Which way is 'right'?: Uncovering limitations of Vision-and-Language Navigation models

Extended Abstract

Meera Hahn
Google
United States
meerahahn@google.com

Amit Raj
Google
United States
amitrajs@google.com

James M. Rehg
Georgia Institute of Technology
United States
rehg@gatech.edu

ABSTRACT

The challenging task of Vision-and-Language Navigation (VLN) requires embodied agents to follow natural language instructions to reach a goal location or object (e.g. 'walk down the hallway and turn left at the piano'). For agents to complete this task successfully, they must be able to ground objects referenced into the instruction (e.g.'piano') into the visual scene as well as ground directional phrases (e.g.'turn left') into actions. In this work we ask the following question – to what degree are spatial and directional language cues informing the navigation model's decisions? We propose a series of simple masking experiments to inspect the model's reliance on different parts of the instruction. Surprisingly we uncover that certain top performing models rely only on the noun tokens of the instructions.

KEYWORDS

Instruction Following; Navigation; Embodied AI; CV; NLP

ACM Reference Format:
Meera Hahn, Amit Raj, and James M. Rehg. 2023. Which way is 'right'?: Uncovering limitations of Vision-and-Language Navigation models: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Vision Language Navigation (VLN) is the task of having a robot navigate a visual 3D environment via following human generated natural language instructions such as *'Leave the bathroom and walk to the right'*. VLN is a popular task with multiple benchmark datasets [2, 4, 8, 12, 14] which are largely conducted in simulated indoor environments such as Matterport3D (MP3D) [3] and contain thousands of human annotated instructions. MP3D is constructed of a set panoramic nodes connected via navigational ability to form a navigation graph. The task is challenging as it requires accurate visual grounding of objects and visual descriptions provided in the instructions into the environment. Furthermore it requires the model to understand spatial language to ground instructions such as *'walk to the right'* into actions. The action space at a given time step for the navigating agent is to move to a neighboring node or to cease navigation.

Solutions to the VLN task can be divided into two distinct settings: a discriminative path-ranking setting and a generative path selection setting. In the path-ranking setting, using beam search from the given starting location, up to 30 [5, 10] possible paths are generated and a path is selected using a discriminative path selection model. In the generative setting the agent is placed at the starting location and the model sequentially selects the next node to navigate to, until the model selects the stop action.

Both sequential and path-ranking settings have seen large success in modeling by using multi-modal transformer models which leverage large-scale pre-training [5, 7, 10]. The data used to pre-train these models consists of large scale web data [6, 9, 11, 13] containing image-text pairs to learn visual grounding as well as large text corpora [15] to learn linguistic semantics.

In many instances the episode's instructions refer to the spatial layout of objects in the environment and the agents position to these objects. In this paper we seek to understand to what degree the model is learning spatial and directional words and how these words impact the performance of the model. To this end we outline a simple method via token masking to understand how different types of part of speech and object vs direction tokens are used by path-ranking models. The experiments uncover path-ranking models rely almost exclusively on nouns and object tokens to make navigation decisions, while disregarding direction tokens and other parts of speech. This is a large limitation as the models are not utilizing large amounts of the available information to inform action prediction.

This abstract aims to provide answers to the following questions:

(1) Are there a specific set of linguistic cues in the instructions which more heavily inform navigational decisions?
(2) Can we measure the degree to which instruction following agents attend to the spatial and directional cues present in the instructions?

2 MASKING EXPERIMENTS

2.1 Methods

Ablation Experiment Design.
To answer these questions, we create a set of ablation experiments over the navigation instructions and evaluate on standard trained SOTA VLN models. Specifically we modify the navigational instructions by removing (via masking) or replacing tokens of a specific linguistic cue set which fall into a particular part of speech (POS) or if they are a object/spatial/numeric token.

By testing the models performance while it doesn't have access to a specific type of token, we gain insight into the degree to which that type of token informs the model's predictions. We examine 5 different masking criterion: nouns, verbs, adjectives, left-right,

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Instruction	Walk straight to the bar with the chairs. Turn left and go straight until you get to 3 tables with chairs. Turn left and wait at the couch.
Mask Directions	Walk [MASK] to the bar with the chairs. Turn [MASK] and go [MASK] until you get to 3 tables with chairs. Turn [MASK] and wait at the couch.
Mask Nouns	Walk straight to the [MASK] with the [MASK]. Turn left and go straight until you get to 3 [MASK] with chairs. [MASK] left and [MASK] at the [MASK].
Mask Objects	Walk straight to the bar with the [MASK]. Turn left and go straight until you get to 3 [MASK] with [MASK]. Turn left and wait at the [MASK].
Mask Numbers	Walk straight to the bar with the chairs. Turn left and go straight until you get to [MASK] tables with chairs. Turn left and wait at the couch.
Swap	Walk straight to the bar with the chairs. Turn right and go straight until you get to 3 tables with chairs. Turn left and wait at the couch.

