视觉语言导航研究进展

司马双霖[1, 2] 黄 岩[1, 3] 何科技[1] 安 东[1] 袁 辉[1] 王 亮[1, 2, 3, 4, 5]

摘 要 视觉语言导航, 即在一个未知环境中, 智能体从一个起始位置出发, 结合指令和周围视觉环境进行分析, 并动态响应生成一系列动作, 最终导航到目标位置. 视觉语言导航有着广泛的应用前景, 该任务近年来在多模态研究领域受到了广泛关注. 不同于视觉问答和图像描述生成等传统多模态任务, 视觉语言导航在多模态融合和推理方面, 更具有挑战性. 然而由于传统模仿学习的缺陷和数据稀缺的现象, 模型面临着泛化能力不足的问题. 系统地回顾了视觉语言导航的研究进展, 首先对于视觉语言导航的数据集和基础模型进行简要介绍; 然后全面地介绍视觉语言导航任务中的代表性模型方法, 包括数据增强、搜索策略、训练方法和动作空间四个方面; 最后根据不同数据集下的实验, 分析比较模型的优势和不足, 并对未来可能的研究方向进行了展望.

关键词 视觉语言导航, 视觉语言理解, 跨模态匹配, 具身智能

引用格式 司马双霖, 黄岩, 何科技, 安东, 袁辉, 王亮. 视觉语言导航研究进展. 自动化学报, 2023, **49**(1): 1−14

DOI 10.16383/j.aas.c210352

Recent Advances in Vision-and-language Navigation

SIMA Shuang-Lin[1, 2] HUANG Yan[1, 3] HE Ke-Ji[1] AN Dong[1] YUAN Hui[1] WANG Liang[1, 2, 3, 4, 5]

Abstract Vision-and-language navigation means that an agent in an unknown environment, starting from a starting location, dynamically generates a series of actions by making analysis with language instructions and the visual environment, and finally navigates to the goal location. And due to the widespread application prospect, in recent years, it has received increasing attention from researchers especially in multi-modal research. It is different from traditional multi-modal tasks such as vision question answer and image captioning, vision-and-language navigation is more challenging in terms of dynamic reasoning and multi-modal fusion. However, with the limitations of imitation learning and the phenomenon of data scarcity, the model is faced with the problem of insufficient generalization. In this paper, we review the current advances in the research of vision-and-language navigation. Firstly, we briefly introduce data sets in visual-and-language navigation. Then, we comprehensively introduce the representative models in vision-and-language navigation, including data augmentation, search strategies, training methods and action spaces. Finally, from the experiments under different data sets, we analyze the advantages and disadvantages of the existing models, and prospect some future and possible research directions.

Key words Vision-and-language navigation, vision-and-language comprehension, cross-modal matching, embodied artificial intelligence

Citation Sima Shuang-Lin, Huang Yan, He Ke-Ji, An Dong, Yuan Hui, Wang Liang. Recent advances in vision-and-language navigation. *Acta Automatica Sinica*, 2023, **49**(1): 1−14

近年来, 越来越多研究人员意识到单模态分析技术在现实中处理信息的局限性, 对于自然语言、音频信息以及视觉等多模态融合方面的研究投入日益增加. 视觉语言导航[1] 是智能体在第一视角下, 基于真实环境下的全景图, 综合处理指令和视觉信息并进行推理的多模态任务, 也是智能管家等应用的核心技术之一. 视觉语言导航尝试使用多模态融合的方式, 为室内导航任务的研究提供了一个新的方向. 如图 1 所示, 智能体需要结合指令信息和视觉信息, 在模拟器中完成一系列的决策, 最终到达目标位置. 其中主要难点在于如何学习理解指令和视觉的信息, 从而完成导航过程中的每一步决策.

Anderson 等[1] 于 2018 年首先提出视觉语言导航任务, 并公开了与任务相对应的基于真实环境的

收稿日期 2021-04-22 录用日期 2022-06-16
Manuscript received April 22, 2021; accepted June 16, 2022
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 中国科学院自动化研究所智能感知与计算研究中心 北京 100190 2. 中国科学院大学人工智能学院 北京 100049 3. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190 4. 中国科学院自动化研究所脑科学与智能技术卓越创新中心 上海 200031 5. 中科人工智能创新技术研究院 胶州 266300
1. Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049 3. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 4. Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Shanghai 200031 5. Artificial Intelligence Research, Chinese Academy of Sciences, Jiaozhou 266300

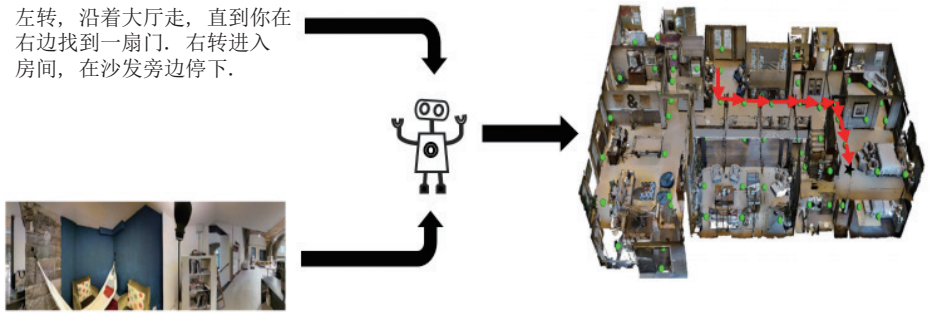


图 1 视觉语言导航过程示意图

Fig. 1 The process of vision-and-language navigation

Room-to-Room (R2R) 数据集, 并在 Matterport-3D[2] 模拟器完成了导航任务的仿真. 视觉语言导航一经提出便引起了广泛的关注. 随着研究的不断深入, 研究人员提出了很多拓展任务, 如室外视觉语言导航 Touchdown[3]、结合导航和指称表达 (Remote embodied visual referring expression in real indoor environments, REVERIE)[4−5] 以及视觉对话导航任务[6]. 除此之外, 研究人员也发现了一些视觉语言导航亟需解决的问题, 如数据量级还远不能满足实际需求, 数据稀缺问题愈发突出、以及模型的泛化能力低. 这些问题一直阻碍着视觉语言导航的发展. 另外现有的方法全是基于模拟器的仿真环境, 该设定与现实场景下的导航仍存在很大差异. 如何将现有的模型应用到实际环境中, 也是视觉语言导航的一大难题.

为了解决以上问题, 一系列的改进模型方法被相继提出. 研究人员在指令集扩充、学习策略升级和多模态融合等方面进行较多探索并取得了巨大进展. 本文首先介绍常用的数据集, 然后按照不同的模型改进方式对现有的方法进行分类介绍, 并结合不同数据集的实验结果分析不同模型的优势和不足, 全面介绍目前视觉语言导航的研究现状.

1 视觉语言导航数据集

视觉语言导航的数据集, 主要是针对 MatterPort3D 模拟器的 90 个场景建立的. 为了推动视觉语言导航任务的发展, 研究人员从指令描述粒度、指令长度以及语言种类入手, 收集了大量的人工指令. 这一定程度上扩大了数据量, 对视觉语言导航的发展, 起着非常重要的作用. 本节将按照不同的指令粒度和指令长度的数据集分类介绍.

1.1 R2R 数据集

R2R 数据集是由 Anderson 等[1] 构建, 其中总词汇量约 3 100 个单词, 构成 7 189 条路径下的 21 567 条人工标注的指令, 且每条指令的平均长度为 29 个单词. R2R 数据集在很大程度上覆盖了视觉环境中的大部分细节信息, 具有多样性的特点. 在视觉语言导航中, R2R 数据集被分成训练集、可见环境的验证集、不可见环境的验证集和测试集. 其中训练集和可见环境的验证集共用 61 个真实场景, 但是把相应场景下的数据集分为了两个部分: 用于训练的 14025 条指令和用于可见环境验证的 1 020 条指令. 不可见环境的验证集和测试集中并没有交叉重复的数据, 不可见环境验证集使用 11 个真实场景和 2 349 条指令, 而剩余的 18 个真实场景和 4173 条对应的指令构成测试集.

