Object Goal Navigation using Goal-oriented Semantic Exploration

Winner CVPR 2020 Habitat ObjectNav Challenge
Team Arnold (SemExp)

Devendra Singh Chaplot
Dhiraj Gandhi
Abhinav Gupta
Ruslan Salakhutdinov

Webpage: https://devendrachaplot.github.io/projects/semantic-exploration
Object Goal Navigation
Object Goal Navigation

Object Goal: dining table
Object Goal Navigation

Object Goal: dining table

Geometric Scene Understanding
Understanding navigable space

Semantic Scene Understanding
Object detection and segmentation

Passive
Object Goal Navigation

Object Goal: dining table

- Geometric Scene Understanding: Understanding navigable space
- Semantic Scene Understanding: Object detection and segmentation
- Learning Semantic Priors: Where is ‘dining table’ more likely to be found?
- Episodic Memory: Keeping track of explored and unexplored areas

Passive → Active
Active Neural SLAM

Sensor Pose Reading ($x'_t$)
Observation ($s_t$)
Action ($a_t$)

Neural SLAM ($f_{SLAM}$)
Local Policy ($\pi_L$)

Pose Estimate ($\hat{x}_t$)

Global Policy ($\pi_G$)

Map ($m_t$)
Short-term goal ($g^{st}_t$)

Long-term goal ($g^l_t$)

Short-term goal ($g^{st}_t$)

Map

$f_{Plan}$
Active Neural SLAM

Sensor Pose Reading ($x'_t$)

Observation ($s_t$)

Action ($a_t$)

Neural SLAM ($f_{SLAM}$)

Global Policy ($\pi_G$)

Long-term goal ($g^l_t$)

Local Policy ($\pi_L$)

Map ($m_t$)

Short-term goal ($g^s_t$)

Pose Estimate ($\hat{x}_t$)

$\text{Chaplot et al. ICLR-20}$
Incorporating Semantics

Obstacle Map Representation
(Active Neural SLAM)

Obstacle Map \((2 \times M \times M)\)

- Obstacles
- Explored Area
Incorporating Semantics

Obstacle Map Representation (Active Neural SLAM)

Obstacle Map \( (2 \times M \times M) \)

Semantic Map Representation (SemExp)

Semantic Map \( (K \times M \times M) \)

\[ K = C + 2 \]
Semantic Mapping

RGB ($I_i$)

Depth ($D_i$)
Semantic Mapping

RGB ($I_t$) → Mask RCNN → First-person Semantic Predictions

Depth ($D_t$)
Semantic Mapping

RGB ($I_t$) → Mask RCNN → First-person Semantic Predictions

Depth ($D_t$) → Point Cloud
Semantic Mapping

RGB ($I_t$) → Mask RCNN → First-person Semantic Predictions

Depth ($D_t$) → $X \ Y \ Z \ C_1 \ C_2 \ C_3$

Point Cloud → Semantic Labels
Semantic Mapping

RGB ($I_t$)

Depth ($D_t$)

Mask RCNN

First-person Semantic Predictions

Point Cloud

Semantic Labels

Voxel

$(C + 1) \times H \times M \times M$
Semantic Mapping

RGB ($I_t$) and Depth ($D_t$) inputs are processed through a Mask R-CNN model to obtain first-person semantic predictions. These predictions are then used to create a projection map of size $(C + 2) \times M \times M$, where $C$ represents semantic categories and $M$ is the map resolution.

The projection map is further processed to create a semantic mapping, represented as a voxel grid of size $(C + 1) \times H \times M \times M$, where $H$ is the height of the scene.

