# VLN BERT: A Recurrent Vision-and-Language BERT for Navigation

Yicong Hong Qi Wu Yuankai Qi Cristian Rodriguez-Opazo Stephen Gould

**摘要**

许多视觉语言学任务的准确性已经显著受益于视觉和语言（V&L）BERT 的应用。然而，它在视觉和语言导航（VLN）任务上的应用仍然有限。其中一个原因是，BERT 架构难以适应 VLN 中存在的部分可观察的马尔可夫决策过程，这需要与历史相关的关注和决策。在本文中，我们提出了一个用于解决 VLN 任务的循环 BERT 模型。具体地说，我们为 BERT 模型配备了一个循环函数，该函数可以维护代理的跨模态状态信息。通过对 R2R 数据集的实验，我们证明了我们的模型可以实现最先进的结果。

**关键词：**VLN；BERT；多模态

## 1 引言

人工智能领域一直有一个目标：让机器人根据人类的指令执行任务。而让机器人根据人类的指令导航 (也就是到达指令指定的位置) 是其中重要的一步。本论文研究的是 Vision-and-Language Navigation (VLN) 就是导航任务的一种，在这种任务中，机器人需要根据视觉和文本信息进行导航。最近许多方法使用大量图像-文本对 Transformer 进行预训练，学习通用的跨模态表示，这些方法一般被叫做 V&L BERT，这些方法启发了本论文的工作。本论文希望使用这种预训练过的 V&L BERT[1] 用来处理下游的 Vision-and-Language Navigation 任务。由于 Vision-and-Language Navigation 任务与 Vision-and-Language 的任务不同，它需要用到历史信息来辅助当前的动作决策，因为本文设计了一种循环结构：用一个 state 记录历史信息，使用 state 来辅助动作决策，每一步循环使用并更新这个 state。

## 2 相关工作

### 2.1 MatterPort3D 模拟器和 R2R 数据集

MatterPort3D 模拟器[2]是一个强化学习的模拟环境，它基于 MatterPort3D 数据集。其中，MatterPort3D 数据集是目前最大的 RGBD 数据集，它是从许多房间的许多位置中拍摄的。MatterPort3D 模拟器使用了这些数据集，根据拍摄的点位做了许多导航图，如图 1 所示，让机器人可以在导航图的顶点中移动，并给机器人提供每个顶点对应的全景图。R2R 数据集是有许多指令-路径对组成的数据集，机器人可以利用该数据集，在 MatterPort3D 模拟器中训练自己完成 VLN 任务。

