VISITRON: Visual Semantics-Aligned Interactively Trained Object-Navigator

Ayush Shrivastava[1]*, **Karthik Gopalakrishnan**[2], **Yang Liu**[2], **Robinson Piramuthu**[2], **Gokhan Tür**[2], **Devi Parikh**[1], **Dilek Hakkani-Tür**[2]
[1]Georgia Tech, [2]Amazon Alexa AI
{ayshrv, parikh}@gatech.edu
{karthgop, yangliud, robinpir, gokhatur, hakkanit}@amazon.com

Abstract

Interactive robots navigating photo-realistic environments face challenges underlying vision-and-language navigation (VLN), but in addition, they need to be trained to handle the dynamic nature of dialogue. However, research in Cooperative Vision-and-Dialog Navigation (CVDN), where a navigator interacts with a guide in natural language in order to reach a goal, treats the dialogue history as a VLN-style static instruction. In this paper, we present VISITRON, a navigator better suited to the interactive regime inherent to CVDN by being trained to: i) identify and associate object-level concepts and semantics between the environment and dialogue history, ii) identify when to interact vs. navigate via imitation learning of a binary classification head. We perform extensive ablations with VISITRON to gain empirical insights and improve performance on CVDN. VISITRON is competitive with models on the static CVDN leaderboard. We also propose a generalized interactive regime to fine-tune and evaluate VISITRON and future such models with pre-trained guides for adaptability.

1 Introduction

Vision-and-language navigation (VLN) is a challenging cross-modal research task in which agents need to learn to navigate in response to natural language instructions in photo-realistic environments. VLN has been studied extensively with the advent of the Room-to-Room (R2R) dataset (Anderson et al., 2018) and there has been growing interest recently in pushing the pre-train/fine-tune paradigm towards VLN, with work on leveraging disembodied corpora (Majumdar et al., 2020) to learn cross-modal pre-trained representations that can improve embodied VLN performance. The Cooperative Vision-and-Dialog Navigation (CVDN) dataset (Thomason et al., 2020) allows for dialogue with a guide during navigation: a navigator can ask natural language questions to a guide when it needs assistance and the guide responds in natural language by using privileged knowledge of the environment accessible only to it, thus expanding beyond the traditional VLN task towards deployable interactive agents that are more robust and generalizable. But preliminary navigator modeling using CVDN is still VLN-style via the Navigation from Dialog History (NDH) task, treating the dialogue history as a static instruction. The NDH formulation allows for easy transfer and multi-task learning (Hao et al., 2020; Wang et al., 2020) with VLN. However, state-of-the-art VLN models rely on the fully-observable setting when framing the task as *ahead-of-time* path selection (Majumdar et al., 2020), which is fundamentally at odds with the need for dialogue in CVDN: dialogue is aimed at enabling the navigating agent to succeed *while* it makes navigation decisions and decides it needs assistance. The recent Recursive Mental Model (RMM) (Roman et al., 2020) for CVDN attempts to address this by introducing a simulated dialogue

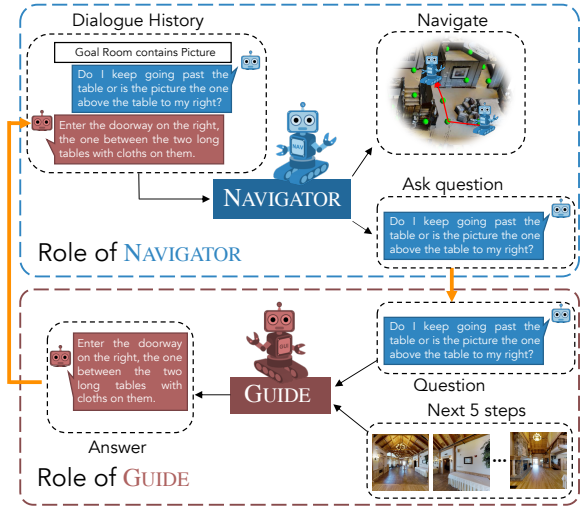

Figure 1: Cooperative Vision-and-Dialog Navigation (CVDN) with Dynamic Question-Asking

