Supplementary for The Road to Know-Where: An Object-and-Room Informed Sequential BERT for Indoor Vision-Language Navigation

Yuankai Qi[1] Zizheng Pan[2] Yicong Hong[3] Ming-Hsuan Yang[4,5,6] Anton van den Hengel[1] Qi Wu[1]
[1]Australian Institute for Machine Learning, The University of Adelaide [2]Monash University
[3]The Australian National University [4]University of California, Merced [5]Google Research [6]Yonsei University
{qykshr, zizhpan}@gmail.com yicong.hong@anu.edu.au mhyang@ucmerced.edu
{anton.vandenhengel, qi.wu01}@adelaide.edu.au

1. Overview

In this supplementary, we present more details about the Reinforcement Learning used in our method and show more illustrations of the learned attentions by our transformers so as to better understand how the model make navigation decisions.

2. Reinforcement Learning

We adopt the same reward as [2] when training with Reinforcement Learning. Concretely, let $\mathbf{S}_t$ be the distance between the current location and the goal location at time step t. Then $\Delta\mathbf{S}_t = \mathbf{S}_{t-1} - \mathbf{S}_t$ represents the relative distance change compared to the previous step. During the navigation, we set the reward to 1.0 if $\Delta\mathbf{S}_t > 0$, which means the agent moves towards to the goal location; otherwise the reward is -1.0. At the last step, the final reward is set to 2.0 if $\mathbf{S}_t < 3.0$, which means the agent successfully arrives at the goal location; otherwise the final reward is -2.0. At the end of a navigation, the discounted reward as well as the estimated reward Z_t are used to perform the A2C algorithm [1].

3. Learned Attention by Transformers

Here we show more detailed attentions for the navigation step 6 in Figure 3 of our Main paper.

Our model has 12 transformer layers and each transformer has 12 heads. We observe that some heads are able to learn object-to-word, word-to-object, word-to-word, and object-to-object relationships. These attentions facilitate the model make its navigation action. Specifically, the blue rectangles in Figure 1 show object-to-word attentions, where attentions mainly focus on related words, such as "wait by the sink in the adjacent bathroom", which helps the model be aware of the current navigation progress. The green rectangle in the left panel of Figure 1 shows the word-to-object attentions. These attentions mainly focus on the ['drawer#sink#table'] object, which is closely related to the current goal room "bathroom". This indicates the model is able to match words to corresponding objects and to learn the co-occurrence of object and room. Figure 2 shows the word-to-word attentions. The left panel of Figure 2 shows that each word token has attentions on its following several words, indicating the model is aware of local textual context. The right panel of Figure 2 shows that the attentions of the current sub-instruction mainly distribute on the target "adjacent bathroom", indicating the model is able to tell the most important word token. Figure 3 shows the object-to-object attention. It shows that each object has attentions on other objects and the object that is the most related to the target room receives the most attention. This indicates the model is able to be aware of local context among objects and to filter out the most important one.

References

[1] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016. 1

[2] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL-HLT*, pages 2610–2621, 2019. 1

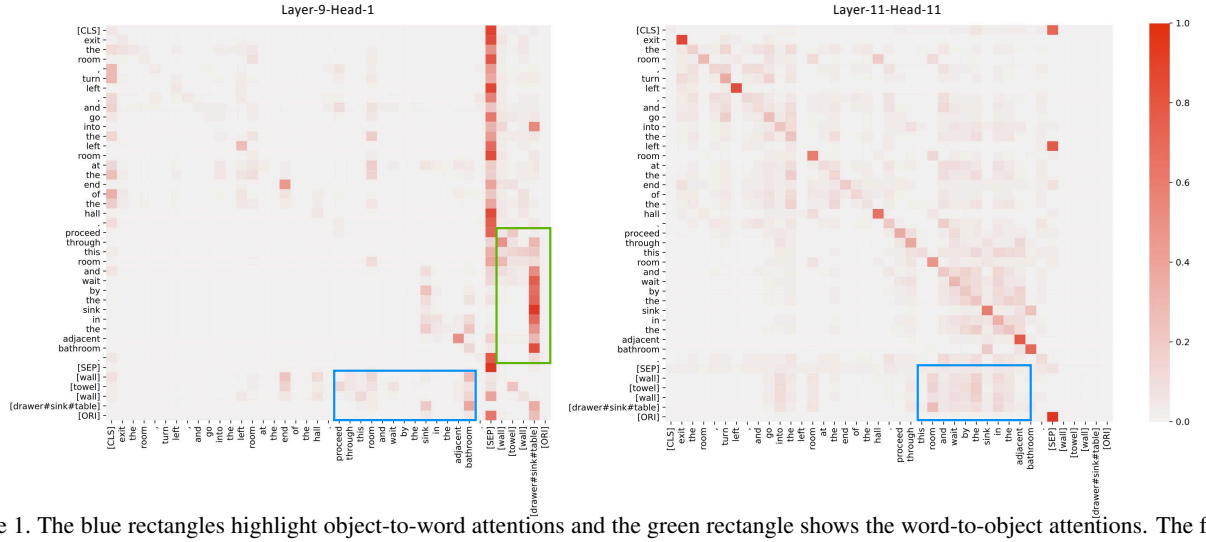

Figure 1. The blue rectangles highlight object-to-word attentions and the green rectangle shows the word-to-object attentions. The former helps the model to be aware of the navigation progress, and the latter helps determine the navigation action. Each row is normalised.

