Understanding Task-Oriented Language with Deep Learning

Ozan Arkan Can

April 10, 2018

Koç University, Artificial Intelligence Laboratory

Table of contents

Navigational Instruction Following

Navigational Instruction Following

Figure 1: The first person view of the agent.

- Example: “Turn right at the easel.”
- Goal: generate the right sequence of *MOVE*, *RIGHT*, *LEFT*, *STOP*

SAIL Dataset

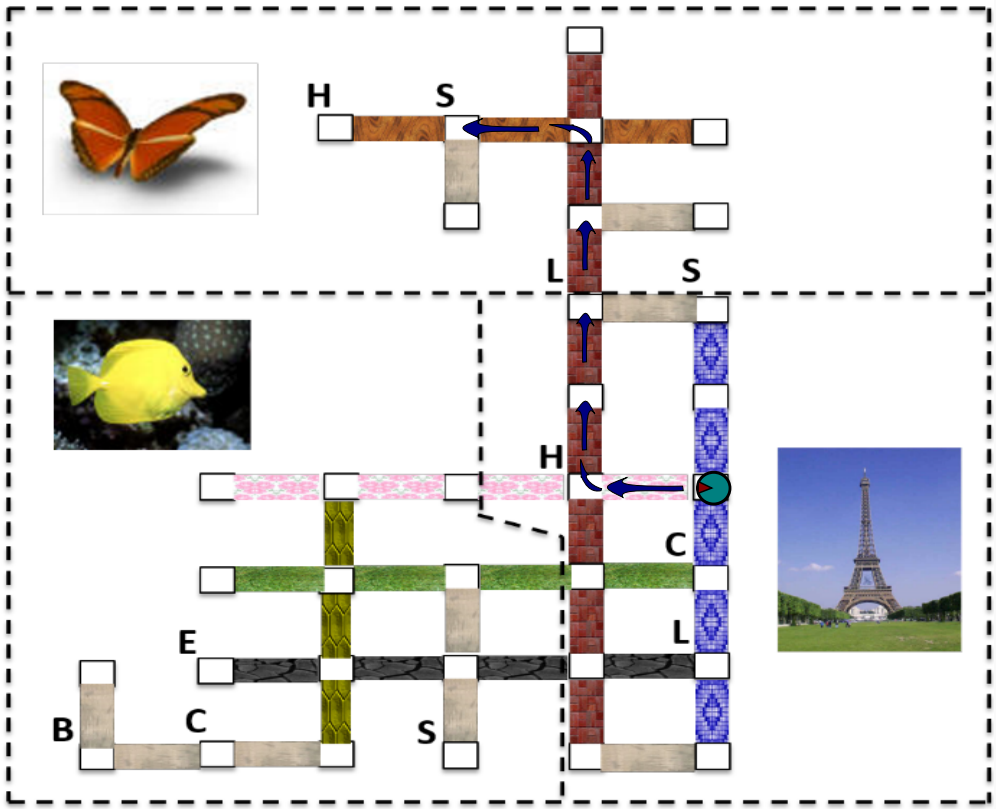


Figure 2: An example instruction with the corresponding path.

- ”Take the pink path to the red brick intersection. Go right on red brick. Go all the way to the wood intersection. Go left on wood. Position one is where the sofa is.”

Language

- Instructors and Followers
  - free-form language
  - syntactic and semantic errors
- A sequence of instructions for a (start, goal) pair
- Chen and Mooney (2011) split them into single sentences
- First version: *Paragraph*
- Second version: *Sentence*

