# **视觉语言导航综述**

## 1. VLN研究背景

一个真正具有智能的robot，应该使用自然语言和人沟通，而不仅是接收机器指令。最近，随着BERT、GPT等大模型的发展，大家开始意识到可以把自然语言和传统的视觉导航融合到一起，形成视觉语言导航（Vision-and-Language Navigation），简称VLN。

VLN是具身型人工智能（embodied AI），它将vision、text和action三种模态融合到一起，并且可以运用到实际生活中。具身型人工智能需要理解来自不同维度的信息，例如vision、text、audio甚至是knowledge base或symbolic language。推进embodied AI的研究可以从以下几个方面来做：首先提出合理的benchmark，这种benchmark带有真实的环境来测试智能体。进一步地，需要提出模型和方法来构造相关的VLN agent。VLN agent需要实时接收环境的观测画面以及来自人的指令，从而在环境中移动来完成任务。

## 2. VLN任务定义

VLN具有三个要素，即Environment、Agent、Oracle。

Environment就是VLN agent运行需要的实际环境。在创建数据集的时候，需要有一个虚拟环境取测试模型，甚至需要打造一个真实环境。对于Environment的要求，第一是尽可能真实，其次是3D环境让agent自由探索。Agent是我们需要打造的可以处理实际任务的模型。Oracle是能够给Agent提供自然语言指令指导的智能体，可以理解为人类。

Agent需要和Environment交互。Agent不选移动产生新的观察，新的观察会指导Agent的下一步动作。VLN还要求Agent与Oracle之间进行自然语言的交互。此外，Oracle和Environment也要有交互。简而言之，VLN要求Oracle和Agent互相合作，在Environment中完成移动任务。

## 3. VLN数据集

目前VLN研究者提出了多个数据集，这些数据集从不同角度反映了VLN的核心问题。这些数据集根据两个维度划分，第一个是Communication Complexity（交互复杂度），第二个是Task Objective（任务目标）。

Communication Complexity对agent的语言能力有要求，而Task Objective表示任务本身的复杂程度。

先介绍Communication Complexity。分为三个层次：

1）第一个层次是最低的Initial Instruction：Oracle一开始的时候给agent一句Instruction，告诉它要去完成什么任务，这是最简单的一种。

2）第二种是Oracle Guidance，即在接收到初始的任务之后，例如find the apple，如果agent感觉到困惑，它找了很久仍然找不到苹果，这时候，我们告诉它应该向右转。Agent需要有后续继续理解自然语言的能力。

3）第三种是最复杂的，要求它不仅能够后续理解，还需要主动用自然语言去问问题，就是Dialogue的形式，也会更像人类一样。

Task Objective也分为三个层次：

1）第一个是Fine-grained，即开始给agent的指令是非常详细的，它只要按照给指令迁移就可以完成任务。

2）第二个是Coarse-grained。这里不会给agent详细的指令，只给它一个很高层次的目标来自己去找到这个物体。例如告诉它find the mirror in the bedroom，这种情况下它就需要自己去找到bedroom在哪。

3）第三个是Navigate and Object Interaction。为了完成任务，agent不仅仅需要Navigate，也需要做Object Interaction。例如我们要求agent去厨房里拿一些切开的苹果，但是它没有找到切开的苹果，只找到了完整的苹果，这时候，光靠Navigation是没有办法完成这个任务的，它需要拿一把刀，去执行和环境交互的动作，把苹果切开。

接下来介绍几个重要的DataSet。

1）Room-to-Room（R2R）。R2R给一个很长详细的Instruction，然后agent可以完全根据详细的Instruction去完成这个任务。R2R现在也是经常打榜的Benchmark。

2）TOUCHDOWN是第一个户外的VLN数据集，也是目前研究者最常测试的户外VLN数据集。目前几乎所有的户外数据集都是基于Google Street View，因为谷歌有丰富的街景，方便研究者提取使用。

3）REVERIE会让agent初始化在房间当中的任一个位置，但是并不会给出Agent的具体路线，所以agent需要对房间有更强地理解能力，甚至需要一些Prior Knowledge。例如，需要知道卧室一般在什么位置，如果需要去找刀子，一般就是在厨房，如果需要找台灯，可能在卧室。

