Exploring the effectiveness of Small-Scale Vision and Language Models for Vision and Language Navigation Tasks in Continuous Environments

Wesley Chiu, Abdulrahman Altahhan

University of Leeds, School of Computing, ODL MSc in AI, UK.

Abstract. Vision-and-Language Navigation (VLN) is a rapidly evolving field of research that aims to enable an embodied agent to follow textual instructions given in natural language to navigate through an unseen environment to a goal position. Existing approaches to this task fall into two main categories: "specialist" models that have been constructed and trained specifically to solve this task, and zero-shot or few-shot models that aim to leverage the implicit knowledge within Large Language Models (LLMs) and Vision Language Models (VLMs). Prior work in the latter approach have used the most powerful models (70B+ parameters). This work aims to evaluate the suitability of using a lighter-weight variant of an existing open-sourced model (Qwen2-VL) as the primary VLM, which would allow it to be run "on-device" rather than being connected to a broader network. The experiments in a simulated environment demonstrates that the current ability of the lighter-weight model is not yet fit for purpose, failing to reach the success rates seen approaches that use the full-sized variants.

Keywords: Vision-and-Language Navigation, Vision Language Model, Large Language Model, Prompt Engineering, Prompting

1 Introduction

This is the introduction to your project. This is nice intro. In [4]

2 Literature Review

[2] has established a new TD(0) method that replays all past experiences. On the hand, [3] has taken this further to include a target that incorporates all past updates via TD(λ). [1] applied conjugate gradient update on TD.

3 Methodology

In this section, we lay out the methodology. Note how in Fig. 2 we have shown the boundary.

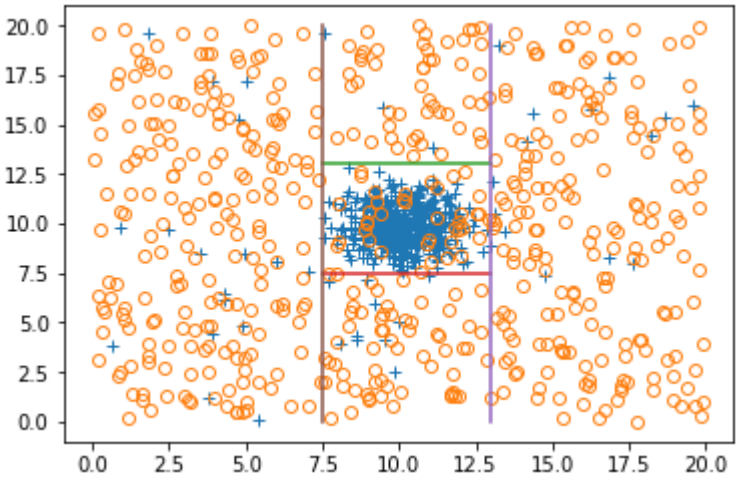

Fig. 1. Caption

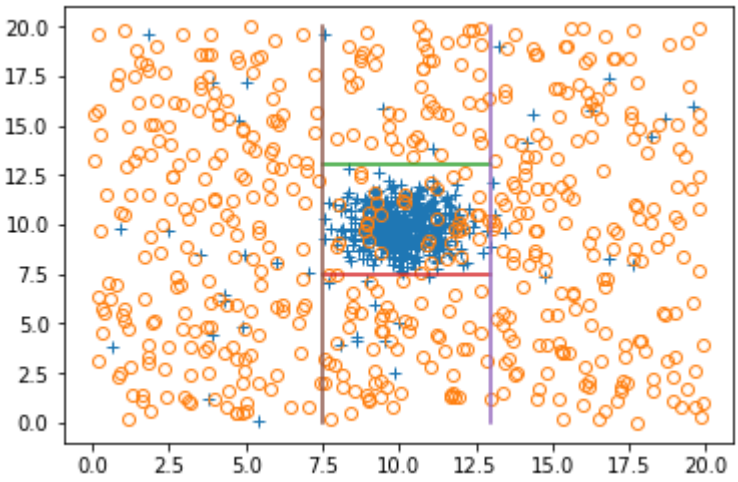

Fig. 2. Caption

In Fig. 2 that $\alpha = n^2$. In eq. (1) we have shown that the $1 + 2 + 3 + 4 = 4 \times 5/2 = 10$.

$$\sum_{i=1}^{n} y = 10$$
$$M = \beta^2$$
$$\boldsymbol{B}^\top = \boldsymbol{\Lambda}^2$$

$$\sum_{i=1}^{n} i = \frac{(n+1)n}{2} \tag{1}$$
$$\beta^2 = \alpha_i^n$$

4 Experiment Results

4.1 Experiment 1

Exp Note that the figure and the tables might be laid out on another page. Do not worry about that, and do not attempt to change it. Leave this to Latex.

Table 1. This is a table

Year	World	Duration
8000 B.C.	5,000,000	10
50 A.D.	200,000,000	20
1650 A.D.	500,000,000	30

5 Conclusion and Future Work

We have conducted a study on ...

References

[1] Abdulrahman Altahhan. "Robot visual homing using conjugate gradient Temporal Difference learning, radial basis features and a whole image measure". In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 2010, pp. 1–10. DOI: 10.1109/IJCNN.2010.5596784.

[2] Abdulrahman Altahhan. "TD(0)-Replay: An Efficient Model-Free Planning with full Replay". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018, pp. 1–7. DOI: 10.1109/IJCNN.2018.8489300.

[3] Abdulrahman Altahhan. "True Online TD(λ)-Replay An Efficient Model-free Planning with Full Replay". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9206608.

[4] Amine Naimi et al. "Dynamic Neural Network-based System Identification of a Pressurized Water Reactor". In: *2020 8th International Conference on Control, Mechatronics and Automation (ICCMA)*. 2020, pp. 100–104. DOI: 10.1109/ICCMA51325.2020.9301483.