Map	Sentence	Paragraph
Grid	874	224
Jelly	1293	242
L	1070	236

Table 1: Number of instances

Previous State of the Art

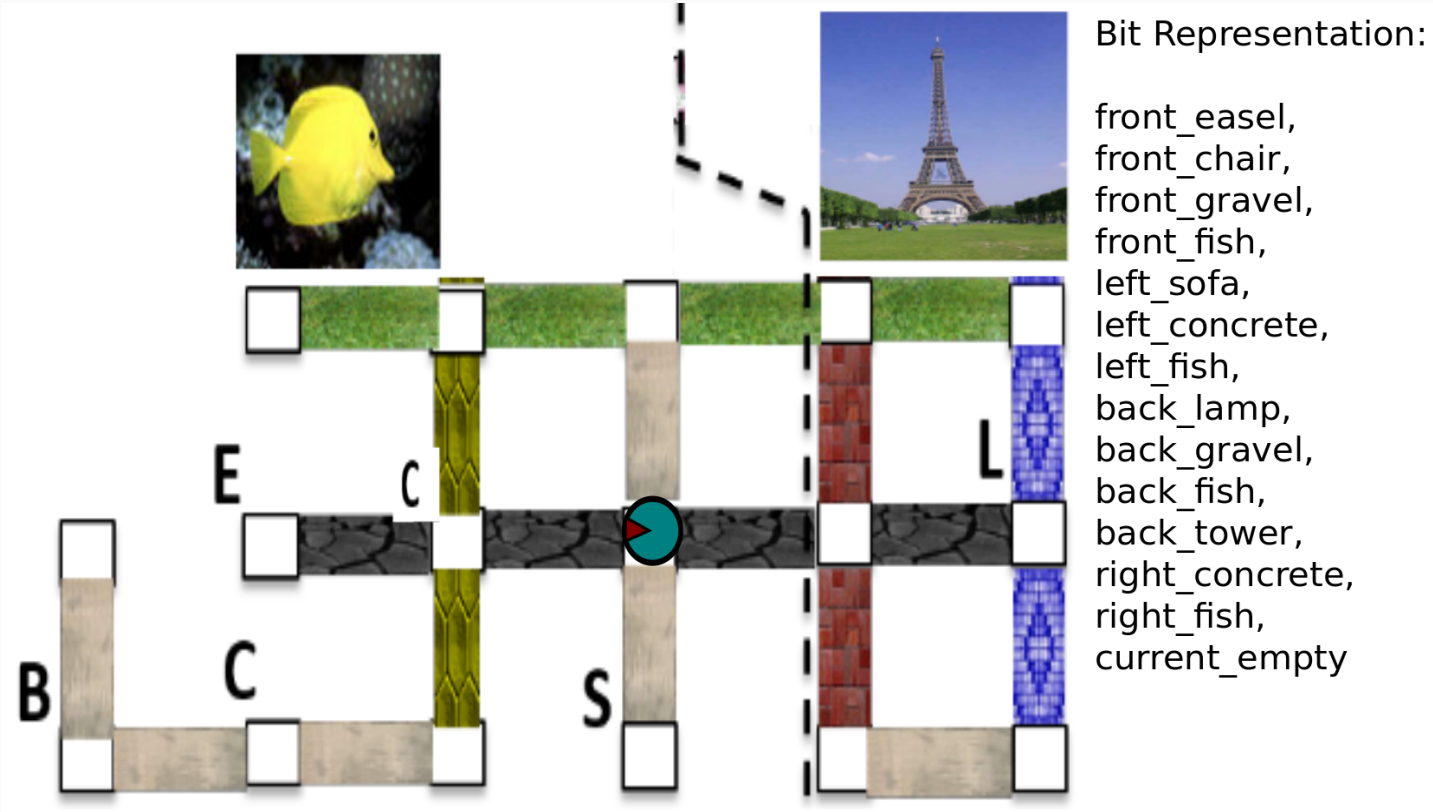
Comparison of previous studies

Method	Sentence	Paragraph
Chen and Mooney (2011)	54.40	16.18
Chen (2012)	57.28	19.18
Kim and Mooney (2012)	57.22	20.17
Kim and Mooney (2013)	62.81	26.57
Artzi and Zettlemoyer (2013)	65.28	31.93
Artzi et al. (2014)	64.36	***35.44***
Andreas and Klein (2015)	59.60	-
Mei et al. (2015) (vDev)	69.28	26.07
Mei et al. (2015) (vTest)	***71.05***	30.34
Human (MacMahon et al. 2006)	-	69.64

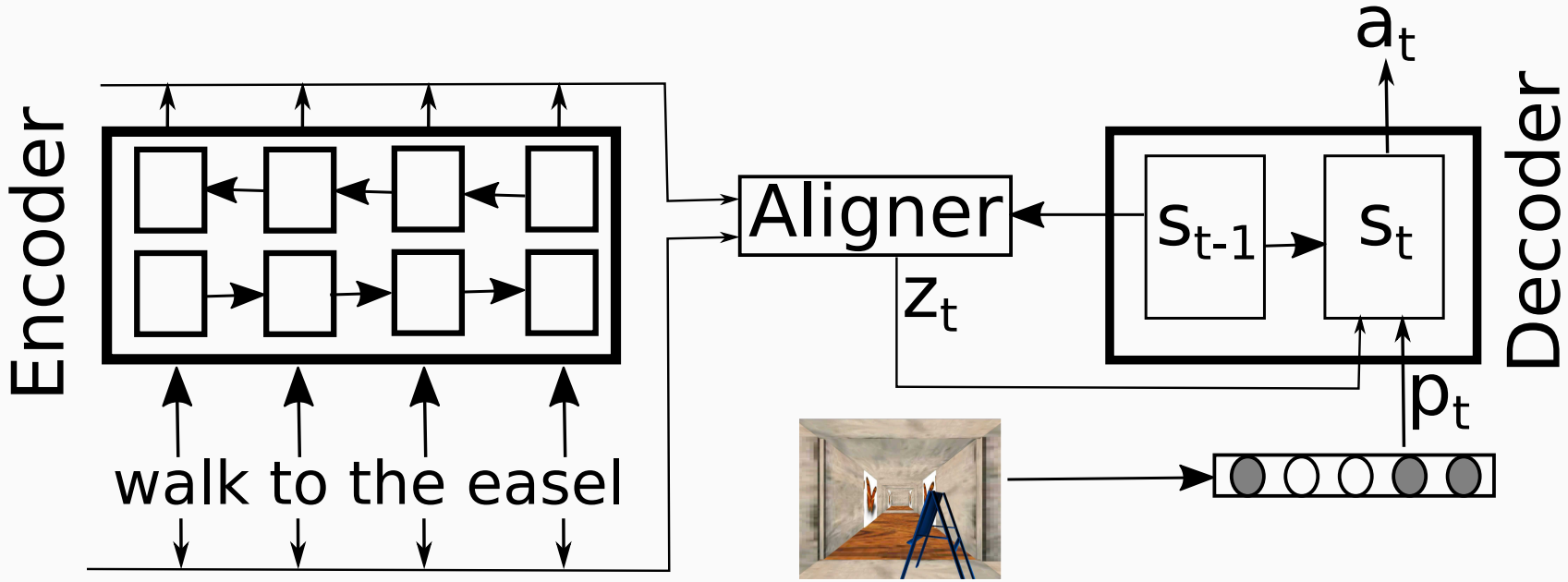
Table 2: Accuracy of reaching the final position.

