1. Causal Induction from Visual Observations for Goal Directed Tasks

标题:目标定向任务的视觉观察因果归纳

作者：Suraj Nair、Yuke Zhu、Silvio Savarese、Li Fei-Fei

论文地址：https://arxiv.org/abs/1910.01751

在人类和其他聪明的动物与物质世界进行交互的过程中，因果推理能力一直不可或缺。在本篇论文中，李飞飞、朱玉可等研究者提出赋予人工智能体因果推理能力，以完成特定的任务。他们利用基于学习的方法来生成有向无环图格式的因果知识，这些知识可以用于将学习到的目标条件策略（goal-conditional strategy）置于语境中，从而在具有潜在因果结构的新环境中执行任务。研究者在他们的因果判断模型和目标条件策略中利用到了注意力机制，从而可以逐渐从智能体的视觉观察（visual observation）中生成越来越多的因果图，并在行动决策中有选择性地使用这些生成的因果图。实验结果表明，这种基于学习的方法可以有效地在具有未知因果结构的新环境中完成新任务。

2. 介绍

因果推理是自然智能不可分割的一部分。

作为一种生存手段，人类和其他智能动物都具有对因果关系进行推理的能力

这种能力在幼儿与物质世界的互动中起着至关重要的作用。

正如行为心理学研究所表明的，幼儿从与世界的游戏中发现潜在的因果机制，

而他们对因果关系的了解反过来又有助于他们随后对物体、概念、语言和物理的学习。

如今，人工智能中以数据为中心的方法，如深度网络，已经在从大量数据的输入和输出中学习关系的方面（如大样本图像分类标签任务）取得了巨大成功。

然而，经验证据表明，这些方法中缺乏正确的因果模型对泛化造成了重大威胁，导致图像字幕模型生成不现实的字幕，在新的问题实例中深度强化学习策略失败，并使学习模型适应新的分布变慢。

Empirical evidence indicates that the absence of correct causal modeling in deep networks has posed a major threat to generalization, causing image captioning models to generate unrealistic captions, deep reinforcement learning policies to fail in novel problem instances, and transfer learning models to adapt slower to new distributions.

图像标注–看图说话（Image Captioning） 图像标注是一项引人注目的研究领域，它的研究目的是给出一张图片，然后用一段文字描述它。

在这项工作中，我们提出赋予一个基于学习的交互网络因果推理能力，使其在视觉环境中完成目标导向任务。

想象一下一个家用机器人第一次进入一户新人家。

在事先不知道家庭灯路布线配置的情况下，它需要通过尝试各个开关来整理出各盏灯和各开关之间的对应关系，使得收到命令打开厨房灯或浴室灯的时候能够准确控制。

我们将机器人尝试切换开关的第一阶段称为因果归纳，在这一阶段，智能体通过执行动作并观察动作结果来发现潜在的因果关系。

我们将打开特定灯光的第二阶段称为因果推理，智能体使用学习获得的因果关系来指导其完成特定任务的行动。

为了建立一个有效的因果归纳和因果推理计算模型，我们必须在测试时解决智能网络对新的因果关系和新的任务目标的泛化问题，这两个任务在训练过程中是无法预测的。

我们可以把这看作是继Dasgupta等人之后的两阶段的元学习问题。

在第一阶段中，我们使用因果归纳模型，从智能体干预环境的反馈观察结果中构建因果框架——即随机变量的有向无环图。

在第二阶段，我们使用学习得到的因果架构将目标条件策略文本化，以执行给定目标的任务。

但是，与Dasgupta等人的研究相反。

我们显式地构造了因果框架，而不是利用潜在特征进行编码表征，这使得在长期任务中，相对于简单的一步查询，对新问题实例有了更好的泛化能力。

为此，我们提出两项技术：

1）注意力机制的迭代因果归纳模型，该模型学会为环境中的每一个观察到的相互作用递增地更新预测到的因果图

2）基于注意力机制的图编码的目标条件策略，迫使它在每一步都关注得到的因果图上的相关部分

我们发现通过使用因果图分解归纳和推理过程，它可以很好地泛化到低于50个训练因果结构的未知因果结构。

To this end, we propose two technical contributions: 1) an iterative causal induction model with attention, which learns to incrementally update the predicted causal graph for each observed interaction in the environment, and 2) a goal-conditioned policy with an attention-based graph encoding, forcing it to focus on the relevant components of the causal graph at each step.

