FedVLN: Privacy-preserving Federated Vision-and-Language Navigation

Kaiwen Zhou
University of California, Santa Cruz
1156 High St, Santa Cruz
kzhou35@ucsc.edu

Xin Eric Wang
University of California, Santa Cruz
1156 High St, Santa Cruz
xwang366@ucsc.edu

Abstract

Data privacy is a central problem for embodied agents that can perceive the environment, communicate with humans, and act in the real world. While helping humans complete tasks, the agent may observe and process sensitive information of users. In this work, we introduce privacy-preserving embodied agent learning for the task of Vision-and-Language Navigation (VLN), where an embodied agent navigates house environments by following natural language instructions. We propose a novel federated vision-and-language navigation (FedVLN) framework to protect data privacy during both training and pre-exploration, where we view each house environment as a local client. Experiment results show that, under our FedVLN framework, the decentralized VLN model achieve comparable results with centralized training while protecting seen environment privacy, and federated pre-exploration significantly outperforms other pre-exploration methods while preserving unseen environment privacy.

1. Introduction

Real-world embodied agent interacts closely with humans and environments. Thus, the agent might receive sensitive information during training and inference. For example, in the task of Vision-and-Language Navigation (VLN) [1, 3, 5, 6], the training and inference data may include private information such as what the user's house looks like and what the user has said. Data privacy is a central problem for building trustworthy embodied agents but seldomly studied before, so in this work, we introduce privacy-preserving embodied agent learning for the task of vision-and-language navigation.

VLN models are typically trained on seen environments with ground-truth instruction-trajectory pairs and then deployed to unseen environments without any labeled data. After deployment, the agent may explore the unseen environment and adapt to the new environment for better performance, which is known as pre-exploration [2, 9, 10]. However, most of the existing methods assemble all the data in a server to train a navigation agent for both seen environment training and unseen environment pre-exploration, which is not practical.

In this work, we propose a novel Federated Vision-and-Language Navigation framework (FedVLN), to address the aforementioned data privacy issues and improve the adaptation performance on unseen environments at the same time. Specifically, on the seen environment training stage, we treat each environment as a client. The client's local models will first be trained on local private data, and then the model updates will be sent to the server for model aggregation. During pre-exploration, we will train the clients on seen environments and unseen environments simultaneously under federated learning paradigm, and the clients upload only the language encoder to the server for aggregation. Under our FedVLN framework, users do not need to share their data with any other party, thus the privacy of training data and inference data is protected.

2. The FedVLN Approach

2.1. Decentralized Training

We first divide the VLN dataset by environments. We treat each environment as a client, then assign a local navigation agent w_i^0 on each environment, which is initialized as the same as global navigation agent w^0. At each communication round between clients and server, a certain percentage of clients will be randomly selected for training, the local agent on each selected client will be trained for a certain number of epochs on their own data d_i:

$$w_i^t = \text{ClientUpdate}(w^{t-1}, d_i) \tag{1}$$

Where ClientUpdate is the local training process. Then each selected client will send the update $\Delta w_{i,t} = w_i^t - w^{t-1}$ of their model to the server, and the server will aggregate all the models with a server learning rate η:

$$w^t = w^{t-1} + \eta \sum_{i \in \phi_t} \frac{n_j}{\sum_{j \in \phi_t} n_j} \Delta w_i^t \tag{2}$$

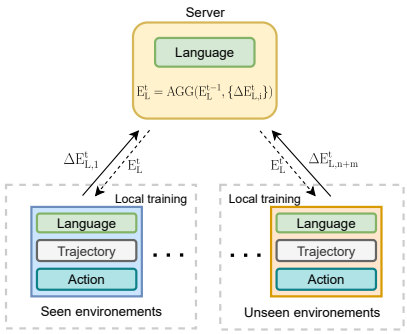

Figure 1. The framework of Federated pre-exploration.

Here $\frac{n_j}{\sum_{j \in \phi_t} n_j}$ is the proportion of the use's sample in the total training sample of this communication round.

