* 方向取决于历史位置信息

针对轨迹推理问题，本文使用了监督学习的方法。仅考虑可用的源和目标，训练判别模型以通过输入源节点来预测目标节点。

本文中利用了不完整的网格的图结构来解决问题，因为这个问题中包含的很强的推理，关系（连续节点的序列）和非关系（隐方向函数取决于时间，位置和历史信息）。

考虑到许多候选路径将源链接到目标，其实轨迹推理问题是很困难的，可以观察到的唯一线索是由移除的节点或边形成的块引向的目标节点。

在本文中，并没有简单地寻找最短路径，而是寻找了一些潜在动力学支配的真实轨迹模式。本文的评估标准基于预测的准确性和解释的可读性。



**图1 双流模型架构**

**3.引入注意力流模型**

**建模图上的注意力流**

在本文中，给定源预测目标的问题被视为在给定输入注意力分布的情况下预测节点上的输出注意力分布，**注意力分布表示参加节点的概率分布**。 对于每对（Vsrc，Vdst），输入注意力分布全部集中在源节点上，然后在对图上进行一系列计算之后，得到的输出注意力分布预测最可能的节点即是目标。 注意力从源传递到目标，意味着通过图上存在一个模仿潜在的因果链的流。

接下来文章尝试从三个不同的角度对图上的注意力流进行建模。

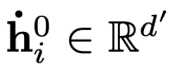
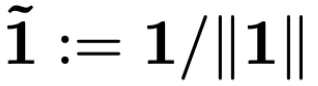
**1.图网络中的隐注意力流**

图网络主要采用由神经网络构建块来实现消息传递机制，例如MLP和GRU模块。 图网络中的表示包括节点级状态hi∈Rd，边级别消息mij∈Rd，以及图级别全局状态g∈Rd。

图网络框架有三个阶段：初始化阶段，传播阶段和输出阶段。

为了建模注意力流，首先要好定义初始状态

https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYWxhralKrFepryPNW4Y1ibscMXGtQsussU3FbfYWk7tvEKOKm2iaAiaaiag/640?wxfrom=5&wx_lazy=1&wx_co=1

其中是注意力状态，https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYJib0p0ottuMibCIWAquQCRcuBdaq1cicT9Q5ev8lJzNYSzCJt6Z2GxJ7g/640?wxfrom=5&wx_lazy=1&wx_co=1是辅助状态。研究者初始化源节点为1，其余节点为0。而作为用于计算注意力分布的参考向量。

传播阶段的模型包括：

* 信息函数 https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKY5uEqzLw7AFzE73YIgtjOlWZYibokmxZBopTTwZjpeLKaKh759Lq6NbQ/640?wxfrom=5&wx_lazy=1&wx_co=1
* 信息整合操作
* 节点更新函数
* 全局更新函数 https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYgCkX9KvrsPIB4duWPlpuFlqRL0FYSETRadC0eP1KHJquEic0LNVbtsw/640?wxfrom=5&wx_lazy=1&wx_co=1

在输出阶段，通过计算softmax作为输出的注意力分布。

这个模型将注意力流包装在开头的消息传递过程中，并在结束时将其取出。 基于神经网络的计算使得传播模型成为黑盒子，缺乏描述注意力流的明确方式，对解释目标无能为力。

**2.基于随机游走的显式注意力流（explicit attention flow）**

为了显式化地模拟注意力流，研究者使用随机游走（random walks）中可学习的转移矩阵T。注意力流动的动力学由https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYYejskERcuE4HlX1pOoUbFuwCFx3icWnoSdMlyNo9paibPwf5LC3favYg/640?wxfrom=5&wx_lazy=1&wx_co=1驱动，其中https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYFIrYvlPnsXYt4K5u3cgIDWy5mWeMLYpUpyibXCuTx34snqW3mskKsLA/640?wxfrom=5&wx_lazy=1&wx_co=1和https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYNmuzMhOOD0FYGj8w9LVicOiaLicYgjELXcyia0fNrvTZtN3vTxF8OZiaDKw/640?wxfrom=5&wx_lazy=1&wx_co=1代表两个连续的注意力分布。  在这里，研究者采用两种模型设置T：

