Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation

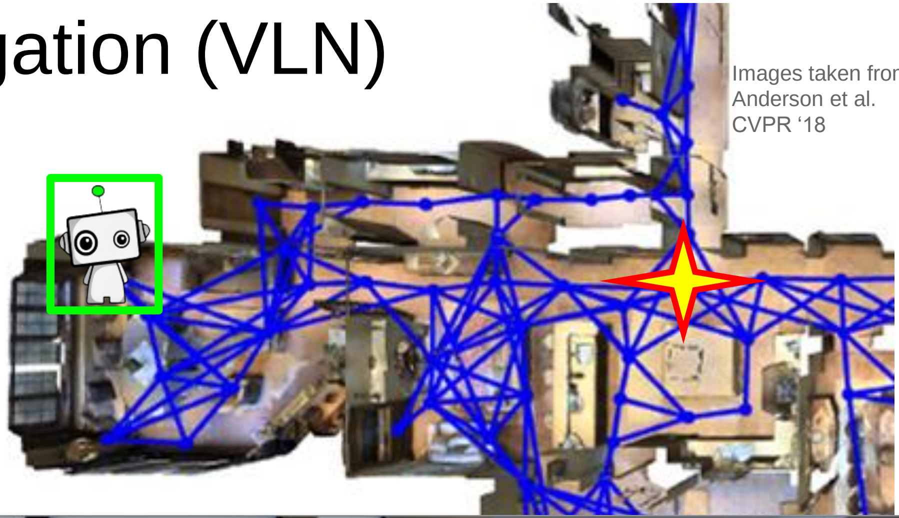
Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky

Princeton University

Key Contributions

- Identify bias in action space of Room-to-Room (R2R) dataset for Vision-and-Language Navigation (VLN) task, dubbed “action priors”.
- Remedy scarce amount of annotated data and these priors through random walk sampling.
- Validate our approach to show improved performance and agent generalizability.

Vision-and-Language Navigation (VLN)



Images taken from Anderson et al. CVPR '18

An agent is…

- in an environment represented by a graph.
- given panoramic egocentric view of scene, navigable locations, and natural language instructions.
- Tasked to navigate through environment to goal.

Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by moose antlers hanging on wall.

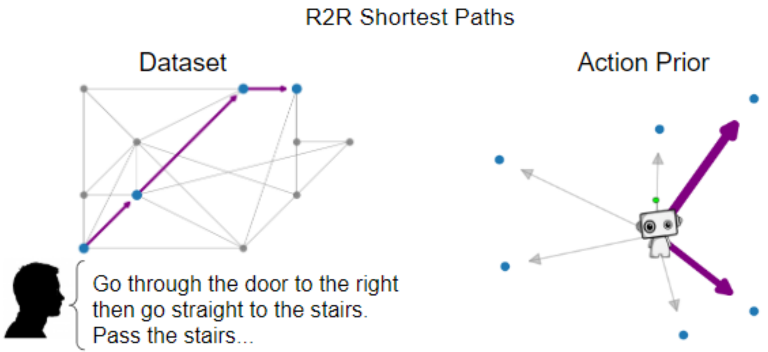
Difficulties in VLN

- **Lack of Data**
  - Time-consuming to obtain multiple sentence-long annotations.
  - Room-to-Room Dataset [1] only contains 21,567 samples for 7,189 paths

[1] Anderson, Peter, et al. (CVPR ‘18) "Vision-and-language navigation”

Difficulties in VLN

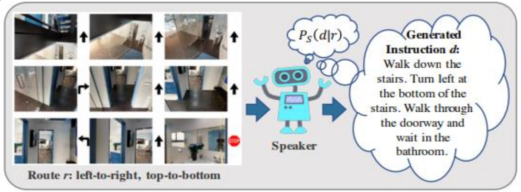
- **Lack of Data**
- **Poor generalizability** to unseen environments.
    - Action priors exist in **shortest path sampling**



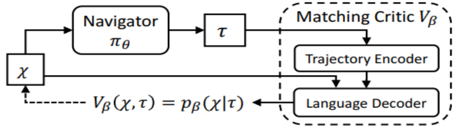
Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Related Works

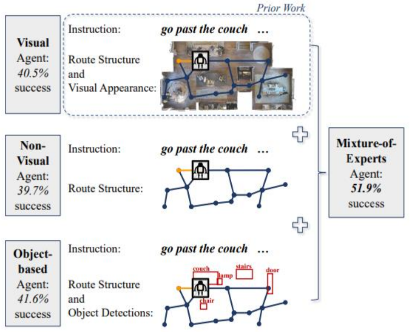
- Lack of Data
- Poor generalizability to unseen environments.
    - Action priors exist in shortest path sampling



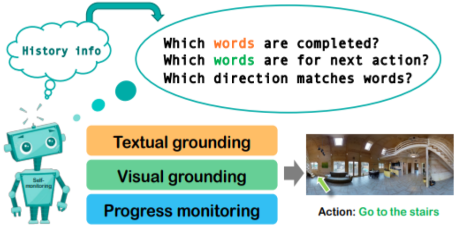
Fried et al. (NeurIPS ‘18) “Speaker-Follower”



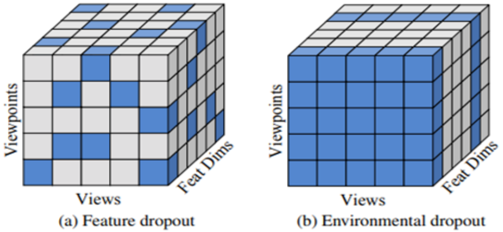
Wang et al. (CVPR ‘19) “Cross-Modal Matching”



Hu et al. (ACL ’19) “Are You Looking?”