4）Alfred也是非常火的一个Dataset，它是第一个结合了Interaction和Navigation的数据集。

5）VNLA & HANNA是Oracle Guidance的代表数据集。

6）CVDN：Agent会在解决Navigation Task时主动用自然语言问问题，这对Agent技能有更高的语言能力要求，需要agent能够主动去识别什么时候问问题，应该问什么样的问题。

7）TEACh：要求Agent有主动问问题的能力，还要求能和环境去交互。

## 4. VLN方法

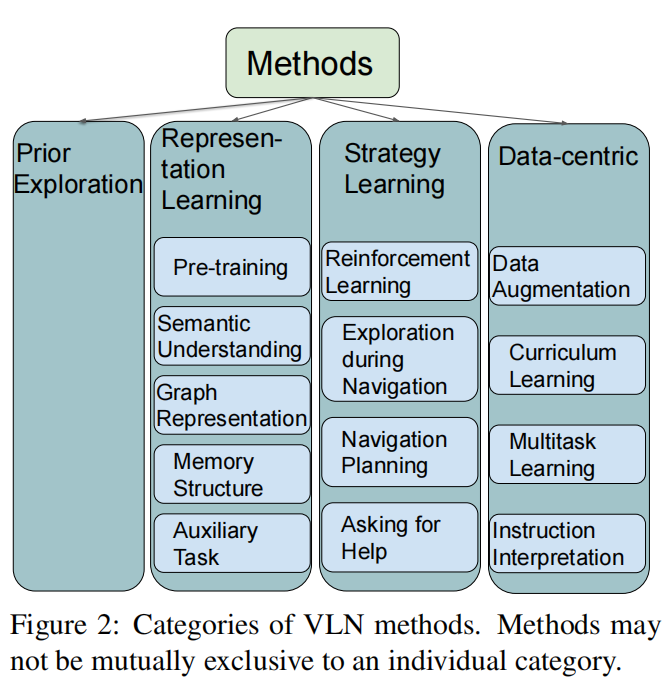
根据下图的这个框架，我们把方法划分为4类方法：Prior Exploration，Representation Learning，Strategy Learning，Data Centric Learning。

### 4.1 Prior Exploration

提前探索环境可以提升性能。例如，智能体提前对环境进行SLAM或三维重建形成地图，之后在这个地图的基础上进行路径规划和探索，会取得很好的性能，但这不是目前主流VLN研究的重点。

### 4.2 Representation Learning

Representation Learning可以帮助Agent获得更好的对信息的理解。



**（1）Pre-training**

Pre-training可以在多个维度使用。例如在language上使用预训练好的BERT，多模态可以使用ViLBERT，等。

**（2）Semantic Understanding**

Semantic Understanding是从语义的角度提升对输入的理解。分为两个方向， 1）Intra-Modality，从模态内部去提升，发现模态内部有什么规律。例如从画面中抓取重要的地点信息，进而指导agent下一步该往哪走。

2）Inter-Modality是两个模态之间的信息，例如说，在画面发现了楼梯、厨房、桌子这些重要的地点信息的同时，在语言中也发现了相关的名词词汇，把它们匹配起来。

**（3）Graph Representation**

VLN涉及很多物体、地点和动作之间的显性信息，如果把这些信息用一个图来表示出来，那么就可以提供一些指导。另一个方向是建造一个迁移图，agent在环境探索的过程中，每走一步，就发现新的画面，用图把看到的重要信息都记录下来，也可以提升性能。

**（4）Memory Structure**

可以使用一个额外的模块来处理迁移历史信息。VLN了涉及太多的信息，例如过去的所有Action、对话交互，以及所看到的图像。

**（5）Auxiliary Task**

VLN Task拥有很多额外信息可以帮助agent更好地理解自身和环境，这些信息都可以用来作为辅助任务。

### 4.3 Strategy Learning

Strategy Learning，即理解了输入信息之后，如何去做一个正确的决策。Strategy Learning主要方法是利用一些VLN 相关的特征来设计方法。例如在迁移当中如何去探索，如何主动问问题，这些都是和这个任务本身紧密相关的。而之前说到Representation Learning，主要是如何积极地学习其他模态的信息，如何更好地理解输入。

**（1）Reinforcement Learning**

VLN是一个动态过程，给定初始状态，agent需要执行后面的很多个动作，直到任务成功或者失败。强化学习中非常重要的一个问题就是如何定义Reward，最基本的是使用现有的Evaluation Metric，以这样的metric来作为reward。另外，也可以专门训练一个模型得到reward。每走一步，这个模型会端到端的告诉我们当前action的reward。