1.2 Fine-Grained R2R 数据集

由于 R2R 参考路径是由初始位置到目标位置间的最短路径构成, 这在一定程度上影响了路径与指令的耦合度, 同时缺少细粒度指令和视觉场景的对应关系. Hong 等[7] 提出了一种细粒度的子指令形式, 对原先的 R2R 指令使用启发式算法生成相应的子指令, 构造了细粒度的 Fine-grained R2R (FGR2R) 数据集. FGR2R 训练集和验证集的每条指令平均可拆分为 3.6 条子指令, 且每条子指令平均包含 7.2 个单词和 2.6 个对应的导航点. 例如这条简单的指令: “左转, 走上楼梯, 进入卫生间”, 对智能体, 准确无误地理解它十分困难, 必须对指令分解逐一理解每个词语的意思. 这意味着将导航任务简化为多个子任务, 每个子任务都有与其对应的子指令. 此外, 智能体不仅需要理解指令信息, 而且需要对环境中的视觉物体进行识别. 比如 “走上楼梯”, 直到检测到楼梯匹配到指令信息, 才可以执行后续的动作.

1.3 R4R、R6R 和 R8R 数据集

由于 Room-for-Room (R4R)、Room-6-Room (R6R) 和 Room-8-Room (R8R) 数据集构建的思路

一致, 本节将介绍这 3 个数据集. 在 R2R 数据集中的路径普遍需要 4 ~ 6 个步骤完成, 并且利用最短路径的方式到达目的位置. 这样不利于评估指令和路径的匹配程度, 因此需要一个包含更长路径的数据集来对导航中的动作与指令的一致性进行评价. 文献 [8−9] 提出拼接 R2R 数据集指令的方式, 由此形成更长指令长度的 R4R、R6R 和 R8R 数据集. 由于指令长度和参考路径变得更长, 从而增加了训练模型的难度. 同时, 先前的评价指标仅仅关注是否到达目标位置, 对指令和路径是否匹配并不敏感. 因此针对长指令数据集, 研究人员提出了一些新的评价指标和方法, 来衡量和提高模型的泛化能力.

1.4　RxR 数据集

对于现有的 R2R 数据集中存在偏差、指令和路径的数量少等问题, Ku 等[10] 提出新的 Room-across-Room (RxR) 数据集, 分别从指令语言种类、数据集规模、路径和指令的匹配粒度和姿态跟踪 4 个方面对数据集进行了拓展和改进. RxR 数据集包括 16 500 条路径, 且每条路径对应 3 条不同语种的指令, 总词汇量高达 980 万条, 构成 126 000 条指令. 此外, He 等[11] 利用标志物信息, 将 en-RxR 划分成短指令的形式, 构建了 Landmark-RxR 数据集. 相较于 R2R 数据集, RxR 数据集中指令对应的路径长度更长, 并且在指令和路径的匹配程度上更为一致. 同时 RxR 采用对三种语言指令进行测试, 可以避免对单一语种产生过拟合的现象. RxR 数据集中首次引入姿态跟踪的方式, 即对比与人执行指令时所采取的动作和经过的位置. 后续的工作将 RxR 数据集引入到连续环境的模拟器, 以寻找更具有更加实用的模型.

除了以上的指令数据集, 还有一些视觉语言导航拓展工作的数据集, 如 REVERIE[4] 和 Bilingual Room-to-Room (BL-R2R)[12] 数据集. 表 1 介绍了不同数据集的各项属性.

2　视觉语言导航模型

目前视觉语言导航所面临的两大难题: 数据稀缺和模型的泛化性低, 一直阻碍着该领域的发展. 但随着越来越多研究人员投入到视觉语言导航中, 这些问题都不同程度地得到解决. 我们将视觉语言导航模型分为数据增强、搜索策略、动作空间、训练策略 4 个方面来进行介绍.

2.1　基于数据增强的视觉语言导航模型

视觉语言导航是根据真实场景下的照片所构成的仿真环境和人工指令, 进行一系列推理的过程. 专业人员标注的指令不仅成本高, 且数量十分有限, 例如常用的 R2R 数据集中仅含有 21 567 条语言指令. 因此, 数据稀缺是视觉语言导航中的先天问题, 不仅使得学习跨模态匹配更加困难, 还在很大程度上限制了模型的性能. 当前很多领域的研究已经证明了数据增强的有效性, 特别是提升模型的性能有很大帮助. 接下来, 本节将从合成新指令和拼接旧指令两个方面介绍视觉语言导航中的数据增强方法.

2.1.1　基于生成新指令的数据增强模型

Fried 等[13] 首先提出了一种数据增强的方式, 如图 2 所示, “说话者”模型可以从视觉轨迹合成新的指令, 拓展当前有限的训练指令集. 而 “跟随者”模型来检验生成指令, 产生的轨迹可以作为 “说话者”的输入, 从而达到数据增强的目的. 使用数据增强指令训练导航模型的方法, 不仅可以显著提升模型性能, 而且对提升泛化能力很有帮助. 除此之外, “说话者”模型可以用来评价智能体导航路径的好坏. 很多后续工作都在这种数据增强的基础上, 再做相应的方法改进. Fu 等[14] 指出现有的数据增强模型性能表现并不理想, 究其原因在于困难样本的导航成功率不高. 针对这一问题, 他们提出一种对

表 1　视觉语言导航不同数据集的对比

Table 1　The comparison of different datasets in vision-and-language navigation

数据集	训练集 (条)	可见验证集 (条)	不可见验证集 (条)	测试集 (条)	平均指令长度 (单词个数)	语言种类
FGR2R[7]	51 377	3 775	8 481	15 385	7.2	英语
REVERIE[4]	10 466	4 944	3 573	6 292	18.0	英语
BL-R2R[12]	14 025	1 020	2 349	4 188	20.6	英语/中文
R2R[1]	14 039	1 021	2 349	4 173	29.4	英语
R4R[8]	233 613	1 035	45 162	—	58.4	英语
R6R[9]	89 632	—	35 777	—	91.2	英语
R8R[9]	94 731	—	43 273	—	121.6	英语
RxR[10]	79 467	8 813	13 625	24 164	77.8	英语/印地语/泰卢固语
Landmark-RxR[11]	133 602	13 591	19 547	—	21.0	英语

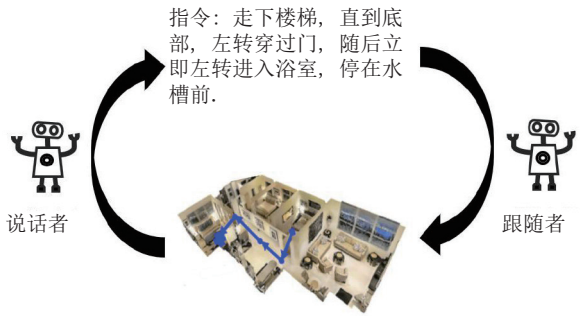


图 2 “说话者”和“跟随者”[13] 模型的数据增强过程

Fig. 2 The data augmentation process in “speaker-follower”[13]

抗训练的方式. 该方法通过模型动态更新路径生成器, 旨在提高困难样本的成功率, 以进一步提升整体的泛化效果. 而文献 [15] 通过分析 “说话者”模型生成的合成指令, 发现大部分合成指令存在语句逻辑问题, 并没有建立好和视觉环境之间的联系, 反而引入了更多的误差. 因此 Huang 等[15] 提出一种生成高精度指令的数据增强方式, 通过设置一个判别器来评价生成数据的质量, 同时引入一些负样本, 以提高训练的鲁棒性. 尽管模型的表现不错, 但是依然没有解决有效的生成指令数量少的问题. 由于缺少指令的评估指标, Zhao 等[16] 提出一个不需要参考指令的指令轨迹亲和模型.