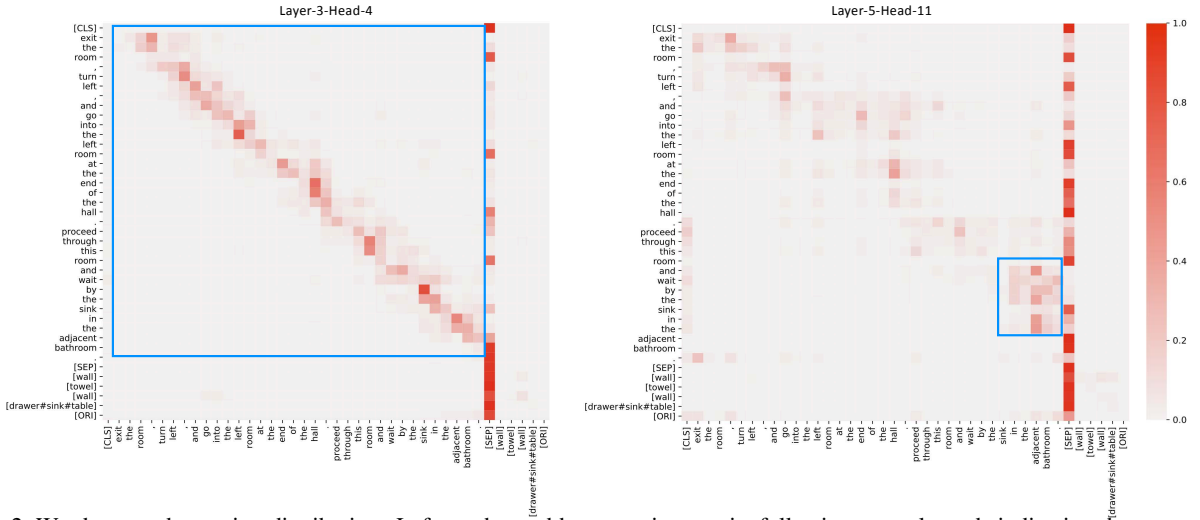

Figure 2. Word-to-word attention distribution. Left: each word has attentions on its following several words indicating the awareness of local context. Right: attentions have focused on the target room, indicating awareness of the current navigation goal.

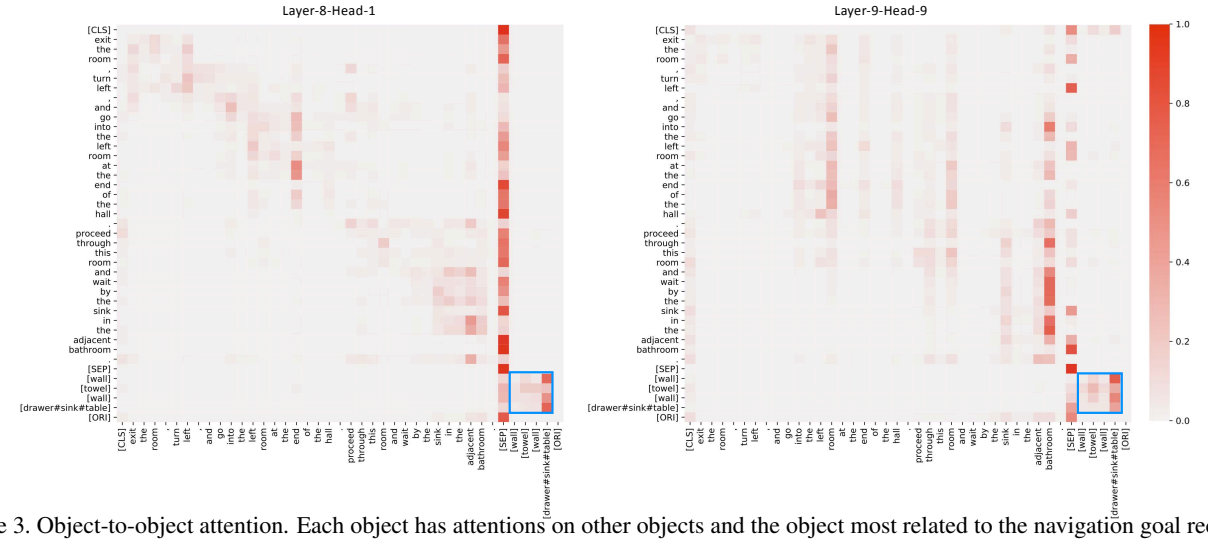

Figure 3. Object-to-object attention. Each object has attentions on other objects and the object most related to the navigation goal receives the most attention.