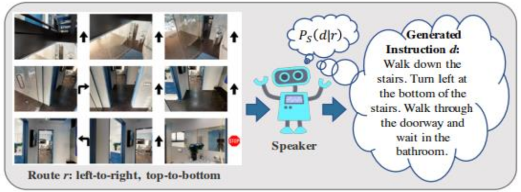
Ma et al. (ICLR ‘19) “Self-Monitoring Agent



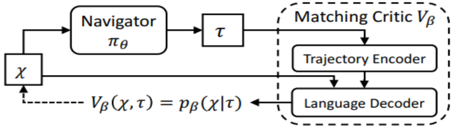
Tan et al. (NAACL ‘19) “Environmental Dropout”

Related Works

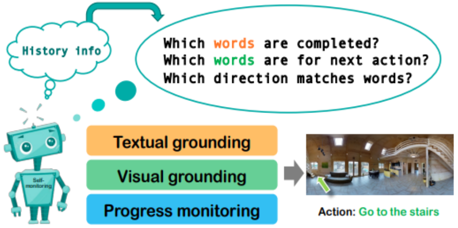
- **Lack of Data**
- **Poor generalizability** to unseen environments.
  - Action priors exist in **shortest path sampling**



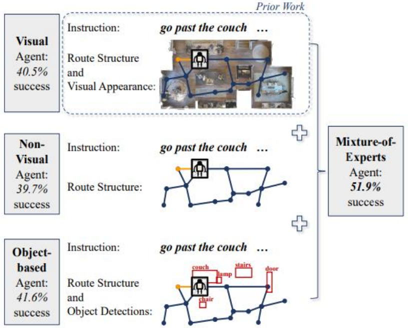
Fried et al. (NeurIPS ‘18) “Speaker-Follower”



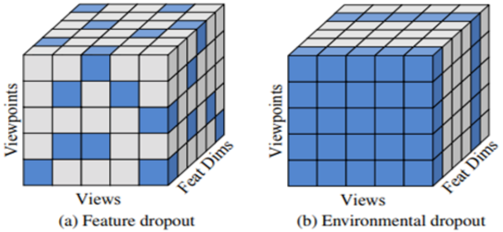
Wang et al. (CVPR ‘19) “Cross-Modal Matching”



Ma et al. (ICLR ‘19) “Self-Monitoring Agent



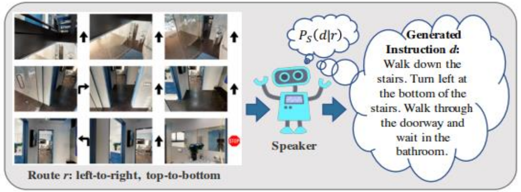
Hu et al. (ACL ’19) “Are You Looking?”



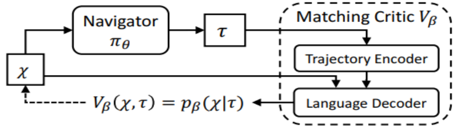
Tan et al. (NAACL ‘19) “Environmental Dropout”

Related Works

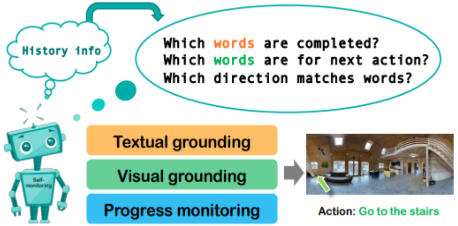
- **Lack of Data**
- **Poor generalizability** to unseen environments.
  - Action priors exist in **shortest path sampling**



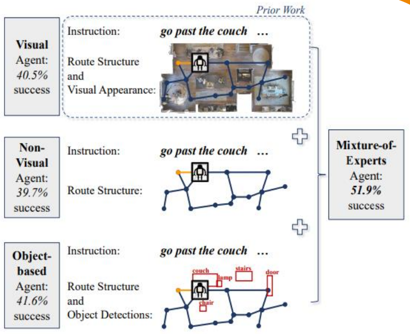
Fried et al. (NeurIPS '18) "Speaker-Follower"



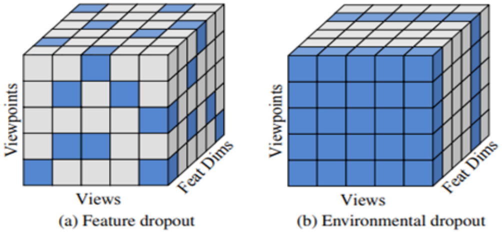
Wang et al. (CVPR '19) "Cross-Modal Matching"