Bag-of-features world representation (Mei et al. 2015)

- Agent Centric
- The concatenation of bag-of-features representation of each direction and the current position
- Spatial relations are not preserved
  - distance, order, relative position



Mei et al., 2015. ”Neural Mapping of Navigational Instructions to Action Sequences



Looking into the SAIL Dataset

Action Statistics

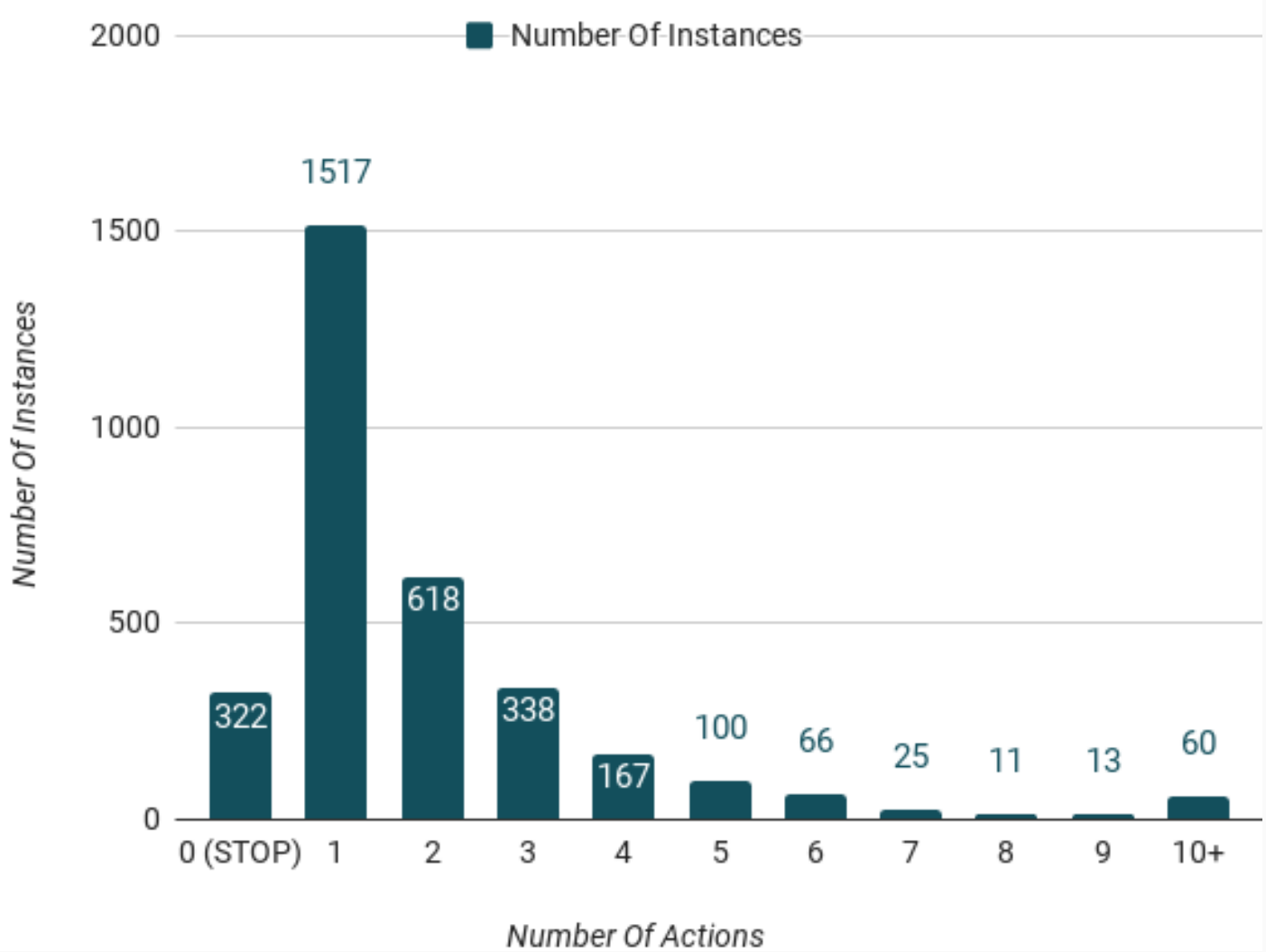


Figure 3: The distribution of the length of action sequences.

Instruction categories

- **Language Only (32%)**
  Turn (17%) : turn left, turn around, turn right twice
  Move (13%) : move one step, go straight two blocks
  Combination (2%) : walk forward once then take a left

  Visual (68%)
  Turn to X (7%) : face toward the hall with fish on the walls
  Move to X (13%) : move until the wall, move to the chair
  Turn and Move to X (2%): turn and move to the chair
  Orient (5%) : turn so the wall is on your back
  Describe (10%) : there should be the brick path on your right
  Move until (9%) : move until you see the green path on your left
  Combination (22%) : turn and move to the sofa, go towards the lamp on the brick road and take a right onto the grass, at the chair turn right

