Landmark-RxR: Solving Vision-and-Language Navigation with Fine-Grained Alignment Supervision

许瑞

2023年11月7日

Solved questions

In this study, the authors address the challenge of cross-modal alignment at a fine-grained level. Firstly, to mitigate the limitations of weak cross-modal alignment supervision observed in coarse-grained data, a human-annotated, landmark-based, fine-grained Vision-and-Language Navigation (VLN) dataset is introduced, known as Landmark-RxR. Secondly, in order to further enhance local cross-modal alignment under fine-grained supervision, the research explores focal-oriented rewards in both soft and hard forms, with a particular focus on the crucial points selected from the fine-grained Landmark-RxR dataset. The focal-oriented rewards outperform the commonly used goal-oriented reward and fidelity-oriented reward. Furthermore, to provide a comprehensive evaluation of the navigation process, a re-initialization mechanism is proposed to render the metrics less susceptible to challenging points, which have the potential to lead the agent astray from the correct trajectories.

Direction

This dataset serves the purpose of mitigating the limitations associated with weak cross-modal alignment supervision, particularly in the context of coarse-grained data. The research experiments are meticulously designed to showcase how the combination of supervision from fine-grained and coarse-grained data can mutually reinforce the cross-modal alignment capabilities of the model. Furthermore, the study presents two distinct forms of focal-oriented rewards that utilize fine-grained supervision signals to enhance local cross-modal alignment. These rewards prioritize critical points, thereby facilitating the navigation process to a greater extent.

Method

In the described process, annotators were instructed to engage with the 3D environments based on comprehensive instructions and corresponding trajectories. Upon identifying landmarks within these environments, annotators were required to mark them and subsequently segment the corresponding sub-instructions utilizing the provided collection tool.