Deep Learning with Knowledge Graphs for Pedestrian Intent Prediction: A literature Review

Muhammad Saud Ul Hassan

Contents

Abstract

Pedestrian intent prediction is a task critical to the success of autonomous cars. However, the work been done so far on the problem has treated the spatial and temporal aspects of the problem in isolation. Liu. et. al. [1] reason that for reliable intent prediction both these components need to be considered in mutuality. They have designed a Graph Neural Network (GNN) based approach to reason on the spatiotemporal relationships in the driving scene and have reported better and faster prediction on the Stanford-TRI Intent Prediction (STIP) as well as the Joint Attention for Autonomous Driving (JAAD) datasets.

In our work, we strive to improve on [1] using the ideas suggested in the paper as well as by employing better panoptic segmentation and tracking methods. However, given the limited time, I have so far only been able to take up a brief literature survey on the problem.

1 Introduction

Reasoning about how objects interact in a complex visual scene is one of the central problems in computer vision. Among other applications, it is crucial to autonomous driving systems, especially to the end of pedestrian intent prediction. Previous work on predicting the pedestrian intent has typically formulated the problem as one of trajectory prediction [2], however, predicting trajectories long enough into the future is a challenging task [1]. Other approaches [3] have relied on pedestrian location, velocity, and other such features to infer the future position of the pedestrian, but these approaches have typically ignored the evolution of the pedestrian—scene-objects relationship through time, which is a crucial component to reliably predicting the pedestrian intent.

Recently, [1] proposed a framework based on Graph Convolutional Networks to reason about object relationships in a driving scene, and they have claimed their model to have outperformed the previous work on pedestrian intent prediction. Our work aims to improve on [1] by employing better panoptic segmentation and tracking methods, and by incorporating the suggestions the authors have put forward in the paper, for example, incorporating a probability calibration method [4] to obtain better confidence scores on the predictions.

It has not been long since the work on the project started, and so far, I have only managed to study the literature on the problem and briefly analyse the models for panoptic segmentation and deep learning on graphs. The literature I have surveyed so far has been briefly reviewed below.

2 The Challenge of Autonomous Driving

Grigorescu et. al. [5] categorize autonomous driving into four main subproblems:

1. Driving Scene Understanding and Localization
2. Path Planning
3. Behaviour Arbitration
4. Control

Researchers have long been striving to solve the problem of autonomous driving. Even though the classical methods for perception, planning and control of autonomous vehicles have managed to achieve good performance in the lab setting, it is in unstructured and stochastic real-world environments that these approaches start to fall apart. In fact, in the DARPA Grand Challenge of 2004, none of the 15 entrants could complete the race [6]. Stanely [7], in 2005, finally completed the DARPA Challenge, and that vehicle leveraged machine learning under the hood.

3 Panoptic Segmentation for Driving Scene Understanding

The first task of an autonomous vehicle is to understand the surrounding scene. Much of the work on deep learning for visual scene understanding has for a long time treated semantic and instance segmentation as two separate problems [8]. The joint problem was just recently formulated as *Panoptic Segmentation* by [8], and the work on the problem, though still in its infancy, has already started yielding impressive results. Panoptic segmentation provides greater detail for visual perception than bounding-box object detectors, which has led to improved models for driving scene understanding [9].

3.1 Pedestrian Intent Prediction

Pedestrian intent prediction is a sub-problem of driving scene understanding, albeit a very important and challenging one. Important because most road accidents happen when pedestrians are crossing the road [10], and challenging because pedestrians are also very difficult to predict (compared to cars and other objects in a visual scene) because of their high dynamic range [3]. Several major auto manufacturers now integrate a Collision Warning and Emergency Braking system in their suite of Advanced Driver Assistance Systems (ADAS), however, these systems alert the driver or initiate automatic emergency braking only when the pedestrian in front is under a certain threshold distance from the car, and by then, it might already be too late to avoid a collision [11]. This necessitates the early prediction of a pedestrian's intent to cross the road.

Much of previous work on pedestrian intent prediction for road crossing utilized bounding-box-like detectors, or instance or segmentation methods, however, now the trend is shifting towards panoptic segmentation. Panoptic segmentation can provide a more detailed pedestrian shape and silhouttee and allow for better pedestrian tracking [9], which allows the downstream model to make a more informed decision as to whether the pedestrian would cross or not.

