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 ask all 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

Vision-and-Language Navigation (VLN) is a rapidly evolving field of research, which tasks an embodied agent to follow a set of textual instructions given in natural language to navigate a previously unseen 3D indoor environment. Whilst initial research leveraged models such as RNNs and CNNs to process visual and natural language data, the introduction of Transformer-based models [10] has accelerated capabilities in Natural Language Processing (NLP) and Computer Vision (CV) - both key components in the VLN task. The rapid development of Large Language Models (LLMs) and Vision-Language Models (VLMs) has further intensified the volume and pace of research in VLN, with each advance in LLM and VLM competency directly benefiting the VLN task. The traditional VLN task, as proposed by [2], has the agent navigating between pre-defined nodes in the environment (sometimes referred to as a navigation graph), which essentially reduces the VLN task to a vision-based graph-search problem; models trained in this manner have difficulty translating to performance in the real-world

[1]. To help combat this gap, the VLN Continuous Environment (VLN-CE) task was introduced, and has no such predefined navigation points, allowing agents to take any action to reach a navigable point in the environment; this introduces new challenges for the models to overcome, such as avoiding getting stuck on obstacles and dealing with distances [7].
Approaches to the VLN task (whether traditional VLN or VLN-CE) that leverage existing LLMs and VLMs typically fall into two broad categories: zero-shot approaches or 'specialist' models. Zero-shot models [8, 3, 11] aim to leverage the implicit knowledge, NLP ability, and strong reasoning skills within LLMs and VLMs [**CITATION˙NEEDED**] to navigate the environment. These models rely on specific prompting to generate historical trajectories, make navigation decisions, and to monitor progress. However, these models typically rely on translating visual data into purely textual information, which risks a loss of information in the process [9]; as a result, these models typically underperform against specialist models. The strength of these models lie in the fact that no pre-training is required, and as LLMs and VLMs continue to improve, this increase in competency can directly translate into improved VLN performance without the need for pre-training or adaptation of a complex model structure.
Specialist models [6, 12, 4, 5] are those that either take an off-the-shelf LLM or VLM and finetune the model for the VLN task, or are models (usually consisting of multiple sub-models) specifically finetuned and constructed to tackle the VLN task. These models exhibit much improved performance in VLN tasks, but cannot take advantage of more competent LLMs or VLMs as they are released.
Against this background, this work proposes to use a finetuned open-source model as its primary VLM, using QLoRA [**citatation˙required**] to finetune the very capable Qwen2-VL-7B variant [**citation˙required**].

2 Literature Review

[**TD0-Replay**] has established a new TD(0) method that replays all past experiences. On the hand, [**TD-Replay**] has taken this further to include a target that incorporates all past updates via TD(λ). [**ConjugateTD**] 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.

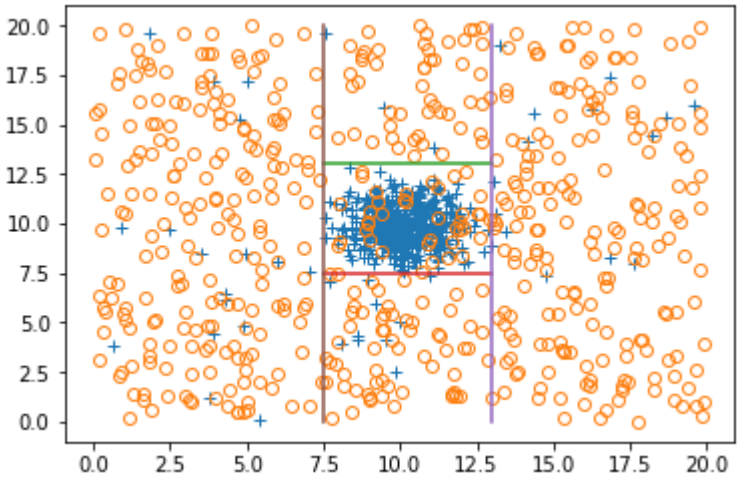

Fig. 1. Caption

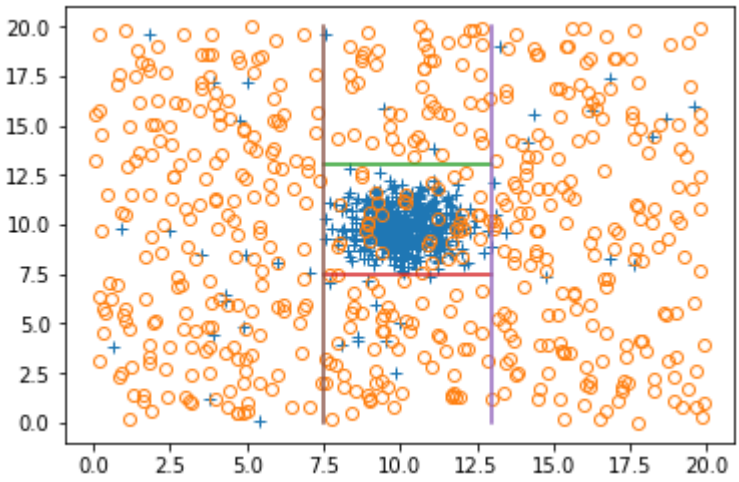

Fig. 2. Caption

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] Peter Anderson et al. “Sim-to-Real Transfer for Vision-and-Language Navigation”. In: *Proceedings of the 2020 Conference on Robot Learning*. Ed. by Jens Kober, Fabio Ramos, and Claire Tomlin. Vol. 155. Proceedings of Machine Learning Research. PMLR, 16–18 Nov 2021, pp. 671–681. URL: https://proceedings.mlr.press/v155/anderson21a.html.

[2] Peter Anderson et al. "Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3674–3683. DOI: 10.1109/CVPR.2018.00387.

[3] Jiaqi Chen et al. *MapGPT: Map-Guided Prompting with Adaptive Path Planning for Vision-and-Language Navigation*. 2024. arXiv: 2401.07314 [cs.AI]. URL: https://arxiv.org/abs/2401.07314.

[4] Shizhe Chen et al. "History aware multimodal transformer for vision-and-language navigation". In: *Advances in neural information processing systems* 34 (2021), pp. 5834–5847.

[5] Keji He et al. "Memory-Adaptive Vision-and-Language Navigation". In: *Pattern Recognition* 153 (2024), p. 110511. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2024.110511. URL: https://www.sciencedirect.com/science/article/pii/S0031320324002620.

[6] Yicong Hong et al. "Vln bert: A recurrent vision-and-language bert for navigation". In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2021, pp. 1643–1653.

[7] Jacob Krantz et al. *Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments*. 2020. arXiv: 2004.02857 [cs.CV]. URL: https://arxiv.org/abs/2004.02857.

[8] Yuxing Long et al. "Discuss before moving: Visual language navigation via multi-expert discussions". In: *arXiv preprint arXiv:2309.11382* (2023).

[9] Bowen Pan et al. *LangNav: Language as a Perceptual Representation for Navigation*. 2024. arXiv: 2310.07889 [cs.CV]. URL: https://arxiv.org/abs/2310.07889.

[10] Ashish Vaswani et al. *Attention is All you Need*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[11] Gengze Zhou, Yicong Hong, and Qi Wu. "NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 7. 2024, pp. 7641–7649. DOI: 10.1609/aaai.v38i7.28597.

[12] Gengze Zhou et al. "NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models". In: *Computer Vision – ECCV 2024*. Ed. by Aleš Leonardis et al. Cham: Springer Nature Switzerland, 2025, pp. 260–278. ISBN: 978-3-031-72667-5.