* 稳定转换设置： https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYIianI9r44tnEYQRVGKXRz7yrWCR1M0lvwFK8KLQq2PicD3nAKfZHFSicA/640?wxfrom=5&wx_lazy=1&wx_co=1

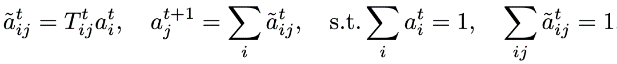
并对每行计算softmax ，转移矩阵T在输入和每一步之间是静止的。

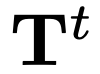
* 动态转换设置：https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYnBibl7icyhqfn9rDfPsWAUhpiax8G5ic7ibqsNKxslu2KCMgjUZgiaragY4w/640?wxfrom=5&wx_lazy=1&wx_co=1

**请注意，这里没有任何消息传递（message passing）**。 全局状态的更新，基于受影响的节点状态的加权和。 研究者强调注意力分布不仅可以作为内部状态的输出，还可以有效地影响内部状态。在没有利用消息传递的情况下，想以这种方式捕获图的前后关系（context）仍然受到限制。

**3.基于图网络的显式注意力流**

在享受图网络的表达性时，研究者也希望可以明确注意力流的可解释性。研究者通过引入节点级注意力https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYPxNKm6lQOO7T5aYUK90xsO8uKyNuF1ceWpVSezxXicQu2LAj92YXjgA/640?wxfrom=5&wx_lazy=1&wx_co=1（node-level attention，又被成为集中注意力 focused attention ）和连边级注意力https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYgqABpYIXpCEmypoZ5z10vibPdnf3d3DPibFZuBrd8ibPm94qNt1sw8s3A/640?wxfrom=5&wx_lazy=1&wx_co=1（ edge-level attention，又被称为流注意力 flowing attention ）来呈现注意力流动机制。 在上述的t时刻，动力学由以下因素驱动：



其中，转移矩阵被称为信息流（information flow），它依赖于在图网络中传递的消息所携带的丰富上下文。参见图1中的双流模型架构。

**很明显，信息流决定了注意力流，但研究者更关注注意力流如何影响信息流**。我们已经看到集中注意力中的节点级的向后传播行为主要关注节点状态之和。 这篇文章研究了在连边级别（edge level）上注意力流是如何影响信息流的。

首先通过定义**消息参与函数（message attending function）**https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYJYgAFZSMQ50ULjhL7s5lASStPZbZQFkveib6jpZyC0TpeHJdKDqjPNw/640?wxfrom=5&wx_lazy=1&wx_co=1，以替换原始消息https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKY18S8Dw6SdzTFdTuU6d3RrfJ2BicUpLibDpEKiaEaTv4j4VHO19ic1SicqaA/640?wxfrom=5&wx_lazy=1&wx_co=1。为了设计一个有意义的https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYM4mOzoRN1u48VGCc5bmuGC4hqEJcsLIa7YKIMGgIuDIIf1CL2xLFhg/640?wxfrom=5&wx_lazy=1&wx_co=1，当https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYmVAf4MrIn82LrB0Aw652UFgZiamias2Pib6MksSInE5HZgyNJpy8NicWfw/640?wxfrom=5&wx_lazy=1&wx_co=1时，使https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYveErN8Fa1gogXXtD1Rhjd3TjC2HVSAIB9j0NQqWGPQhbHxphKT19Kw/640?wxfrom=5&wx_lazy=1&wx_co=1独立于https://mmbiz.qpic.cn/mmbiz_png/wibWV1DB7tWK19yNEJGSG3tMzZVChianKYg9mWR6hPWgMpbPAuKEmibSwFROChJ6UPRMGicqcdXpCE9bYzJfo27wAA/640?wxfrom=5&wx_lazy=1&wx_co=1，也就意味着当没有注意力在这条消息上时，消息参与函数不一定是0。研究者找到了消息在大多数情况下，参与函数何时表现最佳，这不仅揭示了向后作用的重要性，而且还揭示了“即使没有参与，也要保持信息流动”的必要性。

