HOP: History-and-Order Aware Pre-training for Vision-and-Language Navigation – Supplementary Material

Yanyuan Qiao[1] Yuankai Qi[1] Yicong Hong[2] Zheng Yu[1] Peng Wang[3] Qi Wu[1*]
[1]The University of Adelaide [2]The Australian National University
[3]Northwestern Polytechnical University
{yanyuan.qiao,qi.wu01}@adelaide.edu.au, {qykshr,william.zhengyu}@gmail.com
yicong.hong@anu.edu.au, peng.wang@nwpu.edu.cn

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 pretraining 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.

[*] Corresponding author

	HOP (proposed)	**PREVALENT [3]**	**VLN-BERT [7]**	**Airbert [2]**
Dataset	Augmented R2R dataset Processed BnB dataset	Augmented R2R dataset	Conceptual Captions [9] Wikipedia and BookCorpus R2R dataset	Conceptual Captions [9] BnB dataset
Visual Input	Trajectory	Panoramic view (single step)	Trajectory	Trajectory
Objectives	Masked Language Modeling Action Prediction with History Trajectory-Instruction Matching Trajectory Order Modeling Group Order Modeling	Masked Language Modeling Action Prediction	Masked Language Modeling Image-Caption matching Trajectory-Instruction matching	Masked Language Modeling Image-Caption matching Trajectory-Instruction matching (shuffling loss)
Downstream task	R2R, REVERIE, NDH, RxR	R2R, NDH, HANNA	R2R	R2R, REVEIRIE

Table 1. Comparison with related works.

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.

Methods	R2R Validation Seen		R2R Validation Unseen	
	nDTW ↑	CLS	nDTW ↑	CLS
Self-Monitoring [6]	65.4	64.1	43.7	41.5
EnvDrop [10]	67.2	67.2	56.7	57.0
Syntax [5]	70.0	70.0	59.0	59.0
HOP	**76.1**	**73.3**	**65.8**	**64.5**

Table 2. Performance of HOP on R2R. HOP denotes finetuned model pre-trained on data of both PREVALENT and our processed data from BnB.

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

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 1

[2] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *ICCV*, pages 1634–1643, 2021. 2

[3] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, pages 13134–13143, 2020. 2

[4] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, pages 4392–4412, 2020. 1

[5] Jialu Li, Hao Tan, and Mohit Bansal. Improving cross-modal alignment in vision language navigation via syntactic information. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *NAACL*, pages 1041–1050, 2021. 2

[6] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *ICLR*, 2019. 2

[7] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *ECCV*, pages 259–274, 2020. 2

[8] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den

Hengel. REVERIE: remote embodied visual referring expression in real indoor environments. In *CVPR*, pages 9979–9988, 2020. 1

[9] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 2

[10] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL-HLT*, pages 2610–2621, 2019. 2

[11] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *CoRL*, pages 394–406, 2019. 1