Extended Abstract: Improving Vision-and-Language Navigation with Image-Text Pairs from the Web

Arjun Majumdar [1] **Ayush Shrivastava** [1] **Stefan Lee** [2] **Peter Anderson** [1 3] **Devi Parikh** [1 4] **Dhruv Batra** [1 4]

Abstract

Following a navigation instruction such as *'Walk down the stairs and stop near the sofa'* requires an agent to ground scene elements referenced via language (e.g. *'stairs'*) to visual content in the environment (pixels corresponding to *'stairs'*).

We ask the following question – can we leverage abundant 'disembodied' web-scraped vision-and-language corpora (e.g. Conceptual Captions (Sharma et al., 2018)) to learn visual groundings (what do *'stairs'* look like?) that improve performance on a relatively data-starved embodied perception task (Vision-and-Language Navigation)? Specifically, we develop VLN-BERT, a visiolinguistic transformer model that scores the compatibility between an instruction (*'...stop near the sofa'*) and a sequence of panoramic images. We demonstrate that pretraining VLN-BERT on image-text pairs from the web significantly improves performance on VLN – outperforming the prior state-of-the-art in the fully-observed setting by 4 absolute percentage points on success rate. Ablations of our pretraining curriculum show each stage to be impactful – with their combination resulting in further synergistic effects.

1. Introduction

Consider the navigation instruction in Figure 1 (right), *'Walk straight and pass the couches then pass the white table with the four chairs and stop by the brick wall.'* In vision-and-language navigation (VLN) (Anderson et al., 2018), agents must interpret such instructions to navigate through photo-realistic environments. In this instance, the agent needs to select a path that passes *'the couches'*, *'the white table'*, *'the four chairs'*, and ends at *'the brick wall'*. As such, the ability to ground references to these objects and scene elements is central to success. Existing work has focused on learning this grounding solely from a task-specific training dataset of path-instruction pairs (Fried et al., 2018; Ke et al., 2019; Ma et al., 2019; Tan et al., 2019; Wang et al., 2019) – which are expensive, laborious, and time-consuming to collect at scale and thus tend to be relatively small – e.g. the Room-to-Room dataset (Anderson et al., 2018) contains around 14k path-instruction pairs for training. As an alternative, we propose learning visual grounding from webly-supervised internet data, such as the images and captions captured in the Conceptual Captions dataset (Sharma et al., 2018), containing around 3.3M image-text pairs.

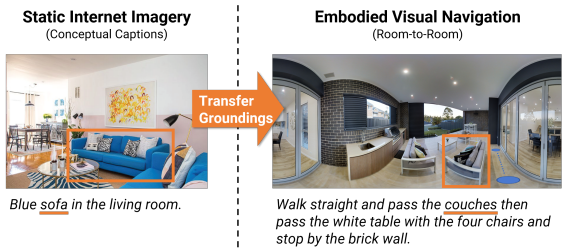

Figure 1. We propose a model architecture and training curriculum specifically designed to transfer visual grounding learned from image-text pairs from the web (left) to the embodied AI task of vision-and-language navigation (VLN) (right).

Conceptually, transfer learning from large-scale web data to embodied AI tasks such as VLN is an attractive alternative to collecting more data. Empirically, however, the effectiveness of this strategy remains open to question – would such a transfer even work? Unlike web images, which are highly-curated and stick closely to aesthetic biases, embodied data contains content and viewpoints that are not widely published online. For example, as shown in Figure 2, an embodied agent may perceive doors via a close-up view of a door frame rather than as a carefully composed image of a (typically closed) door. In VLN, image framing is a result of the agent's position rather than choices made by a photographer. Consequently, we investigate the question – to what degree can webly-supervised visual grounding learned on static images be transferred to the embodied VLN task? Put more succinctly, can 'disembodied' web data be used to improve visual grounding for embodied agents?

To answer this question, we introduce VLN-BERT, a visi-

[1]Georgia Institute of Technology [2]Oregon State University [3]Now at Google [4]Facebook AI Research (FAIR). Correspondence to: Arjun Majumdar <arjun.majumdar@gatech.edu>.

Accepted to the 1st Workshop on Language in Reinforcement Learning, ICML 2020, Copyright 2020 by the author(s).