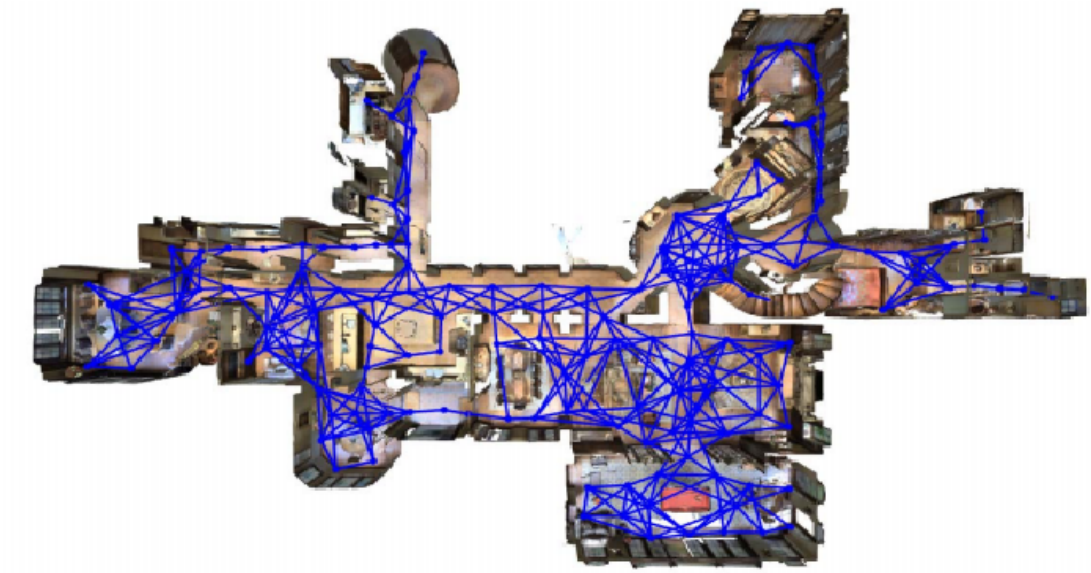

图 1: 方法示意图

## 2.2 传统做 VLN 的方法

对于 VLN 任务，传统的方法使用 seq2seq 结构的模型[2]来做导航，如图 2 所示，首先会输入一串文本指令，随后文本指令被编码成一个隐状态，在每个时间步中，模型会被输入当前机器人所在位置的全景图，并输出一个动作，让机器人前往下一个位置，如此类推。最后，当机器人认为它已经到达目的地的时候，可以输出一个“stop”的动作结束导航。训练的时候通常使用监督学习 + 强化学习，这可以让机器人在学习到正确路径的同时充分地探索地图，学习到更多的信息。

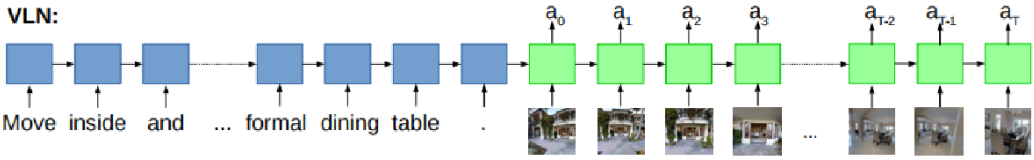


图 2: 方法示意图

## 2.3 BERT

BERT[3]是谷歌在 2018 年提出的一个模型，我们可以把它视作是一个 transformer 的编码器，如图 3 所示。BERT 最初用在自然语言处理中，它的输入是一串文本，输出是这串文本对应的编码。后来也被应用到了计算机视觉中，因此出现了多模态的 V&L BERT[1]，它可以接受混合模态的输入，把视觉和语言当作同一种模态进行编码，可以解决很多多模态的问题。

