Convolutional Neural Network Information Fusion for Urban Scene Understanding

Masha (Mikhal) ITKINA

mitkina@stanford.edu

Supervisor: Professor Mykel Kochenderfer

Introduction

Autonomously accounting for obstacle occlusion is an open problem for self-driving cars. Dempster-Shafer Theory (DST) provides a scene-understanding and information fusion strategy that addresses occlusion by modeling both lack of information and conflicting information directly [3]. DST can combine sensor and digital street-level map occupancy grids to discern cells that contain potential hazards from cells that are navigable by the vehicle. The perception system can then anticipate areas where occluded hazards may appear. DST has been previously used as a pre-processing step to a CNN in both semantic segmentation and object detection [9, 5].

Problem Statement

The approach in [3] is sensitive to parameters that require manual tuning to discern static and moving obstacles. This work focuses on merging the FCN semantic segmentation approach in [4] with the information fusion algorithm presented in [3], to increase the latter's robustness in discerning occupancy grid cells containing static and moving objects from navigable space as compared with DST alone. An offset is learned between the perception grid output of the DST algorithm and the expected semantic labels. The effectiveness of the approach is measured in reference to the DST baseline using the intersection over union (IU) metric for each of the classes.

Dataset

The KITTI tracking dataset [1] was augmented for use in information fusion. Four driving sequences were selected for training (140 examples), two for validation (48 examples) and two for testing (64 examples). The augmented dataset consists of a geographic information system (GIS) grid [6, 7], a sensor grid containing HDL-64E Velodyne LIDAR data, and the labeled perception grid segmentation. These grids are created for each ego vehicle GPS coordinate, which is obtained every 1 s or 2 s depending on the sequence [1]. The datasets contain an imbalanced class distribution with only 0.09% static cells and 0.26% moving cells in the training set.

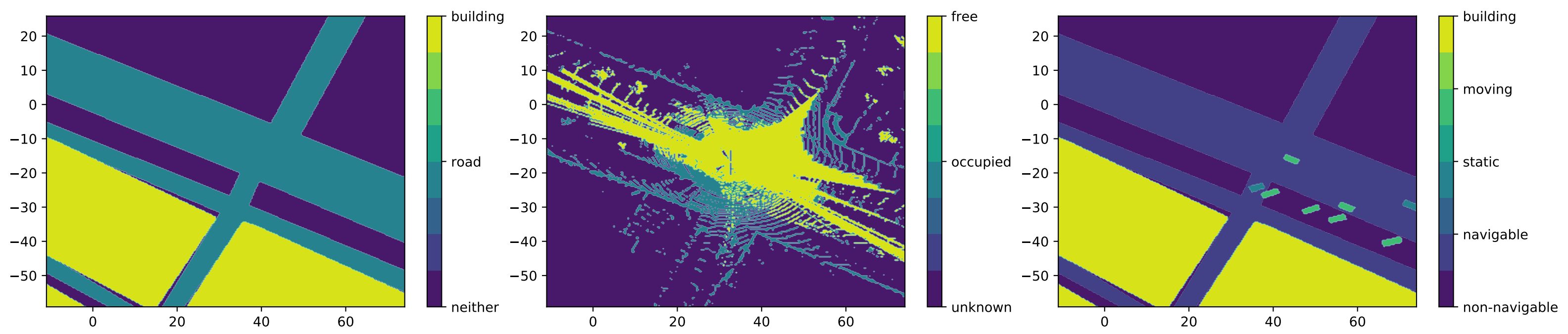

Figure 1: Example from the training dataset. From left to right, the figures are: GIS grid, sensor grid, and the perception grid labels. The grids are of dimension $42.7m$ x $42.7m$, with the ego vehicle in the center. Given a discretization of $0.33m$ per grid cell, each grid is of size 256×256 cells.

Methodology

The inputs to the baseline DST algorithm in [3] are the current sensor and GIS grids as well as the perception grid at the previous time-step. The DST framework is incorporated into an FCN architecture to optimize grid sensitivity. The input to the FCN is the set of probabilistic perception grids generated by the DST algorithm at the current and previous time-step stacked in channels (10 channels total). By providing the current and previous DST perception grids as inputs, the FCN should be able to learn the appropriate temporal information to classify moving and static cells effectively. The architecture is based on [4]; it consists of 16 pre-trained convolutional layers of VGG-19 interspersed with dropout and pooling operations, followed by 2 convolutional layers and 3 deconvolutional layers. To make the 10-channel DST occupancy grids compatible with the VGG model, an additional convolutional layer was added at the start to reduce the input channel number to three. Since the segmentation output is of occupancy grid dimension, a modified cross-entropy loss is used, where the loss is weighted to resolve class imbalance and summed over all grid cells. The Adam solver [2] is used to optimize the FCN parameters.

