Neural Topological SLAM for Visual Navigation

CVPR-2020

Webpage: https://devendrachaplot.github.io/projects/Neural-Topological-SLAM
Semantic Priors and Common-Sense

• Humans use semantic priors and common-sense to explore and navigate everyday

• Most navigation algorithms struggle to do so
Semantic Priors and Common-Sense

- Humans use semantic priors and common-sense to explore and navigate everyday
- Most navigation algorithms struggle to do so
Image Goal Task

Source Image ($I_S$)

Goal Image ($I_G$)
Image Goal Task

- Agent observations are panoramic images
Image Goal Task

• Agent observations are panoramic images
• Take actions to navigate to the goal location
Image Goal Task

- Agent observations are panoramic images
- Take actions to navigate to the goal location
- Take the `stop’ action at the goal location
Image Goal Task

- Agent observations are panoramic images
- Take actions to navigate to the goal location
- Take the `stop’ action at the goal location
- Sequential goals
Prior work
Prior work

End-to-end Reinforcement or Imitation Learning

Observations → Neural Network → Actions

End-to-end Learning

- High sample complexity
- Ineffective in large environments
Prior work

End-to-end Reinforcement or Imitation Learning

- High sample complexity
- Ineffective in large environments

Modular Metric Maps

- Can not learn semantic priors
- Pose error accumulation
Topological Maps
Topological Maps
Topological Graph Representation
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas

- **Agent’s Current Node**
- **Regular Nodes**
- **Ghost Nodes**
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas

**Observation**

**Goal Image**

- Agent’s Current Node
- Regular Nodes
- Ghost Nodes

Selected Ghost Node
Topological Graph Representation

- **Nodes**: areas
- **Regular nodes**: Explored areas
- **Ghost nodes**: Unexplored areas
- **Edges**: Spatial relationship between nodes

---

**Observation**

- **Agent’s Current Node**
- **Regular Nodes**
- **Ghost Nodes**

**Goal Image**

- **Selected Ghost Node**
Four learnable functions
Four learnable functions

\( \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \)

\( \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \)

\( \mathcal{F}_L(I_1, I_2) = \text{Localization} \)

\( \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \)
Geometric Prediction
Geometric Prediction

\( \mathcal{F}_G(I_1) = \) Geometric Prediction: Free directions
Semantic Prediction
Semantic Prediction

$$\mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target}$$
Localization
\( \mathcal{F}_L(I_1, I_2) = \text{Localization} \)
Relative Pose Prediction
Relative Pose Prediction

\[
\mathcal{F}_R(I_1, I_2) = \text{Relative Pose}
\]
Neural Topological SLAM
Neural Topological SLAM

\( \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \)

\( \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \)

\( \mathcal{F}_L(I_1, I_2) = \text{Localization} \)

\( \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \)
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]
\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]
Neural Topological SLAM

$\mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions}$

$\mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target}$

$\mathcal{F}_L(I_1, I_2) = \text{Localization}$

$\mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction}$
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]

\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]

\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]

\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]

Graph \( G_{i-1} \)

Graph Update \( \ell_{GU} \)

Graph \( G_i \)

Global Policy \( \ell_{GP} \)

Local Policy \( \ell_{LP} \)

Image Obs \( I_i \)

Goal Image \( I_G \)

Subgoal Directions \( \Delta p \)

Navigation action \( a_i \)
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]
\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) \] = Geometric Prediction: Free directions
\[ \mathcal{F}_S(I_1, I_2) \] = Semantic Prediction: Closeness to target
\[ \mathcal{F}_L(I_1, I_2) \] = Localization
\[ \mathcal{F}_R(I_1, I_2) \] = Relative Pose Prediction
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]
\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]
\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]
Neural Topological SLAM

\[ \mathcal{F}_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ \mathcal{F}_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ \mathcal{F}_L(I_1, I_2) = \text{Localization} \]
\[ \mathcal{F}_R(I_1, I_2) = \text{Relative Pose Prediction} \]
Single supervised learning model
Single supervised learning model

- No reinforcement learning, no interaction needed
- Can be trained completely with static data
Demo video

Observation

Goal Image

Topological Map and Pose

- Goal Location
- Ghost nodes
- Node Locations
- Selected Ghost node
Demo video
Demo video

Observation

Goal Image

Topological Map and Pose
- Goal Location
- Ghost nodes
- Node Locations
- Selected Ghost node
Demo video

Observation

Goal Image

Topological Map and Pose
- Goal Location
- Ghost nodes
- Node Locations
- Selected Ghost node
Learning Semantic Priors

Observation

Goal Image

Topological Map and Pose

- Goal Location
- Node Locations
- Ghost nodes
- Selected Ghost node
Learning Semantic Priors

Observation

Goal Image

Topological Map and Pose

Goal Location
Ghost nodes
Node Locations
Selected Ghost node

0.20
0.17
0.13
0.27
0.76
0.56
Learning Semantic Priors

Observation

Goal Image

Topological Map and Pose
- Goal Location
- Ghost nodes
- Node Locations
- Selected Ghost node
Learning Semantic Priors