Ma et al. (ICLR '19) "Self-Monitoring Agent



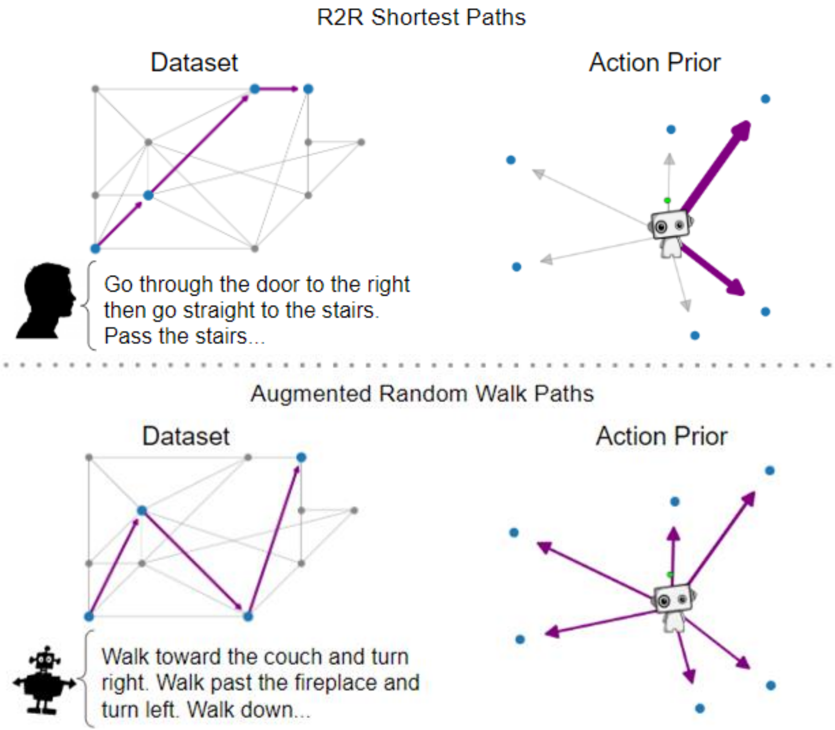
Hu et al. (ACL '19) "Are You Looking?"



Tan et al. (NAACL '19) "Environmental Dropout"

Our Method: Augmenting with Random Walks

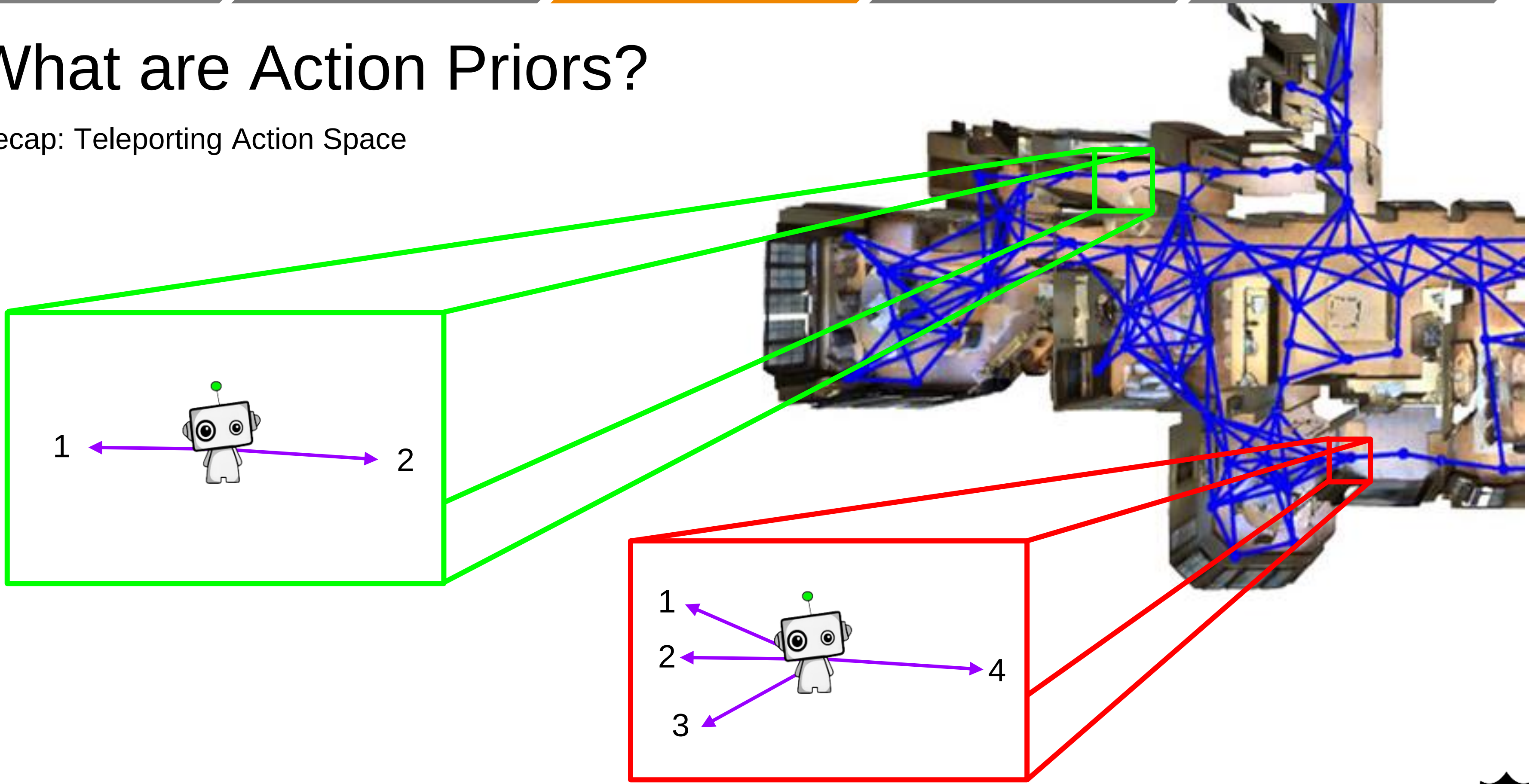
- **Lack of Data** → Data Augmentation
- **Poor generalizability** to unseen environments. → Reduce Priors
  - Action priors exist in **shortest path sampling**
    → Random Walk Sampling



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

What are Action Priors?