The map includes categories such as obstacles, explored area, and category-wise projections.
Semantic Mapping

RGB ($I_t$) and Depth ($D_t$) inputs are fed into a Mask R-CNN model for first-person semantic predictions. The predictions are then converted into a 3D point cloud and semantic labels. The category-wise semantic predictions are then combined with a denoising network to produce a semantic map prediction. The projection map is generated by summing across the height of all obstacles and exploring areas, resulting in a $(C + 2) \times M \times M$ dimensional map.
SemExp Model Overview

Sensor Pose Reading \((x_t)\)

Observation \((s_t)\) (RGBD)

Object Goal \((G = \text{“chair”})\)
SemExp Model Overview

Sensor Pose Reading \( (x_t) \)

Observation \( (s_t) \) (RGBD)

Object Goal \( (G = \text{“chair”}) \)

Semantic Mapping

Semantic Map \( (m_t) \)
SemExp Model Overview

Sensor Pose Reading ($x_t$)

Observation ($s_t$) (RGBD)

Object Goal ($G$ = "chair")

Semantic Mapping

Semantic Map ($m_t$)

Long-term goal ($g_t$)

Goal-Oriented Semantic Policy

Object Goal ($G = \text{"chair"}$)
SemExp Model Overview

Sensor Pose Reading ($x_t$)

Observation ($s_t$) (RGBD)

Object Goal ($G =$ “chair”)

Semantic Mapping

Semantic Map ($m_t$)

Long-term goal ($g_t$)

Goal-Oriented Semantic Policy

Deterministic Local Policy ($\pi_L$)

Action ($a_t$)
Demo Video

Observation (Goal: bed)

Predicted Semantic Map

Ground Truth

Navigable Area
0: chair
1: couch
2: potted plant

3: bed
4: toilet
5: tv
6: dining-table

7: oven
8: sink
9: refrigerator
10: book

11: clock
12: vase
13: cup
14: bottle

https://youtu.be/h56dA2uxpGU
ObjectGoal Navigation Results
ObjectGoal Navigation Results

ObjectGoal Navigation Results

Success Rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Success Rate</th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>RGBD + RL [1]</td>
<td>0.037</td>
<td>0.027</td>
</tr>
<tr>
<td>RGBD + Semantics + RL [2]</td>
<td>0.031</td>
<td>0.049</td>
</tr>
<tr>
<td>Classical Map + FBE</td>
<td>0.311</td>
<td>0.124</td>
</tr>
<tr>
<td>Active Neural SLAM [3]</td>
<td>0.321</td>
<td>0.145</td>
</tr>
<tr>
<td>SemExp</td>
<td>0.36</td>
<td>0.199</td>
</tr>
</tbody>
</table>

## Habitat Challenge Leaderboard

<table>
<thead>
<tr>
<th>Method</th>
<th>Test-standard</th>
<th></th>
<th></th>
<th>Minival</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPL</td>
<td>Success</td>
<td>Dist</td>
<td>SPL</td>
<td>Success</td>
<td>Dist</td>
</tr>
<tr>
<td>Arnold (SemExp)</td>
<td>0.071</td>
<td>0.179</td>
<td>8.818</td>
<td>0.246</td>
<td>0.467</td>
<td>3.334</td>
</tr>
<tr>
<td>Active Exploration</td>
<td>0.041</td>
<td>0.089</td>
<td>9.461</td>
<td>0.108</td>
<td>0.167</td>
<td>5.079</td>
</tr>
<tr>
<td>DD-PPO</td>
<td>0.021</td>
<td>0.062</td>
<td>9.316</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Blue Ox</td>
<td>0.017</td>
<td>0.060</td>
<td>8.903</td>
<td>0.083</td>
<td>0.133</td>
<td>4.254</td>
</tr>
<tr>
<td>SRCB-robot-sudoer</td>
<td>0.002</td>
<td>0.004</td>
<td>10.276</td>
<td>0.124</td>
<td>0.233</td>
<td>4.848</td>
</tr>
<tr>
<td>PPO RGBD</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>6.055</td>
</tr>
<tr>
<td>Random</td>
<td>0.000</td>
<td>0.000</td>
<td>10.330</td>
<td>0</td>
<td>0</td>
<td>6.379</td>
</tr>
</tbody>
</table>
Real-world Transfer

See video at https://devendrachaplot.github.io/projects/semantic-exploration
Thank you

Devendra Singh Chaplot

Webpage: http://devendrachaplot.github.io/
Email: chaplot@cs.cmu.edu
Twitter: @dchaplot