**与强化学习和概率隐模型的联系**

如果注入噪声，然后在每一步中选择表现最好的节点，那么该过程就会以某种类似于强化学习的方式出现。如果在Gumbel-Softmax或Concrete分布中应用噪声并且保持噪声的柔和，它就会变成概率隐模型（probabilistic latent model）。注意流可以被视为，在概率空间中而非离散的样本空间，直接地用数字操作完成图级计算（graph-level computation）。

**4.实验及结果分析**

**数据集生成和统计**

文中生成了许多数据集组，每个数据集表示一个带有多种不同轨迹的随机网格世界，轨迹由特定的节点序列组成，由特定的潜在动力学驱动。 更具体地，对于每个数据集组，构建了不完整的N×N网格世界，一般为32×32或64×64，然后从每个节点开始根据潜在的动力学函数绘制出了多个轨迹。

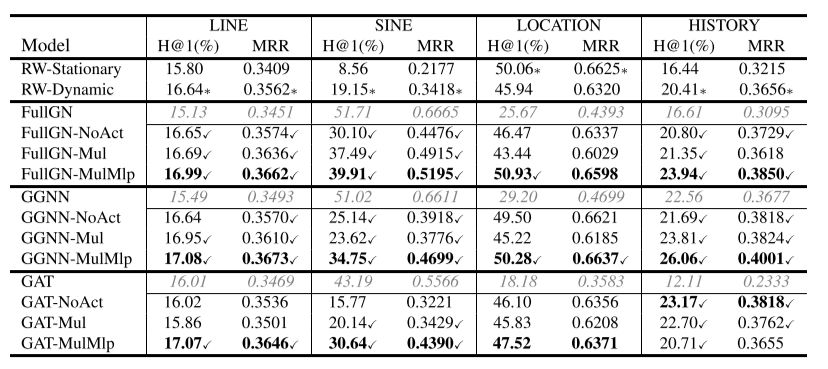
一种生成了24个数据集组。其中，每个数据集的可观察部分包括包含所有边信息的网格图以及用于训练、验证和测试的源-目标对。通过在源节点上进行8：1：1分割来制作训练、验证和测试集，以便可以根据处理不可见源节点对的性能来评估模型。

**比较模型结果**

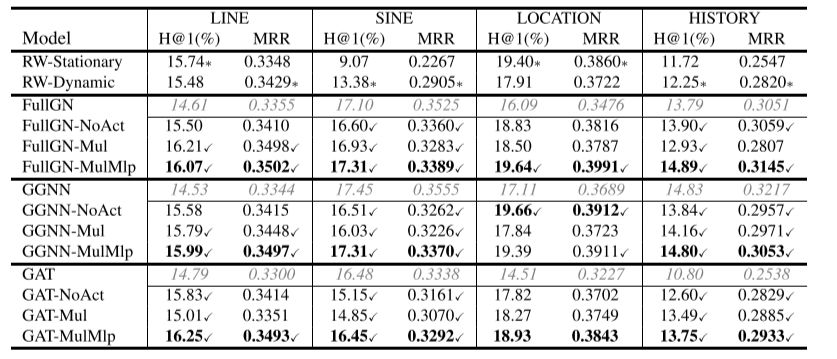
为了全面评估注意力流动机制，研究者选择三种类型的图形网络和两种基于随机游走的模型来设置基准。 图形网络包括完整的图网络（FullGN），门控图神经网络（GGNN）和图注意网络（GAT）。

请注意，这些图网络的常规版本无法明确地模拟注意力流并实现轨迹推理目标，但能够进行预测。 文中分别用注意力流机制对它们进行了改造，这种机制以三种不同的方式实现，关注流动的注意力如何作用于消息传递。

**表1**



**表2**



**首先**，研究者比较了基于随机游走和基于图网络的显式模拟注意力流的模型。

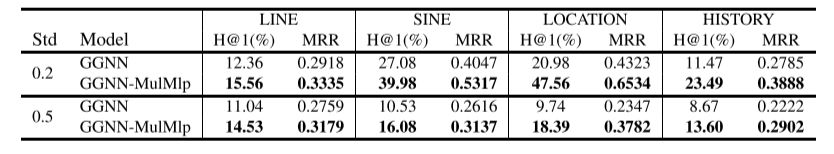
从表1和表2可以看出，在大多数情况下，基于图网络的模型超过了基于随机游走的模型，通常是大幅度的。虽然有一些例外，RW-Stationary在依赖于位置的情况下表现强劲，可能是由于除当前位置信息之外所需的上下文很少，但最好的**基于图网络的模型，例如FullGN-MulMlp**，仍然可以击败它。