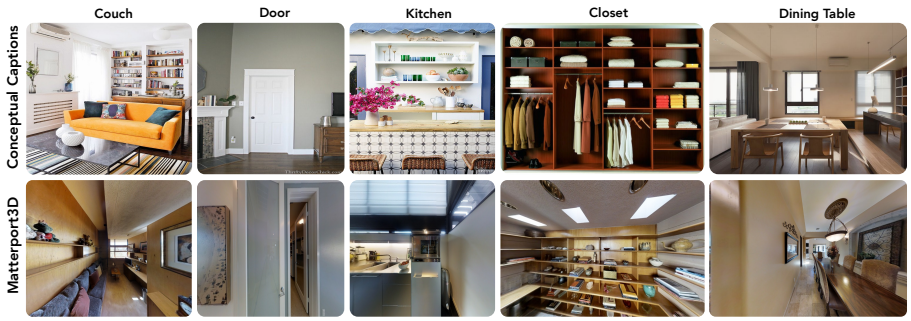

Figure 2. Images from Conceptual Captions (CC) (Sharma et al., 2018) (top) and Matterport3D (MP3D) (Chang et al., 2017) (bottom) illustrate the differences between the two domains such as viewpoint and lighting.

olinguistic transformer-based model for scoring the alignment between an instruction and an agent's observations along a path. We structure VLN-BERT to enable straightforward transfer learning from a model from prior work on general visiolinguistic representation learning (Lu et al., 2019), and explore a training curriculum that incorporates both large-scale internet data and embodied path-instruction pairs. VLN-BERT is sequentially trained using 1) language-only data (Wikipedia and BooksCorpus (Zhu et al., 2015) as in BERT (Devlin et al., 2018)), 2) web image-text pairs (Conceptual Captions (Sharma et al., 2018) as in ViLBERT (Lu et al., 2019)), and 3) path-instruction pairs from the Room-to-Room dataset (Anderson et al., 2018). Following this protocol the model learns to represent language, then to ground visual concepts, and finally to ground visual concepts alongside action descriptions. We evaluate VLN-BERT on path selection in VLN, demonstrating that this training procedure leads to significant gains over prior work (4 absolute percentage points on leaderboard success rate).

2. Approach

We describe path selection in VLN (Sec. 2.1), our model (Sec. 2.2), and transfer learning curriculum (Sec. 2.3).

2.1. Vision-and-Language Navigation as Path Selection

In Vision-and-Language Navigation (VLN) (Anderson et al., 2018), agents traverse a path τ within a navigation-graph G to follow the natural language instructions x. Following prior work (Fried et al., 2018; Tan et al., 2019), we consider the previously explored environment setting, in which an agent can consider arbitrarily many paths before selecting one to follow. In this setting, navigation involves identifying the path that best aligns with the instructions. Concretely, given a set of valid paths $\mathcal{T}$ originating from the same starting position and an instruction x, the problem of navigation is to identify a path τ^* such that

$$\tau^* = \underset{\tau \in \mathcal{T}}{\operatorname{argmax}} \quad f(\tau, x) \tag{1}$$

for some compatibility function f that determines if the path follows the instruction and ends near the goal. To focus on transfer learning, this work addresses learning the function f given a set of paths $\mathcal{T}'$ generated with beam-search on the follower agent from (Tan et al., 2019).

2.2. Modeling Path-Instruction Compatibility

We model $f(\tau, x)$ as a visiolinguistic transformer-based model denoted as VLN-BERT (illustrated in Figure 3). The architecture of VLN-BERT is structurally similar to ViLBERT (Lu et al., 2019), which is composed of two BERT-like (Devlin et al., 2018) processing streams that that operate on visual and textual inputs, respectively. The two streams are connected using co-attention transformer layers, which attend from the vision stream over language stream and vice versa. By design, VLN-BERT reuses large parts of the ViLBERT architecture to enable straight-forward transfer of visual grounding learned from large-scale web data.

Representing Trajectories and Instructions. Predicting path-instruction compatibility requires jointly reasoning over a sequence of panoramic images and a sequence of instruction words (Figure 1 right). We represent each panorama as a set of image regions $\{r_1^{(i)}, \ldots, r_K^{(i)}\}$ (generated by an object detector). Thus, the inputs to VLN-BERT are "visual tokens" representing each region from each panorama and "language tokens" for each word in the instruction. Special `IMG` tokens are used to separate panoramas and a `CLS` token separates the two modalities.