* Work done as an intern at Amazon Alexa AI. Code available at: www.github.com/alexa/visitron

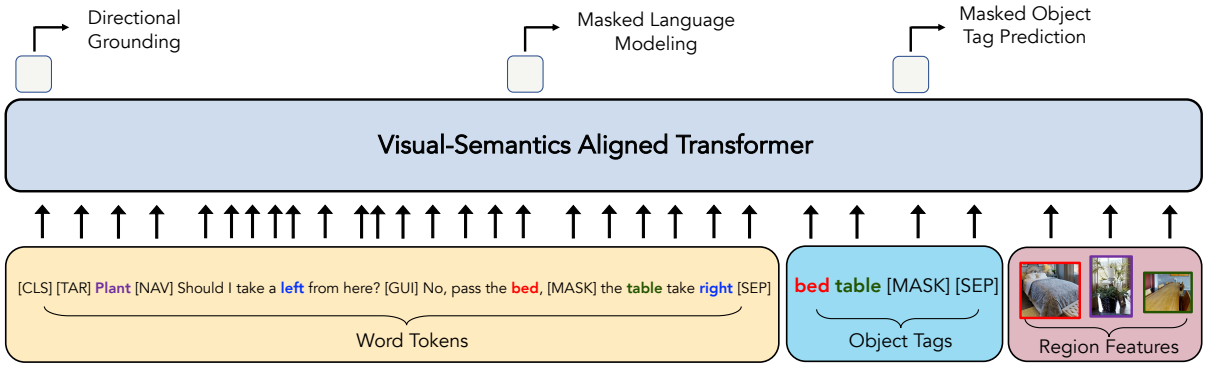

Figure 2: Semantics-aligned navigation pre-training of visio-linguistic representations using a Transformer

game-play regime, where a navigator is fine-tuned jointly with a pre-trained guide and evaluated in such a regime. But the RMM navigator relies on a rigid heuristic of asking questions after every 4th navigation step instead of doing so dynamically.

In this paper, we present work on training a navigator (which we call VISITRON) with a focus on tackling challenges unique to CVDN: i) moving beyond rote memorization to associative learning in order to learn to identify and acquire visio-linguistic concepts and semantics while interacting in new environments, and ii) learning when to ask questions in the first place (Chi et al., 2020). VISITRON builds off the recent cross-modal object-semantics aligned pre-training (OSCAR) strategy and uses object-tags as explicit anchor points during training to learn to associate the environment's visual semantics with the textual dialogue history, thus allowing for interaction/experience-grounded (Bisk et al., 2020) visio-linguistic concepts and semantics identification and acquisition. VISITRON is trained in a data-driven fashion to identify when to engage in dialogue, i.e., ask questions, vs. when to navigate, thus providing the first known empirical baselines for this task. We also present empirical results from various first-principles modeling ablations performed with VISITRON. We demonstrate that viewpoint selection is a better formulation than discrete turn-based action prediction for CVDN, akin to what has been seen on VLN with R2R. We observe that multi-task learning with the recent RxR dataset (Ku et al., 2020) leads to significant CVDN performance gains relative to training on CVDN alone. VISITRON is competitive with models on the leaderboard for the *static* NDH task on EvalAI (Yadav et al., 2019). Given VISITRON's design and ability

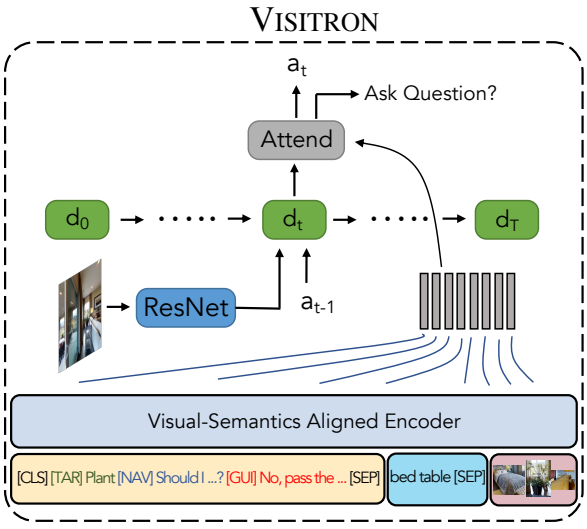

Figure 3: NAVIGATOR predicts navigation actions, given dialogue history and visual observations. The same stack decides when to ask the GUIDE a question. A similar setup can be used for question generation.

to identify when to engage in dialogue, we propose to generalize the heuristic-based game-play regime introduced with RMM by allowing dynamic question-asking as depicted in Figure 1 and fine-tune and evaluate VISITRON in such an interactive regime with pre-trained guides in future work.

