Review of paper – History Aware Multimodal Transformer for Vision-and-Language Navigation

Xuan Chen - Student #82286956

Paper link: https://arxiv.org/pdf/2110.13309.pdf

Summary of the paper (3-4 sentences)

HAMT consists of a unimodal transformer for text, history and observation encoding, and a multimodal transformer for capturing long-range correlations of historical sequences, where the history contains all previously observed sequences at the cost of high computational cost. To reduce the cost, this paper proposes a hierarchical visual transformer, which progressively learns the representation of individual vision, the spatial relationships between views in a panorama, and the temporal dynamics in a historical panorama. To learn better visual representations, this paper also proposes agent tasks for end-to-end training. These tasks include single-step action prediction based on imitation learning, self-supervised spatial relationship inference, masked language and image prediction, and instruction trajectory matching.

Main contributions (2-3 bullet points)

1. Introduction of HAMT to efficiently model long-horizon history of observed panoramas and actions via hierarchical vision transformer.
2. Trained HAMT with auxiliary proxy tasks in an end-to-end fashion and use RL to improve the navigation policy.
3. Validated the model and outperform state of the art in a diverse range of VLN tasks, which demonstrating larger gains for long-horizon navigation.

Positive and negative points (2-3 points each)

1. Positive:
   - HAMT achieves new state of art on a broad range of VLN tasks, including VLN with fine-grained instructions (R2R, RxR), high-level instructions (R2R-Last, REVERIE), dialogs(CVDN) as well as long-horizon VLN(R4R, R2R-Back).
   - HAMT is particularly effective for navigation tasks with longer trajectories.
2. Negative:
   - It improves the performance at the cost of high computation.

Unclear (2-3 points)

1. Is it better to training HAMT in 2 steps than training the whole model once directly?
2. Maybe try to implement experiments on SOON task (Scenario Oriented Object Navigation), in which agent navigates from an arbitrary position in a 3D embodied environment to localize a target object following an instruction.