**（2）Exploration during Navigation**

在Navigation过程中，agent可以探索多个角度，把周围环境都看了一圈之后，再决定 Agent究竟是往哪一步路移动比较好。

**（3）Navigation Planning**

提前计划后面的移动方式，plan可以从视觉和语言两个方向得到。例如从视觉方面，可以先预测画面当中有哪几个点是一定要走的地方，并依次迁移。

**（4）Asking For Help**

主动去向人来寻求帮助可以帮助agent走出困境。这里具体涉及两个能力，第一个，agent要知道什么时候去问问题，第二个是如何用自然语言问出问题。

### 4.4 Data-centric Learning

目前VLN设计的信息太多，包含视觉、语言、动作，这些信息都要求理解，进而输出一个动作。目前在这种数据极度稀缺的情况下，Data-centric Learning发挥的作用非常大。目前Data-centric Learning中主要方法是Data Augmentation。

**（1）Data Augmentation**

数据增强分为两种。一种是Trajectory-Instruction Augmentation，直接生成一条一条的路径，然后配上语言Instruction；另一种是Environment Augmentation，先增强环境，当有更多的虚拟环境之后，再在其中建造更多的路径。

**（2）Curriculum Learning**

VLN当中，先从简单的任务开始做，在简单的任务上训练，接着在复杂的任务上训练。我们可以使用迁移场景中的房屋数目或者自然语言指令长度来决定任务难度。

**（3）Multitask Learning**

因为不同的VLN Task可以相互帮助，实现起来也很简单，一个agent同时用在两个数据集上面，然后在这两个数据集上面同时去训练。

**（4）Instruction Interpretation**

Instruction Interpretation，是换一个角度来解释Instruction。例如把过长的Instruction拆解成很多细致的步骤，Agent就能更好地去理解并执行。

## 5. VLN评估

现有的评估方法可分为两种：

1）Goal-oriented Metrics，即目标是否完成，完成是1，否则为0，类似分类任务的正确率。

2）Path-fidelity Metrics，即要求agent所走路径与Ground Truth路径尽可能相似，通过相似度来评价。

## 6. VLN未来方向

未来方向有如下几个：

**1）协同导航**

当前的VLN主要集中于只有一个agent导航的任务，但是复杂的现实世界场景可能需要多个机器人协作。未来的工作可能针对多个agent之间的协作VLN。

**2）Simulation to Reality**

如何把在仿真器中建立的模型，大规模地在现实世界中实现。因为现实世界不像仿真中那么理想，现实世界是噪声的，图像是模糊或具有视差的。基于3D模型和真实图像的仿真可以改善虚拟传感器（仿真中）和真实传感器之间的匹配。另外，仿真没有考虑到其他动态因素，例如人行走或物体移动，也没有考虑到白天的照明条件。具有概率转移函数的VLN环境也可以缩小仿真与现实之间的差距。

**3）Ethics & Privacy**

VLN代理可能会观察并存储可能被泄露或滥用的敏感信息。

**4）多元化VLN**

VLN数据集缺乏多样性，不能完全代表现实世界。未来的工作可以探索跨多种文化和地区的更加多样化的环境。

## 参考文献(references)

Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. 2021. Neighbor-view enhanced model for vision and language navigation. arXiv preprint arXiv:2107.07201.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. 2020. Sim-to-real transfer for vision-andlanguage navigation. In Conference on Robot Learning (CoRL).

Shurjo Banerjee, Jesse Thomason, and Jason J. Corso. 2020. The RobotSlang Benchmark: Dialog-guided robot localization and navigation. In Conference on Robot Learning (CoRL).

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. 2020. Learning to map natural language instructions to physical quadcopter control using simulated flflight. In Proceedings of the Conference on Robot Learning, volume 100 of Proceedings of Machine Learning Research, pages 1415–1438. PMLR.

Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. 2021a. Topological planning with transformers for vision-and-language navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11276–11286.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021b. History aware multimodal transformer for vision-and-language navigation. arXiv preprint arXiv:2110.13309.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In AAAI Conference on Artifificial Intelligence.

Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. 2020. Evolving graphical planner: Contextual global planning for vision-and-language navigation. Advances in Neural Information Processing Systems, 2020-December.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. arXiv preprint arXiv:2202.13330.

Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain pretraining for vision-and-language navigation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1634–1643.