DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Robotics, Cognition, Intelligence

Anomaly Detection for the behavior of drivers based on Structural Temporal Graph Neural Networks

Hanxi Jiang

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Robotics, Cognition, Intelligence

Anomaly Detection for the behavior of drivers based on Structural Temporal Graph Neural Networks

Erkennung von Anomalien im Verhalten von Autofahrern auf der Grundlage struktureller temporaler neuronaler Netze

Author:	Hanxi Jiang
Supervisor:	Supervisor
Advisor:	Advisor
Submission Date:	Submission date

I confirm that this master's thesis in informatics: robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, Submission date　　　　Hanxi Jiang

Acknowledgments

Abstract

Kurzfassung

Contents

1. Introduction

Despite the fact that High Driving Automation(SAE Level 4) is achievable in the foreseeable future[1], the majority of drivers nowadays would still prefer comprehensive control over their own vehicles. Therefore, increasingly more studies have been focusing on driving safety[2, 3]. According to these papers, driver behavior represents the majority cause of accidents while driving. In that case, several methods have been developed to avoid potential danger. Such methods involve either improving monitoring of the vehicle's inside situation, which analyzes the driving parameters as well as the driver him or herself to determine whether there's no abnormality[4], or proposing a vehicle detection and tracking system from an outer view that estimates time-to-collision (TTC) and warn the driver for a possible collision[5].

On the other hand, only few research focus on driving behavior prediction, which may be due to the lack of application for the undetermined decision model in behavior prediction, specifically under the topic of autonomous driving. Scientists have succeeded in years of generating dynamic graphs based on videos. However, there is still a gap in the rational use of these forms for behavioral prediction.

In this work, we would like to focus on constructing a comprehensive behavior prediction model based on graph neural networks (GNNs). A Graph Neural Network is a novel type of neural network architecture that can be applied to graph-like inputs. As we are now expecting to train the behavior model for the participators, the training data would contain interaction between several objects and participators while they are driving. GNN is, therefore, prioritized due to its unique structure. Along with that, we would specifically focus on building dynamic graphs as training datasets, as our model would be trained based on sequences of behaviour descriptions extracted from videos. Besides, to ensure the compatibility of the dataset and training model, we made some adaptions based on JODIE[6], which is the model based on the dynamic evolution of users and items. In order to make the predictions as detailed as possible, we expanded the output catalog to ensure that not only the interaction itself but also the type of it would be described.

In a nutshell, We would first gather all the necessary information with the help of the Large language model (LLM), convert it into dynamic graphs, and insert these graphs into the model we have adapted from model JODIE. These graphs could contain either one specific participant or all the concerned people. By learning how dynamic graphs change under time series, like the appearance and vanish of all these edges and nodes in the graph, the model should be able to absorb the feature behind it and come up with the prediction for behaviour in the future. Therefore, anomaly detection could also be achieved by comparing the predicted graph with the real one and alerting once when any unsuitable behaviour during driving is detected. we believe that this model provides a new perspective for the prediction of driving behavior detected from videos and allows for more diversified and targeted forecasting.

1.1. contribution

The main contributions of this thesis are summarized as:

Dataset Collecton From the Dataset *drive & act*, we acquire the hierarchical activity labels of given video data and rewrite them in the form of time sequences. We also cluster the behavior types into several categories with the help of the Large Language Model.

Dynamic Graph Construction By reassembling the nodes and edges in the video data, we construct time-sequenced dynamic graphs that could be used as training data for the model. Such graph captures complex dependencies and offers hierarchical representation abstracted from the raw data.

Learning Model Adaption To enrich the diversity of the prediction, we expand the output of the dynamic graph based learning model from binary to catalog description, to ensure that not only the interaction itself but also the type of it would be described.

anormaly detection By comparing the predicted model with the real one, we could detect any unsuitable behavior during the driving and alert the driver. Differ from the research before, our

1.2. Structure

This thesis is structured as follows. In Chapter2 we would introduce and explain concepts and definitions concerned this document. Chapter3 reviews related work in the field of anomaly detection and dynamic link prediction models, the dataset this work refers to and the model we have adapted. Chapter4 describes the methodology of this work, including the dataset collection, dynamic graph construction and model adaption. Chapter5 presents the evaluation of the model and the results of the prediction. Chapter6 discusses the potential future work that could be done based on this work. Chapter7 concludes the thesis and gives a summary of the work done.

2. Background

2.1. Large Language Model

Large Language Models (LLMs) are highly complex artificial intelligence systems that can learn from the vast amounts of available text data[7]. Thanks to the attending of *Transformer* [8], a deep learning architecture, these language models which employed self-supervised pre-training have demonstrated improved efficiency and scalability in many fields. Based on self-attention mechanisms and feed-forward module,*Transformer* has overwhelming advantages in computing representations and global dependencies.

In concrete terms, the Large Language Models equipped with *Transformer* are capable of diverse tasks raised by Natural Language Processing[9], such as textual entailment, question answering, semantic similarity assessment, and document classification. Take BERT[10] and GPT [7, 11, 12] as two examples, the former utilizes transformer encoder blocks to predict missing words in a given text, and the latter has been enjoying a tremendous reputation for generating diverse and human-like responses, showcasing its potential in various domains.