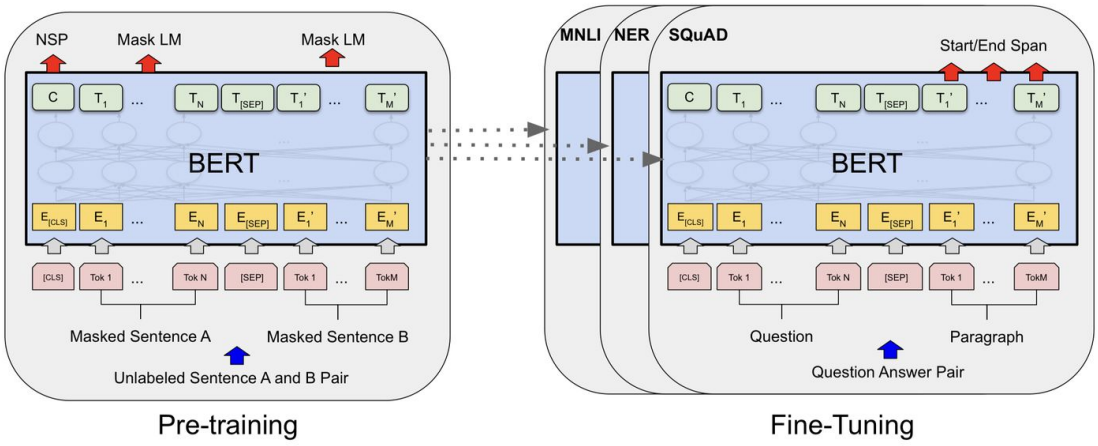


图 3: 方法示意图

# 3 本文方法

## 3.1 本文方法概述

此部分对本文将要复现的工作进行概述，图的插入如图 4所示：如图 4 所示，本文提出的方法主要有三个部分组成，分别是语言部分，视觉部分，状态部分。

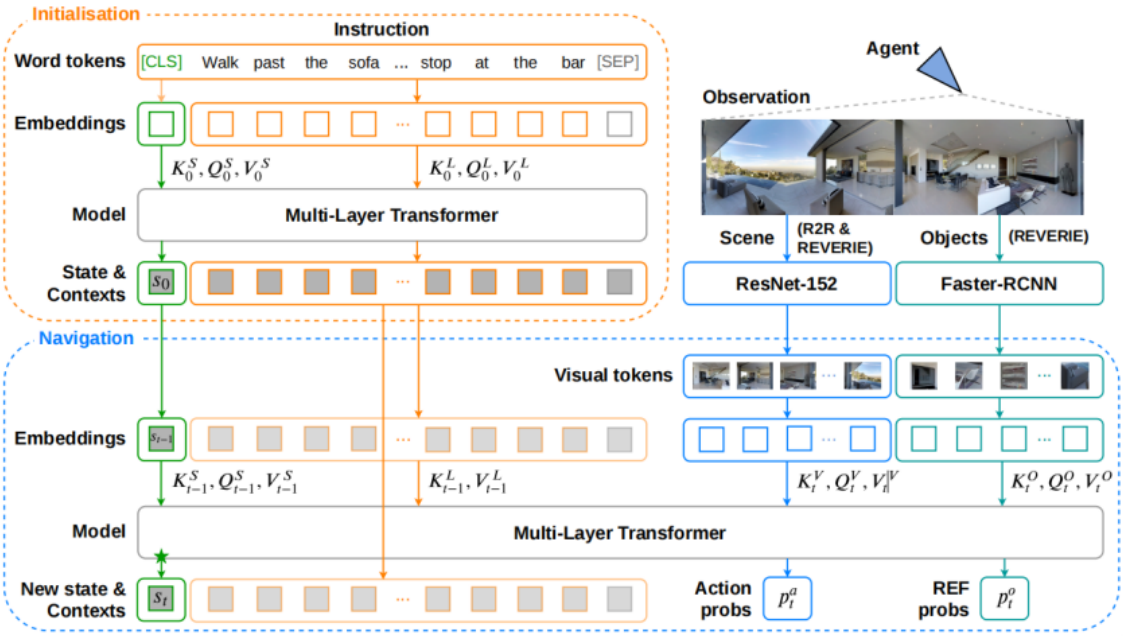


图 4: 方法示意图

## 3.2 语言部分

在初始化阶段，使用 BERT 对输入的语言指令进行编码，随后在导航阶段，该编码会被一直复用，不会更新。原因是如果在每个时间步都更新语言编码，会消耗大量的计算资源，其次语言指令的语义信息不受时间影响，在任何时间步应该都是一样的，因此在 VLNBERT 的设置中，语言编码不需要在导航阶段更新。

## 3.3 视觉部分

在导航阶段，使用预训练好的 ResNet 提取视觉图像的编码，并使用一个投影器投影到 VLNBERT 的输入维度中，再和语言编码拼在一起输入 VLNBERT 中。在每个时间步中，模拟器会给机器人提供其所在位置的全景视觉图像，因此图像编码在每个时间步都会更新。

### 3.4 状态部分

VLNBERT 维护一个状态向量，用于记录历史信息辅助决策。在初始阶段，使用 [cls] 作为状态，再导航阶段，状态会收集指令信息，视觉信息，动作信息，并更新自己。再下一个时间步中继续循环使用。

### 3.5 动作决策

在每个时间步中，VLNBERT 需要输出当前的动作决策，它会使用状态向量与每个视角的视觉图像编码求一个相似度，并把相似度作为选择该方向前进的概率，最后 sample 一个方向前进。

### 3.6 训练

使用监督训练 + 强化学习的方式训练。根据每个动作是否缩短了机器人与目的地的距离来评估奖励的大小，同时使用 A2C 的方式训练。

## 4 复现细节

### 4.1 与已有开源代码对比

本论文提供了代码，因此下面的结果是在源代码的基础上改进的。在跑通源码的基础上，本次实验还做了另外一种组合的尝试，就是用 GPT[4]来做 VLN 任务。在自然语言处理中，一直存在两个劲敌，分别是 BERT 和 GPT，它们一个取自 transformer 的编码器，一个取自 transformer 的解码器。它们都是预训练模型，BERT 的预训练做的是理解的，对齐的任务，GPT 的预训练做的是推理的，生成的任务。在多模态领域，由于对齐的多模态数据是比较好找的，因此现在的多模态模型一般都使用对齐的预训练任务，而这与 BERT 的特性比较匹配，因此出现了 V&L BERT。但是 VLN 比起对齐任务，它更偏向一个推理任务，因此它的本质是根据语言和视觉推理出下一步的动作。然而，由于推理任务的多模态数据不太好收集，因此基本没有出现专门使用推理任务做预训练的多模态模型。因此本次实验尝试使用单模态 (语言) 预训练的 GPT 来解决 VLN 任务，观察其效果，具体流程如下图所示：