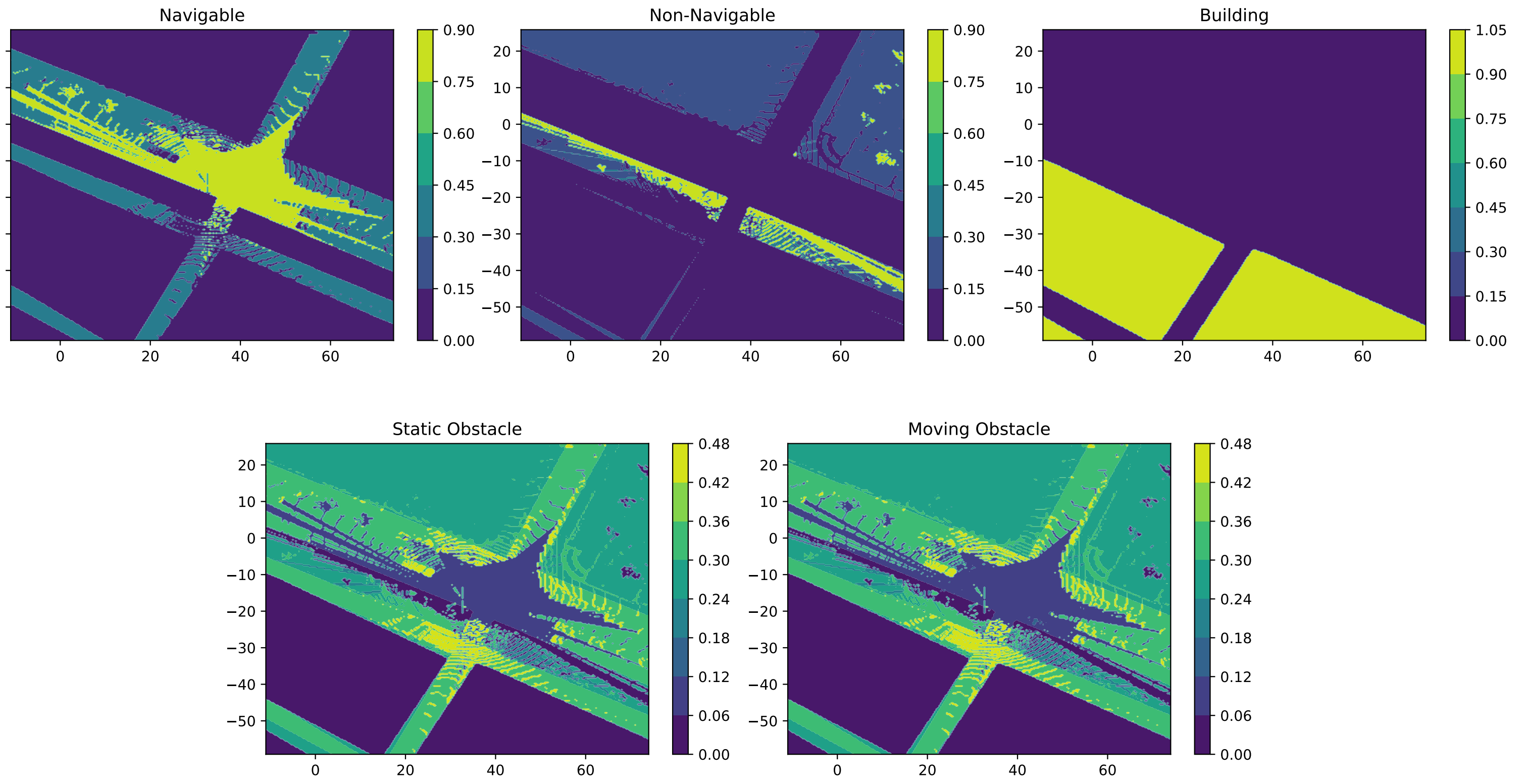

Figure 2: Example of the DST output; each plot shows a probabilistic occupancy grid associated with the semantic class label.

Experimental Results

	learning rate	keep probability	batch size
FCN	1e-4	0.85	32

Table 1: Tuned model hyper-parameters based on IU metric.

The metric used to evaluate the performance of the proposed algorithm is IU:

$$\mathrm{IU} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP} + \mathrm{FN}}.$$

IU is often used for semantic segmentation to directly account for class imbalance.

	Navigable	Non-Navigable	Building	Static	Moving	Mean
FCN-DST IU val	0.923	0.946	0.884	0.00737	0.0116	**0.555**
DST IU val	0.85	0.83	1.00	0.00135	0.0140	**0.539**
FCN-DST IU test	0.885	0.923	0.886	0.0	0.000649	**0.539**
DST IU test	0.775	0.786	1.00	0.000558	0.00120	**0.512**

Table 2: IU evaluations for each class on the validation and test sets.

