Environment-agnostic Multitask Learning for Natural Language Grounded Navigation

Xin Eric Wang[1,2⋆ ⋆⋆], Vihan Jain[3⋆], Eugene Ie[3], William Yang Wang[2], Zornitsa Kozareva[3], and Sujith Ravi[3,4⋆⋆]

[1] University of California, Santa Cruz
[2] University of California, Santa Barbara
[3] Google Research
[4] Amazon

A Appendix

A.1 Multitask Learning with Total Training Paths Fixed

To verify whether multitask learning helps only due to implicit data augmentation which increases the number of training paths for both the tasks, we conducted an experiment by fixing the total number of training paths. We fix the size of training dataset to be exactly $4,742$ paths (which is the same as the number of paths in the NDH task's training dataset) and replace a fraction of the paths by the paths from the VLN task's training dataset. The results in Table 1 show that the agent's performance on previously unseen environments in NDH task improves significantly when trained jointly on NDH paths mixed with a small fraction of VLN paths. Since the total training paths are fixed, there is no benefit due to data augmentation which furthers the argument that multitask learning on NDH and VLN tasks complements the agent's learning. As expected, the agent's performance on NDH task degrades when trained on datasets containing smaller fractions of NDH paths but larger fractions of VLN paths.

Table 1: Comparison of agent's performance on NDH task when trained on fixed number of paths. The paths belong to either of the two tasks to support multitask learning.

	Fraction of VLN paths (%)						
	0	10	20	30	40	60	80
Progress (Val Seen)	6.49	6.21	5.69	5.72	5.82	5.74	3.66
Progress (Val Unseen)	2.64	3.13	3.09	3.31	2.80	2.86	2.48

⋆ Equal contribution.
⋆⋆ Work done at Google.

A.2 Detailed Ablation on Parameter Sharing of Language Encoder

Table 2 presents a more detailed analysis from Table 4 (main paper) with access to different parts of dialog history. The models with shared language encoder consistently outperform those with separate encoders.

Table 2: Comparison of agent's performance when language instructions are encoded by separate *vs.* shared encoder for VLN and NDH tasks.

Fold	Language Encoder	NDH Evaluation: Inputs for NDH (t_0 A_i Q_i $A_{1:i-1}; Q_{1:i-1}$)				NDH Evaluation: Progress ↑	VLN Evaluation: PL	NE ↓	SR ↑	SPL ↑	CLS ↑
Val Seen	Shared	✓				**3.00**	11.73	4.87	54.56	52.00	65.64
		✓	✓			**5.92**	11.12	**4.62**	54.89	**52.62**	66.05
		✓	✓	✓		**5.43**	10.94	4.59	54.23	52.06	66.93
		✓	✓	✓	✓	**5.28**	10.63	5.09	**56.42**	49.67	**68.28**
	Separate	✓				2.85	11.43	4.81	54.66	51.11	65.37
		✓	✓			4.90	11.92	4.92	53.64	49.79	61.49
		✓	✓	✓		5.07	11.34	4.76	55.34	51.59	65.52
		✓	✓	✓	✓	5.17	11.26	5.02	52.38	48.80	64.19
Val Unseen	Shared	✓				1.69	13.12	5.84	42.75	38.71	53.09
		✓	✓			**4.01**	11.06	5.88	42.98	40.62	54.30
		✓	✓	✓		**3.75**	11.08	5.70	44.50	39.67	54.95
		✓	✓	✓	✓	**4.36**	10.23	**5.31**	**46.20**	**44.19**	**54.99**
	Separate	✓				**1.79**	11.85	6.01	42.43	38.19	54.01
		✓	✓			3.66	12.59	5.97	43.45	38.62	53.49
		✓	✓	✓		3.51	12.23	5.89	44.40	39.54	54.55
		✓	✓	✓	✓	4.07	11.72	6.04	43.64	39.49	54.57

A.3 VLN Leaderboard Submission

Table 3 shows the performance of our *MT-RCM + EnvAg* agent on the test set of the R2R dataset for the VLN task. Our *MT-RCM + EnvAg* agent outperforms the comparable baseline *RCM* on the primary navigation metrics of SR and SPL which proves the effectiveness of multitask and environment-agnostic learning. It is worth noting that the baselines scoring high on the test set use additional techniques like data augmentation and pre-training which were not explored in this work but are complementary to our techniques of multitask learning and environment-agnostic learning.

Table 3: VLN Leaderboard results for R2R test dataset.

Model	**PL**	**NE** ↓	**SR** ↑	**SPL** ↑
Human	11.85	1.61	86	76
Random	9.93	9.77	13	12
Seq2Seq [1]	8.13	7.85	20	18
Look Before You Leap [10]	9.15	7.53	25	23
Speaker-Follower [2]	14.82	6.62	35	28
Self-Monitoring [6]	18.04	5.67	48	35
RCM [9]	11.97	6.12	43	38
The Regretful Agent [7]	13.69	5.69	48	40
ALTR [4]	10.27	5.49	48	45
Press [5]	10.77	5.49	49	45
EnvDrop [8]	11.66	5.23	51	47
PREVALENT [3]	10.51	5.30	54	51
MT-RCM + EnvAg (Ours)	13.35	6.03	45	40

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3683 (2018)
2. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: Neural Information Processing Systems (NeurIPS) (2018)
3. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
4. Huang, H., Jain, V., Mehta, H., Ku, A., Magalhães, G., Baldridge, J., Ie, E.: Transferable representation learning in vision-and-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision178(ICCV) (2019)
5. Li, X., Li, C., Xia, Q., Bisk, Y., Çelikyilmaz, A., Gao, J., Smith, N.A., Choi, Y.: Robust navigation with language pretraining and stochastic sampling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 1494–1499. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1159, https://doi.org/10.18653/v1/D19-1159
6. Ma, C.Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., Xiong, C.: Self-monitoring navigation agent via auxiliary progress estimation. arXiv preprint arXiv:1901.03035 (2019)
7. Ma, C.Y., Wu, Z., AlRegib, G., Xiong, C., Kira, Z.: The regretful agent: Heuristic-aided navigation through progress estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6732–6740 (2019)

8. Tan, H., Yu, L., Bansal, M.: Learning to navigate unseen environments: Back translation with environmental dropout. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2610–2621. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1268
9. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6629–6638 (2019)
10. Wang, X., Xiong, W., Wang, H., Yang Wang, W.: Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 37–53 (2018)