另外, 不少研究人员在导航环境方面做出新的尝试, 并指出 “说话者”模型训练时的有限环境数量, 限制了指令的多样性. 首先, Tan 等[17] 在 “说话者”和 “跟随者” (Speaker-follower, SF) 模型的基础上, 提出了基于环境的数据增强模型, 即通过遮挡环境中的同类物体, 进而产生新的环境; 从这些环境收集新的路径, 然后通过 “说话者”模块生成新的指令; 最后利用这些数据微调模型. 此外, An 等[18] 认为当前视角可能缺失指令中的关键物体信息, 进而导致错误决策, 于是提出邻近视角增强模型 (Neighbor-view enhanced model, NvEM). 该模型使用当前视角的图像特征和相邻视角的图像特征, 以扩大智能体的感受野. 无论是从指令或环境入手, 这些方法均是基于最短路径的原则来导航, 这样会导致学习过程中出现依赖于训练时所做过的动作, 从而出现忽略重要语言信息和视觉信息的问题. 为了解决该问题, 文献 [19] 提出基于随机路径方式的数据增强. Yu 等[19] 基于 “说话者”和 “跟随者”模型, 额外设置路径选择器动态地采样随机路径, 并用 “说话者”模块为这些路径生成相应的指令, 然后再使用生成数据训练 “跟随者”模块, 最终达到随机路径形式的数据增强目的. 这些方法都基于自主合成新指令的方式, 但合成的指令与人类指令之间仍存在较大差异. 主要原因是合成指令的细节不足和逻辑不通, 从而导致了合成新的有效指令比较匮乏的问题.

2.1.2 基于拼接旧指令的数据增强模型

除了生成新指令的方式外, 文献 [8] 提出拼接 R2R 数据集, 来构成 R4R 数据集的方法, 进而达到数据增强的目的. 由于直接训练较长路径的模型比较困难, Jain 等[8] 提出模型先在较短路径下训练, 然后再将模型迁移到较长路径的导航任务中. Zhu 等[9] 进一步将 R2R 数据集拓展到 R6R 和 R8R 长指令数据集, 并提出一个记忆缓存来保存历史子指令和子轨迹对, 同时使用模仿学习和课程强化学习进行两个阶段的训练.

尽管当前视觉语言导航的工作已经取得一定的进展, 但是在提高视觉信息和指令耦合度方面, 并没有很多突破性的工作. 以往的研究验证了使用循环神经网络训练会存在长期依赖的问题, 即当前状态会受一段时间之前的状态影响, 这在长指令集训练过程中是无法避免的. 无论是哪种数据增强手段产生的指令都存在偏差, 以及有效指令和路径的数量少等问题. 因此 Ku 等[10] 提出了新的 RxR 数据集, 从路径轨迹采样方式、路径和指令的数量、路径和指令的粒度、语言种类四个方面对 R2R 数据集进行拓展和改进.

在视觉语言导航任务中, 数据增强作为一种提升模型泛化能力的方法, 一定程度上缩小模型在可见环境和不可见环境的表现差距. 但是视觉语言导航中仍存在导航成功率低和过于依赖拓扑结构的问题, 纯粹依赖数据增强不能根本缓解以上问题.

2.2 基于改进搜索策略的视觉语言导航模型

早期的视觉语言导航任务采用的搜索策略是贪婪解码[20]. 因 MatterPort3D 平台将真实环境简化成离散点集, 而导航过程需要连续地推理决策, 来得到全局最优解, 所以贪心算法的效果并不理想. Fired 等[13] 发现这一问题, 提出在全景动作空间中将导航任务简化为加权无向图搜索方法. 如图 3 所示, 通过采用波束搜索[21] 的方式, 能够选择多条备选全局路径进行打分来选择最优路径. 这种搜索策略大幅提升导航成功率, 但是存在路径过长、搜索效率低的缺点. 为了改进以上的方法, 研究人员提出带回溯的前沿搜索 (Frontier aware search with back tracking, FAST)[22] 和基于回溯机制的后悔模型[23], 旨在降低搜索成本. 回溯机制是在每次决策后及时评估, 如果打分低, 则选择回溯上一步, 否则

图 3　视觉语言导航任务中的不同搜索策略[22]

Fig. 3　Different search strategies in vision-and-language navigation[22]

选择邻近未探索的节点. 而 FAST 则是在此基础上, 提出了一种局部信息和全局信息相结合的方式. 该方法通过比较不同长度的局部路径, 结合全局信号, 利用异步搜索的方式来实现有效回溯.

由于存在指令语义模糊和局部视觉不确定性的问题, Wang 等[24] 提出一种动态决定探索方向、允许对下一步节点探索并进行预测的方法. 但探索过多会导致轨迹长度过长和基于路径长度加权的成功率降低的问题, 整体效果并不理想. 而 Chi 等[25] 提出当智能体不知选择哪个方向时, 采用辅助解惑的手段. 当学习到的融合信息让智能体感到 “疑惑” 时, 交互学习方式能帮助智能体解决导航中下一步移动的问题. 而为了缩小训练和测试之间的差距, Deng 等[26] 提出可变图规划器模型 (Evolving graphical planner, EGP), 这是一种使用原始图像高效生成全局规划的方法. 其通过动态地构建一个图的表示 (包括动作空间), 以便于更好决策. Hong 等[27] 则通过构建语言和视觉实体关系图模型 (Language and visual entity relationship graph, Relgraph), 更好地利用不同模态间的关系. 同时他们提出一种消息传递算法, 在图中的语言元素和视觉实体之间传播信息, 然后将这些信息组合起来以确定下一步要采取的动作. 为了更好地长期规划决策, Wang 等[28] 提出一种结构化场景记忆的模型 (Structured scene memory, SSM), 允许智能体对已探索的区域保留访问权力, 然后通过这种持久性的空间表示, 智能体在细粒度指令的辅助下, 在全局决策方面表现出色.

现有的导航策略都是尽可能地找到每步的最佳决策, 寻找一条从起始位置到目标位置的有效路径. 虽然一定程度上会造成导航路径长度过长, 却对导航成功率的提升十分明显. 此外, 随着研究的深入, 记忆机制和图结构的引入, 为视觉语言导航策略提供了不同的思路和方向.

2.3　基于不同动作空间的视觉语言导航模型

文献 [29] 首次按照动作空间划分不同的模型, 将依赖于模拟平台和导航图的模型称为高级动作模型, 对于直接预测下一个基础动作的模型称为低级动作模型. 如图 4 所示, 图 4(a) 表示低级动作空间的 5 个基础动作, 图 4(b) 表示高级动作空间基于导航点的动作. 本节根据动作空间的划分, 分别介绍高级动作空间和低级动作空间的视觉语言导航方法.

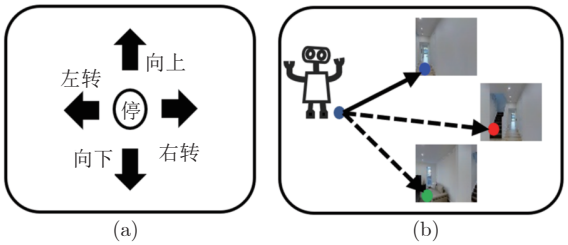


图 4　低级动作空间和高级动作空间表示[29]

Fig. 4　Low-level action space and high-level action space[29]

2.3.1　基于高级动作空间的视觉语言导航

早期 Fried 等[13] 提出将 36 张不同仰角和水平偏角的图像合成一张全景图的方法, 后来该形式被通称为高级动作空间. 在此高级动作空间中, 智能体只需选择邻近节点移动. 高级动作空间不仅可以简化导航过程, 并且能显著提升导航成功率. 在高级动作空间下, 文献 [30] 发现导航结果反馈模糊的问题, 即导航成功产生的反馈结果, 并不能反应指令和路径是否匹配. 由此, Wang 等[30] 提出强化跨模态匹配 (Reinforced cross-modal matching, RCM) 的方法来解决上述问题. 利用推理导航器在局部区域内进行跨模态对齐, 再使用匹配评判器促进路径和指令之间的全局匹配, 进一步强化模态融合效果和提高导航成功的标准. Ma 等[31] 提出自我监控智能体模型 (Self-monitor agent, SMNA). 他们根据模态匹配的关系, 认为 “下一个动作的执行常常是由上一个动作完成与否决定的”, 并相应提出了视觉和语言联合对齐模型, 来监控导航进度. 另外, 由于指令中含有丰富的实体描述和方向信息, Qi 等[32] 提出物体和动作可知模型 (Object-and-action awared model, OAAM), 分别对视觉特征和方向特征使用注意力机制, 最后再融合两部分特征. 该方法充分利用指令中实体和方向信息, 来与视觉场景进行匹配, 最后设置路径损失来限制智能体仅沿着最短路径移动.