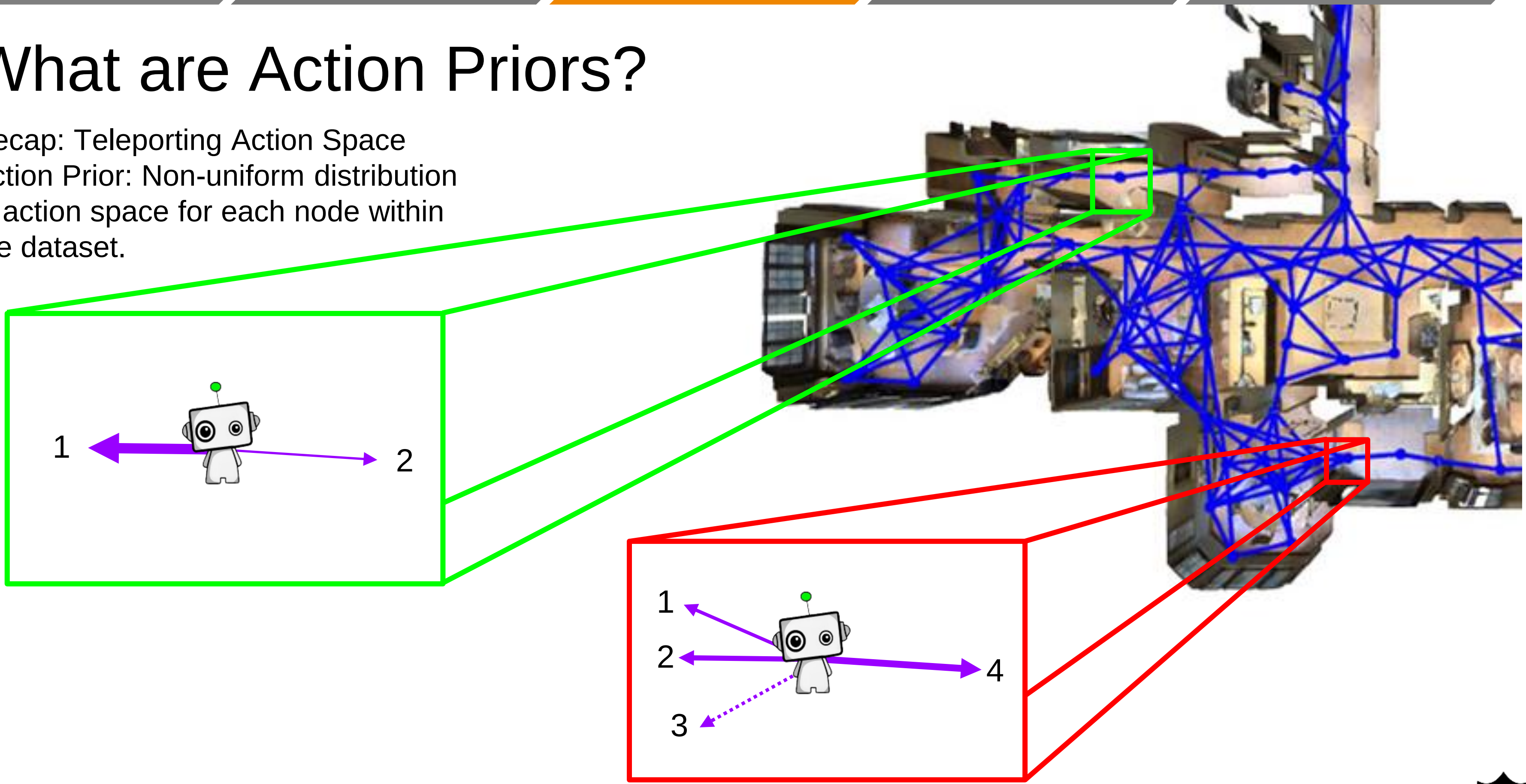
Recap: Teleporting Action Space



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

What are Action Priors?

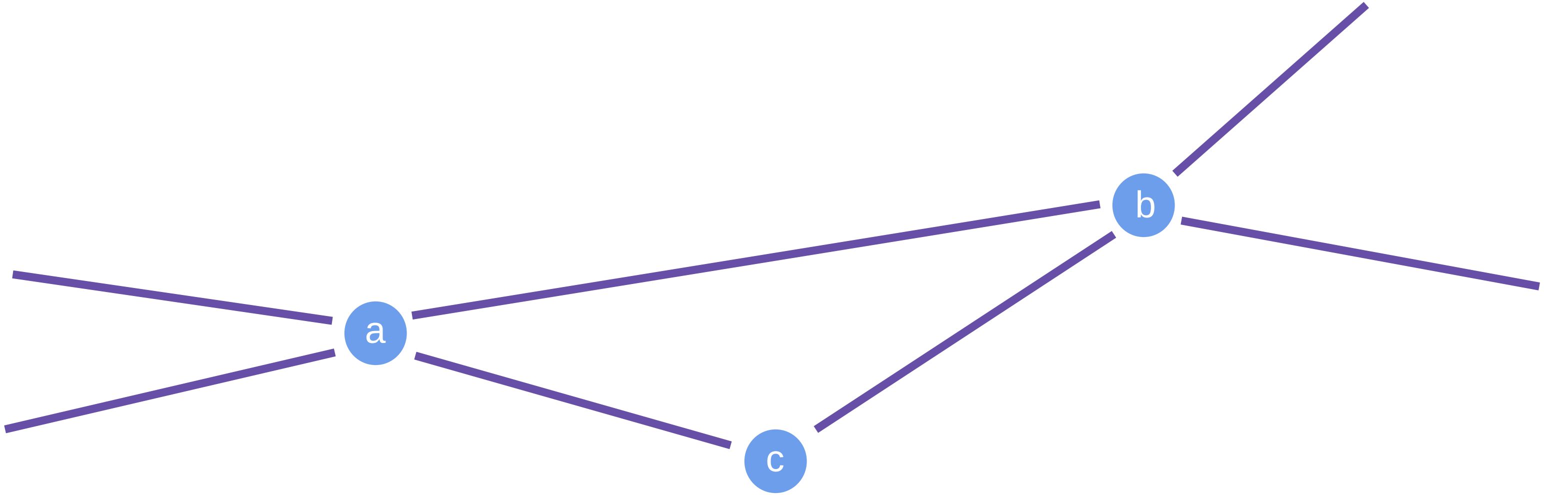
Recap: Teleporting Action Space
Action Prior: Non-uniform distribution in action space for each node within the dataset.



