Long-term Vision Language Navigation Based on Actor-Critic for Recurrent Visual Attention

Junhua Liu

April 1, 2022

1 Introduction

Navigation is a fundamental and critical ability for agents to operate in real environments. Traditional navigation tasks mainly give point-goal, image-goal or object-goal directly. But it is not convenient for us to interact with robots. Visual Language Navigation(VLN) is a kind of new-style navigation. The intelligent agent follows the natural language instructions to navigate. This task needs to understand both the natural language instructions and the visible image information in the visual Angle, and then make corresponding actions to its own state in the environment, and finally reach the target position.

There has been a classic application of reinforcement learning method on computing vision, called the recurrent attention model. It's been developed in 2014 and has over 1700 papers to follow up. It is one of the few applications of reinforcement learning who has a deeper connection between them rather than straightforwardly implementing an existing reinforcement method under the context of a practical task.

The computation of neural network is very large, especially with the increase of resolution, the computation increases at least linearly. The nature of human vision is not to process the whole scene at once, but to extract key location information and then combine to construct the whole scene information. Therefore, Recurrent Visual Attention uses RNN to process the image sequence, obtain the key position of the image and extract the image information.

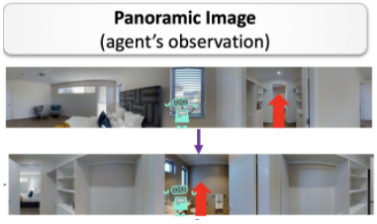

Figure 1: Panoramic image

VLN faces some new challenges compared to traditional visual-text multimodal tasks. First, agents are in a dynamically changing environment. They constantly observe new environments and make decisions. The observed environmental information and actions performed in the past are important for agents' next actions. Secondly, the observation of robts is a panoramic image on real world, a 360-view of image and a bigger field of view (FOV), which means the resolution and size of picture is big.[GSW+22]

2 Motivation

The existing monitoring data of training VLNS are relatively limited, which makes the model easy to overfit. In real life, the navigation of robot is long-term and the training data is huge. In seen and unseen environments, the existing model has worse performance compared to other environment in Room-to-Room dataset.

Previous study mainly uses a recurrent unit to turn all previous historical information into a fixed-size vector. But it is hard for model training to converge. It is also easy to lose key historical information during long distance navigation tasks.

Recurrent visual attention model can obtain the key position of the image and extract the image information.[MHG+14] The initial approach task was a direct application of policy gradient and recurrent neural networks. The former is known to solve reinforcement learning, while the later address the partial information setting. The vanilla method involves many drawbacks, most notably the delayed reward. The actions thus cannot be evaluated timely, causing significant increase in sample complexity and instability in training. Model training is difficult to converge.

One way to improve on this is the actor-critic framework, which is a very standard approach to continuous decision making. In an actor, it takes observations as input and outputs the action to be performed next. In the comments, it takes Take states as inputs and estimate the cumulative rewards that actors would have received if they had followed the same policy. [LWT+17]

So we propose a long-term vision language navigation based on Actor-Critic for recurrent visual attention. Considering the dynamic and unstable nature of environment and the huge computation, for reducing the amount of training data, it more suitable in the long-term nagivation.

Acknowledgments

This is an example of an unnumbered section.

References

[GSW+22] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

[LWT+17] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[MHG+14] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.