2.1.1. Text Classification

In our project, a bunch of hierarchical activity labels would be acquired from the dataset *drive & act* to depict every detail of the participant's movements in the driving behavior recorded in the video. These labels, however, are too trivial for the construction of learning data, as considering each activity individually will be tedious in such a vast and complex model training process. Therefore, labels should be classified according to the object on which this behavior operates or the specificity of the moment in which the action takes place. For example, the fastening of a seat belt should occur shortly after entering the vehicle, and all behavior related to eating or drinking should be classified into the same group.

In our task, we use zero-shot text classification. This is a task where a model is trained on a set of labeled examples and then classifies new examples from previously unseen classes. This method, which leverages a pre-trained language model, can be thought of as an instance of transfer learning which generally refers to using a model trained for one task in a different application than what it was originally trained for. This is particularly useful for situations where the amount of labeled data is small, for example, our work with 39 different behaviors to be classified.

The model used in the pipeline is *BART* [13]. According to the paper, BART is trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes. With all these prerequisites it is quite obverious that *BART* is a very practical model for our classification task.

2.2. graph

In graph theory, a graph is a structure made up of a collection of objects, where certain pairs of these objects are connected in a specific way[14].As a powerful tool for modeling and analyzing complex systems, researchers have been combined graph theory with deep learning to solve various problems in different fields, such as social networks[15], complex physic system[16] and Protein-protein interactions (PPIs)[17].

A Graph is $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a pair of sets, where nodes $\mathcal{V}$

2.2.1. Graph Neural Networks

2.2.2. Dynamic graph

Dynamic graphs are a type of graph, the input feature or topology of which changes over time.

2.3. behavior prediction

3. Related Work

3.1. Dynamic Scene Graph for Video

The scene graph is a structured representation of a scene that can clearly express the objects, attributes, and relationships between objects in the scene[18]. Accompanied by the development of computer vision technology, simply detecting and recognizing objects in images no longer satisfies the researchers, as they would expect some higher level of understanding and reasoning for image vision tasks. In this way, an intuitive idea comes up about adding up the relationship between the detected objects(See example in figure 3.1). The earliest research could dated back to 2017, when some objects and relations of a given image could be inferred and a scene graph would be produced as a result[19]. Other research like Neural Motifs[20] also shows the possibility of predicting the most frequent relation between object pairs with the given labels and object detections. Later in 2018, videos came into discussion, and both spatial and temporal relations would be concerned in the dynamic graph researchers propose to represent[21]. Meanwhile, the accuracy of Scene Graph Detection tasks has significantly improved thanks to the application of unbiased SGG[21], fueled by the **Detection2** [22], a library that contains various state-of-the-art detection and segmentation algorithms. In a nutshell, representing videos as dynamic scene graphs including the detection of objects and the relations in between has been realized in the past years. And our work would utilize such technique and extract our driver-oriented dynamic scene graphs for further learning.

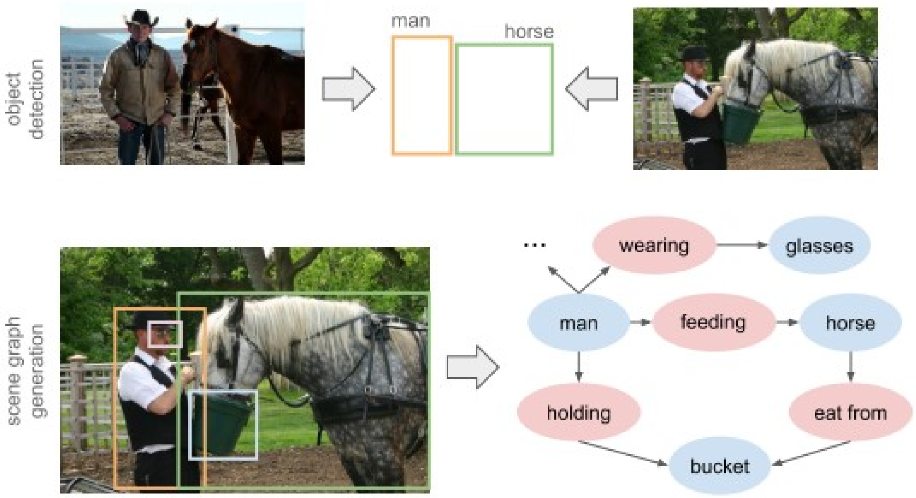

Figure 3.1.: Scene Graph Generation by Iterative Message Passing

4. Methodology

This work aims to construct a graph neural network-based architecture for predicting, analyzing, and detecting any potentially abnormal behavior regarding the driver during the whole driving process. In particular, The model extracts a description graph, the so-called scene graph, of the driver from the video filmed inside the vehicle and trains itself with these data to learn for future behavior prediction. The result will be used to compare and detect any abnormal behavior. Here we would lay most emphasis on the construction of the training model. To make precious anomaly detection we aim to predict not only if there is a behavior between humans and a specific kind of object but the type of behavior as well, which will cause several adaptions based on existing model *JODIE*.