在视觉语言导航的设定中, MatterPort3D 模拟器是将场景划分为离散的可导航位置点集. 这一做法简化导航过程为一个无向图的探索过程, 即每步移动都从邻近的有限点集中选择下一个目标节点. 这在一定程度上减少了视觉信息对任务的影响.

文献 [33] 指出在视觉语言导航和问答任务中, 不利用视觉信息的单模态模型的表现好于多模态模型, 模态融合反而造成了性能衰减. 针对以上问题, Hu 等[34] 提出在不同模态融合条件下, 对比 "说话者" 和 "跟随者"模型[13] 和自我监控智能体模型[31] 的性能表现, 发现模型更容易利用几何拓扑结构信息, 而忽略了大量的视觉模态信息的现象. 模型对于拓扑结构的依赖一定程度降低了指令和视觉信息的耦合度. 为了解决这个问题, Yu 等[19] 提出改变最短路径为随机路径的移动策略, 旨在消除对于路径结构的依赖, 更多地专注语言和视觉之间的信息匹配. 针对不同环境中的性能差异问题, Zhang 等[35] 设计新的环境划分和特征替换的方案, 研究环境偏差的影响.

除了改变路径采样的方式之外, 另一个思路是回到低级动作空间. Anderson 等[36] 尝试转移模拟环境训练的智能体到现实场景中, 并提出一个子目标模型来识别临近可达的节点. 他们使用即时定位与地图构建和路径规划的方法, 建立智能体学习的高级动作和智能体的低级动作的变换联系, 将模型性能损失控制在可接受的范围内. 但高级动作空间方面的迁移工作, 还是受 Matterport3D 模拟器不能支持低级动作的影响, 需要通过特定算法转换为低级动作, 因此很多模型不便于直接由模拟环境转移现实场景中.

2.3.2 基于低级动作空间的视觉语言导航

相对于高级动作空间中选择邻近节点的移动方式, 低级动作空间只包括六种基础动作: 向上、向下、左转 30 度、右转 30 度、前进和停止. 在这种动作空间下, 模型在对于环境拓扑结构未知的情况下, 直接预测智能体的动作. 如图 5 所示, 在视觉语言导航任务中, 基于编码 _ 解码的方法首先通过长短期记忆网络 (Long short-term memory, LSTM)[37] 编码器将指令编码 $[x_1, x_2, \cdots, x_l]$ 和真实图像 $[v_1, v_2, \cdots, v_t]$, 映射到一个上下文的动作序列, 再通过 LSTM 解码器融合编码后的语言特征和视觉特征预测每一步的动作 $[a_0, a_1, \cdots, a_T]$. 解码过程中额外加入了注意力机制, 这一机制选择性关注视觉感知和当前指令中相关联的内容, 帮助智能体结合环境选择相应的基础动作. 通过建立紧密的模态间的信息联系, 模型生成一系列的低级动作命令来指导完成导航任务.

高级动作空间中存在过度依赖已知的路径拓扑结构的问题, 不利于未来部署在现实场景中. 因此, 不少研究人员开始关注更具有现实意义的模型, Landi 等[38] 提出使用动态卷积滤波器的方法, 模型基于

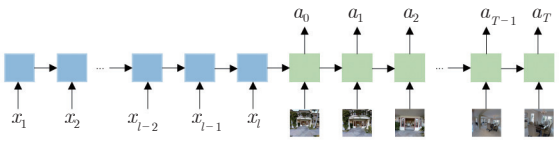


图 5 视觉语言导航中的 seq2seq 模型

Fig. 5 The seq2seq model in vision and language navigation

当前指令信息动态地从视觉信息中提取相关信息, 并输出低级动作空间下的动作概率. 在此基础上, Landi 等[29] 进一步提出类似 Transformer[39] 结构的感知转化移动模型 (Perceive, transform and act, PTA), 通过多次利用注意力机制的形式来融合模态间信息. 实验证实了该模型同时兼容高级动作空间和低级动作空间. 由于 MatterPort3D 平台环境本身的约束, 新型的模拟环境平台相继被提出. 首先是 FacebookAI 实验室在 2019 年提出的 Habitat 平台[40]和 Shen 等[41] 在 2020 年提出的 iGibson 平台. 这些支持连续环境模拟器的推出, 大大推动了基于低级动作空间方面的研究. 文献 [42] 提出基于 Habitat 平台的连续环境下的视觉语言导航任务. 不同于以往的高级动作空间方法, 存在传送移动、依赖几何结构和精准定位的问题, Krantz 等[42] 通过构建一个跨模态注意力机制的连续环境的视觉语言导航模型, 验证了数据增强、数据聚合和进度控制对模型的积极作用. 同时对比高级动作空间的视觉语言导航模型, 他们发现先前的视觉语言导航模型中存在过多理想化的条件, 在真实环境中的可行性有待验证. Chen 等[43] 在基于连续环境的视觉语言导航模型的基础上, 将其分解为两个阶段: 计划和控制, 在探索过程中, 拓扑地图被建立用于导航规划. 然后局部控制器接受导航规划并生成低级动作来完成导航任务.

无论在高级动作空间和低级动作空间中, 现有的方法并未详细解释模型在模态融合后性能提升的原因. 文献 [33] 对以往的模态融合方式提出了质疑, 并建议以后的模型增加模态消融实验以佐证效果. 为了更好地融合模态间的信息, Zhu 等[44] 提出辅助推理导航模型 (Auxiliary reasoning navigation, AuxRN). 该模型通过四个辅助任务: 动作解释、估计进度、预测方向和轨迹一致性评价, 来提高模型的推理和环境感知的能力. 由于指令间信息差异和指令中语义模糊的问题, Xia 等[45] 编码相同轨迹的所有指令, 其中每条指令互作补充, 去提高模型的文本理解能力. 在视觉语言导航中, 模态间的联系并不是简单地合并指令和视觉信息, 而是需要建立

互为补充的关系, 进一步提升模型的性能, 并通过合理的实验证明在不同的动作空间下模态融合方式的有效性.

2.4　基于训练方法的视觉语言导航模型

视觉语言导航任务中常用的两种模型学习方式: 监督学习和强化学习. 监督学习是通过 R2R 数据集中的最短路径标注数据, 学习得到一个优化的模型, 进而预测不可见环境中的路径序列. 而强化学习是把视觉语言导航任务看作一个马尔可夫决策过程. 智能体在导航过程中观察周围环境并进行分析和反馈, 并通过特定的奖励函数, 尝试将学习到的经验知识应用到导航任务中, 不断地进行试验, 以达到完成视觉语言导航任务的目的. 尽管这些方法的有效性得到很好的验证, 但是各自的局限性也被揭露出来. 研究人员在训练方法的选择上进行了更为深入的研究, 当前预训练模型通过大规模数据训练, 对视觉语言导航的性能提升十分明显. 下面将对传统训练模型和预训练模型展开介绍.