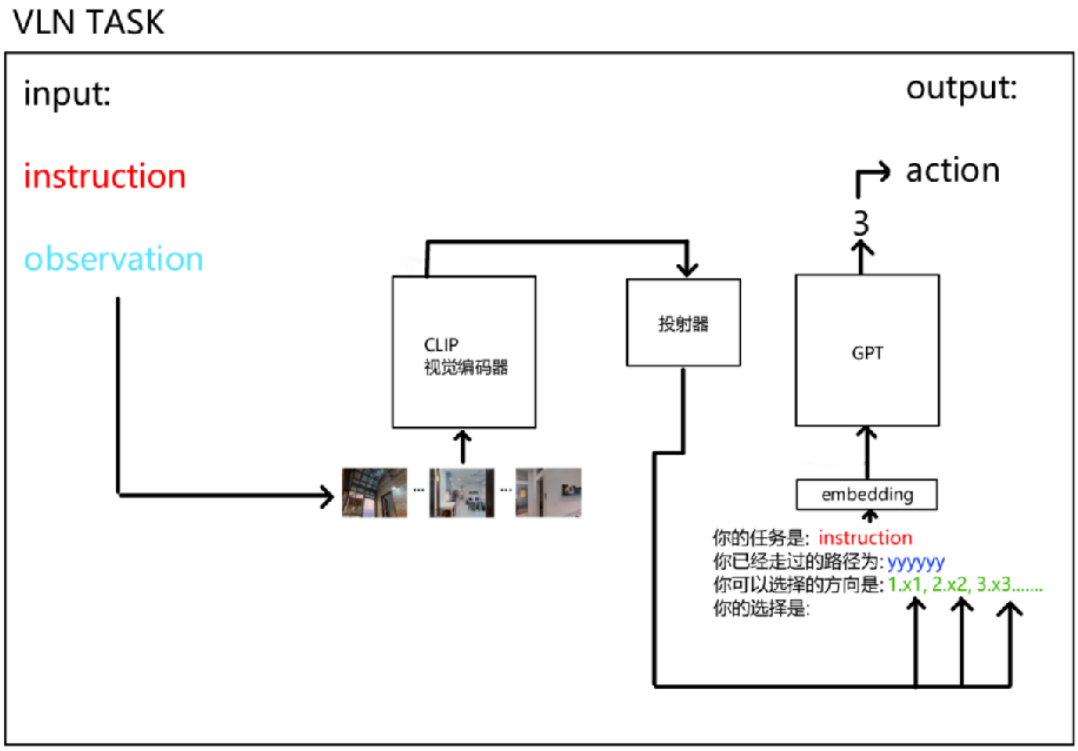


图 5: 方法示意图

首先会输入语言指令与视觉图像，图像经过 clip 编码[5]，投射到 GPT 的输入空间中，把语言指令和图像编码填入模板里面，然后一起输入 GPT，让 GPT 去生成一个数字，这个数字就表示机器人所选择的前进方向。使用监督学习的方式训练。

如需伪代码，采用如下的写作方式进行描述

### 4.2 实验环境搭建

首先配置 MatterPort3D 模拟环境：

1. 下载 MatterPort3D 模拟环境项目提供的 docker 环境；
2. 下载 MatterPort3D 数据集，放到项目指示中指定的位置，根据提示设置数据集路径；
3. 进入 docker 容器，对 MatterPort3D 项目进行编译；
4. 对数据集进行预处理，并运行测试文件，测试通过即配置成功。

然后配置 VLNBERT 项目：

1. 通过 git clone 下载 VLNBERT 项目；
2. 下载 R2R 数据集以及其他所需要的数据放到对应的文件夹下；
3. 下载预训练模型权重放到对应的文件夹下；
4. 运行测试文件，测试通过即配置成功。