Language-only model

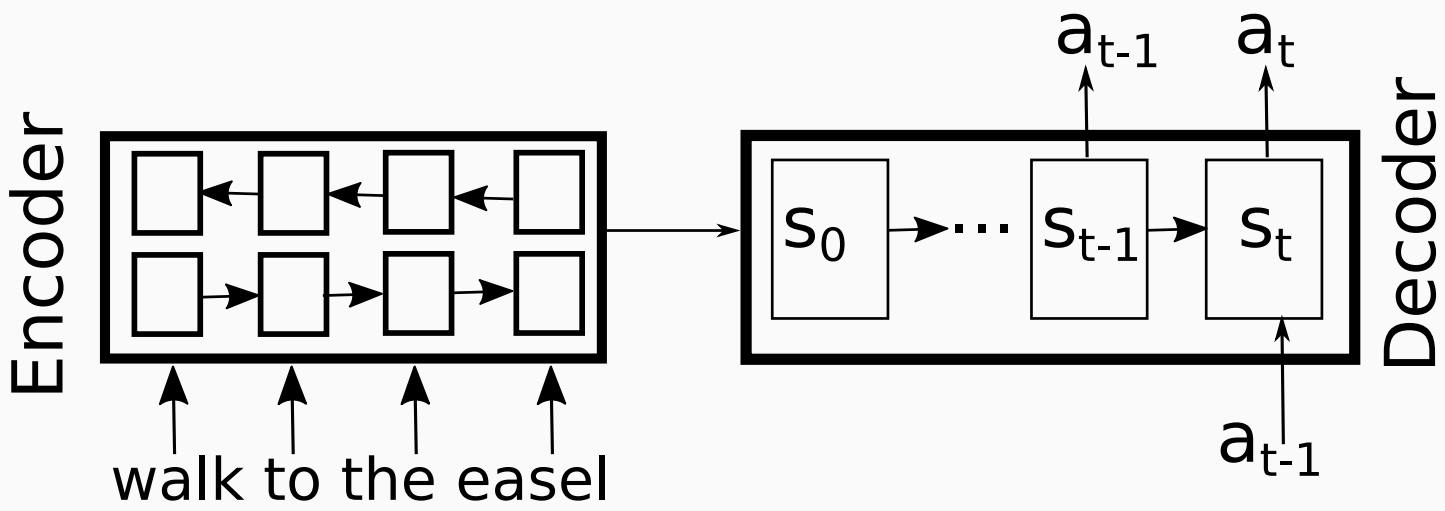


Figure 4: Sequence to sequence model without the perceptual information.

Language-only model

Category	Frq (%)	L.O. (%)
Language only	32	86.65
Turn to X	7	85.02
Move to X	13	62.82
Turn and Move to X	2	44.64
Orient	5	93.41
Describe	10	86.22
Move until	9	38.3
Combination	22	20.16
Overall	100	63.61

Table 3: The performance of the language only model

Bag-of-features model

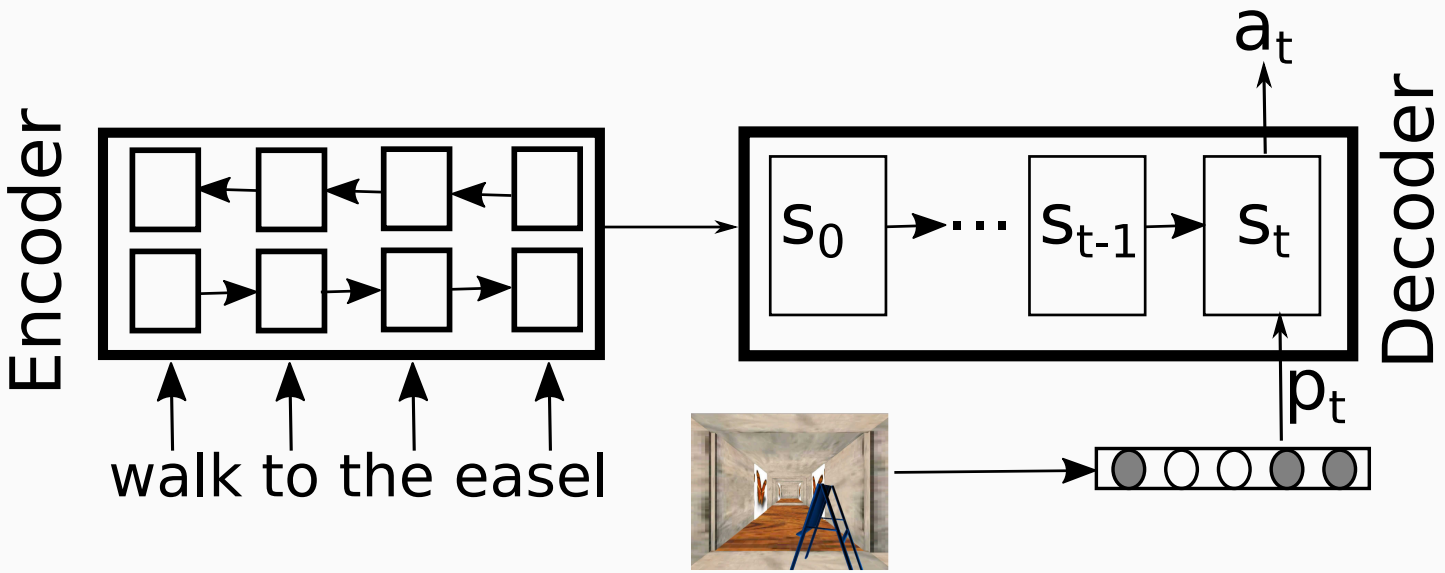


Figure 5: Sequence to sequence model with the perceptual information.