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Why do Action Priors Exist in VLN

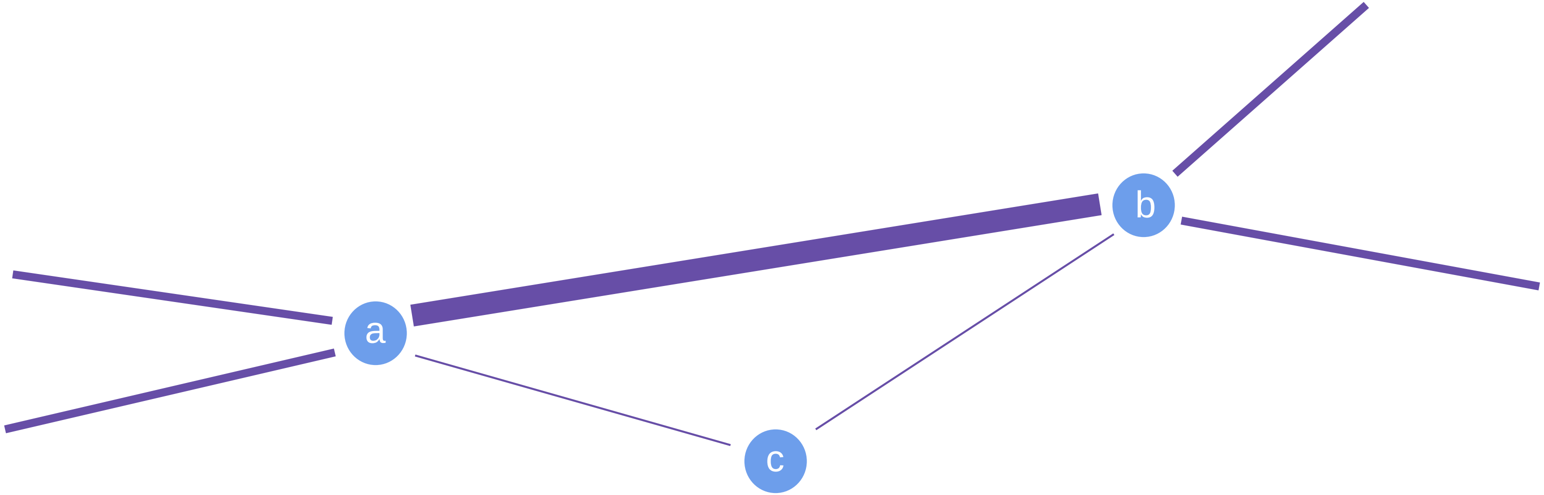
Intuitive Explanation



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Why do Action Priors Exist in VLN

Intuitive Explanation

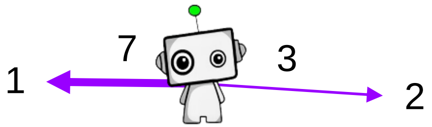


Many paths go through a ↔ b.
For a ↔ c or b ↔ c to appear in data, path must begin or end at c. **Less likely!**

Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

How Much can Action Priors be Exploited?

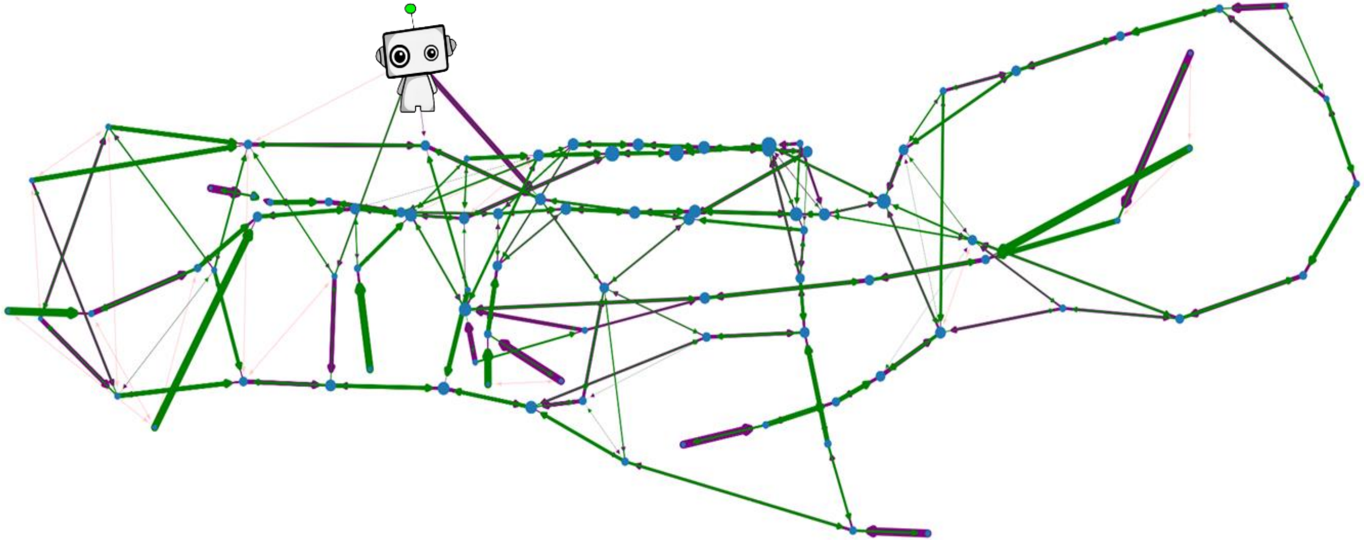
1. Calculate transition matrices from **Shortest** Path Augmented Data (Fried et al. NeurIPS '18)