Observation

Goal Image

Topological Map and Pose
- Goal Location
- Node Locations
- Ghost nodes
- Selected Ghost node
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>RGBD</th>
<th>RGBD (No Noise)</th>
<th>RGBD (No Stop)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>End-to-end Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM + Imitation</td>
<td>0.10</td>
<td>0.14</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>LSTM + RL</td>
<td>0.10</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>Modular Metric Maps</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupancy Maps + FBE + RL</td>
<td>N/A</td>
<td>0.26</td>
<td>0.31</td>
<td>0.24</td>
</tr>
<tr>
<td>Active Neural SLAM</td>
<td>0.23</td>
<td>0.29</td>
<td>0.35</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>Topological Maps</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural Topological SLAM</td>
<td>0.38</td>
<td>0.43</td>
<td>0.45</td>
<td>0.60</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>RGB</td>
<td>RGBD</td>
<td>RGBD (No Noise)</td>
<td>RGBD (No Stop)</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----</td>
<td>------</td>
<td>----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>LSTM + Imitation</td>
<td>0.10</td>
<td>0.14</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>LSTM + RL</td>
<td>0.10</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>Modular Metric Maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupancy Maps + FBE + RL</td>
<td>N/A</td>
<td>0.26</td>
<td>0.31</td>
<td>0.24</td>
</tr>
<tr>
<td>Active Neural SLAM</td>
<td>0.23</td>
<td>0.29</td>
<td>0.35</td>
<td>0.39</td>
</tr>
<tr>
<td>Topological Maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural Topological SLAM</td>
<td>0.38</td>
<td>0.43</td>
<td>0.45</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Robustness to Pose Noise

NTS is better than occupancy map models, captures and uses semantic priors.
Sequential Goals and Ablations

<table>
<thead>
<tr>
<th>Model</th>
<th>Easy</th>
<th>Med.</th>
<th>Hard</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGBD - No stop RGB - No Noise</td>
<td>0.76</td>
<td>0.28</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>ResNet + GRU + IL</td>
<td>0.71</td>
<td>0.18</td>
<td>0.06</td>
<td>0.32</td>
</tr>
<tr>
<td>Target-driven RL</td>
<td>0.89</td>
<td>0.45</td>
<td>0.21</td>
<td>0.52</td>
</tr>
<tr>
<td>Metric Spatial Map + RL</td>
<td>0.89</td>
<td>0.45</td>
<td>0.21</td>
<td>0.52</td>
</tr>
<tr>
<td>Metric Spatial Map + FBE + RL</td>
<td>0.92</td>
<td>0.46</td>
<td>0.29</td>
<td>0.56</td>
</tr>
<tr>
<td>Active Neural SLAM (ANS)</td>
<td>0.93</td>
<td>0.50</td>
<td>0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>Neural Topological SLAM (NTS)</td>
<td>0.94</td>
<td>0.70</td>
<td>0.60</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 2: No stop and no noise. Success rate of the proposed model NTS and the baselines without stop action (left) and without motion noise (right) in the RGBD setting.

6.1. Ablations and Sequential Goals

In this subsection, we evaluate the proposed model on sequential goals in a single episode and study the importance of the topological map or the graph and the Semantic Score Predictor ($F_S$). For creating a test episode with sequential goals, we randomly sample a goal between 1.5 m to 5 m away from the last goal. The agent gets a time budget of 500 timesteps for each goal. We consider two ablations:

- NTS w/o Graph. We pick the direction with the highest score in the current image greedily, not updating or using the graph over time. Intuitively, the performance of this ablation should deteriorate as the number of sequential goals increases as it has no memory of past observations.
- Neural Topological SLAM w/o Score Function. In this ablation, we do not use the Semantic Score Predictor ($F_S$) and pick a ghost node randomly as the long-term goal when the Goal Image is not localized in the current graph. Intuitively, the performance of this ablation should improve with the increase in the number of sequential goals, as random exploration would build the graph over time and increase the likelihood of the Goal Image being localized.

We report the success rate and SPL of NTS and the two ablations as a function of the number of sequential goals in Figure 8. Success, in this case, is defined as the ratio of goals reached by the agent across a test set of 1000 episodes. Firstly, the performance of NTS is considerably higher than both the ablations, indicating the importance of both the components. The performance of all the models decreases with an increase in the number of sequential goals because if the agent fails to reach an intermediate goal, there is a high chance that the subsequent goals are farther away. However, the performance gap between NTS and NTS w/o Score Function decreases and the performance gap between NTS and NTS w/o Graph increases with increase in the number of sequential goals as expected. This indicates that the topological map becomes more important over time as the agent explores a new environment, and while the Semantic Score Predictor is the most important at the beginning to explore efficiently.

7. Discussion

We designed topological representations for space that leverage semantics and afford coarse geometric reasoning. We showed how we can build such representation autonomously and use them for the task of image-goal navigation. Topological representations provided robustness to actuation noise, while semantic features stored at nodes allowed the use of statistical regularities for efficient exploration in novel environments. We showed how advances made in this paper make it possible to study this task in settings where no prior experience from the environment is available, resulting in a relative improvement of over 50%.

In the future, we plan to deploy our models on real robots. Acknowledgements

This work was supported by IARPA DIV A D17PC00340, ONR Grant N000141812861, ONR MURI, ONR Young Investigator, DARPA MCS, Apple and Nvidia.

Gibson License: http://svl.stanford.edu/gibson2/assets/GDS_agreement.pdf
Sequential Goals and Ablations

Semantic score function improves efficiency when no prior experience with environment is available.

As experience in environment increases, utility of semantic function decreases.

But, at the same time, importance of the topological representation increases.
Neural Topological SLAM for Visual Navigation
Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, Saurabh Gupta
CVPR 2020

Webpage: https://devendrachaplot.github.io/projects/Neural-Topological-SLAM

Thank you

Devendra Singh Chaplot

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