4.1. scene graph generation

4.1.1. video data extracting

4.1.2. graph generating

4.2. model architecture

After comparing all the training results of the below models we would find that **JODIE** is one coming up with the best prediction. However, the model jodie still fail to predict the state of the predicted edge. In my masterwork I would like to rewrite the embedding function and the loss function of **JODIE to make the state prediction possible.

- function from **JODIE**:
embedding function

$$\mathbf{u}(\mathbf{t}) = \sigma(W_1^u \mathbf{u}(\mathbf{t}^-) + W_2^u \mathbf{i}(\mathbf{t}^-) + W_3^u f + W_4^u \Delta_u)$$

$$\mathbf{i}(\mathbf{t}) = \sigma(W_1^i \mathbf{i}(\mathbf{t}^-) + W_2^i \mathbf{u}(\mathbf{t}^-) + W_3^i f + W_4^i \Delta_i)$$

loss function(BCE)

$$L = -(j_{pos} \log \tilde{j} + j_{neg} log(1 - \tilde{j}))$$

where

$$\tilde{j}(t + \Delta) = W_1 \hat{u}(t + \delta) + W_2 \bar{u} + W_3 i(t + \Delta^-) + W_4 \bar{i} + B$$

- functions adapted in my work:

embedding function

$$\mathbf{u}(\mathbf{t}) = \sigma(W_1^u \mathbf{u}(\mathbf{t}^-) + W_2^u \mathbf{i}(\mathbf{t}^-) + W_3^u f + W_4^u s + W_5^u \Delta_u)$$

$$\mathbf{i}(\mathbf{t}) = \sigma(W_1^i \mathbf{i}(\mathbf{t}^-) + W_2^i \mathbf{u}(\mathbf{t}^-) + W_3^i f + W_4^i s + W_5^u \Delta_i)$$

we will change it from BCE to CE for predictiing state.

$$\tilde{j}(t + \Delta) = W_1 \hat{u}(t + \delta) + W_2 \bar{u} + W_3 i(t + \Delta^-) + W_4 \bar{i} + W_5 s + B$$

5. Evaluation

6. Future Work

7. Conclusion

A. General Addenda

If there are several additions you want to add, but they do not fit into the thesis itself, they belong here.

A.1. Detailed Addition

Even sections are possible, but usually only used for several elements in, e.g. tables, images, etc.

B. Figures

B.1. Example 1

✓

B.2. Example 2

✗

List of Figures

List of Tables

Bibliography

[1] T. Inagaki and T. B. Sheridan. “A critique of the SAE conditional driving automation definition, and analyses of options for improvement”. In: *Cognition, technology & work* 21 (2019), pp. 569–578.

[2] J. D. Lee. “Driving safety”. In: *Reviews of human factors and ergonomics* 1.1 (2005), pp. 172–218.

[3] J. D. Lee. “Fifty years of driving safety research”. In: *Human factors* 50.3 (2008), pp. 521–528.

[4] M. Karrouchi, I. Nasri, M. Rhiat, I. Atmane, K. Hirech, A. Messaoudi, M. Melhaoui, and K. Kassmi. “Driving behavior assessment: a practical study and technique for detecting a driver's condition and driving style”. In: *Transportation Engineering* 14 (2023), p. 100217.

[5] B. Aytekin and E. Altu. “Increasing driving safety with a multiple vehicle detection and tracking system using ongoing vehicle shadow information”. In: *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. 2010, pp. 3650–3656.

[6] S. Kumar, X. Zhang, and J. Leskovec. “Predicting dynamic embedding trajectory in temporal interaction networks”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 1269–1278.

[7] A. Radford. “Improving language understanding by generative pre-training”. In: (2018).

[8] A. Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).

[9] K. Chowdhary and K. Chowdhary. “Natural language processing”. In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[10] S. Alaparthi and M. Mishra. “Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey”. In: *arXiv preprint arXiv:2007.01127* (2020).

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[12] T. B. Brown. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: `1910.13461 [cs.CL]`. URL: `https://arxiv.org/abs/1910.13461`.

[14] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. "Graph neural networks: A review of methods and applications". In: *AI open* 1 (2020), pp. 57–81.

[15] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen. "Graph convolutional networks with markov random field reasoning for social spammer detection". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 1054–1061.

[16] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. "Graph networks as learnable physics engines for inference and control". In: *International conference on machine learning*. PMLR. 2018, pp. 4470–4479.

[17] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur. "Protein Interface Prediction using Graph Convolutional Networks". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper_files/paper/2017/file/f507783927f2ec2737ba40afbd17efb5-Paper.pdf`.

[18] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann. "A Comprehensive Survey of Scene Graphs: Generation and Application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2023), pp. 1–26. DOI: `10.1109/TPAMI.2021.3137605`.

[19] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. "Scene graph generation by iterative message passing". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5410–5419.

[20] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. "Neural motifs: Scene graph parsing with global context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5831–5840.

[21] X. Wang and A. Gupta. "Videos as space-time region graphs". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 399–417.

[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. *Detectron2*. `https://github.com/facebookresearch/detectron2`. 2019.