随后对 VLNBERT 做出自己的改进：

1. 把 VLNBERT 的架构改成如图 5 所示的架构；
2. 使用监督学习训练模型；
3. 测试模型效果，记录并与原模型对比。

### 4.3 导航过程示意

每次导航的过程如下图所示：

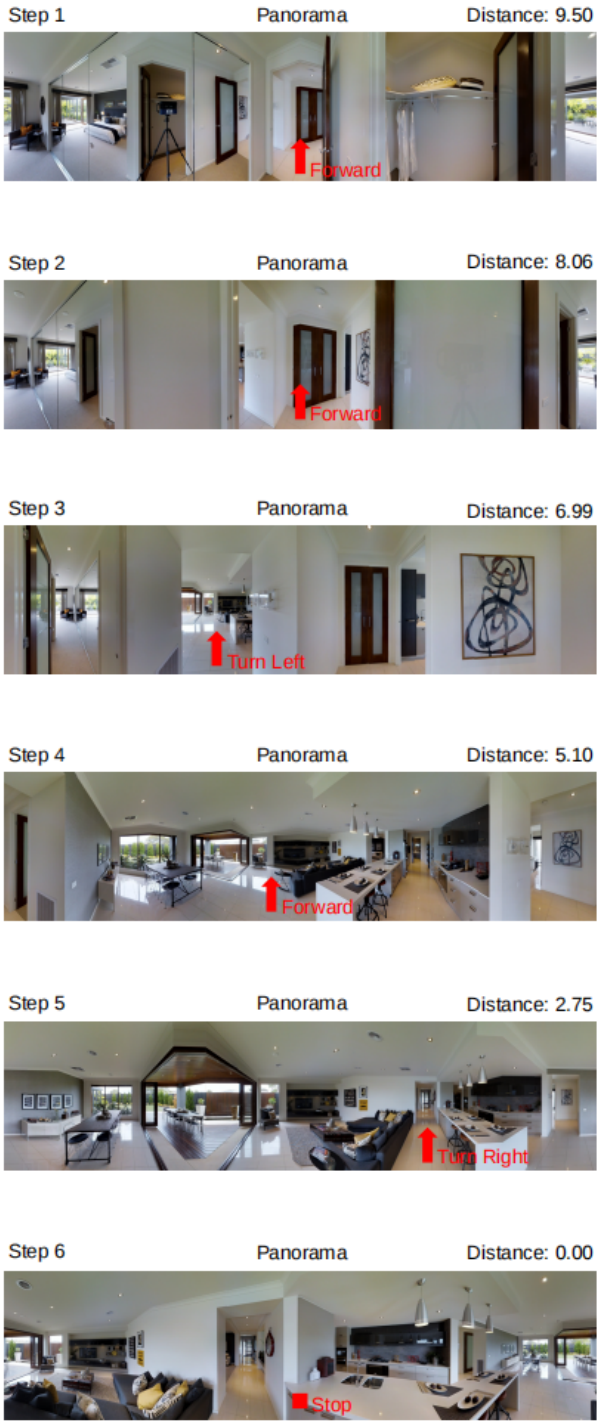


图 6: 导航过程示意图

### 4.4 创新点

尝试使用 GPT 来解决 VLN 问题，使用单模态预训练的模型来解决多模态任务。

## 5 实验结果分析

VLNBERT 和 VLNGPT 的实验结果如下图所示：

| Method | R2R Validation Seen | | | | R2R Validation Unseen | | | |
|---|---|---|---|---|---|---|---|---|
| | TL ↓ | NE ↓ | SR ↑ | SPL ↑ | TL ↓ | NE ↓ | SR ↑ | SPL ↑ |
| VLNBERT | 11.03 | 2.9 | 72 | 68 | 12.01 | 3.93 | 63 | 57 |
| VLNGPT | 12.06 | 4.8 | 54 | 49 | 13.97 | 6.59 | 38 | 30 |

图 7: 实验结果示意

可见，使用单模态预训练的 GPT 执行 VLN 任务的效果远远不如 VLNBERT。

## 6 总结与展望

本次实验尝试了利用单模态预训练的 GPT 模型，使用监督学习，完成 VLN 任务。结果完全比不上使用多模态预训练，使用强化学习 + 监督学习的 VLNBERT。我认为理由有如下几点：

1. 单模态和多模态的数据还是有着很大的鸿沟，即使使用投射器把视觉模态投射到语言模态空间上，也未必能表达出确切的语义，并让模型完全理解。

2.GPT 模型的预训练数据来自于互联网上的文本，但是互联网上的文本与 VLN 所使用的指令文本也有很大的差距，指令文本通常会反映一些空间信息，带有一些指代，引用的表达等等，这在预训练数据中很少出现。

3. 只使用监督学习的效果不如强化学习。

这次实验反映了，对于多模态的一些复杂任务，单模态模型很难解决，虽然有像 ClipCap 这样的先例，但是它解决的问题是根据图像生成文本描述这样相对简单的多模态问题。对于导航问题来说，它与单模态的预训练数据差距太大，以后还是得依靠多模态模型。

## 参考文献


[1] HONG Y, WU Q, QI Y, et al. Vln bert: A recurrent vision-and-language bert for navigation[J]., 2021: 1643-1653.

[2] ANDERSON P, WU Q, TENEY D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[J]., 2018: 3674-3683.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[4] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[5] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.