Bag-of-features model

Category	Frq (%)	L.O. (%)	BOF (%)
Language only	32	86.65	87.23
Turn to X	7	85.02	87.66
Move to X	13	62.82	73.21
Turn and Move to X	2	44.64	64.29
Orient	5	93.41	93.41
Describe	10	86.22	86.22
Move until	9	38.3	47.87
Combination	22	20.16	33.24
Overall	100	63.61	69.54

Table 4: The performance of the b.o.f model

A new architecture

Our model

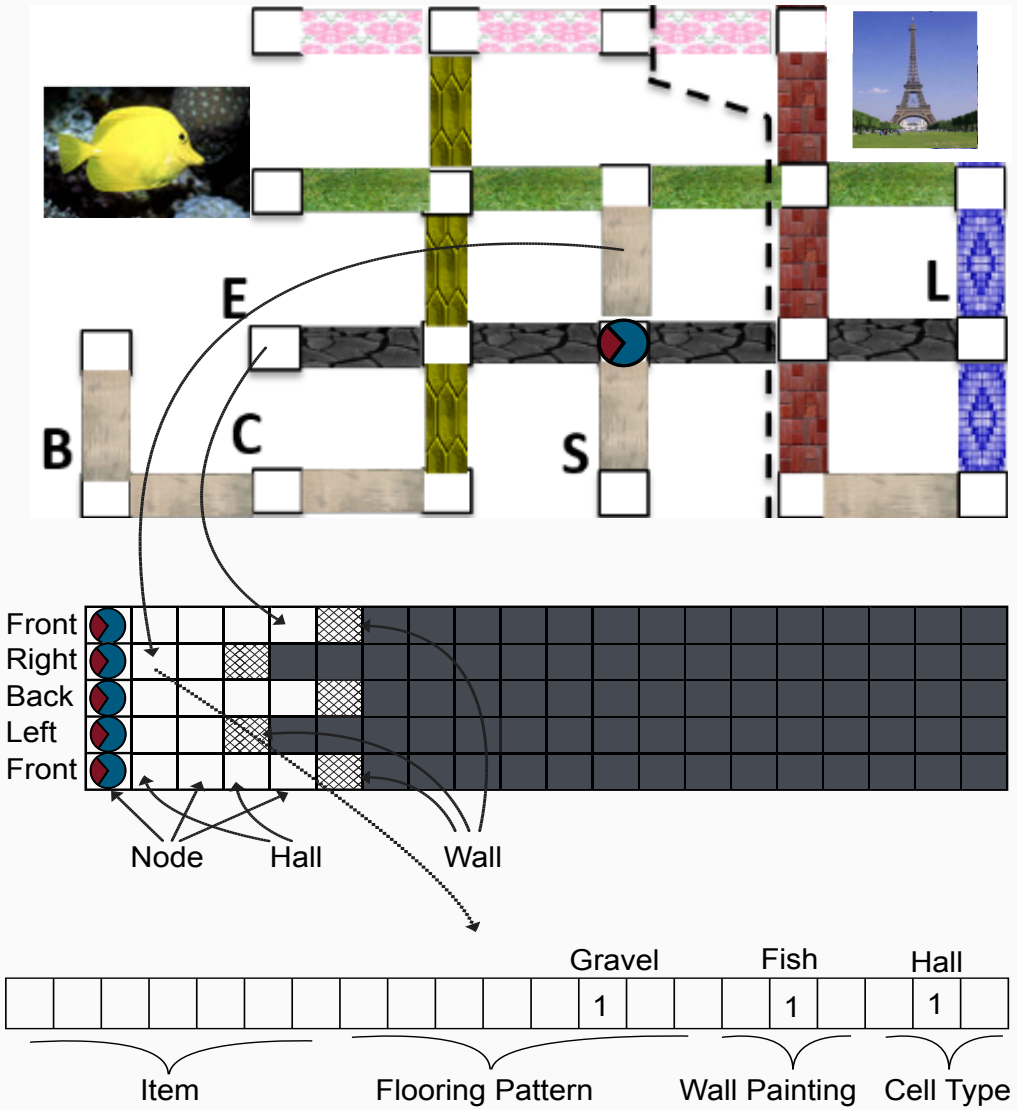


Figure 6: An example grid representation for the perceptual information.

