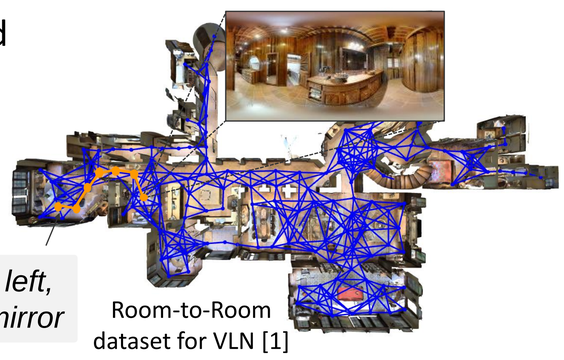
Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation

Ronghang Hu[1] Daniel Fried[1] Anna Rohrbach[1]
Dan Klein[1] Trevor Darrell[1] Kate Saenko[2]
[1]University of California, Berkeley [2]Boston University

Vision & Language Navigation (VLN) Task

Given visual observations and a language instruction, take actions to navigate to the described target location:



Room-to-Room dataset for VLN [1]

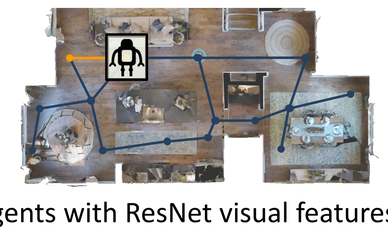
go down the second hallway on the left, enter the bedroom and stop by the mirror

Room-to-Room (R2R) dataset[1]: real images + discrete locations

Visual features are not helping agents generalize!

Surprisingly, we find that state-of-the-art models do not benefit from their visual inputs in new environments.

Compare agents **with** and **without visual features** using two state-of-the-art architectures: Speaker-Follower [2] and Self-Monitoring [3], that use pre-trained ResNet features:



Agents with ResNet visual features

Agents without visual features (only route structure)

Agents without vision (relying just on discrete route structure) are comparable or better in new, *Unseen*, environments:

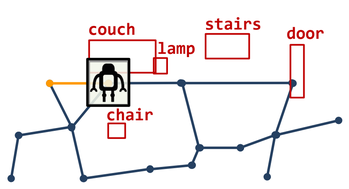
model architecture	training approach	visual features	success rate on Seen envs.	success rate on Unseen envs.
Speaker-Follower [2]	student-forcing	(none)	29.7	**31.7**
		ResNet	53.3	29.0
	teacher-forcing	(none)	34.1	**35.2**
		ResNet	40.4	29.0
Self-Monitoring [3]*	student-forcing	(none)	36.1	39.7
		ResNet	62.8	**40.5**
	teacher-forcing	(none)	34.3	32.2
		ResNet	44.0	**32.8**

(* Self-Monitoring results are based on our implementation; teacher-forcing: sampling actions from shortest paths to the goal; student-forcing: sampling actions from the agent's prediction)

Will higher-level visual features generalize better?

Sometimes: using object-based visual features generalizes better than using ResNet features in one model, and generalizes comparably in a second model.

Object detections from Faster R-CNN [4], trained on Visual Genome:



