# Tactical Rewind:

## Self-correction via Backtracking in Vision-and-language Navigation

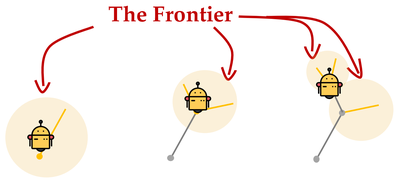
***Liyiming "Kay" Ke (kayke@uw.edu)***, *Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, Siddhartha Srinivasa*

## INTRO

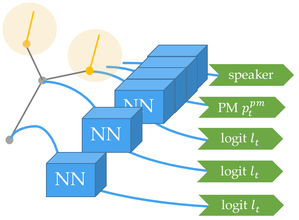

- The Vision-and-Language Navigation task, **VLN**, builds robots to navigate in houses from language instructions.
- Many approaches formulate it as a ***seq2seq*** task, build neural network and use either ***greedy*** or ***beam search*** decoding.
- We invented a ***plug-n-play*** procedure, **FAST**, that lets us reuse all the existing or new neural models – but more successfully and efficiently.

## METHOD

- ***Frontier-Aware*** Search with backtracking (FAST)



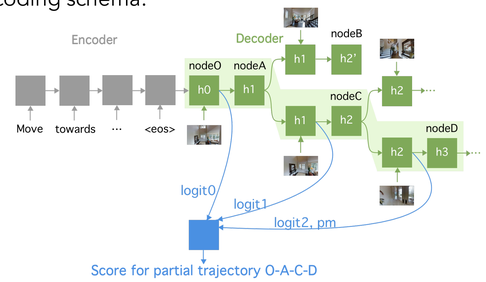

- As the agent explores, it asks:
  - Did I reach the target?
  - Am I lost?
  - Should I backtrack?
  - Where to backtrack to?




NN jointly scores the partial or completed trajectory

- To answer, we leverage existing neural models but only change the decoding schema.

- View the search running:




*Search* as a **global** framework,

*Neural Network* as a **local** heuristic.

Scan for code & paper

## RESULT

- The simple change to decoding enables immediate gains ***without modifications*** to existing models.

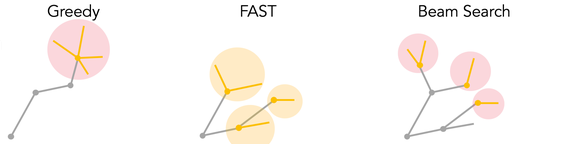
| Validation Unseen | SR (%) | SPL (%) | TL |
|---|---|---|---|
| SPEAKER-FOLLOWER | 37 | 28 | 15.32 |
| + FAST | 43 (+6) | 29 (+1) | 20.63 |
| SMNA | 47 | 41 | 12.61 |
| + FAST | 56 (+9) | 43 (+2) | 21.17 |

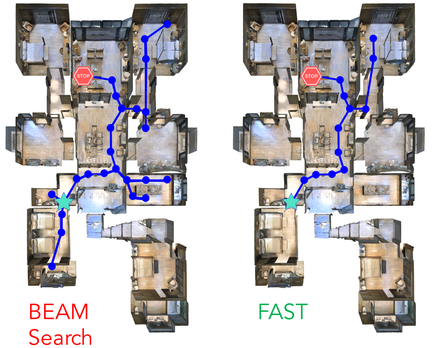
- We took the previous SoTA and achieved the new ***state-of-the-art*** in SPL at the time of submission. Shown on test unseen.

| | Model | TL | NE | SR | SPL |
|---|---|---|---|---|---|
| Greedy | SMNA | 18.04 | 5.67 | 0.48 | 0.35 |
| Greedy | SPEAKER-FOLLOWER | **14.82** | 6.62 | 0.35 | 0.28 |
| Greedy | + FAST (short) | 22.08 | **5.14** | **0.54** | **0.41** |
| Beam | SMNA | 373.09 | 4.48 | **0.61** | 0.02 |
| Beam | SPEAKER-FOLLOWER | 1,257.30 | 4.87 | 0.53 | 0.01 |
| Beam | + FAST (long) | **196.53** | **4.29** | **0.61** | **0.03** |
| | Human | 11.85 | 1.61 | 0.86 | 0.76 |

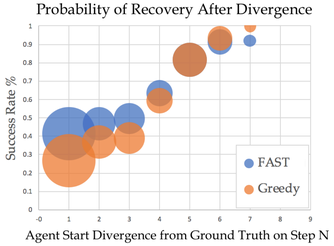
## DISCUSSION

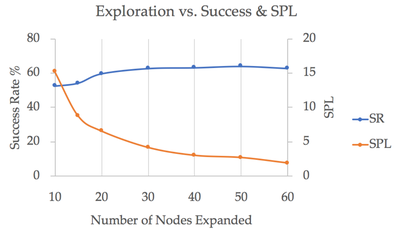
- FAST ***generalizes*** between greedy decoding and beam search.




- Compared with Beam Search, FAST is more ***efficient***.
- Compared with Greedy decoding, FAST is more ***robust***.




- FAST also offers a simple knob that one can tune to trade between success rate and efficiency.