我们将我们的方法同使用了ground-truth的因果架构(which provides oracle performance)的方法（直接预测因果结构的非迭代框架）进行了比较，并将观察到的数据编码进策略的LSTM块内存中，与之前的工作类似。

实验表明我们的方法在恢复因果图和完成目标条件任务的成功率方面都优于基线，并且在多个任务大小、类型和训练因果结构的数量方面都达到接近oracle的性能。

3. 问题陈述

在一个由六元组(S, A, p, G, r,γ)定义的基于目标条件的马尔可夫决策过程（MDP）中，我们描述了智能体与环境的交互作用, 其中S是状态空间，A是行为空间，p是状态转移概率，G是目标空间，r是回报函数，γ是折现因子。

我们的目标是习得一个目标条件政策, 它最大化回报的期望和：

马尔可夫决策过程的特点是目前状态s向下一个状态s′转移的概率和回报仅仅和当前状态的s和选择的a有关，和历史无关，因此MDP的转移P和立即回报r也只取决与当前状态和选择的动作，与历史状态和动作无关，公式表达：



π(s, a)表示在给定状态s下选择动作a的概率 ，且π不随时间变化。

在这项工作中，我们不仅希望G能很好地泛化到G中的各个目标，而且考虑一个更宏远的目标：使G能在一组MDPs中进行泛化推广。

我们认为M = {M(1), M(2), ... , M(K)}是具有相同状态空间S和行为空间A但不同的状态转移概率p的K个MDPs的整体集合，其中M(k)由(S, A, p(k), G, r,γ)定义。

状态转移概率p(k)表示状态和行为之间的潜在因果关系。

在同一状态s下采取相同的行动a，在不同的状态转移概率p(k)下可能导致不同的下一个状态s’。

我们期望我们的智能体以其第一人称的眼光行事，而无法接触到潜在的因果关系。

它接受高维RGB观测数据，并且自主从观测数据中归纳出一个因果模型。

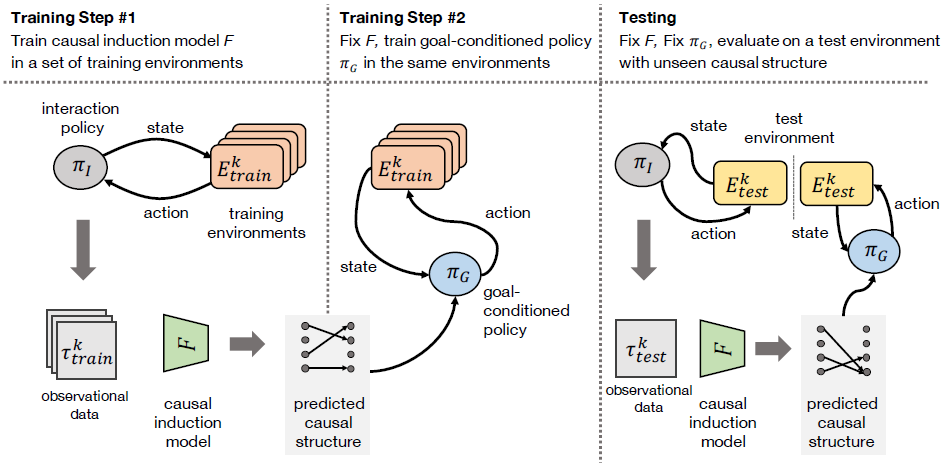


图1. 因果归纳和推理流程概述。

在训练过程中，每一次都从K个训练环境Mtrain中抽取一个作为样本，使用交互策略πI于环境交互并收集视觉观察到的迹象τ。利用监督学习训练因果归纳模型F，它以观测数据的迹象τ作为输入，并构造潜在的因果模型，即的估计，它捕获了潜在的因果结构。