4 Graph Neural Networks for Pedestrian Intent Prediction

As already described, much of the work on pedestrian intent prediction has so far focused on either the temporal evolution of the pedestrian in a visual scene by formulating it as a trajectory prediction problem [2] or on some sort of spatial characteristics of the scene [3] without any regard to the temporal dimension. This disparate treatment of the spatial and temporal aspects of the problem has, according to Liu et. al. [1], limited the capability of the current pedestrian intent prediction models. To address this, Liu et. al. [1] employ graph convolution to reason about the spatial relationship between the pedestrian and other objects in the scene and connect the important nodes of the graph through time to model the temporal dynamics of the scene.

4.1 Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) [12] belong to a family of neural networks used to learn on non-euclidean data, called *Graph Neural Networks (GNNs)*. More specifically, according to the recently proposed taxonomy for classifying GNNs in [13], GCNs are a type of Convolutional Graph Neural Networks (ConvGNNs), that are, GNNs that employ convolution to learn on graph data.

Shortly after the GCNs [12] paper came out, Ferenc Huszar published a critical review of it on their blog citing several of its shortcomings [14]. However, an accomplishment of GCNs, which are spectral-based ConvGNNs, is that they managed to show the equivalency between the spectral and spatial based methods, and since then spatial-based ConvGNNs have taken off in reputation among the deep learning community because of their efficiency and applicability [13].

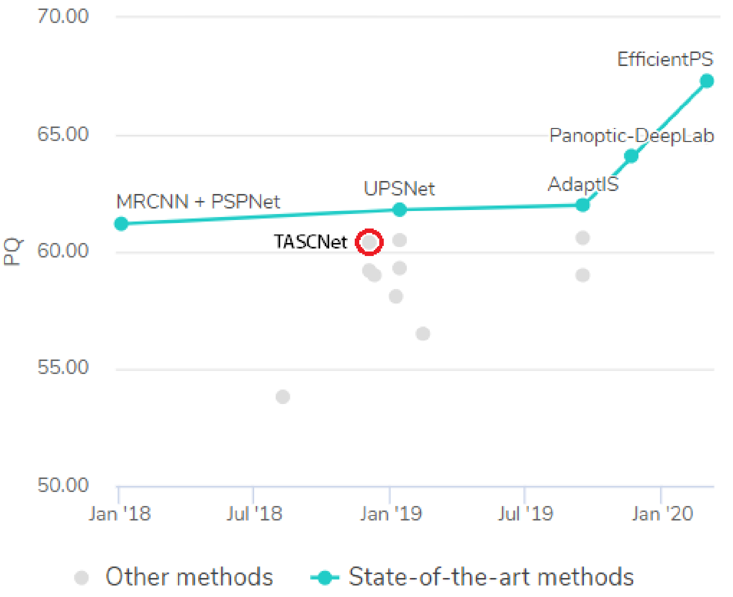

Figure 1: Panoptic Quality (PQ) of the various panoptic segmentation models on the Cityscapes dataset. Source: PapersWithCode [15]

4.2 GCNs for Predicting the Pedestrian Intent

Liu et. al [1] consider the application of GCNs to the problem of pedestrian intent prediction from two different point of views: using pedestrian-centric graphs and using location-centric graphs. In the former, a star graph centered at each pedestrian instance is connected to the other object instances in the scene and a context node (which encodes the contextual information in the visual scene). In the latter, a graph is centered at the car with edges going to all the objects of interest in the visual scene, and the problem of intent prediction is formulated as how likely a pedestrian is to cross the trapezoidal region in front of the car in the next T seconds. This setting is beneficial for crowded scenes where building a graph for each pedestrian instant would be computationally expensive.

5 Possible improvements in Liu et. al. [1]

Some possible directions of improvement include improving the panoptic segmentation model and the graph network used in Liu et. al. [1]. They em-

ploy TASCNet [16] for panoptic segmentation, however, the recently developed panoptic segmentation models such as Panoptic-DeepLab [17] and EfficientPS [18] do a significantly better job in terms of the Panoptic Quality (PQ) score [8] at segmenting the visual scene. Comparison of the various panoptic segmentation methods on the Cityscapes [19] and Mippillary [20] datasets can be viewed at [15] and [21], respectively. The plots in [15] and [21] have also been reproduced below in Fig. 1 and Fig. 2 for convenience.
As for the graph neural network used in the paper, i-e, the GCN network, it has also been outperformed by the newer models, in-particular by those based on Spatial-ConvGNNs.