2.4.1　基于传统训练的视觉语言导航模型

由于 R2R 数据集提供了参考路径, 通过匹配预测动作分布和最佳路线, 最初的视觉语言导航方法大多采用的是监督学习方式. 文献 [1] 使用基于注意力机制的 LSTM 的序列到序列模型 (Sequence-to-sequence, seq2seq)[46], 并结合 “学生自学”[47] 的训练方法, 对于先前的分布采用动作输出序列预测下一步动作, 这是初期流行的一种基础方法. 该方法使用交叉熵损失函数, 学习标注数据的特征信息, 泛化到未知环境中. 但是由于人工标注的数据成本过于昂贵, 最优路径并不容易获取. 在后续的研究中, 为了进一步提升在不可见环境下的泛化能力, Wang 等[48] 提出使用强化规划 (Reinforced planning ahead, RPA) 的方法, 将模型无关和基于模型两种强化学习联合在一起. 其中展望模型结合了环境模型和策略模型, 在 R2R 数据集上取得了不错的效果. 最近的研究提出了很多新颖的学习方法, 诸如 Wang 等[30] 提出一种自监督模仿学习的方法. 通过训练, 智能体可以根据过往的决策, 学习产生多条可能的轨迹. 模型利用最佳匹配的轨迹辅助训练, 并优化轨迹的生成. 文献 [17] 改进以往的方法, 提出将模仿学习和强化学习的损失结合作为一个损失函数, 并用半监督学习的方式进行反向翻译和环境消除 (Environmental dropout, Envdrop), 分别为了训练额外的数据和生成未知环境. 这种方法对模型的泛化能力进一步提升, 如图 6 所示, 模仿学习和强化学习结合的方法通过结合两种学习策略的优

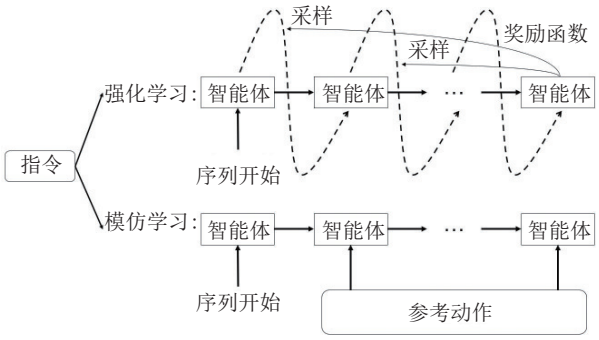


图 6　融合强化学习和模仿学习的过程

Fig. 6　The mixture of reinforcement learning and imitation learning

势, 能够有效地提升模型的性能. Wang 等[49] 对于模仿学习的错误累积和强化学习中的奖励设计成本问题, 尝试使用蒸馏的方法减少过拟合, 提出软专家奖励学习模型 (Soft expert reward learning, SERL). 具体地, 通过设置软专家蒸馏模块让模仿过程减少错误, 同时利用自我感知模块让智能体一直向目的位置移动. 研究人员在学习方法上的研究仍在继续, 不断地优化方法策略.

在视觉语言导航中, “学生自学”和 “老师指导”[50] 训练策略的选择, 是影响生成路径序列的一大因素. 在视觉语言导航的设定中, 因为导航路径长, 并且采用离散动作的问题, 所以早期的动作抽样工作是基于 “学生自学”的方式. 但由于全景动作空间的提出, 路径长度被缩短了一大半, “老师指导”的方式开始流行起来. 但是两者都存在曝光偏差[51] 的问题. 一旦出现误差, 则会导致大量累计误差, 从而偏离正确路线. 针对 “学生自学”引入偏差的问题, 文献 [9] 提出了基于 “学生自学”的模仿学习, 保证轨迹和指令的一致性. 为了充分发挥两者的优势, Li 等[52] 从课程抽样方式中得到灵感, 提出随机动作采样的方式. 具体地, 基于伯努利分布的抽样策略通过随机选择每一步的动作抽样方式, 借此来保证利用 “学生自学”和 “老师指导”的优势, 进而得到一种相对偏差较小的动作抽样方法.

目前, 很多工作通过结合模仿学习和强化学习的方式, 取得了较大性能提升. 模仿学习学习老师的动作, 而强化学习通过从奖励中采样动作, 使智能体可以探索环境和提升泛化能力. 如何更好地选择和利用训练策略, 是提高导航模型泛化能力的一个关键因素.

2.4.2　基于大规模预训练的视觉语言导航模型

近年来, 研究人员在如何提升智能体对不可见环境的泛化能力方面进行了不少尝试, 包括预探索、

数据增强和分析模态融合关系等方式. 当前使用预训练模型提取特征的方式已经应用到各种任务中, 其可以有效地提升下游任务的性能. 受此启发, 研究人员提出了在视觉语言导航中使用预训练模型, 以解决泛化能力不足的问题. 首先, 文献 [52] 提出了使用 Bidirectional encoder representations from transformers (BERT)[53] 等大规模预训练语言模型, 来丰富指令表达. Hao 等[54] 提出使用一种通用的预训练视觉语言导航智能体 (Pre-trained vision-and-language based navigator, PREVALENT), 并利用图像–语言–动作信息来进行预训练. 实验证明预训练模型对提升模型的泛化能力很有帮助. 后来 Huang 等[55] 在 PREVALENT 的基础上, 使用参数共享的方法来减低预训练模型的参数量. 文献 [56] 中指出模型可以学习更多的语言知识, 来提高推理的效率. 此外, Hong 等[57] 提出 Recurrent vision-and-language bert for navigation (RecBERT). 这是一个多模态 BERT 模型, 搭配时间感知递归函数, 为智能体提供更丰富的信息. 针对 RecBERT 会存在历史信息丢失的问题, Chen 等[58] 提出 History-aware multi-modal transformer (HAMT), 将完整的历史信息编码保存, 并设计了层次化的历史编码方法, 降低计算复杂度. 实验结果显示使用预训练的语言模型分别在可见环境和不可见环境中的导航成功率高达 76% 和 66%, 不可见环境中的基于路径加权的成功率为 60%. 进一步地证明预训练模型可以提高模型的泛化能力.

相较于传统训练方法, 预训练模型引入了额外的知识表达, 对视觉语言导航模型的提升十分显著. 正因为高效的性能和强大的模态融合能力, 如今预训练模型已经成为视觉语言导航模型的重要研究方向.

3 视觉语言导航方法的实验分析

第 2 节和第 3 节主要介绍了视觉语言导航的数据集和当前主要的模型方法, 本节将对视觉语言导航的评价指标进行全面介绍, 并结合 R2R 数据集、R4R 数据集和 RxR 数据集对比分析视觉语言导航模型.

3.1 视觉语言导航的评价指标

对于不同模型的评判, 评价指标发挥着重要的作用, 是衡量模型性能的关键性指标. 随着视觉语言导航任务的发展, 新的模型评价指标相继被提出. 表 2 给出了视觉语言导航任务的评价指标, 包括其定义和计算公式. 这为第 3.2 节视觉语言导航模型性能比较提供帮助. 视觉语言导航的评价指标不仅关注导航成功率 (Success rate, SR) 和路径长度 (Path length, PL), 而且需要对导航过程中路径轨迹和指令之间的一致性程度进行相应的度量评估. 接下来将主要介绍目前的核心评价指标, 其中基于路径加权的成功率 (Success weighted by path length, SPL) 的主要思想是将成功率和路径长度融合处理, 来衡量导航的好坏. 早期视觉语言导航模型的目标是尽可能地提高基于路径加权的成功率, 来评估模型的性能. 但它仅关注是否成功到达目标位置, 而忽略了预测路径和参考路径的一致性问题. 后续工作中提出的长度加权的覆盖分数 (Coverage weighted by length score, CLS)[8] 和基于动态时间规整加权成功率 (Success rate weighted normal-