**其次**，研究者比较了无动作（No acting），乘法动作（Multiplying acting）和非线性动作（non- linearly acting）之间的向后作用机制。

非线性动作在几乎所有情况下都表现最佳。令人惊讶的是，简单地进行乘法运算可能会降低性能，使其比不采取行动更糟糕。如何设计有效的向后作用机制在今后的工作中值得进一步研究。

**最后**，研究者将改造后的图形网络与针对常规图形网络的显式注意力流进行比较。

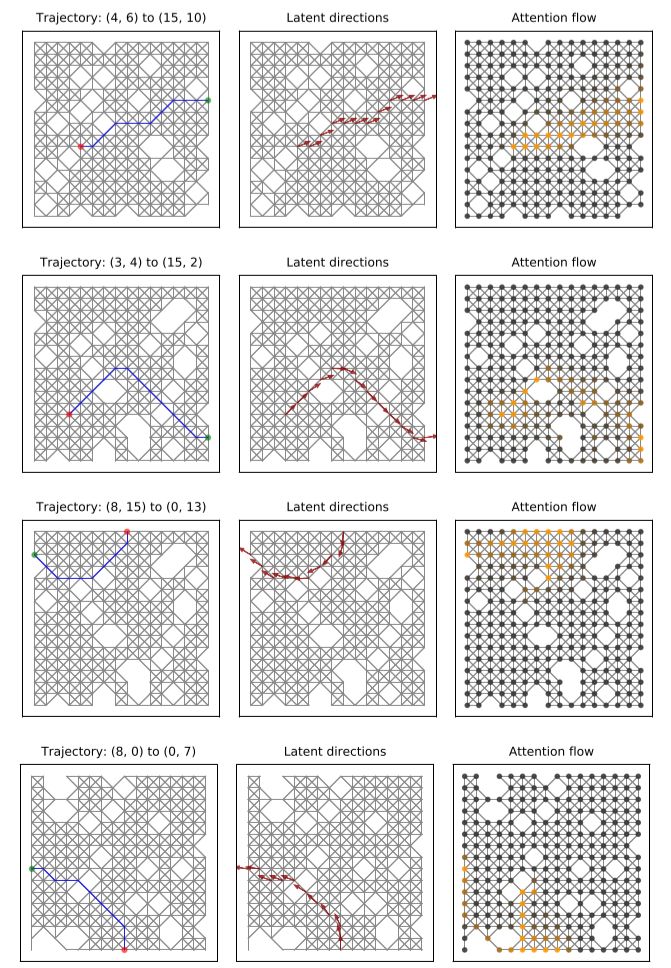
对于较大的64×64数据集，研究者测试GGNN和GGNN-MulMlp，**并发现GGNN-MulMlp(一种图网络模型)在每个评估指标上的表现都明显优于GGNN**，如表3所示。



**5.结果可视化**

研究者可视化了16×16的不完整的网格图中的学习注意力流，通过对每个方向设置采用一个示例与真实轨迹和潜在方向进行比较，如图所示。

乍一看，在T步骤上绘制的注意力流使得从源节点到目标节点的链带，几乎与真实轨迹相匹配，特别是如图2中的第一行和第二行所示。



**图2**

仔细观察，我们发现注意力流可能不一定遵循单一路径模式，而是会改为分支，以扩大探索区域（更可能包含目标节点），尤其是在间隙附近，如最后两行所示。

**6.小结**

在本文中，作者提出了基于图网络的“注意力流机制”，设计了包含“源和目标”的推理任务，通过实验，发现注意力流机制的预测准确性比随机游走算法更好，并且对推理轨迹有良好解释。

真实世界往往是因果关系被隐藏的复杂系统，需要的不仅仅是对结果的预测，更是有理有据的推理过程。基于图网络的注意力流机制，在这方面表现不俗。

作者：辛茹月

编辑：王怡蔺