Grid-based model

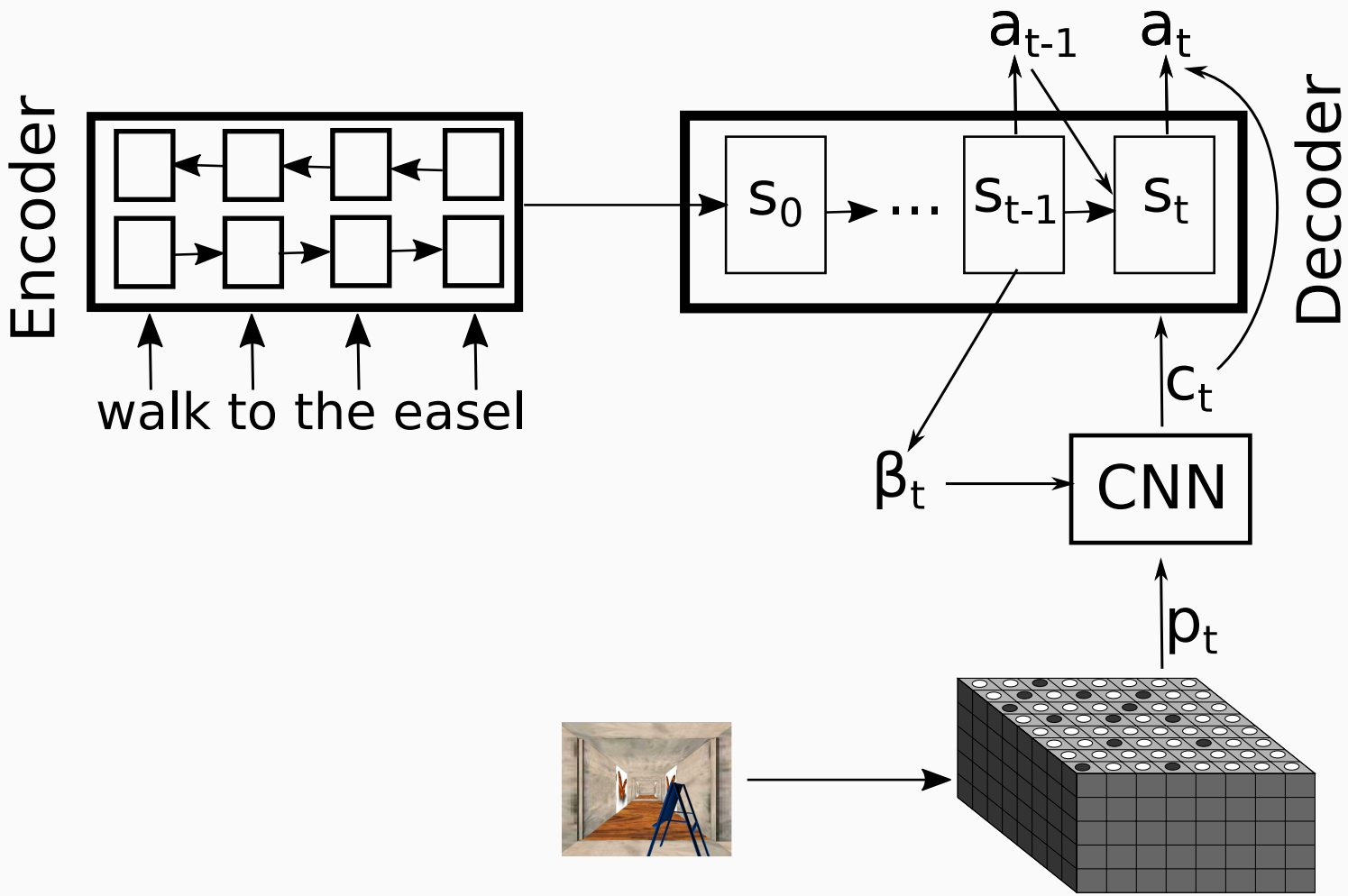


Figure 7: Sequence to sequence model with the perceptual information.