然后，将预测结构作为目标g条件下的策略πG的输入，πG学习如何利用因果模型在训练环境中有效地完成指定的目标。

在测试时，F和πG是固定的，并且在具有不可见因果结构的新环境中对该智能体进行评估。

如图1所示，整个过程分为两个阶段：

1) 我们执行一个交互策略πI来收集转换序列τ={(s1，a1), (s2，a2),…}，将它输入一个归纳模型F可以构造出潜在的因果模型

2) 我们使用因果模型将目标条件策略πG文本化，使其能够在新的因果关系的新MDP中执行任务。

我们称它为元学习问题。

我们将所有MDPs M的集合划分为两个不相交集Mtrain和Mtest。

在训练过程中，我们利用Mtrain学习了我们的归纳模型F和目标条件策略πG。

在测试过程中，我们从Mtest中评估归纳模型F是否能够从一个新的MDP中,通过执行交互策略 πI获得的观测数据中学习，并以此构建一个因果模型，πG可以通过它来执行这个新的MDP中的各种任务。

原始像素空间中因果关系的直接建模因其维数大而难以实现。

根据Chalupka等人的研究成果，我们可以假设，我们问题中的因果关系可以定义在几个因果宏观变量上。

例如，厨房的亮度（智能体的视觉观察）由厨房灯的开与关（宏变量）决定，这是由控制灯的开关按钮状态（另一个宏变量）决定的。

这一假设使我们能够构造一个有向非循环因果模型来表示，这些宏观变量上的各种行为的因果影响。

给定一组宏观变量，归纳模型F根据视觉观察到的信息预测各变量之间的有向边。

这里的一个主要挑战是，在智能体的视觉感知中，由于部分观测和虚假相关而产生的混淆。

例如，厨房的光照度变化可能是由于打开/关闭厨房灯或客厅灯所致。

因此，它要求智能体能从视觉输入中分离出正确的因果关系。

4. 方法概述

我们的方法的目标是使一个策略能够在不可见因果结构的环境中完成基于目标条件视觉的控制任务，只给定环境中观测数据的一个短轨迹。先前Dasgupta等人使用基于LSTM的策略（将交互信息编码到策略的记忆中）在解决简单的一步查询问题上已经取得了较好的结果。我们的假设是，要推广到复杂的多步控制问题，需要一个更结构化的归纳和策略方案。

Prior work (Dasgupta et al., 2019) has shown promising results on simplistic one-step querying problems using an LSTM-based policy which encodes the interaction into the policy’s memory. Our hypothesis is that to generalize in complex multi-step control problems, a more structured induction and policy scheme will be required.

为了解决这个问题，我们分别在归纳模型和策略中提出了迭代更新和注意瓶颈。我们证明了这两种方法显著地改进了对不可见因果结构的泛化性能。

5. 迭代因果感应(因果归纳)网络

从原始的感官观察中归纳出因果结构，需要我们准确地捕捉每个动作对环境的独特影响，并同时需要考虑其他动作的混杂影响(比如大厅的灯的开关对厨房亮度的影响)。

我们假设泛化能力最强的因果推断网络将能够分离开各个动作及其相关影响，并且每次只更新因果图中的相关部分。

我们在我们的迭代模型中实现了这一想法，在迭代模型中，我们从对因果结构的初始猜测开始，该结构的所有边权重都为0（意味着我们假设没有因果关系）。

然后，我们使用一个观测编码器将观测数据的每个图像映射到一个编码s，并计算后续步骤之间的状态残差R。

捕获状态变化的R然后与相应的动作a连接，然后反馈进入边解码器模块。

边解码器模块的输出是对因果图的边强度的更新。

此更新应用于每个观察的状态转移，即，并在最后一层对整个图进行编码，以便在预测因果图之前进行最后的边更新(见图2)。

边解码器要么输入已编码的R和a，要么输入已编码的的边矩阵，然后输出一个1×N 的soft attention向量α和1×N边权变化e，其中N是环境中交互的动作总数。

