Boosting Outdoor Vision-and-Language Navigation with On-the-route Objects

Yanjun Sun[1,2], Yue Qiu[2], Yoshimitsu Aoki[1], Hirokatsu Kataoka[2]
[1]Keio University
[2]National Institute of Advanced Industrial Science and Technology(AIST)
sunyanjun@aoki-medialab.jp, aoki@elec.keio.ac.jp, {qiu.yue, hirokatsu.kataoka}@aist.go.jp

Abstract

Outdoor Vision-and-Language Navigation (VLN) is a challenging task that requires an agent to navigate using real-world urban environment data and natural language instructions. Current outdoor VLN models tend to overlook crucial navigation roles, such as objects that serve as landmarks for accurate turn and stop locations. This occurs because they primarily focus on panoramas and instructions, while disregarding objects that provide essential information for accurate decisions, such as identifying correct turn and stop locations, which humans naturally use as landmarks in unfamiliar places. In this paper, we propose the Object-Attention VLN (OAVLN) model, inspired by human navigation, which focuses on relevant on-the-route objects. Our model outperforms previous methods across all evaluation metrics on two benchmark datasets, Touchdown and map2seq.

1. Introduction

Enabling a robot to navigate real-world environments using natural language instructions has been a longstanding goal in AI research. The vision-and-language navigation (VLN) field has proposed various ways to achieve this [1, 4, 8, 18].

Recent outdoor VLN models [3, 4, 17, 19, 21] concatenate instruction and panoramic features and to predict a sequence of actions. However, these models lack the ability to learn specific semantics and ignore objects, tend to learn data biases and have a poor understanding of the environment. In a preliminary experiment, we examined the attention of existing methods on navigation texts by plotting a heatmap of attention weights for instructions. We discovered that outdoor VLN agents insufficiently focus on object tokens, causing erroneous turns and stops. The average attention weight of object tokens from the test set was only 0.128, highlighting the insufficient attention given to object tokens compared to other components of the instructions. Furthermore, DiagnoseVLN [20] found that outdoor VLN agents prefer to use directional information and ignore objects from the instructions, which is counterintuitive to humans who use landmarks like buildings and objects to navigate unfamiliar places [2]. Therefore, landmarks like buildings and objects are crucial for providing helpful clues in outdoor VLN.

Inspired by using landmarks to navigate unfamiliar places, we propose a simple yet effective Object-Attention VLN (OAVLN) model that allows the agent to focus more on objects described in the navigation instructions and better understand the environment. Our method leverages object information to enhance the agent's environmental awareness and improve navigation performance.

We extensively experimented with our OAVLN on the Touchdown [4] and map2seq [16] dataset, and comparing it with four outdoor VLN models [3, 4, 17, 21]. The experimental results demonstrate that our model outperforms existing methods by effectively utilizing objects as navigation landmarks and accurately guiding the agent to turn or stop at suitable locations.

2. Proposed Method: OAVLN

The proposed Object Attention VLN (OAVLN) model, as shown in Fig. 1, takes in four inputs: navigation instructions, panorama features, object features, and scene texts. At each decoding timestep, the model computes a panorama visual representation of the current agent state in the environment based on previously predicted actions. The first layer encodes metadata and visual representations, while the second layer encodes contextualized text to predict the next action.

The Object Attention VLN model consists of four encoders. The Instruction Encoder embeds and encodes navigation instructions using a bidirectional LSTM [5]. The Panorama and Object Encoders extract visual features from panoramas and objects, while the Scene Text Filter and Encoder handle scene text detection and recognition. To obtain higher-quality scene text from low-quality panorama images, we use the Object Encoder to detect the entire panorama and identify the 'sign' regions. Then, we apply scene text recognition only to these 'sign' regions using the MMOCR [9] model and the SAR [10] model for text recog-

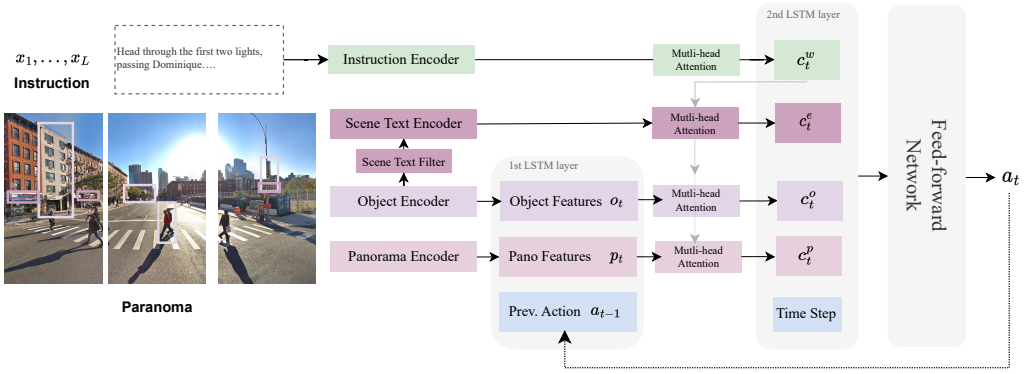

Figure 1. Overview of the proposed model.

Table 1. Navigation results on Touchdown and map2seq for the seen scenario.