A cnn architecture with the channel attention

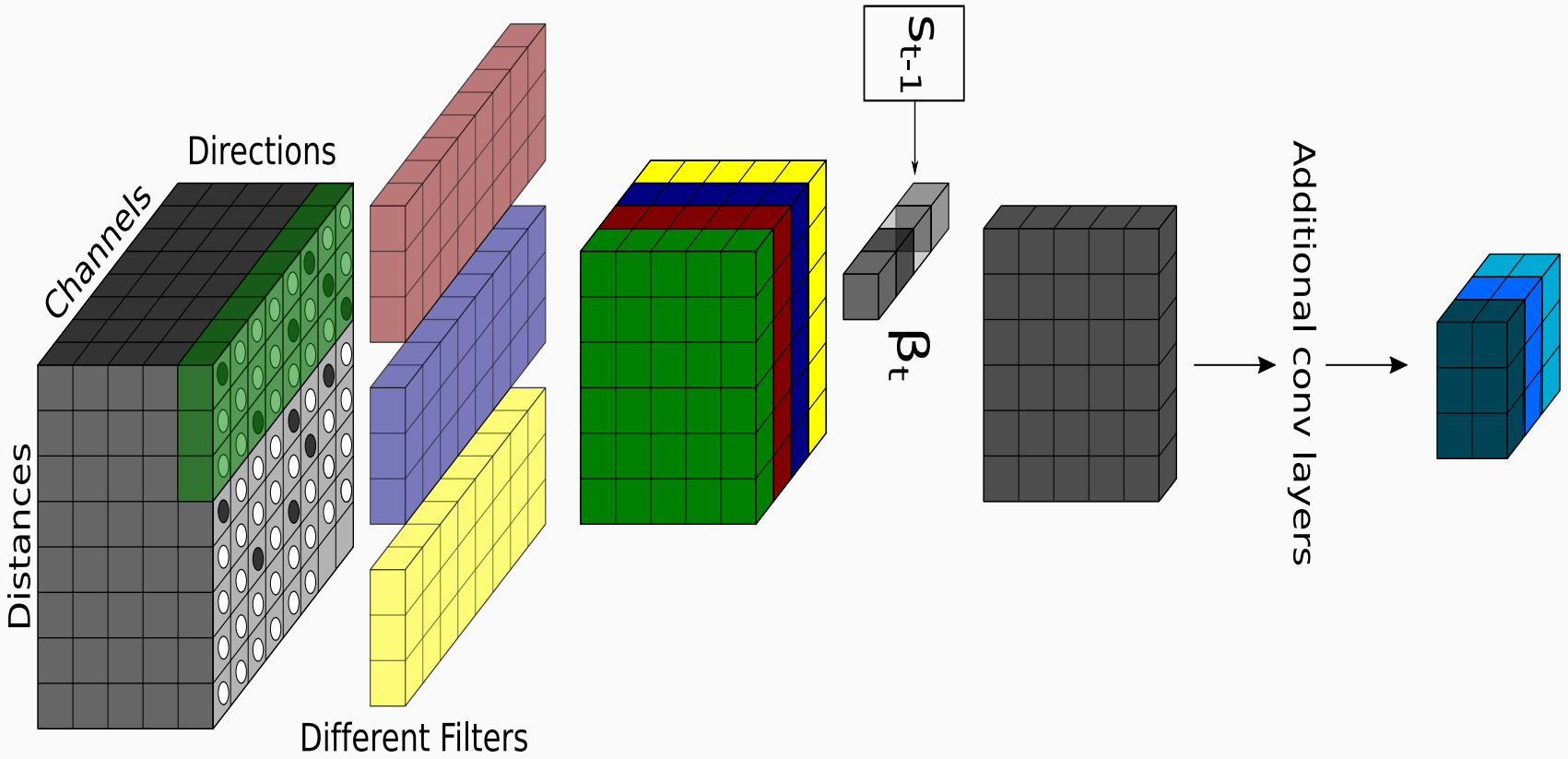


Figure 8: A cnn architecture with the channel attention

Our results

Category	Frq (%)	L.O. (%)	BOF (%)	Ours (%)
Language only	32	86.65	87.23	89.38
Turn to X	7	85.02	87.66	88.11
Move to X	13	62.82	73.21	76.91
Turn and Move to X	2	44.64	64.29	73.21
Orient	5	93.41	93.41	92.81
Describe	10	86.22	86.22	89.42
Move until	9	38.3	47.87	46.45
Combination	22	20.16	33.24	35.56
Overall (Ens=3)	100	63.61	69.54	71.58
Overall (Ens=10)				**72.82**
Mei et al. (Ens=10)				71.05

Table 5: The performance of the grid-based model

A new dataset

Motivation

- Solution for Data Sparsity
- Controllable Tasks
  - Language Complexity
  - World Complexity

Method - World Generation

- Generate a map randomly
- Decorate the map with floor and wall patterns
- Distribute items randomly
- Generate random start and goal positions
- Find a path from start to goal position

Method - Language Generation

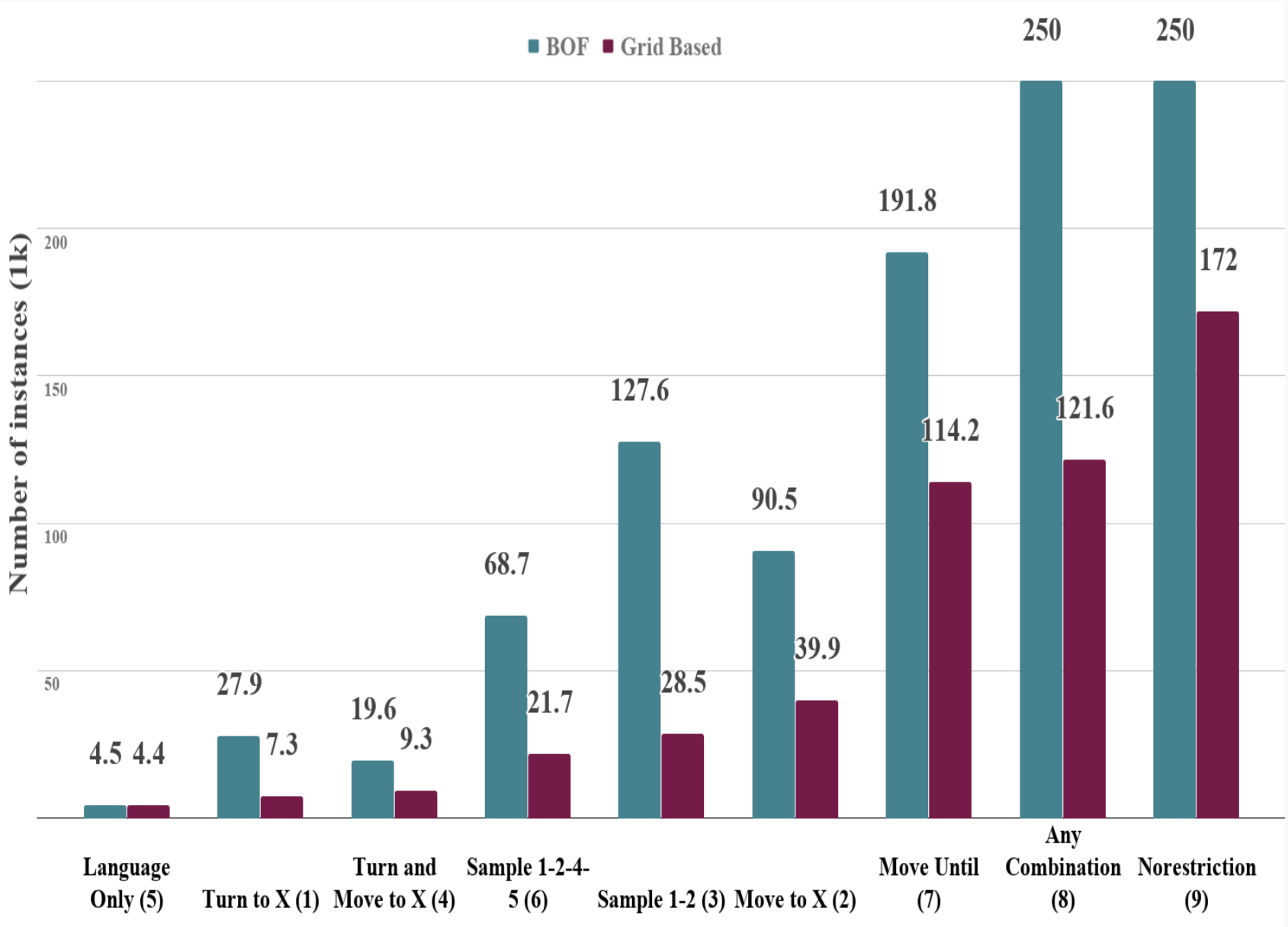
- Segment the path into turning and moving parts
- For each segment
  - Generate possible instructions using templates conditioned by the task category and the world configuration
  - Sample an instruction
- Task Patterns
  - reaching a corner, turning to a unique item, turning to a unique floor pattern
  - $\sim$ 5 patterns for each task
- Sentence Templates
  - move to the ”corner” / ”end of the path/hall/corridor”
  - ”turn” / ”turn your face” to the ”sofa” / ”bench”
  - $\sim$ 8 templates for each pattern

