Distilling Vision-Language Models for Real-Time Traversability Prediction

Will Huey, Sean Brynjólfsson, Don Greenberg

{wph52, smb459, dpg5}@cornell.edu

Introduction

We propose a visual traversability predictor by distilling a large open source open-vocabulary semantic segmentation network, achieving a 400x speedup with only a 20% decrease in mIoU. Additionally, we demonstrate zero-shot long-horizon visual robotic exploration in an unseen simulation environment using the distilled predictor

Background

Visual (image-based) navigation has become more popular in recent years due to advancements in foundation models. However, existing approaches require expert demonstrations of correct navigation to transfer to environments [2, 3].

By leveraging the reasoning capabilities of large transformer-based vision models, traversability prediction can be performed without heuristics or demonstrations, enabling transferable visual navigation models. However, these models are too slow for real-time traversability prediction. Knowledge distillation [cite] aims train a lightweight student model to match the predictions of a larger teacher model.

Our approach distills a large pretrained model to achieve fast and transferable traversability prediction.

References

[1] https://github.com/willh003/cut
[2] Shah, et al. ViNT: A Foundation Model for Visual Navigation, 2023.
[3] Frey et al. Fast traversability estimation for wild visual navigation, 2023.
[4] Poudel, et al. Fast-SCNN: Fast Semantic Segmentation Network, 2019.
[5] Liang et al. Open-vocabulary semantic segmentation with mask-adapted clip, 2023.
[6] Per Bellersen. Lake Shore Drone Scan. https://skfb.ly/oDrNw , 2022. Licensed under Creative Commons Attribution 4.0 International.
[7] Hatamizadeh, et al. FasterViT: Fast Vision Transformers with Hierarchical Attention, 2024.
[8] Hirose, et al. Sacson: Scalable autonomous control for social navigation. RA-L, 2023.
[9] Shah, et al. Rapid exploration for open-world navigation with latent goal models, 2021.
[10] Geiger, et al. Vision meets robotics: The kitti dataset, 2013.
[11] Clement, et al. Learning matchable image transformations for long-term metric visual localization, 2020.

Approach

Using OVSeg [5], an open-vocabulary segmentation model, we distill a lightweight traversability model by fixing its prompt.

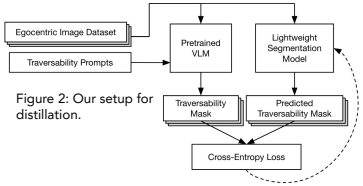

Figure 2: Our setup for distillation.

We assemble a dataset from 6 sources containing ego-centric images from robots navigating through a wide variety of environments. Unlike prior work [2, 3], we only assume access to robot observations, and not ground truth actions.

Dataset	Platform	Samples	Env.
SACSoN [8]	TurtleBot2	12512	office
RECON [9]	Jackal	3912	off-road
KITTI [10]	Car	3301	self-driving
UTIAS [11]	Grizzly	22	off-road
RSL Lab	Anymal D	275	busy office
Schulstrasse 44	Anymal D	161	construction
Total		20183	

Table 1: the sources used in our traversability distillation dataset. RSL Lab refers to the Robotic Systems Lab at ETH Zürich, which is a large, busy open-floor office space. Schulstrasse 44 refers to a construction site in Zürich, CH.

The dataset includes ~1 hour of data collected from an ANYmal D robot navigating a construction site and an office space.

Figure 1: the Anymal robot setup used for data collection. The dataset was obtained over 4 stories of an ongoing construction project. It contains images of varying brightness, with a large amount of clutter and obstacles.

The prompts "something a robot could walk on" and "other"are input to the VLM to obtain traversability predictions.

Results

The performance of two existing architectures [4, 7] distilled on the traversability dataset are shown in Table 2. Evaluation is performed using the ANYmal datasets, ensuring that the approach effectively transfers to new environments.

Model	Dataset	mIoU	Latency (ms)	Speedup
OVSeg	-	0.99	1651.9	1x
FasterViT	full	0.75	18.21	91x
FasterViT	partial	0.80	17.68	93x
Fast-SCNN	full	0.79	4.11	402x
Fast-SCNN	partial	0.72	4.03	409x

Table 2: Distilled model performance. "full" indicates training on all non-Anymal datasets, and "partial" indicates that SACSoN was held out from the training set.

We integrate our traversability predictor with a simple exploration heuristic, which navigates to the most traversable area. Fig. 3 shows example robot paths from simulation.

Figure 3: Several rollouts of the ANYmal D robot exploring freely in simulation [6]. We observed emergent path-following and obstacle avoidance with our heuristic.

Conclusions

We apply model distillation techniques to train a smaller traversability prediction network capable of real time inference, and demonstrate a heuristic that uses this distilled network to perform obstacle avoidance when roaming freely.

Additionally, we demonstrate our traversability prediction network running in real time on an ANYmal D robot using a custom ROS (Robotic Operating System) package [1]

Future Work:

- Extend the traversability scenarios to be robot-specific or task-specific.
- Integrate our method with elevation mapping and a global planner to achieve long-horizon navigation instead of just exploration.

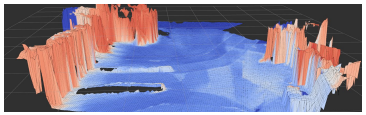

Figure 4: An example of traversability predictions projected onto a real elevation map from the Schulstrasse 44 construction site. In the future, this type of map could be used for visual robot navigation.

Acknowledgements

We thank the Rawlings Cornell Presidential Research Scholars Program and the Cornell Greenberg Lab for supporting this research. Additionally, we thank Marco Hutter, David Hoeller, Jonas Frey, and the Robotic Systems Lab at ETH Zurich for their collaboration.

For any questions, please contact us at:
wph52@cornell.edu smb459@cornell.edu
Additionally, if you are interested in robotics, VR, or graphics, our lab may have projects available for next semester.