Action	# Occurrence	Transition Probability
<Move to 1>	7	0.7
<Move to 2>	3	0.3

How Much can Action Priors be Exploited?

1. Calculate transition matrices from **Shortest** Path Augmented Data (Fried et al. NeurIPS '18)
2. Feed transition matrices to greedy agent.
3. Evaluate greedy agent.



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

How Much can Action Priors be Exploited?

Results

Input Modality	MTM		V + L
	Greedy	Random	Follower
Success Rate	0.35	0.12	0.66
Fraction of times agent stops within three meters of goal	Agent which follows Transition Matrices	Agent which performs random walks.	Agent described in Speaker-Follower (Fried et al. NeurIPS '18).

How Much can Action Priors be Exploited?

Results

Input Modality	MTM		V + L
	Greedy	Random	Follower
Success Rate	0.35	0.12	0.66

Points of interest:

- Having action priors allows agent to increase success rate by a factor of 3.

How Much can Action Priors be Exploited?

Results

Input Modality	MTM		V + L
	Greedy	Random	Follower
Success Rate	0.35	0.12	0.66

Points of interest:

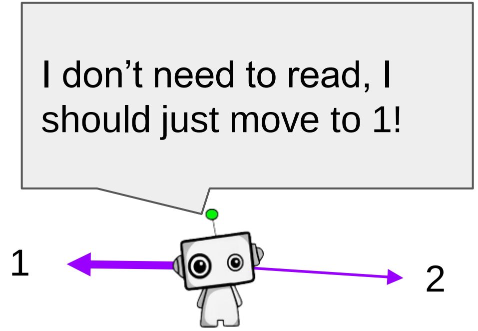
- Having action priors allows agent to increase success rate by a factor of 3.
- Even with no language language information, greedy agent is able to achieve over half the success rate of Follower.

Note: Input modality between Greedy agent and Follower are not the same, so direct comparison can't be made.

How do Action Priors Affect Generalization?

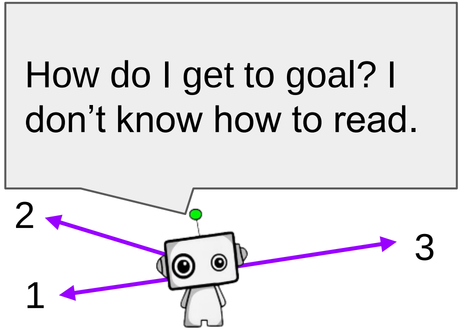
- Models localize through visual features.
- Choose action according to priors, rather than instructions.
- In new environment, no such priors exist.

Seen Environments



Move toward the...

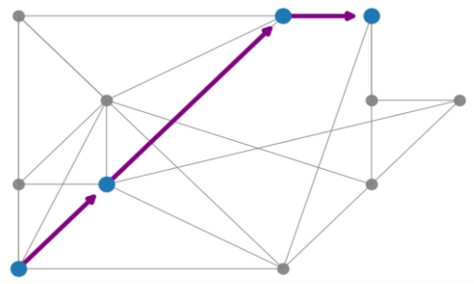
Unseen Environments



Exit through the...

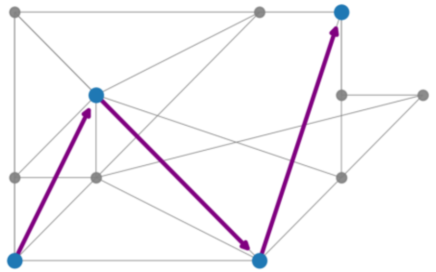
Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR ‘20.

Random Walk Sampling



Shortest Path Sampling:
- Pick two points in environment.
- Calculate shortest path.
- If # steps between 4-6, keep.

Contains Action Priors



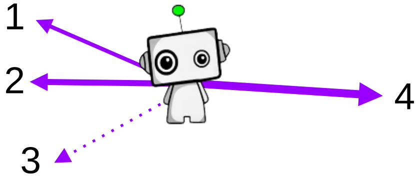
Random Walk Sampling:
- Randomly pick starting location.
- Sample # of steps.
- Take Random Walk while avoiding cycles.
- If end location > 3 meters from start, keep.

Mitigates action Priors

Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Does Random Walk Sampling Reduce Priors?

1. Calculate Transition Matrices from **Random Walk** Augmented Data.
2. Calculate Skew Factor of each Node. (Ratio between largest transition prob and uniform transition prob).