Coverage statistics of the artificial data

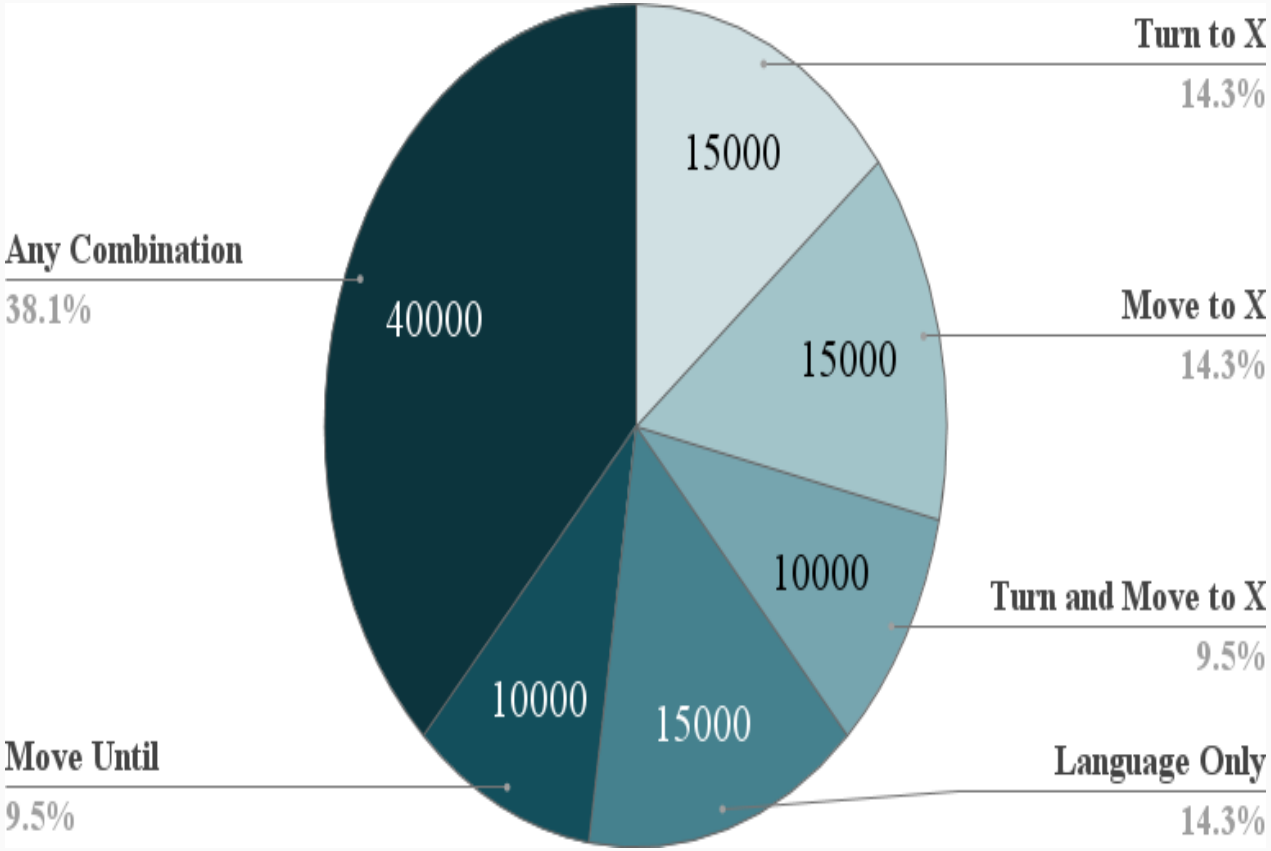
Task	Frequency	Overall	Non-Unique
Language Only	32%	90.25%	98.11%
Turn to X	7%	41.41%	73.75%
Move to X	13%	33.72%	60.95%
Turn and Move to X	2%	50.0%	70.0%
Orient	5%	64.67%	94.5%
Description	10%	14.42%	70.83%
Move Until	9%	2.84%	42.86%
Combination	22%	5.31%	23.86%

Table 6: Coverage statistics of the artificial data.

Convergence results with the new dataset



Fixed Dataset - Task Distribution



Results on the fixed data

	Dev		Test	
	Avg.	Ens.	Avg.	Ens.
L.O	54.85	54..41	55.03	54.64
B.O.F	84.9	86.04	84.61	85.42
Ours	91.93	**93.69**	91.82	**93.48**

Table 7: Performance comparison of baseline models and the proposed model.

Conclusion

We present

- a new world representation
- a grid-cnn based architecture
- a mechanism to connect language and perception
- model comparison by complexity

Questions?