Figure 1: Masking experiment visualization. Input tokens are masked out according to their cue set per ablation.

spatial, object, numerical. Via qualitative analysis over the instructions contained in the standard VLN benchmarks we select the following set of tokens as spatial cues: [*right, left, straight, toward, around, near, front, above, through, down, up, between, past*]. Note in the left-right masking experiment the tokens in the set [*left, right*] are masked out. We add an additional experiment called *swap* in which tokens in the set [*left, right*] are replaced by their antonym. Masking experiments are illustrated in Figure 1.

Selected Models.
We focus our investigation on four SOTA VLN models: VLN-BERT [10], AirBert [5], Recurrent-VLN-BERT [7], PREVALENT [6]. We take the models, trained in their standard practice, and evaluate them using the ablation experiments described above. VLN-BERT and AirBert use the path-ranking approach while Recurrent-VLN-BERT and PREVALENT use the discriminative approach. All four models employ multi-modal transformer architectures and utilize large-scale pre-training and data augmentation techniques. We measure VLN performance in terms of Success Rate (SR) which measures the percentage of selected paths that stop within 3m of the goal [1]. We perform the ablation experiments on the chosen VLN models over the val-unseen split of the R2R dataset. Results are shown in in Figure 2.

2.2 Results and Analysis

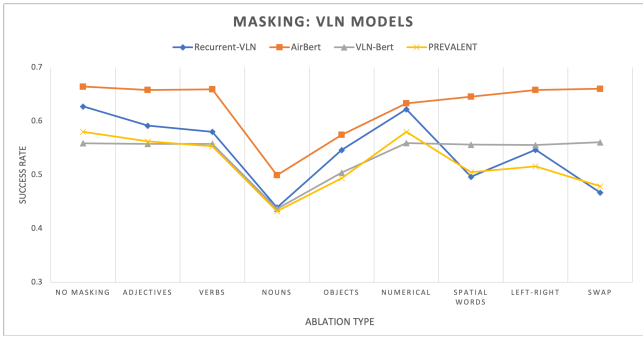

Figure 2: This figure displays the results of the masking experiments. The lines depict performance of each model and each column refers a different ablation experiment.

Path-Ranking models rely almost only on noun tokens.
Surprisingly in Figure 2, we observe that the performance of path-ranking models only suffers in the case that noun or object tokens are masked. Performance drops less than 2% for all other types of token masking. In fact we even observe an increase in performance for VLN-BERT by up to .0034% when the 'right' and 'left' tokens are swapped to their antonym, see the *swap* experiment. These results indicate the models heavily focus on object information while making navigation decisions and seem to ignore directional information. The disregard for spatial words is concerning as they are an integral components of the navigational instructions and replacements (swaps) in directional words should result in very different paths being taken by the agent.

Effects on Sequential Models.
In contrast to the path-ranking methodology, we find that VLN models trained and tested with the sequential procedures take into consideration multiple types of token sets. In Figure 2 we observe that the success rate of the Recurrent-VLN-BERT and PREVALENT model suffers under all masking conditions.

Path-Ranking procedure not sufficiently challenging.
We posit the difference between the path-ranking and sequential inference procedures is the key factor in the difference in results. The inference procedure in path-ranking VLN models allows access to entire navigation paths when predicting alignment with the navigation instruction. We hypothesize that this framework allows the model to do pattern matching across objects in the path and disregard extraneous information such as positional panoramic information of the path. In contrast the inference procedure in sequential VLN models is that the agents only have access to the immediate neighboring environment. This makes the sequential models more susceptible to cascading errors which will compound any drops in performance. Additionally as they are unable to see ahead in the path, they cannot use information about objects the instruction references that they have not seen yet to make navigational decision.

Note that this difference in task set up inherently puts sequential models at a disadvantage (in terms of success rate) compared to path-ranking models. For this reason the two types of models have rarely been compared in terms of accuracy or ablations. This novel comparison exposes shortcuts taken by path-ranking models which might suggest this procedure may need to be treated as a separate task, with a dataset tailored to be more challenging for this task. Based on the results observed in this study we posit that within the current instruction following datasets, especially those built upon the Matterport3D dataset like the one used in this study – the path-ranking VLN procedure is reducing the complexity of the task to such a degree that models can disregard most tokens in the instruction while still achieving a high success rate.

REFERENCES

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757* (2018).

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

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017).

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

[5] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain pretraining for vision-and-language navigation. *ICCV* (2021).

[6] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. *CVPR* (2020).

[7] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. *CVPR* (2021).

[8] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954* (2020).

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS* (2019).

[10] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. *ECCV* (2020).

[11] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Computer Vision–ECCV 2020: 16th European Conference*. Springer, Glasgow, UK, 336–352.

[12] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. *CVPR* (2020).

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *ACL* (2018).

[14] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-Dialog Navigation. *CoRL* (2019).

[15] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *ICCV* (2015), 19–27.