A common practice with BERT-like models is to add positional embeddings to the input representations that encode relationships between tokens. For language, this amounts to an index-in-sequence encoding. For panoramic trajectories the relationship between image regions is significantly more complex. First, each region corresponds to a different heading and elevation relative to the panoramic coordinate system. Further, each panorama within the sequence correspond with different positions in the environment. These geometric relationships are important for language-guided navigation – after all, something on your left going one way is on your right if you go in the opposite direction. Thus, we add a learned panoramic positional embedding that encodes spatial information about each region. Finally, the input

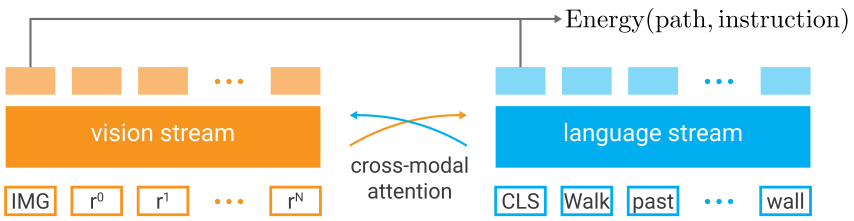

Figure 3. We propose VLN-BERT, a visiolinguistic transformer-based model that extends the model from (Lu et al., 2019) to processes image regions from a sequence of panoramas and word tokens from navigational instruction to solve path selection in VLN.

representation for a region is composed of the element-wise sum of this positional embedding, a panoramic sequence embedding, and the image region features.

Training for Path Selection. To train VLN-BERT, we consider a four-way multiple choice setting. For a given instruction, we sample four paths out of which only one is successful. We pass each path-instruction pair to VLN-BERT and extract the final representations corresponding to the `CLS` and first `IMG` token. The element-wise multiplication of these representations are passed through a linear layer to produce a compatibility score. The scores are normalized via a softmax and supervised with cross-entropy loss.

2.3. Pretraining Curriculum

VLN-BERT was specifically designed to enable transfer learning from language (Devlin et al., 2018) and visiolinguistic (Lu et al., 2019) models trained on large-scale web corpora. This transfer is especially important in the VLN task which is relatively data-sparse. To overcome this challenge, we consider a three stage curriculum focused on learning language, visual grounding, and action grounding.

- **Stage 1: Language.** To develop language understanding, we initialize the language stream of our model with weights from a BERT (Devlin et al., 2018).
- **Stage 2: Visual Grounding.** Starting from a pretrained BERT model, Lu et al. train both streams of ViLBERT on the Conceptual Captions dataset (Sharma et al., 2018) under a masked multimodal language modelling and multimodal alignment objectives (Lu et al., 2019). In this stage, we initialize both the vision and language streams with weights from ViLBERT.
- **Stage 3: Action Grounding.** In the final stage, we pair paths and instructions from VLN and train the model under the masked multimodal modelling objective from (Lu et al., 2019). While the previous stage learns to ground visual concepts, this stage additionally exposes the model to actions and their trajectory-based referents.

After these pretraining stages, we fine-tune VLN-BERT for path selection as described previously.

3. Experiments

We conduct experiments using the Room-to-Room (R2R) navigation task (Anderson et al., 2018) (produced using the Matterport3D dataset (Chang et al., 2017)). To generate a dataset for path selection we run beam search on the follower model from (Tan et al., 2019), producing up to 30 candidate paths for each instruction in R2R. We report results for selecting one path out of these candidates.

Evaluation Metrics. We compare performance using standard VLN metrics – success rate (SR), navigation error (NE), path length (PL), success rate weighted by path length (SPL). For path selection we calculate metrics using only the selected path, however, for the VLN leaderboard results we prepend the exploration path to the selected path (which affects the path length based metrics PL and SPL).

3.1. Results

Pretraining Curriculum Ablation Study. The results in Table 1 demonstrate that in general, each pretraining stage contributes to performance. In particular, pretraining on image-text pairs (stage 2) and path-instruction pairs (stage 3) similarly improve SR (by 4.5 and 4.9 absolute percentage points, respectively). However, when the two stages are combined in series the SR is 9.2 absolute percentage points over the next best setting. This substantial level of improvement suggests that not only does pretraining on webly-supervised image-text pairs from (Sharma et al., 2018) improve performance, but it also constructively supports the action grounding stage (stage 3) of pretraining.