2 Approach

The policy for NDH (and VLN) can be decomposed into an encoder-decoder setup, $\pi_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}_E} \circ f_{\boldsymbol{\theta}_D}$:

- A vision-language encoder $f_{\boldsymbol{\theta}_E} : \{\boldsymbol{s}_{1:t}, \boldsymbol{x}\} \rightarrow \boldsymbol{z}_t$, where $\boldsymbol{s}_{1:t}$ are visual states, $\boldsymbol{x}$ is the dialogue history (or instructions for VLN) and $\boldsymbol{z}_t$ is the joint latent representation at time step t.
- An action decoder $f_{\boldsymbol{\theta}_D} : \{\boldsymbol{s}_t, \boldsymbol{z}_t, \boldsymbol{a}_{t-1}\} \rightarrow \boldsymbol{a}_t$,

where $\boldsymbol{a}_t$ is the next action.

We model $\pi_{\boldsymbol{\theta}}$ by VISITRON, a visio-linguistic Transformer-based model. VISITRON's encoder is structurally similar to OSCAR's Transformer (Li et al., 2020). This is by design to enable easy transfer of visual semantics-aligned representations learned from disembodied image-text data. We make navigation-specific modifications to OSCAR, but they are all structured as augmentations of modules instead of removal of network components, thus enabling us to use the pre-trained weights of OSCAR's Transformer to initialize large portions of our encoder. As with OSCAR, the input to VISITRON's encoder is represented as Word-Tag-Image $(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v})$, where $\boldsymbol{w}$ and $\boldsymbol{q}$ are the sequence of word embeddings of the text and object tags respectively, and $\boldsymbol{v}$ is the sequence of region features of the image. We represent the panorama in 36 views, extract Faster R-CNN (Ren et al., 2015) region features r' from each view and add positional vector p, $r = (r', p)$. To incorporate 3D direction, we add direction embedding d to the region features, $v = r + d$. d is a 128-dimensional orientation vector represented by repeating $[\sin\phi; \cos\phi; \sin\omega; \cos\omega]$ 32 times where ϕ and ω are heading and elevation poses. In addition to the standard [CLS] and [SEP], we also use [TAR], [NAV], [GUI] as delimiter tokens for the initial target hint, NAVIGATOR's questions and the GUIDE's answers respectively. While this input structure is dialogue-specific, it is amenable to instructions-based datasets for multi-tasking.

2.1 VISITRON Pre-Training

We adopt a two-stage pre-training strategy, initializing VISITRON's encoder with weights from OSCAR to begin with web-scale disembodied visio-linguistic representations, followed by facilitating a domain shift to navigation and actions by pre-training on navigation data. For each trajectory in NDH and R2R, we extract $(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v}, \boldsymbol{a})$ tuples where $\boldsymbol{w}$ is the dialogue history/instruction, $\boldsymbol{q}$ is the sequence of object tags from the current panorama, $\boldsymbol{v}$ is the sequence of region features and $\boldsymbol{a}$ is the direction in the 360° panoramic space where the next node in the trajectory is located (Fried et al., 2018). Figure 2 depicts this for an extracted tuple from a sample NDH instance. The objectives are:

1. *Masked Language Modeling:* Like BERT, input tokens are replaced with [MASK] 15% of the time and the masked token x_i is predicted conditioned on surrounding tokens $x_{\setminus i}$.
2. *Masked Object Tag Prediction:* Object tags are replaced with [MASK] 15% of the time. A feed-forward head on top of [MASK] is used to predict the tag from a distribution over Faster R-CNN semantic classes.
3. *Directional Grounding:* [CLS] hidden state goes into a feed-forward head to predict $\boldsymbol{a}$.