attention向量α用于加权因果图中的指定节点的边更新e。

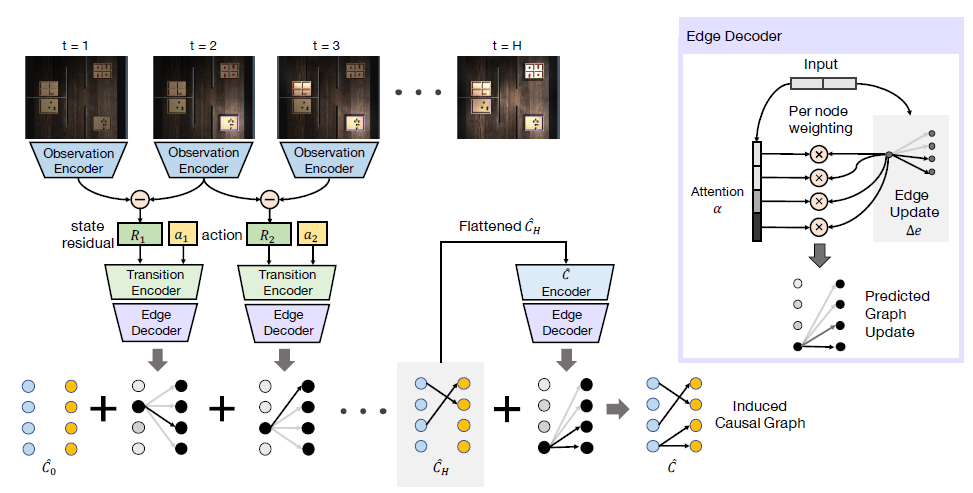


图2. 迭代因果感应网络

我们的迭代网络结构用于从视界为H的观测数据的可视轨迹中归纳出因果结构。

首先将每一帧编码成一个潜在的状态嵌入s。

然后计算跨时间步(t → t+1)的状态嵌入之间的差异（状态残差），并将其与相应的动作连接起来。

这被输入到边解码器模块，该模块预测哪些边更新，并计算一个attention向量，它将说明加权边更新是如何应用于节点的。

在最后一步中，在当前图上再进行一次边更新，并输出最终预测图。

因此，在每次迭代中，更新相当于：



其中φ是转移编码器，一个全连接网络模块（详见附录A）。

使用这种注意机制进一步鼓励模块进行独立的更新，我们观察到这有助于更好的泛化。

6. 学习目标制约政策

该策略的目标是：在给定初始图像s0,目标图像g和预测因果结构的情况下，在T步内达成目标。

此外，策略πG(s, g,)是一个反应性策略，因此它可以学会使用预测因果结构解决目标条件任务。

也就是说，由于策略没有记忆，它无法在推理期间学习归纳内部图，因此必须使用因果图。

我们假设，与因果归纳模型一样，泛化性能最强的策略会学习专门关注因果图中与任务当前步骤相关的边。

为此，我们提出了一个图编码中的注意力瓶颈，它鼓励策略在每一步中选择与一个“效果”相关的边，从而实现更好的泛化。

We hypothesize that like the causal induction model, the policy which best generalizes is one which learns to focus exclusively on the edges in the causal graph which are relevant to the current step of the task. To that end we propose an attention bottleneck in the graph encoding, which encourages the policy to select edges pertaining to one “effect” at each step, which enables better generalization.

