

Peer Review

Review of: "Fine-Grained Alignment in Vision-and-Language Navigation through Bayesian Optimization"

Wenguan Wang¹

1. Zhejiang University of Science and Technology, Hangzhou, China

Summary:

To better align language with visual sequence in vision-language-navigation, the paper proposes a Bayesian Optimization-based adversarial optimization framework. This method creates fine-grained vision negatives as data augmentation to enhance the performance of various VLN algorithms. Empirical experiments on R2R and REVERIE demonstrated the efficacy of the proposed framework.

Strengths:

1. This paper is well-written and highlights good results without grand claims.
2. The concept of the BO-based framework is straightforward, and the experimental results validate its utility in enhancing VLN task performance.

Weaknesses:

1. As shown in Figures 1 and 2, it is not clear what the differences are between negative samples and positive samples of vision. It would be better to directly demonstrate the comparison and pipeline process through the vision replacement proposed in Figures 1 or 2.
2. All equations are not tagged, and it is hard to find the references mentioned in the paper. For instance, Eq. (4) in Section IV.A.2.
3. Limited datasets. The method is only tested on the R2R and REVERIE, and there is a lack of discussion on its broader applicability, such as RxR or R4R.
4. How many negative samples are used in Tables II and III? Could you further provide more implementation details about the experiments in the paper?

5. The literature review is far from complete. Many famous VLN models are missing; the authors should review the following works and make comparisons in the experimental section.

[1] History Aware Multimodal Transformer for Vision-and-Language Navigation. NIPS 2022.

[2] Think Global, Act Local: Dual-Scale Graph Transformer for Vision-and-Language Navigation. CVPR 2022.

[3] Bird's-Eye-View Scene Graph for Vision-Language Navigation. ICCV 2023.

[4] Volumetric Environment Representation for Vision-Language Navigation. CVPR 2024.

Declarations

Potential competing interests: No potential competing interests to declare.