Baseline comparisons. Table 2 compares VLN-BERT with the follower and speaker models from (Tan et al., 2019) (state-of-the-art on the VLN leaderboard). In the single model setting, VLN-BERT (row 3) is 4.6 absolute percentage points better on SR than either baseline. The ensemble model setting demonstrates that when all three models are linearly combined (row 7), their performance is 3.0 absolute percentage points higher on SR than the next best ensemble.

VLN Leaderboard. As shown in Table 3, on the VLN leaderboard our three-model ensemble achieves a success rate of 73%, which is 4 absolute percentage points greater than previously published work (Tan et al., 2019).

4. Conclusion

In this work, we demonstrated internet-to-embodied transfer of visual concept grounding – leveraging large-scale image-text data from the web to improve a discriminative path-

Table 1. Pretraining curriculum ablation study demonstrating the effectiveness of internet-to-embodied transfer of visual grounding.

		Pretraining Stage			Val Seen					Val Unseen				
	#	Language Only	Visual Grounding	Action Grounding	PL	NE ↓	SPL ↑	OSR ↑	SR ↑	PL	NE ↓	SPL ↑	OSR ↑	SR ↑
VLN-BERT	1	(no pretraining)			10.78	6.78	0.35	54.22	37.55	10.29	6.81	0.27	50.62	30.52
	2	✓			10.33	4.89	0.55	69.31	58.73	9.59	5.47	0.41	57.34	45.17
	3	✓	✓		10.42	4.48	0.58	71.57	62.16	9.70	4.96	0.45	62.79	49.64
	4	✓		✓	10.51	4.28	0.60	72.65	63.82	9.81	5.05	0.46	62.75	50.02
	5	✓	✓	✓	10.28	3.73	0.66	76.47	70.20	9.60	**4.10**	**0.55**	**69.22**	**59.26**

Table 2. Baseline comparisons.

	#	Re-ranking Model	Val Unseen PL	NE ↓	SPL ↑	OSR ↑	SR ↑
Single Models	1	follower (Tan et al., 2019)	9.57	5.20	0.49	58.79	52.36
	2	speaker (Tan et al., 2019)	10.71	4.25	0.49	**72.07**	54.66
	3	VLN-BERT	9.60	**4.10**	**0.55**	69.22	**59.26**
Ensemble Models	4	speaker + follower (Tan et al., 2019)	10.10	3.32	0.63	76.63	67.90
	5	speaker + follower + follower	10.12	3.22	0.64	77.56	69.14
	6	speaker + follower + speaker	10.17	2.99	0.65	79.28	70.58
	7	speaker + follower + VLN-BERT	10.00	**2.76**	**0.68**	**81.91**	**73.61**

Table 3. Leaderboard results (with beam search)

Re-ranking Model	Test Unseen PL	NE ↓	SPL ↑	OSR ↑	SR ↑
Speaker-Follower (Fried et al., 2018)	1,257	4.87	0.01	96	53
Tactical Rewind (Ke et al., 2019)	197	4.29	**0.03**	90	61
Self-Monitoring (Ma et al., 2019)	373	4.48	0.02	97	61
RCM (Wang et al., 2019)	358	4.03	0.02	96	63
EnvDrop (Tan et al., 2019)	687	3.26	0.01	**99**	69
Auxiliary Tasks† (Zhu et al., 2019)	41	3.24	0.21	81	71
VLN-BERT	687	**3.09**	0.01	**99**	**73**

†indicates unpublished/concurrent work

instruction alignment model for VLN.

Acknowledgements

The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs, ARO PECASE, Amazon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and van den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., and Darrell, T. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, pp. 3314–3325, 2018.

Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Choi, Y., and Srinivasa, S. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *CVPR*, pp. 6741–6749, 2019.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pp. 13–23, 2019.

Ma, C.-Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., and Xiong, C. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

Tan, H., Yu, L., and Bansal, M. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019.

Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.-F., Wang, W. Y., and Zhang, L. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pp. 6629–6638, 2019.

Zhu, F., Zhu, Y., Chang, X., and Liang, X. Vision-language navigation with self-supervised auxiliary reasoning tasks. *arXiv preprint arXiv:1911.07883*, 2019.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pp. 19–27, 2015.