2.2. Federated Pre-exploration

Pre-exploration allows the agent to explore the newly deployed environment and update itself based on the new information. Fu et al. [2] proposed environment-based pre-exploration, which allows each agent to train on only one environment. It's a private method since no data will be shared with other parties. From the performance point of view, for centralized training, training in all the environments can lead to a more generalized model but may hinder the agent from better adapting to one specific environment. For environment-based pre-exploration, the agent can focus on one specific environment, while the limited data amount and the distribution shift from speaker-generated instruction and human generated instructions may hurt the generalizability on validation data.

Thus, the best solution is to maintain the generalizability to understand language and adapt to a specific visual environment. To this end, we propose federated pre-exploration as in Fig. 1, in which the server will only maintain a global language encoder, which is initialized with the global encoder after decentralized VLN training. During each communication round, the server will send the global language encoder E_L^{t-1} to the selected clients. Then the selected clients will update its language encoder with E_L^{t-1}, and train the full agent on its local data:

$$E_{L,i}^t, E_{T,i}^t, M_i^t = \text{ClientUpdate}(E_{L,i}^{t-1}, E_{T,i}^{t-1}, M_i^{t-1}, \tau, \lambda) \quad (3)$$

Here τ, λ means local training epochs and learning rate.

After local training, the model will send only the language encoder $E_{L,i}^t$ to the server for aggregation. In this way, the encoder will be jointly updated on data from all the participated environments, thus being more generalized. Meanwhile, to further improve the generalizability of the language encoder, we randomly sample a fraction of seen environments at each communication round, where agents will also follow the training process above. The trajectory encoding module and multi-modal decision module will keep training locally, which can help local agents adapt to their own environments.

Stage	Model	Val-Unseen		
		SPL↑	SR↑	nDTW↑
ST	CLIP-ViL	50.7	57.0	46.4
	FedCLIP-ViL	49.8	56.3	46.1
PE	Centralized	61.7	66.1	62.5
	Env-based	65.2	69.2	65.8
	Fed-Pre	**67.3**	**71.0**	**68.7**

Table 1. Here **ST** means seen environment training and **PE** means pre-exploration.

3. Experiments and results

3.1. Experiment settings

We implement our federated learning framework on Room-to-Room (R2R) [1] dataset. The dataset contains 7,189 paths from 90 environments, and each path contains 3 instructions. The environments are split into 61 environments for training and seen validation, 11 for unseen validation, and 18 for testing. We report Success Rate (SR), Success Rate weighted by Path Length (SPL) as goal-oriented metrics, and normalized Dynamic Time Warping (nDTW) [4] to validate the fidelity of navigation paths. We report the results in unseen-validation set.

In our experiment, we adapt CLIP-ViL [8], which uses CLIP [7] to improve vision and language encoding and matching on Envdrop [9] model architecture for vision-and-language navigation.

3.2. Results

Decentralized training As in Table 1, in unseen validation set, our decentralized training achieves comparable performance with centralized training.

Federated Pre-exploration Our federated pre-exploration sharing encoder only across seen and unseen environments achieves the best result and preserves data privacy.

4. Conclusion

In this paper, we study the data privacy problems in vision-and-language navigation on two learning scenarios: seen environment training and unseen environment pre-exploration. We propose a novel federated vision-and-language navigation (FedVLN) framework to preserve data privacy in both learning stages while maintaining comparable navigation performance and even outperform all previous pre-exploration methods. As the first work along this direction, our work does not consider adversarial attacks that can potentially recover data information from shared local model updates, and we believe future work can consider more embodied AI tasks and defend against privacy attacks for more data security.

References

[1] 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 (CVPR)*, June 2018. 1, 2

[2] Tsu-Jui Fu, Xin Eric Wang, Matthew F. Peterson, Scott T. Grafton, Miguel P. Eckstein, and William Yang Wang. Counterfactual vision-and-language navigation via adversarial path sampler. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–86, 2020. 1, 2

[3] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, June 2021. 1

[4] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping, 2019. 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, Nov. 2020. 1

[6] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9979–9988. Computer Vision Foundation / IEEE, 2020. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[8] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? *CoRR*, abs/2107.06383, 2021. 2

[9] Hao Tan, Licheng Yu, and Mohit Bansal. 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)*, pages 2610–2621, June 2019. 1, 2

[10] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1