Technical Report of REVERIE Submission

Chen Gao[2], Jinyu Chen[1], Erli Meng[3], Liang Shi[3], Xiao Lu[3], Si Liu[1]
[1] School of Computer Science and Engineering, Beihang University
[2] Institute of Information Engineering, Chinese Academy of Sciences
[3] XiaoMi NLP Group

gaochen@iie.ac.cn, {chenjinyu,liusi}@buaa.edu.cn, mengerli@xiaomi.com,
shiliang1@xiaomi.com,luxiao3@xiaomi.com

Abstract

For our submission(team CoLabBUAA_MiNLP) of REVERIE[1] Challenge, we improve the baseline method proposed in [1] from two parts. Firstly, we introduce a distance-aware backtrace strategy to balance the navigation length and accuracy. We also find UNITER[2] finetuned on the REVERIE dataset have a strong ability to decide the navigation ending viewpoint. Secondly, we adopt the VilBERT[6] model that is pretrained on multiple vision/language task to replace MattNet[3]. Thus we produce more accurate referring prediction at the ending viewpoint of navigation. Our best results with multi-model fusion strategy achieves RGSPL of 12.96 on validation set and 21.14 on test set.

1 Navigator

At each time step, the baseline navigator FAST[4] predicts the actions probability distribution over all reachable viewpoints to form a confidence score for each action. When the navigator tries to backtrack, it will choose a viewpoint with the highest confidence score from the current untraversed set. In the unseen environment, since the predicted score is not always precise, the agent often wanders to some far and useless viewpoints, which will significantly increase the navigation length and reduce navigation efficiency.

From the above observation, we design a distance-aware backtracking strategy allowing the agent to be lazy. Specifically, we use the reciprocal of backtracking distance to weight the local action probability *logit,* as shown in (1). We adopt $logit'$ to replace original logit to represent the final probability of each action. Therefore the distance-aware strategy will make the navigator considering the distance cost when backtracking, which effectively limits the navigator to choose a far viewpoint for backtracking. The experiment results demonstrate that the navigation length is significantly reduced and the navigation success rate well kept.

$$logit' = \frac{logit}{distance^{w}} \quad (1)$$

Technically, we conduct a MLP after each input RoI feature of UNITER[2] base model to predict the matching score between the image region and the language instruction. We average the RoI score of multiple models to increase the robustness. The object with the max of the score above is always shown at the target viewpoint. We find that the sum of the matching score contains the same object is a very effective indicator for choosing the ending viewpoint of navigation, with which the navigation success rate can be improved significantly.

2 Pointer

ViLBERT shows state-of-the-art results in lots of vision-and-language tasks. Therefore, we adopt ViLVERT model that pretrained with 12-in-1 training strategy[5] to predict the target object at the ending viewpoint of each navigation path.

We use the output of ViLVERT behind each RoI feature as the matching score between the object in this region and instruction. We average the RoI score of the same object to increase robustness. The object with the highest matching score is considered as the final prediction. In order to get better RoI matching score, we also average the score of same RoI feature from three ViLVERT models in different training settings to get an ensembled matching score, which also improves the performance of referring expression comprehension.

3 Exprimental Setup

We train the FAST[4] model with the same training setting as the baseline model. The best backtracking weight w is 1.5.

We train ViLBERT[5] with the learning rate 2e-5, max region num 101. In three different training sets,

End decider	Backtrace weight	ViLbert	Vilbert ensemble	RGSPL	RGS	Nav-Succ	Nav-OSucc	Nav-SPL	Nav-length
				5.49	9.97	16.42	28.79	9.17	32.99
	√, $w = 1.5$			6.62	9.17	15.54	24.94	11.37	16.84
	√, $w = 1.5$	√		7.30	9.83	15.54	24.94	11.37	16.84
		√	√	7.92	10.71	15.54	24.94	11.37	16.84
	√ , w = 2	√		7.53	9.06	14.23	22.86	11.44	14.504
√		√	√	**12.96**	**17.72**	**28.00**	**34.45**	**20.22**	**12.31**

Table 1 Ablation study results.

the warmup proportion is set to 0.10, 0.05 and 0.08. The models are trained for 24, 9 and 35 epochs, respectively.

For UNITER[2], the best configuration for REVERIE finetuning is as follows. The learning rate is set to 5e-6, the warmup iteration is set to 8000, and the model is trained for 45000 iterations.

4 Albation study

We did experiments on the REVERIE[1] dataset unseen validation set. Results are shown on table 1. The result shows that weighting the distance of backtracking can reduce useless exploration, thus reduce the average of navigation length. And the navigation success rate keeps high, so the Nav-SPL increases.

ViLBERT [5] also shows better performance of referring comprehension than MattNet at the ending viewpoint of navigation. Multi-model fusion of different ViLBERT [5] models can also improve the performance.

UNITER[2] endpoint decider Significantly impoves navigation preformances on unseen environments.

5 Insights

Base on our expriments, the navigator performance is the bottleneck of this task. However, for now, the navigator cannot generalize to unseen environment well, the navigation results prove the meaningless wandering is existing. Reducing the meaningless backtracking is a useful way to improve the navigation performance.

Correctly choosing a viewpoint to stop is the key to improve the navigation success rate. Using pretrain model can improve the referring expression comprehension performance. Model ensemble can further enhance the model robustness and improve the referring success rate.

6 References

[1] Qi, Yuankai, et al. "REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments." Proceedings of the Conference on Computer Vision and Pattern Recognition. 2020.

[2] Chen, Yen-Chun, et al. "Uniter: Learning universal image-text representations." arXiv preprint arXiv:1909.11740 (2019).

[3] Yu, Licheng, et al. "Mattnet: Modular attention network for referring expression comprehension." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[4] Ke, Liyiming, et al. "Tactical rewind: Self-correction via backtracking in vision-and-language navigation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[5] Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in Neural Information Processing Systems. 2019.

[6] Lu, Jiasen, et al. "12-in-1: Multi-task vision and language representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.