Technical Report for REVERIE Challenge (CSIG 2022)

Zun Wang Yi Wang Yinan He Yu Qiao
Shanghai AI Laboratory
{wangzun, wangyi}@pjlab.org.cn

Abstract

This report presents the methods of the team BPT in the REVERIE Challenge @ CSIG 2022. Our method improves the current state-of-the-art method, DUET without modifying the model architecture. Specifically, we combine environment-level and instruction-trajectory-pair-level augmentation in both pre-training and fine-tuning. Furthermore, we combine models trained with different augmented environments and pretraining weights for getting better result. Our best single (ensemble) model could achieve 40.20% SPL (43.72% SPL) and 24.47% RGSPL (28.36% RGSPL) on the validation set.

1. Method

1.1. Model

DUET Our model is based on the recent state-of-the-art agent, DUET [4], which maintains a topological map on-the-fly to enable efficient exploration and dynamically combines a fine-scale encoding over local observations and a coarse-scale encoding on the global map via graph transformers for navigation and object grounding. We refer to [4] for detailed model architecture.

Relative GPS sensor Note that in [4], the global GPS sensor is used (absolute world coordinates) for backtracking to a visited viewpoint. In this challenge, GPS isn't allowed so instead, we maintain relative coordinates (where the starting point is origin) for each episode. This could be easily computed using the global heading, elevation, and the relative distance between viewpoints and their neighbors.

1.2. Augmentation

Environment-level augmentation Many works [6–8] explore environment-level augmentation methods for Vision-and-Language Navigation (VLN) on fine-grained instruction-following datasets such as Room-to-Room [2] and Room-across-Room [5]. These methods could be easily applied in the REVERIE task. In our method, we directly use the edited environments in ENVEDIT [6] during training for augmenting environment. Specifically, when agent navigate to a viewpoint, we randomly pick one environment from the original and edited environment, and use the panorama of this environment at this viewpoint as agent's current observation. This helps avoid overfitting limited training environments.

Instruction-trajectory-pair-level augmentation. Augmenting instruction-trajectory pairs has been studied well in previous work. We perform two kinds of instruction-trajectory-pair-level augmentation: (1) We directly use the speaker from [4]. Instead of only using the speaker in pretraining, we empirically found that it improves model performance when fine-tuning with environment-level augmentation. (2) The high-level instructions of the reverie task only describe the surrounding of the target object, thus agent should be able to find the object with different starting points. Therefore, we could augment the training pairs by randomly selecting starting points of an episode. We apply this random-starting-point strategy in pre-training to avoid overfitting the training trajectories. However, we don't apply it in fine-tuning, since we found that this improves agent performance in val-seen split but harms val-unseen even with environmental augmentations.

1.3. Ensemble

We select models trained with different augmented environments, and different pretraining weights. In ensemble, we directly take the average probabilities of object grounding and action prediction from these models.

1.4. Training

Pretraining We use similar pretraining strategy to [4]. Specifically, we use masked language modeling (MLM) [28], masked region classification (MRC), single-step action prediction (SAP) and object grounding (OG) if object annotations are available.

Table 1. Our results.

Methods	Val Seen						Val Unseen					
	Navigation				Grounding		Navigation				Grounding	
	TL	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Ours-Single	13.74	81.58	76.52	67.35	54.48	48.60	21.09	66.04	56.48	40.19	34.20	24.47
Ours-Ensemble	12.48	80.85	78.76	72.24	58.96	54.30	19.04	66.41	58.03	43.72	37.58	28.36

Fine-tuning Similar to [4], agent is fine-tuned with Imitation-learning (pseudo interactive demonstrator (PID) and teacher-forcing). We use cross entropy for learning object grounding. Note that DUET [4] uses visible object as the signal determining whether agent navigates successfully. In the second version, agent could see object 3m away, but we still use 'navigating to the target objects within 3m' as the signal of successful navigation while the GT of object grounding isn't restricted by 3m.

2. Experiments

2.1. Datasets and evaluation metrics

The challenge uses the second version of the REVERIE datasets, which contains 21122 instruction-trajectory pairs. For object grounding, instead of only considering objects within 3 meters to the viewpoint like version 1, in this challenge ALL visible objects are considered for object grounding, making object grounding more challenging. Remote Grounding Succes penalized by Path Length (RGSPL) is the main evaluation metric which measures the efficiency for agent navigating to the object and recognizing it.

2.2. Implementation Details

Features. We adpot the end-to-end fine-tuned ViT-B/16 in [3] as the image-encoder of panoramic views in both original and edited environment. The object features are extracted from Faster-RCNN in BUTD [1]. The orientation feature contains $\sin(\cdot)$ and $\cos(\cdot)$ values for heading and elevation angles.

Model architecture Refer to [4] for detailed model architecture.

Training details We first pretrain the agent with batchsize of 32 on 2 NVIDIA RTX 3090 GPUs, and then fine-tune the agent with learning rate of 1e-5 and batchsize of 64 via AdamW optimizer on 4 NVIDIA Tesla A100 GPUs.

Challenge Submissions We use our full model for submission of channel 2 (own REG model). For channel 1 (pre-defined REG model), we still use our full model but just for navigation.

3. Conclusion

In this challenge, we found that augmentation methods and ensemble methods helps the REVERIE task. We believe general augmentation method for instuction-following tasks and Embodied AI is a good direction to explore.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1

[3] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. *arXiv preprint arXiv:2202.11742*, 2022. 1, 2

[5] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1

[6] Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417, 2022. 1

[7] Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1644–1654, 2021. 1

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