2.2 VISITRON Fine-Tuning

After pre-training the encoder, we leverage it with an attention-based LSTM action decoder. At time-step t, the decoder (cell state d_t) takes the previous action $\boldsymbol{a}_{t-1}$, the panoramic ResNet features extracted from the current location/state and decodes the next action $\boldsymbol{a}_t$, while attending to the VISITRON encoder's cross-modal representation of its input. After this LSTM is trained, the same stack is frozen and used with a randomly initialized two-layer feed-forward head trained with a binary cross-entropy loss to learn to classify when to ask a question. Figure 3 depicts this setup. Note that the decoder's action can belong in either the panoramic space or the low-level visuomotor space (Fried et al., 2018).

3 Experiments

We begin experimenting with cumulative addition of each pre-training stage and objective to obtain an ablative understanding of their effect on the downstream NDH task. Table 1 demonstrates that **our pre-training strategy helps**: best performance on Val Seen (as measured by all metrics) is obtained when using all pre-training stages and objectives. We also see that Goal Progress (GP) is highest on Val Unseen in this setting (an absolute increase of 0.62 relative to no pre-training), with minimal loss in Success Rate weighted by Normalized Inverse Path Length (SPL) and Success Rate (SR) metrics relative to their best setting. Rows 4-5 demonstrate the value of masked object tag prediction as a means towards experience-driven concept and semantics identification and acquisition, with significant increases in all metrics on Val Unseen.

Next, we perform ablations during fine-tuning, leveraging all objectives from Table 1 since our previous analysis demonstrated their effectiveness. For VLN agents, it has been shown that viewpoint selection in the panoramic space is a better formulation than turn-based action prediction in the low-level visuomotor space (Fried et al., 2018). However, it is not immediately obvious or known whether this can be extrapolated to dialogue-based

Table 1: Pre-Training Ablations (Fine-Tuning and Evaluating on NDH)

	#	Semantics-aligned Pre-Training Curriculum: Stage 1: Web (OSCAR) Contrastive+ Masked LM	Stage 1: Web (OSCAR) Object Tags	Stage 2: Navigation Masked LM	Stage 2: Navigation Masked Object Tag Prediction	Stage 2: Navigation Directional Grounding	Val Seen GP (m) ↑	Val Seen SPL (%) ↑	Val Seen SR (%) ↑	Val Seen nDTW (%) ↑	Val Unseen GP (m) ↑	Val Unseen SPL (%) ↑	Val Unseen SR (%) ↑	Val Unseen nDTW (%) ↑
	1	(No pre-training and no object tags)					4.76	36.56	46.07	30.97	2.09	9.96	22.49	6.50
VISITRON	2	✓					4.82	50.73	58.11	47.34	2.67	**24.88**	**34.29**	24.21
VISITRON	3	✓	✓				4.38	45.15	52.09	41.14	2.30	13.03	24.81	8.63
VISITRON	4	✓	✓	✓			5.09	25.92	41.10	17.91	1.90	11.27	23.48	5.62
VISITRON	5	✓	✓	✓	✓		4.83	48.22	56.02	47.01	2.70	24.04	32.86	23.46
	6	✓	✓	✓	✓	✓	**5.34**	**55.16**	**61.78**	**54.83**	**2.71**	24.56	32.52	**24.51**

Table 2: Fine-Tuning Ablations

	#	Action Space	Multi-Task Fine-Tuning NDH+	Val Seen GP (m) ↑	Val Seen SPL (%) ↑	Val Seen SR (%) ↑	Val Seen nDTW (%) ↑	Val Unseen GP (m) ↑	Val Unseen SPL (%) ↑	Val Unseen SR (%) ↑	Val Unseen nDTW (%) ↑
VISITRON	1	Turn-based Action Prediction	✗	1.15	9.66	11.78	26.86	1.60	13.02	14.77	29.28
VISITRON	2	Turn-based Action Prediction	✓(RxR)	1.50	12.30	15.18	19.95	0.97	11.52	15.44	20.49
VISITRON	3	Viewpoint Selection	✗	5.34	55.16	61.78	54.83	2.71	24.56	32.52	24.51
VISITRON	4	Viewpoint Selection	✓(RxR)	5.11	12.33	25.65	4.66	3.25	10.74	27.34	3.78