表 2 视觉语言导航任务中的评价指标

Table 2 The metrics of vision-and-language navigation

评价指标	定义	公式
路径长度	起始位置到停止位置的导航轨迹长度	$\sum_{\boldsymbol{v}_i \in V} d\left(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}\right)$
导航误差	预测路径终点和参考路径终点的距离	$d\left(\boldsymbol{v}_t, \boldsymbol{v}_e\right)$
理想成功率	预测路径中任意节点距离参考路径终点的阈值距离内的概率	$\mathbb{I}\left[\left(\min_{\boldsymbol{v}_i \in V} d\left(\boldsymbol{v}_i, \boldsymbol{v}_e\right)\right) \leq d_{th}\right]$
导航成功率	停止位置与参考路径终点的距离不大于 3 米的概率	$\mathbb{I}\left[NE(\boldsymbol{v}_t, \boldsymbol{v}_e) \leq d_{th}\right]$
基于路径加权的成功率	基于路径长度加权的导航成功率	$SR(\boldsymbol{v}_t, \boldsymbol{v}_e) \cdot \frac{d_{gt}}{\max\left\{PL(V), d_{gt}\right\}}$
长度加权的覆盖分数[59]	预测路径相对于参考路径的路径覆盖率和长度分数	$PC\left(P, R\right) \cdot LS\left(P, R\right)$
基于动态时间规整加权成功率[59]	由成功率加权的预测路径和参考路径的时空相似性	$SR(\boldsymbol{v}_t, \boldsymbol{v}_e) \cdot \exp\left(-\frac{\min_{\boldsymbol{w} \in W} \sum_{(i_k, j_k) \in \boldsymbol{w}} d\left(\boldsymbol{r}_{i_k}, \boldsymbol{q}_{j_k}\right)}{\lvert R\rvert \cdot d_{th}}\right)$

ized dynamic time warping, SDTW)[59] 两个评价指标, 主要是度量轨迹和指令一致性程度. 长度加权的覆盖分数中包括两部分路径覆盖率 (Path coverage, PC) 和路径长度分数 (Length score, LS). 路径覆盖率表示与参考路径的一致程度, 其计算公式如下:

$$PC(P, R) = \frac{1}{|R|}\sum_{\boldsymbol{r}\in R}\exp\left(-\frac{D(\boldsymbol{r}, P)}{d_{th}}\right) \tag{1}$$

式中, R 代表查询路径, P 代表参考路径, r 是查询路径的位置坐标向量, d_{th} 是阈值距离. $PC(P, R)$ 即为所计算的路径覆盖率. 而路径长度分数则是评价预测路径和参考路径的一致性程度, 进而来约束预测路径的长度, 产生与参考路径长度一致的预测路径, 计算公式为:

$$EPL(P, R) = PC(P, R)\cdot PL(R) \tag{2}$$

$$LS(P, R) = \frac{EPL(P, R)}{EPL(P, R) + |EPL(P, R) - PL(P)|} \tag{3}$$

式中, $EPL(P, R)$ 表示导航路径相对于参考路径覆盖范围的期望值, $PL(V)$ 表示路径长度, $PC(P, R)$ 表示路径覆盖率. $LS(P, R)$ 即为所计算的路径长度得分. SDTW 是对预测路径和参考路径在时空相似性上的约束, 由导航成功率和路径一致性合并计算.

第 3.2 节将对比不同数据集下的视觉语言导航模型, 通过以上主要的评价指标进行对比分析.

3.2　视觉语言导航模型的分析对比

表 3 和表 4 分别展示了不同模型在 R2R 数据集和 R4R 数据集上, 基于相应主要评价指标的实验结果. 而表 5 以不同模型的主要创新点来划分模型方法, 包括数据增强、导航策略、动作空间和训练方法 4 个方向. 表 5 中 "√"表示属于对应分类的改进方向, 而—表示不属于对应分类的改进方向.

由表 3 和表 5 可知, 随着引入数据增强和改进导航策略之后, 在 R2R 数据集上, 视觉语言导航模型的 SR 和 SPL, 都较以往得到了不少的提升. 文献 [13] 提出的全景动作空间形式和数据增强方法, 为视觉语言导航模型的快速发展, 提供了有力的支持. 同时文献 [17] 在此基础上提出的融合强化学习和模仿学习的训练方法, 为后续的研究提供了参考模型. 该方法的广泛应用对视觉语言导航任务的发展有重要的意义. 此外, 最新研究发现预训练模型 BERT 和 Transformer 模型使智能体学习到更多有效的知识, 可以进一步提升导航性能. 但值得一提的是, 预训练模型的训练时间和计算成本花销巨大. 如何尽可能地降低计算成本, 设计一个轻量级的模型是亟待解决的问题.

表 3　在 R2R 测试数据集上的视觉语言导航方法对比
Table 3　The comparison of vision-and-language navigation methods on the R2R test dataset

方法	路径长度 (米)	SR (%)	SPL (%)
seq2seq[1]	8.13	20.0	18.0
RPA[48]	9.15	25	23.0
SF[13]	14.82	35	28.0
SMNA[31]	18.04	48	35
PTA[29]	10.17	40.0	36.0
RCM[30]	11.97	43.0	38.0
Regretful[23]	13.69	48.0	40.0
FAST[22]	22.08	54.0	41.0
EGP[26]	—	53.0	42.0
PRESS[52]	10.77	49.0	45.0
SSM[28]	22.10	61.0	46.0
Envdrop[17]	11.66	51.0	47.0
SERL[49]	—	53.0	49.0
OAAM[32]	10.40	53.0	50.0
AuxRN[44]	—	55.0	51.0
PREVALENT[54]	10.51	54.0	51.0
RelGraph[27]	10.29	55.0	52.0
RecBERT[57]	12.35	63.0	57.0
HAMT[58]	12.27	65.0	60.0

表 4　在 R4R 测试数据集上的视觉语言导航方法对比
Table 4　The comparison of vision-and-language navigation methods on the R4R test dataset

方法	SR (%)	CLS (%)	SDTW (%)
seq2seq[1]	25.7	20.7	9.0
SF[13]	23.8	29.6	9.2
FAST[22]	36.2	34.0	15.5
Envdrop[17]	29.0	34.0	9.0
Regretful[23]	30.1	34.1	13.5
RCM[30]	29.0	35.0	13.0
PTA[29]	27.0	35.0	8.0
OAAM[32]	31.0	40.0	11.0
RelGraph[27]	36.0	41.0	34.0
EGP[26]	30.0	44.0	18.0
RecBERT[57]	43.6	51.4	29.9
SSM[28]	32.0	53.0	19.0
HAMT[58]	44.6	57.7	31.8

表 5 视觉语言导航中的不同方法改进的对比
Table 5 The comparison of different improvements in vision-and-language navigation

方法	数据增强	导航策略	动作空间	训练方法	R2R SR (%)	R4R SR (%)
Seq2seq[1]	—	—	—	√	20.0	25.7
RPA[48]	—	—	—	√	25.0	—
SF[13]	√	√	√	—	35.0	23.8
SMNA[31]	—	√	—	√	48.0	—
Regretful[23]	—	√	—	—	48.0	30.1
FAST[22]	—	√	—	—	54.0	—
PTA[29]	—	—	√	—	40.0	24.0
PRESS[52]	—	—	—	√	49.0	29.0
RCM[30]	—	√	—	√	43.0	29.0
Envdrop[17]	√	—	—	√	51.0	—
SERL[49]	—	—	—	√	53.0	—
OAAM[32]	—	—	—	√	53.0	31.0
PREVALENT[54]	—	—	—	√	54.0	—
EGP[26]	—	√	√	—	53.0	30.2
SSM[28]	—	√	—	√	61.0	—
RecBERT[57]	—	√	—	√	63.0	43.6
HAMT[58]	—	√	—	√	65.0	44.6

不同于 R2R 数据集的主要评价指标, R4R 数据集包括更长的轨迹, 更注重指令和轨迹的一致性程度. 因此, R4R 数据集将 CLS 和 SDTW 作为主要评价指标. 由表 4 和表 5 可知, 在导航成功率的评价指标上, R4R 数据集是明显低于 R2R 数据集. 这是因为长指令的影响, 导航的性能降低. 由表 4 可知, 在 CLS 和 SDTW 上, 模型的表现并不尽如人意. 其主要原因是模型过于注重是否到达目标位置, 忽略了指令和轨迹一致性的比较. 尽管模型在基于路径加权的成功率上有不错的表现, 但这不能保证导航轨迹与指令内容一致. 因此, 研究人员开始转向子指令和子轨迹的研究, 通过分段剪切长指令, 对导航过程中的一致性问题展开一系列的研究. 我们可以发现注重指令和轨迹一致性的模型, 在主要评价指标上都有一定的性能提升. 因此, 如何更好地利用指令信息和视觉信息, 是视觉语言导航中的关键问题.

