A. Downstream tasks

A.1. Room-to-Room (R2R)

The task of R2R [1] requires agents to follow detailed instructions to navigate from one room to another. These instructions contain rich linguistic information, such as “Walk out of the room and down the hallway. Walk past the kitchen and stop outside of the door to the dining room”.

The R2R dataset is collected from Matterport3D simulator. It contains 21,567 navigation instructions, 7,189 trajectories and 10,800 panoramic views of 90 real-world building-scale indoor environments. The average length of each instruction is 29 words. The R2R dataset consists of four splits: train, validation seen and validation unseen, test unseen.

A.2. Room-Across-Room (RxR)

The task of RxR [4] is an updated version of R2R and is more challenging than R2R. For example, instructions in RxR are longer and more detailed, describing more landmarks than R2R does. Besides, instructions in RxR no longer describe the shortest path between the starting room to the ending room and the length variance of paths is very large. So, agents cannot simply go directly to the targets and cannot simply use the strong prior of path length to navigate.

The RxR dataset contains 126,069 navigation instructions, 16,522 trajectories. The average length of each instruction is 78 words. The instructions of RxR are in three language (i.e. English, Hindi, and Telugu). Since our pre-training instructions are in English, here we use English monolingual baseline.

A.3. REVERIE

The task of REVERIE [8] gives a concise, high-level instruction referring a remote object, such as “Close the kitchen window”. REVERIE requires the agent to follow instructions to navigate and identify the target object in previous unseen environment.

The REVERIE dataset contains 21,702 instructions. The average length of each instruction is 18 words. The dataset has 4,140 target objects, divided into 489 categories. On average, each target viewpoint has 7 objects with 50 bounding boxes.

A.4. Navigation from Dialog History (NDH)

In the task of NDH [11], the agent is required to find the target location based on the dialog history, which consists of multiple question-and-answer interactions between the agent and its partners. NDH is much more challenging because the instructions from the dialog history are often ambiguous and unspecified. As a result, agents can hardly navigate to the final location directly.

The CVDN dataset is used for NDH task, which contains 2050 human-human navigation dialog and over 7K trajectories. NDH has three settings: (1) Oracle, which utilizes the shortest path as ground truth observed by the Oracle; (2) Navigator, which uses the path adopted by human navigator as ground truth; (3) Mixed, which takes the shortest path or the path of human if the human visits the target location.

A.5. Evaluation Metrics

SPL Success weighted by Path Length trades-off SR (Success Rate) against TL (Trajectory Length).

nDTW Normalized dynamic timewarping penalizes deviations from the reference path.

sDTW Success weighted by normalized Dynamic Time-Warping, constrains nDTW to only successful episodes and effectively captures both success and fidelity.

CLS Coverage weighted by Length measures the overall correspondence between predicted and ground truth trajectories.
B. Comparison with Related Work

As shown in Table 1, we summarize the differences between our method HOP and related VLN pretraining methods, such as PREVALENT [3], Airbert [2] and VLN-BERT [7].

For visual inputs, we use trajectory instead of a static panoramic image of a single step. In addition to the common Mask Language Modeling (MLM) task and Trajectory-Instruction Matching (TIM) task, we propose two tasks to model temporal order information: Trajectory Ordering Modeling (TOM) and Group Ordering Modeling (GOM). Navigation information is enhanced by introducing the Action Prediction with History (APH) task. Finally, we conduct experiments on four downstream tasks to verify agents from different perspectives.

C. Results

As shown in Table 2, we add nDTW and CLS metrics for R2R results.

D. Limitations and Future work

Apart from the issue of computational resources, one other limitation of our work may be that we did not utilize masked image modeling tasks for VLN pre-training, such as Masked Region Classification (MRC) task, Masked Region Feature Regression (MRFR) task, etc. These proxy tasks might further improve the performance and generalization of our proposed pre-training methods, especially on downstream VLN tasks that require an agent to locate the target object such as REVERIE. In the future work, we will consider designing more effective masked image modeling tasks for VLN pre-training.

References