具体地说，该策略对当前图像s和目标图像g进行编码。

基于该编码，它在因果图中的“效果”上输出大小为1×N的注意向量α。

该向量用于对N×N因果图的输出（N个原因、N个效果以及它们之间的边）执行加权和，从而产生所选边e的大小为N向量。

所选边和视觉编码用于输出最终动作：



其中，E与图像编码器F具有相同的架构，但它同时编码当前图像和目标图像，并且φi都是全连接层（见图3）。

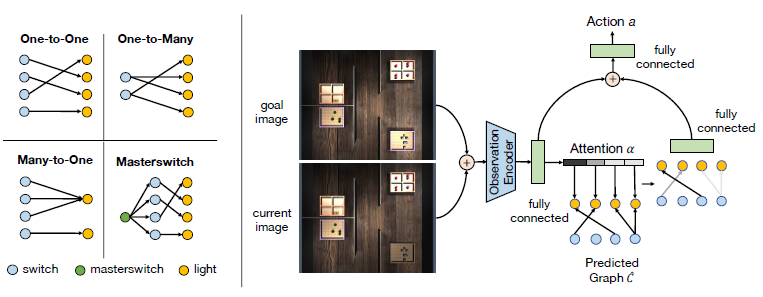


图3. 因果结构类型（左图）

我们探索四种因果结构，一对一，一对多，多对一和主开关。

这些包括标准因果图、共同原因因果模式、共同影响因果模式和因果链。

目标制约政策（右图）。

该策略以当前图像、目标图像和预测的因果图作为输入，将当前图像和目标图像按信道连接并编码。此编码用于预测中“效果”上的注意向量，该“效果”提取相关边，然后与图像编码连接以预测动作。

7. 模型训练

在有限的训练环境中，利用有监督学习训练归纳网络F，特别是最小化ground-truth因果图C和预测因果图之间的“l2重建损失”

策略πG是在训练环境中，利用ground-truth因果图模拟计划者，用DAgger算法训练得到的。

然后在只有视觉输入和目标图像的不可见环境中测试πG。

具体来说，在训练环境中，计划者使用ground-truth图和特权低维状态/目标信息来计算目标的最优计划。

在每个时间步，专家的行为被添加到策略的记忆中，然后使用标准交叉熵损失对策略进行训练，以模拟给定当前图像和目标图像的专家。

该策略在训练期间也会注入-greedy噪声，且=0.3。

网络架构和其他培训细节见附录。

8. 实验

通过我们的实验，我们研究了三个互补的问题：

1）我们的迭代归纳网络是否能够实现更好的因果图归纳？

2）我们在πG图编码中的注意瓶颈是否能够使策略更好地泛化到不可见的因果结构？

3）通过结合我们提出的F和πG，我们是否能够超过当前最先进的Dasgupta等人关于视觉目标导向的任务？

9. 实验设置

**任务定义：**

我们研究了灯开关控制的多步骤任务。

特别是，一个智能体可以控制N个开关（宏变量），它们控制N盏灯光（宏变量）的模式有一些潜在的因果结构。

然而，光的宏观变量表现在有噪声的视觉观测中，其部分可观测性和重迭性导致了混淆因素(比如厨房的光照受大厅灯的影响)。

智能体的目标是从初始状态s0开始，控制开关达到一个指定的照明目标g，其中状态图像和目标图像都是32×32×3图像。

然而，正如在问题设置中所指定的，潜在的因果结构对智能体来说是未知的(即刚来到一个新房子，对一切都清楚)，所提供的只是来自环境的有限观测数据。

因此，智能体必须（1）从观测数据中归纳出开关和灯之间的对应因果结构，然后（2）利用它来达到特定任务(如打开厨房的灯这样一条指令)。

我们探讨了开关和灯之间4种不同的因果模式（见图3）。

第一种因果结构是一对一，其中每个开关对应一盏灯。

第二种因果结构是多对一（公共效应），其中每个开关控制一个灯，但多个开关可以控制同一个灯。

第三种因果结构是一对多（常见原因），即一个灯由一个开关控制，但一个开关可以控制多个灯。

最后，因果结果是主开关域，其中存在因果链，其中只有一个主开关被激活，其他开关因果效应才能被观察到，应用于一对一因果结构之上。

在我们的方法中，我们将因果结构C表示为在环境中的N个开关状态和N个灯状态之间具有有向边的图，其中边强度与开关控制指定灯的可能性相对应，在主开关设置中具有附加的N个边。