Dataset	Touchdown						map2seq					
Model	TC↑	SPD↓	SED↑	CLS↑	nDTW↑	sDTW↑	TC↑	SPD↓	SED↑	CLS↑	nDTW↑	sDTW↑
RCONCAT	8.94	22.48	8.55	43.23	18.20	7.98	14.62	20.61	14.30	54.18	27.43	13.76
GA	9.87	20.34	9.42	47.77	21.51	8.92	17.88	18.25	17.55	58.56	31.46	17.08
VLN Transformer	14.90	21.20	14.60	45.40	25.30	14.00	17.00	-	-	-	29.50	-
ORAR	24.23	17.30	23.70	56.87	37.20	22.87	43.96	6.93	43.09	82.97	60.43	41.78
Ours (+scene text)	24.77	**15.98**	24.14	59.93	37.64	23.14	**50.00**	**6.11**	**49.04**	**84.77**	**65.39**	**47.45**
Ours (+objects)	**25.90**	16.04	**25.40**	**60.84**	**39.00**	**24.47**	49.00	6.40	48.08	84.28	63.38	46.75

nition. Finally, the Decoder predicts the agent's next action using multi-head attention and LSTM layers.

3. Experiments

We conduct experiments on Touchdown [4] and map2seq [16] datasets to assess OAVLN's performance on the outdoor VLN task.

3.1. Experimental Setup

Our framework, implemented in PyTorch [13], uses ResNet50 [6] for panorama features extraction and Faster R-CNN [15], pretrained on Visual Genome datasets [11], for object features extraction. MMOCR [9] recognizes text on signboards, and then we used stanza [14] to summarize the object tokens in the instructions, and to optimize the results of scene text recognition.

We compare our model with RCONCAT [4], GA [3], VLN-Transformer [21], and ORAR [17] on outdoor VLN. These baseline models were selected because they represent widely-accepted or state-of-the-art methods in the field. Our goal is to demonstrate how our proposed method, which focuses on on-the-route object features, can improve upon these existing models by addressing their limitations and achieving better performance in the outdoor VLN task.

The VLN performance was evaluated using six metrics: Task Completion (TC), Shortest-Path Distance (SPD) [4], Success weighted by Edit Distance (SED), Coverage weighted by Length Score (CLS) [7], Normalized Dynamic Time Warping (nDTW) [12], and Success-weighted Dynamic Time Warping (SDTW).

3.2. Experimental Results

We evaluated the impact of object features and scene text in our OAVLN model. Tab. 1 shows a comparison of our model with other studies. Our model outperforms baselines in each metric, highlighting the effectiveness of different datasets. The OAVLN(+scene text) and OAVLN(+objects) share the same structure but differ in the input information. Our model shows significant improvements in path alignment metrics (CLS, sDTW), indicating the effectiveness of object feature attention. The OAVLN(+scene text) model achieves a 6% improvement in goal-oriented metrics (TC, SED) on the map2seq dataset, indicating better object token utilization. The Touchdown boost was insignificant because the map2seq instructions were more focused on the objects. Our results demonstrate that using objects as references significantly improves the ability of our model to navigate effectively (turning and stopping) toward the goal, thereby enhancing its accuracy.

4. Conclusion

Our OAVLN leverages on-the-route objects to improve turning and stopping accuracy. This approach is more intuitive and allows the agent to use surrounding objects as references to reach the goal, even in unfamiliar places. Moreover, our work highlights the importance of leveraging contextual information, such as scene text, in navigation tasks. Our approach could serve as a starting point for future research in this area and inspire the development of more advanced models that better use the contextual information available in real-world environments.

References

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

[2] Edgar Chan, Oliver Baumann, Mark Bellgrove, and Jason Mattingley. From objects to landmarks: The function of visual location information in spatial navigation. *Frontiers in Psychology*, 3, 2012. 1

[3] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *AAAI*, 2018. 1, 2

[4] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, 2019. 1, 2

[5] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In WDuch, Janusz Kacprzyk, Erkki Oja, and SZadrożny, editors, *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, 2005. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2

[7] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*, 2019. 2

[8] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, 2020. 1

[9] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin. MMOCR: A comprehensive toolbox for text detection, recognition and understanding. *CoRR*, abs/2108.06543, 2021. 1, 2

[10] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8610–8617, 2019. 1

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[12] Gabriel Ilharco Magalhaes, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. In *NeurIPS Visually Grounded Interaction and Language (ViGIL) Workshop*, 2019. 2

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 2

[14] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *ACL*, 2020. 2

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[16] Raphael Schumann and Stefan Riezler. Generating landmark navigation instructions from maps as a graph-to-text problem. In *ACL*, 2021. 1, 2

[17] Raphael Schumann and Stefan Riezler. Analyzing generalization of vision and language navigation to unseen outdoor areas. In *ACL*, 2022. 1, 2

[18] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, 2020. 1

[19] Jiannan Xiang, Xin Wang, and William Yang Wang. Learning to stop: A simple yet effective approach to urban vision-language navigation. In *EMNLP*, 2020. 1

[20] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. In *NAACL*, 2022. 1

[21] Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Multimodal text style transfer for outdoor vision-and-language navigation. In *EACL*, 2021. 1, 2