Action	Transition Probability
<Move to 1>	0.1
<Move to 2>	0.2
<Move to 3>	0
<Move to 4>	0.7

$$Skew\ Factor = \frac{0.7}{0.25} = 2.8$$

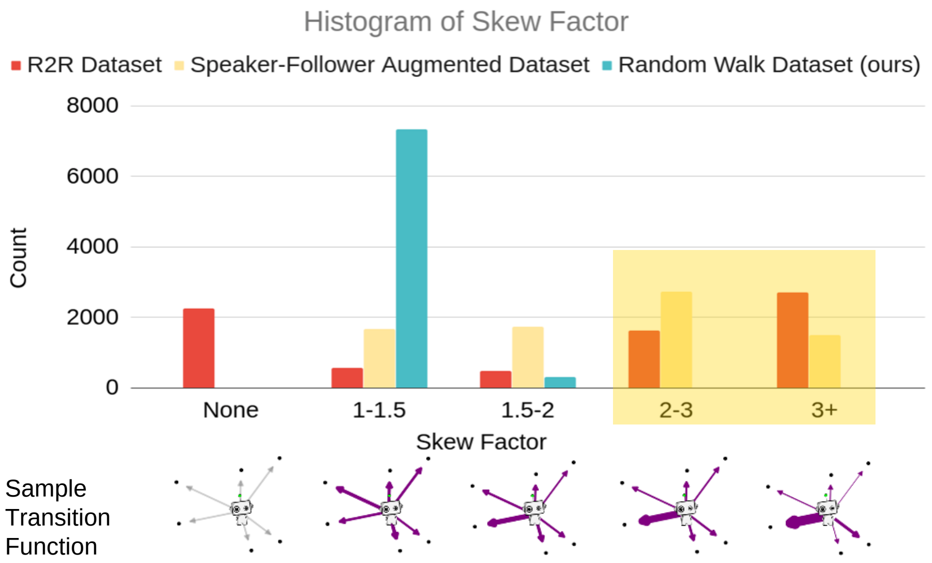
Skew Factors closer to 1 indicate smaller action prior.

Does Random Walk Sampling Reduce Priors?

1. Calculate Transition Matrices from **Random Walk** Augmented Data.
2. Calculate Skew Factor of each Node. (Ratio between largest transition prob and uniform transition prob).
3. Plot Histogram of skew factors.

Points of Interest:

- R2R and Speaker-Follower Augmented have large skew factors.



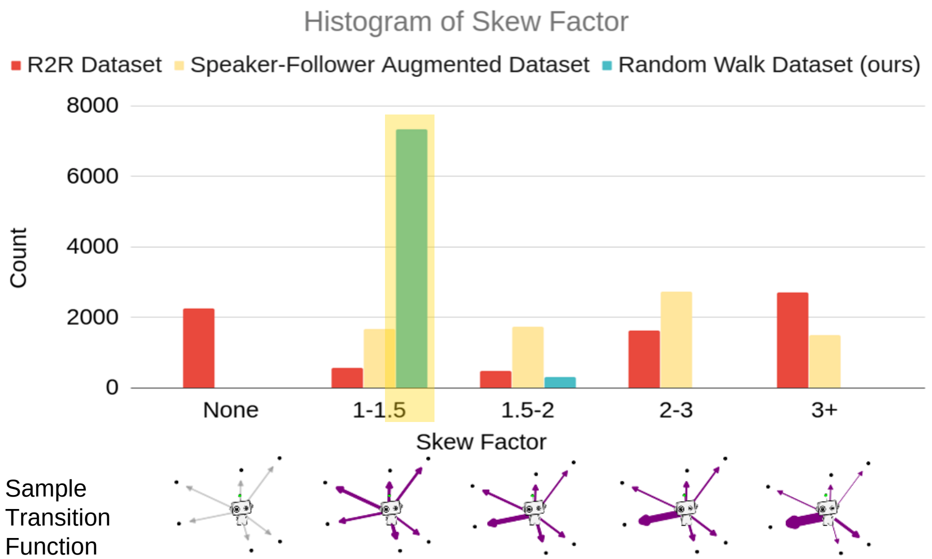
Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Does Random Walk Sampling Reduce Priors?

1. Calculate Transition Matrices from **Random Walk Sampling**.
2. Calculate Skew Factor of each Node. (Ratio between largest transition prob and uniform transition prob).
3. Plot Histogram of skew factors.

Points of Interest:

- R2R and Speaker-Follower Augmented have large skew factors.
- Random Walk skew factors close to 1.



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

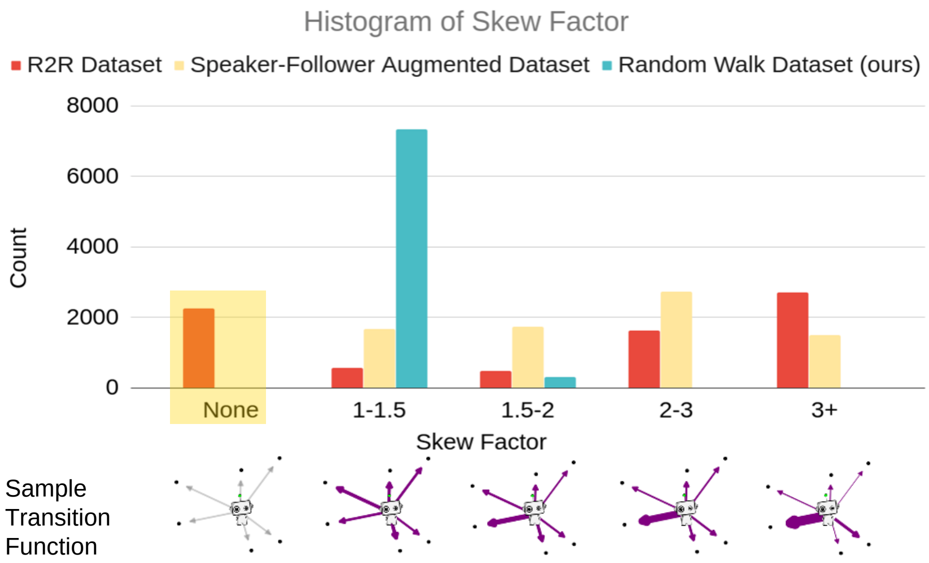
Does Random Walk Sampling Reduce Priors?