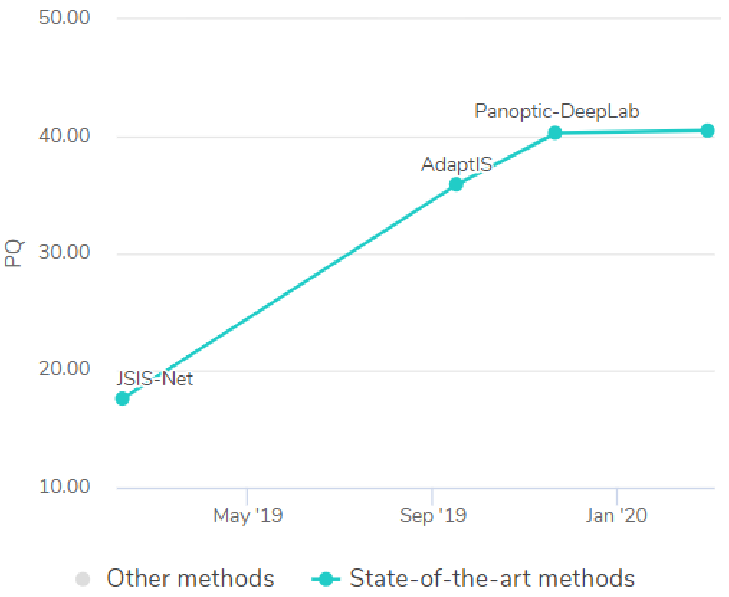

Figure 2: Panoptic Quality (PQ) of the various panoptic segmentation models on the Mapillary dataset. Source: PapersWithCode [21]

Furthermore, some ideas have also been put forth by the authors themselves in the paper, and they would also make up for interesting prospects to explore. Also, further directions for improvement might also become apparent as work on the problem is started.

References

[1] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction. *arXiv:2002.08945 [cs]*, February 2020. arXiv: 2002.08945.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human Trajectory

Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, Las Vegas, NV, USA, June 2016. IEEE.

[3] Sarah Bonnin, Thomas H. Weisswange, Franz Kummert, and Jens Schmuedderich. Pedestrian crossing prediction using multiple context-based models. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 378–385, October 2014. ISSN: 2153-0017.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330, Sydney, NSW, Australia, August 2017. JMLR.org.

[5] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21918.

[6] The DARPA Grand Challenge: Ten Years Later, March 2014.

[7] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, Kenny Lau, Celia Oakley, Mark Palatucci, Vaughan Pratt, Pascal Stang, Sven Strohband, Cedric Dupont, Lars-Erik Jendrossek, Christian Koelen, Charles Markey, Carlo Rummel, Joe van Niekerk, Eric Jensen, Philippe Alessandrini, Gary Bradski, Bob Davies, Scott Ettinger, Adrian Kaehler, Ara Nefian, and Pamela Mahoney. Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, 23(9):661–692, September 2006.

[8] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. *arXiv:1801.00868 [cs]*, April 2019. arXiv: 1801.00868.

[9] Neda Cvijetic. Panoptic Segmentation Helps Autonomous Vehicles See Outside the Box | NVIDIA Blog, October 2019. Library Catalog: blogs.nvidia.com Section: DRIVE Labs.

[10] A. Martin. Factors influencing pedestrian safety: a literature review, June 2008. Library Catalog: trl.co.uk.

[11] Erik Coelingh, Andreas Eidehall, and Mattias Bengtsson. Collision Warning with Full Auto Brake and Pedestrian Detection - a practical example of Automatic Emergency Braking. *13th International IEEE Conference on Intelligent Transportation Systems.*

[12] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. Semi-Supervised Learning With Graph Learning-Convolutional Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11312, Long Beach, CA, USA, June 2019. IEEE.

[13] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[14] Ferenc Huszar. How powerful are Graph Convolutions? (review of Kipf & Welling, 2016), September 2016. Library Catalog: www.inference.vc.

[15] Papers with Code - Cityscapes val Leaderboard, April. Library Catalog: paperswithcode.com.

[16] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to Fuse Things and Stuff. *arXiv:1812.01192 [cs]*, May 2019. arXiv: 1812.01192.

[17] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. *arXiv:1911.10194 [cs]*, March 2020. arXiv: 1911.10194.

[18] Rohit Mohan and Abhinav Valada. EfficientPS: Efficient Panoptic Segmentation. *arXiv:2004.02307 [cs]*, April 2020. arXiv: 2004.02307.

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. *arXiv:1604.01685 [cs]*, April 2016. arXiv: 1604.01685.

[20] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *International Conference on Computer Vision (ICCV)*, 2017.

[21] Papers with Code - Mapillary val Leaderboard, April. Library Catalog: paperswithcode.com.