SMeLt: 2D Semantic Map Prediction with Latent Context

Wang Jiaming (A0040976Y)
Jamie-w@nus.edu.sg
School of Computing, NUS

Introduction

The utilization of a 2D semantic map is prevalent in object goal navigation tasks. SOTA techniques employ a pipeline (fig. 1) to construct the semantic map, which involves the following steps:

(1) Generation of a 3D point cloud,
(2) Object detection using pretrained models
(3) Projection of the 3D point cloud onto a 2D grid map

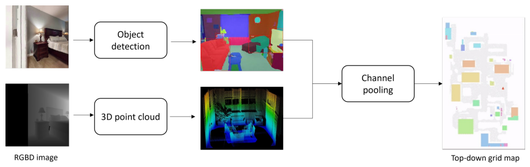

Fig. 1 Classic pipeline to construct the semantic map

However, detecting small objects poses challenges to this classical pipeline due to the presence of noisy sensors and limited resolution, which can impede the object detection module's ability to accurately identify or even detect the object.

To address this issue, we propose the **S**emantic **M**ap Pr**e**diction with **L**atent Contex**t** (SMeLt) framework to tackle the challenges associated with detecting small objects during the 2D semantic mapping procedure. It is inspired by the observation that humans can use contextual information to aid the object searching process as shown in fig. 2.

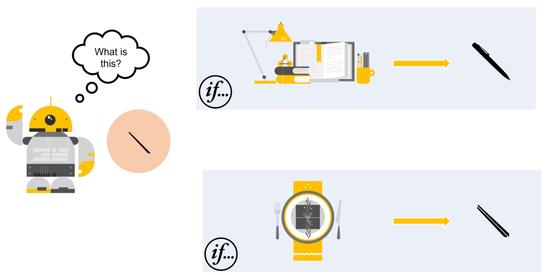

Fig. 2 Human can effectively use context information in the object recognition process.

Methodology

Task definition

The task is to generate a 2D grid map, represented by a tensor of dimensions $C \times H \times W$, from the RGBD observation of an agent. The tensor is structured such that each cell corresponds to a physical location in the world and each entry in the tensor denotes the presence of a specific object within that location, and C represents the total number of distinct object classes.

Formulation

We formulate the problem as a a simplified probabilistic graphical model (PGM) as depicted in fig. 3, Within this model, the latent context is represented by z, and contains high level information about the surroundings, e.g. in a kitchen or a bedroom. The semantic label, denoted as s, is a binary random variable, indicating the existence of a specific object in the given cell. Finally, the RGBD observation is denoted by o. Note we assume conditional independence between different cells in this model: $s_i^k \perp s_j^k | z$, for i $\neq j$

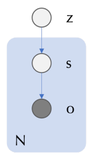

Fig. 3 PGM

$$
\begin{aligned}
\mathcal{L} &= \sum_n log\, p(s,o) \\
&= \sum_n log(\sum_z p(s|o,z)p(o)p(z)) \\
&\propto \sum_n log(\mathbb{E}_z(p(s|o,z)) \\
&\geq \sum_n \mathbb{E}_z\, log\, p(s|o,z) \\
&= \sum_n \mathbb{E}_{z\sim q(z)} \frac{p(z)}{q(z)} log\, p(s|o,z)
\end{aligned}
$$

Given the above PGM, we can derive our training objective by maximizing the observed data log likelihood $\mathcal{L}$ as shown in above derivation. We used Jensen's inequality in step 4, and importance sampling in step 5 to learn from more informative z.