1. Calculate Transition Matrices from **Random Walk Sampling**.
2. Calculate Skew Factor of each Node. (Ratio between largest transition prob and uniform transition prob).
3. Plot Histogram of skew factors.

Points of Interest:

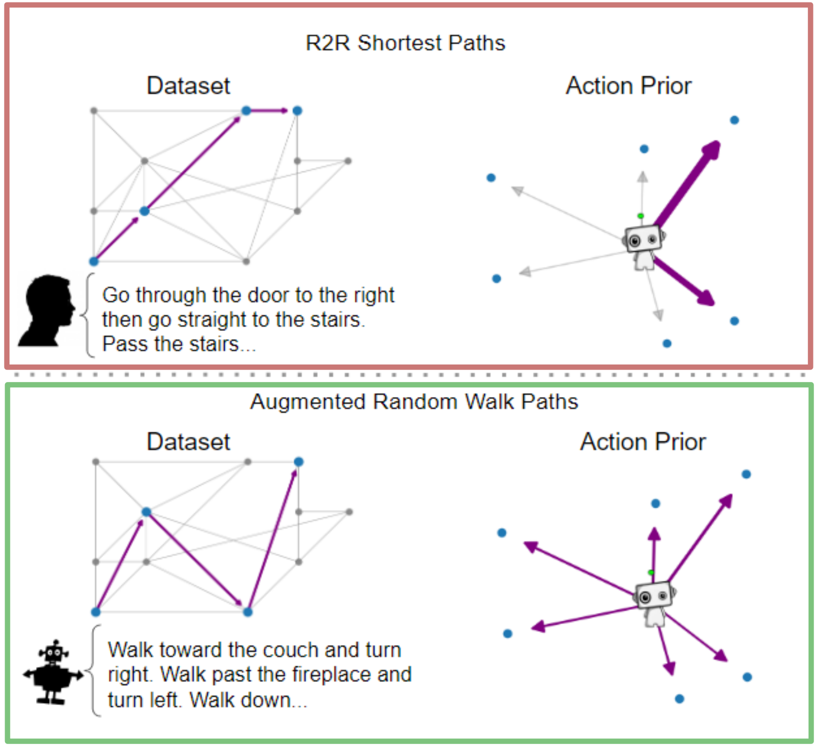
- R2R and Speaker-Follower Augmented have large skew factors.
- Random Walk skew factors close to 1.

Note: ‘None’ means node is never visited in the dataset.



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR ‘20.

Random Walk Sampling



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR ‘20.

Effect of Reducing Action Priors in Practice

Train Follower model from Speaker-Follower (Fried et al. NeurIPS 2018)
Compare between three different augmentation methods.

Room-to-Room Dataset

+

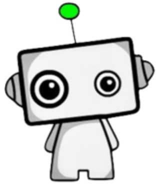
None

or

Shortest Path

or

Random Walk



Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Effect of Reducing Action Priors in Practice

Data Augmentation	Seen Validation Success Rate	Unseen Validation Success Rate	Difference in Success Rate
None	0.571	0.272	0.299
Shortest Path	**0.616**	0.297	0.319
Random Walk (ours)	0.530	**0.389**	0.141

Validation samples in environments seen during training

Validation samples in novel environments

Difference denotes how well agent generalizes to new environments

Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Effect of Reducing Action Priors in Practice

Data Augmentation	Seen Validation Success Rate	Unseen Validation Success Rate	Difference in Success Rate
None	0.571	0.272	0.299
Shortest Path	**0.616**	0.297	0.319
Random Walk (ours)	0.530	**0.389**	**0.141**

Data augmentation always helps in unseen environments.

Effect of Reducing Action Priors in Practice

Data Augmentation	Seen Validation Success Rate	Unseen Validation Success Rate	Difference in Success Rate
None	0.571	0.272	0.299
Shortest Path	**0.616**	0.297	0.319
Random Walk (ours)	0.530	**0.389**	0.141

Data augmentation always helps in unseen environments.
Although shortest path sampling has best seen validation success rate, random walk sampling has best unseen validation success.
Our method allows agents to generalize better to unseen environments.

Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Effect of Reducing Action Priors in Practice

Data Augmentation	Seen Validation Navigational Error	Unseen Validation Navigational Error
None	4.39	6.98
Shortest Path	**3.99**	6.85
Random Walk (ours)	5.03	**6.29**

Trend also hold for navigational error.

Effect of Reducing Action Priors in Practice

SPL = Success Rate Weighted by Path Length

Data Augmentation	Seen Validation SPL	Unseen Validation SPL
None	0.470	0.187
Shortest Path	**0.540**	0.201
Random Walk (ours)	0.504	**0.360**

Trend hold for Success Rate Weighted by Path Length as well.
Suggests counterintuitive notion that it is more effective to train agents on inefficient random paths in order to navigate efficiently during testing.

Felix Yu, Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky. Take the Scenic Route: Improving Generalization in Vision-and-Language Navigation. Learning with Limited Labels Workshop, CVPR '20.

Key Takeaways

- Shortest Path Sampling leads to priors in the action space.
- Using Random Walk data augmentation alleviates these priors while addressing lack of data for the task.
- A model trained using our random walk method generalize better to novel environments.

Lab website: https://visualai.princeton.edu/
Personal Website: https://www.cs.princeton.edu/~felixy
Email: felixy@princeton.edu