**环境设置：**

在H时间步骤轨迹中的每一时间步骤中的一个32 × 32 × 3大小的图像和一个大小为A=N的行为向量组成的观测数据的轨迹，在此视觉空间中，我们可以推断出因果结构(开关与灯的对应关系图)。

视觉场景由安装在有3个房间的MuJoCo（Todorov et al.，2012）模拟房屋中的N个灯组成，灯安装在房间和走廊中，每个灯的效果在打开时照射到地板上，由自上而下的摄像机进行拍摄。

多盏灯的重叠和漫射会引入混淆因子，模型必须将这些因素分离，才能正确预测因果结构。

策略的状态空间（和目标空间）也是同一环境和摄像机的32×32×3图像。

动作空间A有N个动作，每个开关对应一个离散动作。

在策略学习过程中，goals是在环境因果结构（即环境中所有可到达状态的集合，即各种灯各种开关组合的可能结果）下，从可能的照明配置空间中均匀采样的。

**收集观测数据：**

使用启发式交互智能体πI收集观测数据，它执行一个简单的策略。

在除了主开关设置以外的所有情况下，策略每个操作只执行一次（horizon H=N）。

在主开关设置中，探索性智能体按下每个开关，直到其中一个对环境产生影响，然后继续按下其他每个开关（horizon最高可达H=2N-1）。

10. 评价方法

我们评估以下方法和基线，以检验因果归纳模型和目标条件政策的有效性。

首先检查我们的迭代因果归纳网络（iterative causal induction network/ICIN）在归纳因果图方面的性能有多大改进，我们将其与使用时间卷积归纳网络（temporal convolutions induction network /TCIN）的非迭代归纳模型以及使用无注意机制的迭代模型（ICIN（No Attn））的方法进行了比较。我们比较了这些方法，在恢复重建ground-truth因果图的F1分数。

接下来，我们将使用所有变体（ICIN、ICIN（No Attn）、TCIN）来比较目标条件策略的性能，并与Dasgupta等人以前的工作进行比较，他们使用策略的潜在内存（记忆）生成因果图。

具体来说，我们为策略πG提供LSTM记忆内存，在运行DAgger之前，通过策略一步一步地提供交互轨迹。

这种端到端的方法允许将因果图编码到策略的潜在记忆内存中，以便在执行目标导向任务时使用。

我们还比较了记忆内存（RL/Low Dim），Dasgupta等人的一个版本, 它可以访问特权状态信息（ground-truth状态），并使用无模型强化学习进行训练，获得了丰厚回报。

在此设置中，相同的视觉交互轨迹被编码到策略的LSTM记忆内存中，但实际策略接收状态和目标的二进制向量，并使用PPO算法进行训练。

我们还添加了一个比较，将ground-truth因果图（甚至在测试期）作为Oracle（Oracle），它提供了性能的上限。

所有方法都是在看不见的环境中基于成功率进行比较的。

实施细节见附录B。

最后，为了了解目标条件策略中的注意力瓶颈对于泛化性能的重要程度，我们使用由ICIN归纳的因果图比较了目标条件策略中不可见因果结构的成功率，有和没有注意瓶颈。

11. 因果归纳评价

首先，我们检验了我们的方法从观测数据的轨迹中归纳出因果模型的能力。

我们报告了预测因果图和groun-truth因果图边缘之间的F1得分（阈值=0.5）。

我们计算了在5和7个开关任务中，在10、50、100或500个可见因果结构的情况下，多达100个不可见任务的F1分数（见图4）。

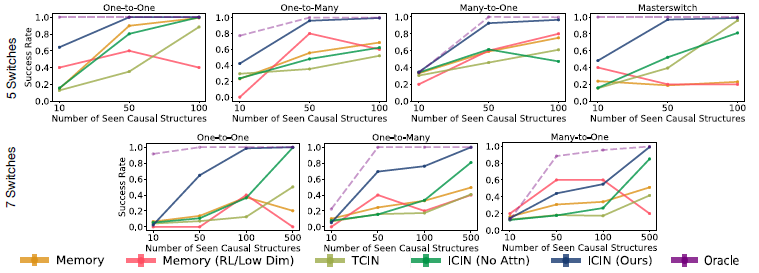


图5. 政策成功率（看不见的因果结构）

每种方法对10、50、100或500个未知因果结构的目标条件策略的最终成功率，对于5个或7个开关的可见因果结构。我们的迭代方法可以在几乎所有的设置中实现对未显示任务的最佳泛化能力。