由表 5 中的对比可以看出, 早先的方法主要研究监督学习和强化学习的选择, 但整体表现并不好. 随着数据增强以及模仿学习和强化学习的结合等方法的提出, 这使模型的性能得到了较大的提升. 但其利用波束搜索会导致路径长度过长, 模型从而丢失现实的应用意义. 导航策略的改进极大地推动了视觉语言导航的应用发展. 随着研究的深入, 研究人员将眼光投入到更贴切现实的低级动作空间, 开拓新的研究方向. 此外, 随着大规模预训练模型的兴起, 研究人员尝试将预训练模型引入到视觉语言导航任务中, 并取得不错的效果, 使导航性能得到大幅提升.

由表 3 ~ 5 可知, 随着各种各样的方法被提出, 视觉语言导航领域的发展更加多元化. 不仅仅是模型在各项评价指标上有较大的提升, 更重要的是在细分研究方向上也得到了更多研究和关注.

4 未来展望

视觉语言导航是近年来在多模态领域中新兴的研究方向, 一经提出就受到大批研究人员的关注. 随着研究的不断深入, 视觉语言导航在导航成功率和泛化能力上, 都得到了巨大的提升. 研究人员通过数据增强手段, 生成新的训练数据. 虽然这种方法对模型性能的提升很有帮助, 但并未解决泛化能力不足的问题. 因此, 后续的研究开始着力于减少过拟合现象, 引入预训练模型. 与此同时, 研究人员不断优化学习方式, 进一步地提出视觉文本对齐和回溯机制等辅助手段, 这些方法显著提升模型的泛化能力, 并取得较为理想的性能.

但是, 目前仍有一些问题亟待解决: 1) 当前提出的一些方法, 受到了仿真环境平台和数据集的限制. 从低级动作空间到高级动作空间的转换, 采用波束搜索的方式, 简化了导航过程. 尽管各方面的性能表现均令人信服, 但脱离现实, 模型难以迁移部署到现实环境中. 如何贴近真实场景、赋予模型更多现实的应用意义, 这需要视觉语言导航研究的重心重新转移到低级动作空间上. 尽管已有基于连续环境的视觉语言导航模型, 但是其导航效果并不理想. 因此视觉语言导航需要结合传统机器人技术, 进一步优化目前在模拟环境中训练的模式. 2) 有实验表明视觉语言导航的模态融合方法会对模型性能产生负面作用. 究其根本是模态信息之间关系模糊, 并未形成良好的互补, 模型不能有效地利用多模态信息. 目前, 大规模预训练模型可以更好地利用多模态信息, 获得不错的性能表现. 但是由于计算能力不足和时间开销过大, 这为视觉语言导航的发展带来新的问题. 因此对多模态信息学习的研究, 仍有很大的发展空间. 3) 在视觉语言导航任务中, 数

据稀缺的问题尤为明显, 这是限制性能的一大阻碍. 尽管研究人员提出了利用机器生成合成指令的方法, 但这些合成指令大部分是有缺陷的, 且不符合人们的语言习惯. 同时不少研究人员重新搜集数据, 从规模和指令长度等方面进行拓展, 获得更接近现实场景的指令集. 由此可见, 数据方面的研究工作一直都是视觉语言导航中的重要内容.

在现实场景中的导航过程是动态连续的, 而非简单的无向图探索过程, 目前不少研究人员重新投入到连续空间的视觉语言导航研究, 即在低级动作空间下, 智能体经过一系列的基础动作, 完成视觉语言导航任务. 当前 Habitat 平台和 iGibson 平台都支持连续的导航. 智能体可以通过低级动作完成导航, 这就为以后应用到现实场景提供了更多的可能性. 此外, 由于多模态任务的输入复杂多变, 机器和人类的理解能力差异较大, 所以 BERT 等预训练模型的引入, 为智能体提供丰富的额外知识, 有助于理解模态信息和模态间的融合. 总体总之, 视觉语言导航任务无论是在现实中的应用, 以及数据获取方面的研究, 未来还有很长的路要走.

5 结束语

视觉语言导航是一种多模态理解任务, 在未来智能家具、娱乐、养老等国计民生领域有较大应用需求. 本文详细介绍了视觉语言导航任务近年来的发展, 首先对于各种主流模型进行了简要介绍, 然后对提升模型泛化能力的方法进行了综述, 分别包括模态间的分析、指令集等拓展方式以及搜索策略、训练方式和预训练模型等辅助策略. 尽管视觉语言导航任务近年来取得了快速的发展, 但是随着研究的深入, 也凸显出各种约束和限制, 这需要更多研究人员投入后续的工作研究. 研究人员不仅要追求性能上的突破, 而且要赋予模型更多的应用价值. 希望通过本文可以让更多人了解视觉语言导航任务, 吸引更多人投入其中, 促进其发展.

References

1 Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderhauf N, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3674–3683

2 Chang A, Dai A, Funkhouser T, Halber M, Niebner M, Savva M, et al. Matterport3D: Learning from rgb-d data in indoor environments. In: Proceedings of the International Conference on 3D Vision. Qingdao, China: IEEE, 2017. 667–676

3 Chen H, Suhr A, Misra D, Snavely N, Artzi Y. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 12538–12547

4 Qi Y K, Wu Q, Anderson P, Wang X, Wang W, Shen C H, et al. Reverie: Remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 9979–9988

5 Gao C, Chen J Y, Liu S, Wang L T, Zhang Q, Wu Q. Room-and-object aware knowledge reasoning for remote embodied referring expression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 3064–3073

6 Thomason J, Murray M, Cakmak M, Zettlemoyer L. Vision-and-dialog navigation. In: Proceedings of the Conference on Robot Learning. Cambridge, USA: 2021. 394–406

7 Hong Y C, Rodriguez C, Wu Q, Gould S. Sub-instruction aware vision-and-language navigation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Virtual Event: 2020. 3360–3376

8 Jain V, Magalhaes G, Ku A, Vaswani A, Ie E, Baldridge J. Stay on the path: Instruction fidelity in vision-and-language navigation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: 2019. 1862–1872

9 Zhu W, Hu H X, Chen J C, Deng Z W, Jain V, Ie E, et al. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: 2020. 2539–2556

10 Ku A, Anderson P, Patel R, Ie E, Baldridge J. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Virtual Event: 2020. 4392–4412

11 He K J, Huang Y, Wu Q, Yang J H, An D, Sima S L, et al. Landmark-rxr: Solving vision-and-language navigation with fine-grained alignment supervision. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: 2021. 652–663

12 Yan A, Wang X, Feng J, Li L, Wang W. Cross-lingual vision-language navigation [Online], available, https://arxiv.org/abs/1910.11301, December 6, 2020

13 Fried D, Hu R H, Cirik V, Rohrbach A, Andreas J, Morency LP, et al. Speaker-follower models for vision-and-language navigation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal Canada: MIT Press, 2018. 3318–3329

14 Fu T, Wang X, Peterson MF, Grafton ST, Eckstein MP, Wang W. Counterfactual vision-and-language navigation via adversarial path sampler. In: Proceedings of the 16th European Conference on Computer Vision. Virtual Event: 2020. 71–86

15 Huang H S, Jain V, Mehta H, Ku A, Magalhaes G, Baldridge J, et al. Transferable representation learning in vision-and-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 7404–7413

16 Zhao M, Anderson P, Jain V, Wang S, Ku A, Baldridge J, et al. On the evaluation of vision-and-language navigation instruc-

tions. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Virtual Event: 2021. 1302−1316

17 Tan H, Yu L C, Bansal M. Learning to navigate unseen environments: Back translation with environmental dropout. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: 2019. 2610−2621

18 An D, Qi Y K, Huang Y, Wu Q, Wang L, Tan T N. Neighbor-view enhanced model for vision and language navigation. In: Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: 2021. 5101−5109

19 Yu F, Deng Z W, Narasimhan K, Russakovsky O. Take the scenic route: Improving generalization in vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020. 920−921

20 Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 7008−7224

21 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: IEEE, 2014. 3104−3112

22 Ke L, Li X J, Bisk Y, Holtzman A, Gan Z, Liu J J, et al. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 6741−6749