Table 3: Question-Asking Classification Performance

Metric (%)	Val Seen	Val Unseen
Accuracy	68.05	67.87
Balanced Accuracy	63.33	61.09

Table 4: NDH Hidden Test Set Performance

#	Method	GP (m) ↑	SPL (%) ↑
1	MT-RCM + EnvAg (Wang et al., 2020)	3.91	17
2	BabyWalk (Zhu et al., 2020a)	3.65	11
3	**VISITRON**	3.11	12
4	Cross-modal Memory Network (Zhu et al., 2020b)	2.95	14
5	PREVALENT (Hao et al., 2020)	2.44	24
6	**VISITRON** (Best SPL)	2.40	**25**

navigation as in CVDN. So we experiment with both formulations for our NAVIGATOR. Given the sparsity of NDH instances ($\sim$ 4k) for fine-tuning, we also study if multi-task fine-tuning with the RxR dataset helps boost performance. Table 2 demonstrates that **viewpoint selection is a better formulation than turn-based action prediction for CVDN**, with Val Unseen GP increasing from 1.6 to 2.71 when switching to viewpoint selection. Further, we observe that **multi-task fine-tuning leads to better CVDN generalization**, with Val Unseen GP increasing from 2.71 to 3.25 when multi-tasking with viewpoint selection. The associated decrease in Normalized Dynamic Time Warping (nDTW) (Ilharco et al., 2019), SPL and SR can be attributed to VISITRON learning from RxR to take paths beyond the next 5 GUIDE steps in the NDH instance which these metrics evaluate against, while GP cares about the final CVDN goal.

Using the best VISITRON model from Table 2 (Row 4), Table 3 shows that imitation learning of the binary classification head is a strong baseline, as measured by accuracy and balanced accuracy (to account for data imbalance) on Val Unseen. We submitted this model to the CVDN leaderboard aimed at the *static* NDH task. We observe in Table 4 that VISITRON's performance is competitive with state-of-the-art models, with its best GP being 3.11. Given the expected decrease in SPL when utilizing RxR, we also trained a version of VISITRON with multi-tasking on NDH, R2R and R4R (Jain et al., 2019) instead of NDH and RxR and observed this model obtains state-of-the-art SPL of 25.

4 Conclusion and Future Work

We presented VISITRON, a navigator designed for interaction-grounded visio-linguistic concepts and semantics identification and acquisition, and decision-making for interactive navigation inherent to CVDN. We demonstrated the efficacy of our approach via experiments and ablations. We proposed generalizing the game-play regime introduced with RMM (Roman et al., 2020) by allowing dynamic question-asking so as to interactively fine-tune and evaluate VISITRON with a pre-trained GUIDE. Future work should delve into Sim-to-Real transfer (Anderson et al., 2020) and robustness in dialogue-based navigation, in presence of speech recognition errors (Gopalakrishnan et al., 2020).

Acknowledgments

Many thanks to Jesse Thomason and Aishwarya Padmakumar for useful technical discussions and actionable feedback on multiple versions of this paper. We would also like to thank the anonymous reviewers for their service and useful feedback.

References

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. 2020. Sim-to-real transfer for vision-and-language navigation. *arXiv preprint arXiv:2011.03807*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. 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.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2459–2466.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 3314–3325.

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tur. 2020. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study. *arXiv preprint arXiv:2008.07683*.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. In *ViGIL@ NeurIPS*.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. *arXiv preprint arXiv:2004.14973*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. Rmm: A recursive mental model for dialog navigation. *arXiv preprint arXiv:2005.00728*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406.

Xin Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020. Environment-agnostic multitask learning for natural language grounded navigation. *arXiv preprint arXiv:2003.00443*.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. 2019. Evalai: Towards better evaluation systems for ai agents. *arXiv preprint arXiv:1902.03570*.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020a. Babywalk: Going farther in vision-and-language navigation by taking baby steps. *arXiv preprint arXiv:2005.04625*.

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, and Xiaodan Liang. 2020b. Vision-dialog navigation by exploring cross-modal memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10730–10739.