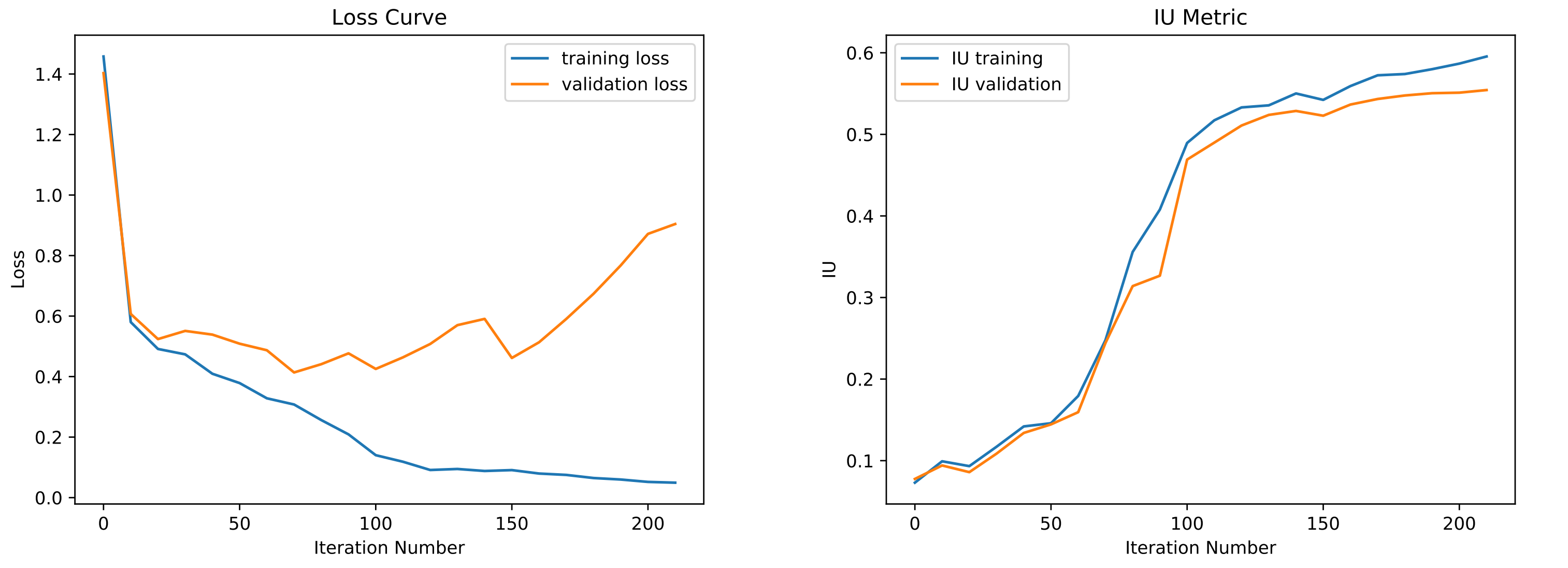

Figure 3: (left) Loss profile over iterations. (right) IU metric profile over iterations.

Table 2 shows that the tuned FCN architecture achieves a higher mean IU than DST alone. However, it undesirably reduces the "static" and "moving" IUs in favor of the more frequent classes. In experiment, increasing the loss weights of these infrequent classes further, in an attempt to counteract this phenomenon, resulted in less effective learning, likely due to insufficient data.

Furthermore, despite the loss curves in Figure 3 showing overfitting, the mean IU continued to increase, indicating that the cross-entropy loss may not be representative of the IU metric. A method to utilize IU directly as a loss in a binary classification problem has been proposed in [8]. Extending this formulation to multi-class segmentation may improve the effectiveness of the proposed approach.

Conclusions and Future Work

The FCN-DST framework outperformed the DST baseline in the mean IU metric. However, little improvement was achieved in discerning moving and static cells due to insufficient data and a loss that was not representative of the evaluation metric. In future work, to improve the accuracy of the model, further pre-processing should be performed on the dataset, including the use of ground segmentation techniques instead of hand-annotated labels [10]. This would significantly expand the size of the dataset, facilitating more effective learning. Additionally, the loss should be restructured such that it is based directly on the IU criterion.

References

[1] Lenz P. Geiger, G. and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *ICLR*, 2015.

[3] Kurdej, M., Moras, D., Cherfaoui, V., and Bonnifait, P. Map-aided Evidential Grids for Driving Scene Understanding. *IEEE Intelligent Transportation Systems Magazine*, pages 30–41, 2015.

[4] Long, J., Shelhamer, E., and Darrell, T. . Fully Convolutional Networks for Semantic Segmentation. *CVPR*, 2015.

[5] Oh, S.-I. and Kang, H.-B. Object Detection and Classification by Decision-Level Fusion for Intelligent Vehicle Systems. *Sensors 2017*, 2017.

[6] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org, 2017.

[7] QGIS Development Team. *QGIS Geographic Information System.* Open Source Geospatial Foundation, 2009.

[8] Rahman, A. Md. and Wang, Yang. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. *ISVC*, 2016.

[9] Yao, W., Poleswkia, P., and Krzystek, P. Classification of Urban Aerial Data Based on Pixel Labelling with Deep Convolutional Neural Networks and Logistic Regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B7, 2016.

[10] Zhang, R., Candra, S.A., Vetter, K., and Zakhor, A. Sensor Fusion for Semantic Segmentation of Urban Scenes. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015.