23 Ma C, Wu Z X, AlRegib G, Xiong C M, Kira Z. The regretful agent: Heuristic-aided navigation through progress estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 6732−6740

24 Wang H Q, Wang W G, Shu T M, Liang W, Shen J B. Active visual information gathering for vision-language navigation. In: Proceedings of the 16th European Conference on Computer Vision. Virtual Event: 2020. 307−322

25 Chi T, Shen M M, Eric M, Kim S, Hakkani-tur D. Just ask: An interactive learning framework for vision and language navigation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. California, USA: 2020. 2459−2466

26 Deng Z W, Narasimhan K, Russakovsky O. Evolving graphical planner: Contextual global planning for vision-and-language navigation. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: 2020. 20660−20672

27 Hong Y C, Rodriguez C, Qi Y K, Wu Q, Gould S. Language and visual entity relationship graph for agent navigation. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: 2020. 7685−7696

28 Wang H Q, Wang W G, Liang W, Xiong C M, Shen J B. Structured scene memory for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 8455−8464

29 Landi F, Baraldi L, Cornia M, Corsini M, Cucchiara R. Perceive, transform and act: Multi-modal attention networks for vision-and-language navigation [Online], available: http://arxiv.org/abs/1911.12377, July 30, 2021

30 Wang X, Huang Q Y, Celikyilmaz A, Gao J F, Shen D H, Wang Y F, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 6629−6638

31 Ma C, Lu J S, Wu Z X, AlRegib G, Kira Z, Socher R, et al. Self-monitoring navigation agent via auxiliary progress estimation. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: 2019

32 Qi Y K, Pan Z Z, Zhang S P, Hengel A V, Wu Q. Object-and-action aware model for visual language navigation. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 303−317

33 Thomason J, Gordon D, Bisk Y. Shifting the baseline: Single modality Performance on visual navigation & QA. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: 2019. 1977−1983

34 Hu R H, Fried D, Rohrbach A, Klein D, Darrell T, Saenko K. Are you looking? grounding to multiple modalities in vision-and-language navigation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: 2019. 6551−6557

35 Zhang Y B, Tan H, Bansal M. Diagnosing the environment bias in vision-and-language navigation. In: Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan: 2021. 890−897

36 Anderson P, Shrivastava A, Truong J, Majumdar A, Parikh D, Batra D, et al. Sim-to-real transfer for vision-and-language navigation. In: Proceedings of the Conference on Robot Learning. Cambridge, USA: 2021. 671−681

37 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

38 Landi F, Baraldi L, Corsini M, Cucchiara R. Embodied vision-and-language navigation with dynamic convolutional filters. In: Proceedings of the 30th British Machine Vision Conference. Cardiff, UK: 2019. 18

39 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: 2017. 6000−6010

40 Savva M, Kadian A, Maksymets O, Zhao Y L, Wijmans E, Jain B, et al. Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 9338−9346

41 Shen B K, Xia F, Li C S, Martín-Martín R, Fan L X, Wang G Z, et al. Igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic: IEEE, 2021. 7520−7527

42 Krantz J, Wijmans E, Majumdar A, Batra D, Lee S. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 104−120

43 Chen K, Chen JK, Chuang J, Vázquez M, Savarese S. Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 11276−11286

44 Zhu F D, Zhu Y, Chang X J, Liang X D. Vision-language navigation with self-supervised auxiliary reasoning tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 10009−10019

45 Xia Q L, Li X J, Li C Y, Bisk Y, Sui Z F, Gao J F, et al. Multi-view learning for vision-and-language navigation [Online], available, https://arxiv.org/abs/2003.00857, March 2, 2020

46 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: 2015.

47 Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduled sampling for sequence prediction with recurrent Neural networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: 2015. 1171−1179

48 Wang X, Xiong W H, Wang H M, Wang W Y. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: 2018. 38−55

49 Wang H, Wu Q, Shen C H. Soft expert reward learning for vision-and-language navigation. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 126−141

50 Lamb A, Goyal A, Zhang Y, Zhang S Z, Courville A, Bengio Y. Professor forcing: A new algorithm for training recurrent networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: 2016. 4601−4609

51 Ranzato M, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: 2016

52 Li X J, Li C Y, Xia Q L, Bisk Y, Celikyilmaz A, Gao J F, et al. Robust navigation with language pretraining and stochastic sampling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: 2019. 1494−1499

53 Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistic. Minneapolis, USA: 2018. 4171−4186

54 Hao W T, Li C Y, Li X J, Carin L, Gao J F. Towards learning a generic agent for vision-and-language navigation via pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 13134− 13143

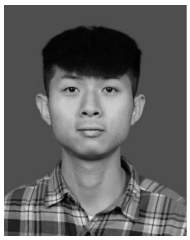
55 Huang J T, Huang B, Zhu L Q, Ma L Y, Liu J, Zeng G H, et al. Real-time vision-language-navigation based on a lite pre-training model. In: Proceedings of the International Conferences on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data and IEEE Congress on Cybermatics. Rhodes, Greece: IEEE, 2020. 399−404

56 Majumdar A, Shrivastava A, Lee S, Anderson P, Parikh D, Batra D. Improving vision-and-language navigation with image-text pairs from the web. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: 2020. 259−274

57 Hong Y C, Wu Q, Qi Y K, Rodriguez-Opazo C, Gould S. Vln bert: A recurrent vision-and-language bert for navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 1643−1653

58 Chen S Z, Guhur P, Schmid C, Laptev I. History aware multimodal transformer for vision-and-language navigation. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: 2021. 5834−5847

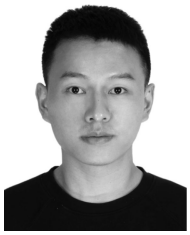
59 Ilharco G, Jain V, Ku A, Ie E, Baldridge J. General evaluation for instruction conditioned navigation using dynamic time warping. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems Workshops. Vancouver, Canada: 2019.



司马双霖 中国科学院自动化研究所智能感知与计算研究中心硕士研究生. 2020 年获郑州大学学士学位. 主要研究方向为视觉语言导航和具身智能.
E-mail: shuanglin.sima@cripac.ia.ac.cn
(**SIMA Shuang-Lin** Master student at the Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Zhengzhou University in 2020. His research interest covers vision-and-language navigation and embodied artificial intelligence.)

黄 岩 中国科学院自动化研究所智能感知与计算研究中心副研究员. 2017 年获中国科学院自动化研究所博士学位. 主要研究方向为计算机视觉和跨模态数据分析.
E-mail: yhuang@nlpr.ia.ac.cn
(**HUANG Yan** Associate professor at the Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2017. His research interest covers computer vision and cross-modal data analysis.)

何科技 中国科学院自动化研究所智能感知与计算研究中心博士研究生. 2019 年获南京邮电大学学士学位. 主要研究方向为视觉语言多模态和机器人. E-mail: keji.he@cripac.ia.ac.cn
(**HE Ke-Ji** Ph.D. candidate at the Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Nanjing University of Posts and Telecommunications in 2019. His research interest covers vision-and-language multi-modality and robot.)



安 东 中国科学院自动化研究所智能感知与计算研究中心博士研究生. 2019 年获北京大学学士学位. 主要研究方向为计算机视觉和具身智能.
E-mail: andong2019@ia.ac.cn
(**AN Dong** Ph.D. candidate at the Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Peking University in 2019. His research interest covers computer vision and embodied artificial intelligence.)

袁 辉 中国科学院自动化研究所智能感知与计算研究中心机器人算法工程师. 2021 年获湘潭大学硕士学位. 主要研究方向为视觉语言理解和机器人导航.
E-mail: hui.yuan@cripac.ia.ac.cn
(**YUAN Hui** Engineer at the Center of Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. He received his master degree from Xiangtan University in 2021. His research interest covers vision-and-language comprehension and robot navigation.)

王 亮 中国科学院自动化研究所研究员. 主要研究方向为模式识别, 计算机视觉, 机器学习和数据挖掘. 本文通信作者.
E-mail: wangliang@nlpr.ia.ac.cn
(**WANG Liang** Professor at Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, computer vision, machine learning and data mining. Corresponding author of this paper.)