SMeLt framework

We amortize $q(z)$ as a neural network f_θ that learns to encode the contextual information into some latent variable z. We parametrize $p(s|o,z)$ as a visual encoder f_ϕ, and a map decoder f_ω, as illustrated in fig. 3

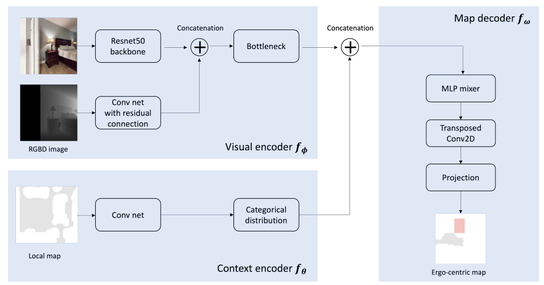

Fig. 4 Overview of the SMeLT framework

The neutral networks can be trained using the following algorithm:

Given f_θ: $q(z)$; f_ϕ and f_ω: $p(s|o,z)$; m: number of samples per observation;

1. Sample $z^m \sim q(z)$ for m=1,...M
2. Approximate $\sum_n \mathbb{E}_{z\sim q(z)} \frac{p(z)}{q(z)} log\, p(s|o,z) \approx \sum_n \frac{1}{m} \sum_m \frac{p(z^m)}{q(z^m)} log\, p(s|o,z^m)$
3. Calculate weight cross entropy loss $\mathcal{L} = \sum_n \frac{1}{m} \sum_m \frac{p(z^m)}{q(z^m)} (f_\theta(o,z^m) log s + (1 - f_\theta(o,z^m)) log(1-s))$
4. Update ϕ, θ with $\frac{\partial \mathcal{L}}{\partial \phi}, \frac{\partial \mathcal{L}}{\partial \theta}$

Dataset

We rendered 127,732 RGB and depth images from randomly sampled view-points in the Matterport 3D semantics scene dataset using Habitat-sim environment., together with calculated ground truth semantic map. The test set is generated using novel scenes that are not presented in the training set. We have published this dataset on Kaggle to foster further research on this task.

Results

We compared our proposed method with a baseline system that does not utilize the contextual information (table 1). We observe a 34% and 11% improvement when using the contextual information, measured by cross-entropy (CE) error on predicted semantic map for all object classes and small object classes respectively.

Model	CE error (all objects)	Improvement	CE error (small objects)	Improvement
W/o context z:	0.055	-	0.078	
With context z	**0.041**	**34%**	**0.070**	**11%**

Table 1. experiment results

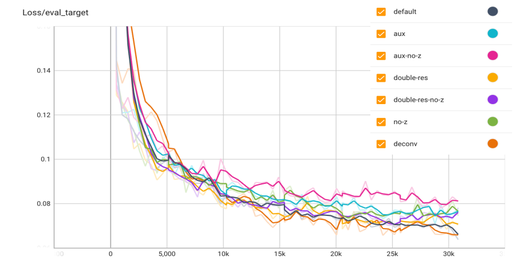

Fig. 5 Evaluation CE loss

In addition to the proposed architecture (fig. 4), we also tested other architectures including:

- MLP decoder (default in fig. 5)
- Using image segmentation as auxiliary loss (aux)
- Double resolution (double res)

The evaluation CE loss curve (Fig. 5) shows that architectures using contextual information consistently outperform those not using the contextual information. A qualitative comparison between the proposed method and baseline that without using contextual information can be found in fig 6.

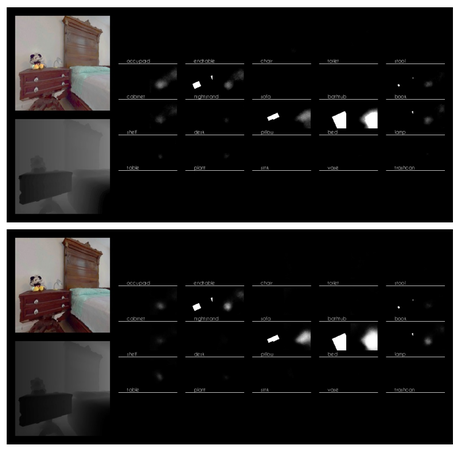

Fig. 6 top: proposed method; bottom: baseline w/o contextual information

Our main contributions are:

- The introduction of a novel framework that uses contextual information to improve the semantic mapping accuracy;
- Analysis of the proposed method;
- The presentation of a challenging benchmark that focuses on small object detection and mapping task.