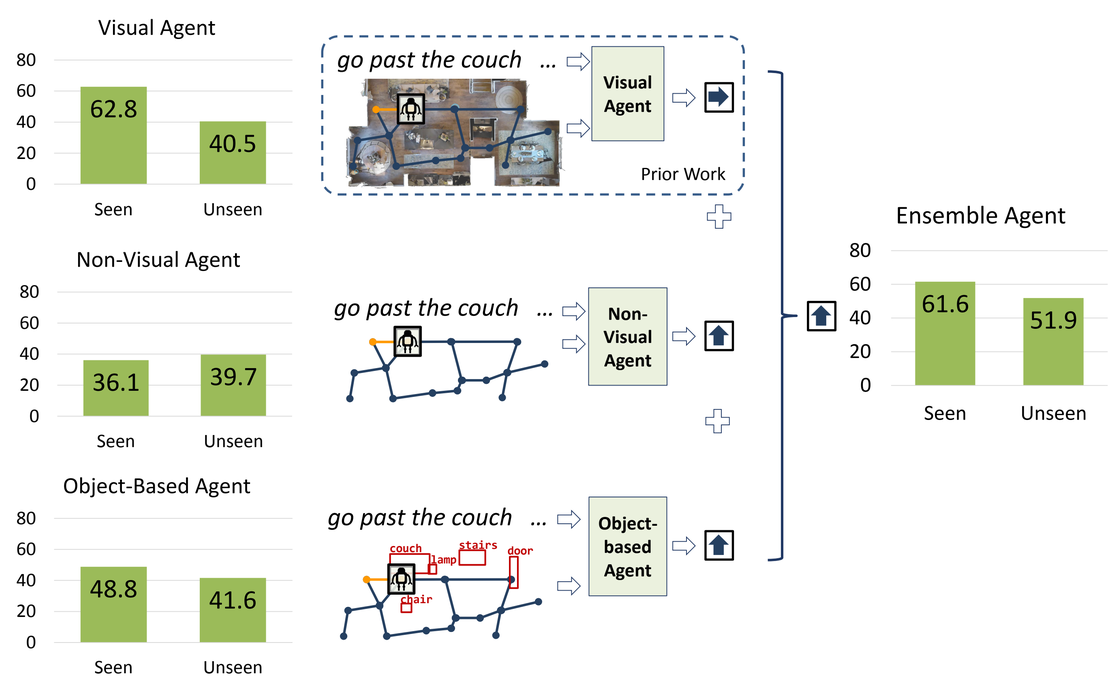
Represent the scene with object detection results, replacing or combining with ResNet visual features:

#	model architecture	visual features	success rate on Seen envs.	success rate on Unseen envs.
1	Speaker-Follower [2]	(none)	34.1	35.2
2		ResNet	53.3	29.0
3		objects	38.5	33.5
4		ResNet + objects	47.8	**39.8**
5	Self-Monitoring [3]	(none)	36.1	39.7
6		ResNet	62.8	40.5
7		objects	48.8	**41.6**
8		ResNet + objects	59.2	39.5

Vision does help if the model is structured carefully

Best overall results from a mixture-of-experts:

- Ensemble a visual agent (Object or ResNet) and a non-visual agent: *better than ensembling two agents of the same modality (both visual or both non-visual)*
- Objects and ResNet features are also complementary
- Further benefits from jointly training agents in the ensemble



#	architecture under Self-Monitoring [3]	mixture-of-experts combination	success rate on Seen envs.	success rate on Unseen envs.
9	mixture of 2 models	(no vis, no vis)	36.8	41.0
10		(ResNet, ResNet)	62.8	43.5
11		(objects, objects)	49.2	45.2
12		(ResNet+objects, ResNet+objects)	63.5	42.2
13		(ResNet, no vis)	63.4	**46.9**
14		(objects, no vis)	44.9	43.4
15		(ResNet+objects, no vis)	60.2	46.4
16	mixture of 3 models	(ResNet, objects, no vis)	60.0	**49.5**
17	joint training	(ResNet, no vis)	63.1	48.3
18		(ResNet, objects, no vis)	61.6	**51.9**

Discussion

- State-of-the-art models have trouble with generalizable visual perception (consistent with [5])
- Higher-level visual features from a pre-trained object detector sometimes generalize better than lower-level ResNet features
- Structuring the agent to encourage it to ground into each modality helps, even by simply ensembling visual- and non-visual models

enter through the sliding door ...

take a right out of the room ...

blue: a non-visual agent's action red: a visual agent's action

References

[1] Anderson et al. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." in CVPR 2018.

[2] Fried, Hu, Cirik, et al. "Speaker-follower models for vision-and-language navigation." in NeurIPS 2018.

[3] Ma et al. "Self-Monitoring Navigation Agent via Auxiliary Progress Estimation." in ICLR 2019.

[4] Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." in NIPS 2015.

[5] Thomason et al. "Shifting the baseline: Single Modality Performance on Visual Navigation & QA." in NAACL 2019.