我们观察到，在几乎所有设置中，我们的注意力迭代方法（ICIN）占主导地位。

虽然我们没有注意的方法通常优于非迭代基线，但两者都明显落后于我们的最终方法，这表明注意提供的模块性在实现泛化方面发挥了很大的作用。

此外，我们还观察到，当训练因果结构较少时，我们的迭代注意方法尤其优于其他方法，这表明通过强迫网络进行独立更新，它能够在有限的训练示例中学习归纳的一般方法。

这可能是因为注意力迫使模型学习在给定单个观察的情况下更新单个边，而对总图不可知，总图不太可能过度适合训练结构。

ICIN的定性例子见附录c。

12. 目标条件政策评估

我们在图5中随机目标的不可见因果结构的500个试验中检验了收敛策略πG的成功率。

在大多数情况下，我们观察到Dasgupta等人基于内存的方法提供了一个强大的基线，优于TCIN和ICIN（No Attn）基线。

我们怀疑它学会了最好地模仿专家，它必须将相关信息从交互轨迹编码到其潜在记忆中，隐式地执行归纳。

虽然这种方法有效，但在不可见结构上它的泛化性能比我们提出的方法差得多，这可能是由于我们的方法的组成结构。

记忆baseline使用低维状态(Memory (RL/Low Dim)),并通过无模型RL进行训练的内存基线也表现良好，事实上在7个开关、多对一设置中的少数情况下优于我们的方法。

然而，总的来说，这种方法的性能比我们的低得多，而且它成功的情况很可能是由于它使用了特权信息（ground-truth状态），而不是视觉观察。

几乎在所有情况下，ICIN都显著优于所有基线。

实际上，在5开关的情况下，我们的ICIN方法几乎与Oracle匹配，这表明它几乎完美地归纳了C。

最后，我们研究了我们提出的注意瓶颈在πG图编码中的重要性，它迫使策略在每个时间步只关注相关的边。

在给定当前图像、目标图像和ICIN预测的因果图的情况下，我们检验了使用πG和注意瓶颈的成功率。

我们发现，使用πG的图编码器中的注意瓶颈，在一对一（1:1）和主开关（MS）的情况下，成功率大约增加10%，在一对多（1:K）和多对一（K:1）的情况下，成功率大约增加40%。

这是因为，通过鼓励策略选择相关的边，它产生了一个模块化策略，

该策略可以1）识别它希望在环境中进行的更改，2）基于因果图预测必要的操作，从而实现更好的泛化。

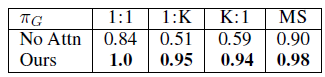


表1：政策关注瓶颈

在5个开关和50个可见结构上进行测试。

我们观察到，在所有设置中，使用策略的图形编码器中的注意瓶颈可以显著提高性能。

13. 总结：

我们提出了两项新的技术：

1）从原始视觉观察进行因果归纳

2）目标条件策略的因果图编码

这两种方法都能更好地泛化到不可见的因果结构。

我们的关键见解是，通过利用迭代预测和注意瓶颈，它有助于我们的因果归纳模型和目标条件策略关注因果图的相关部分。

使用这种方法，我们显示出比以前的有限训练因果结构的工作有更好的对新问题实例的泛化性能。

在这项工作中，我们从启发式策略收集的观测数据中归纳出因果结构。

我们计划探索更复杂的任务，其中探索环境以发现因果结构需要更复杂的策略，并开发算法来共同学习交互策略。