Short-Term Traffic Congestion Forecasting Using Attention-Based Long Short-Term Memory Recurrent Neural Network

Tianlin Zhang [1,2][0000-0003-0843-1916], Ying Liu[1,2][0000-0001-6005-5714] , Zhenyu Cui[1,2], Jiaxu Leng[1,2], Weihong Xie[3], Liang Zhang[4]

[1] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100190 China
[2] Key Lab of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, 100190 China
[3]School of Economics and Commerce, Guangdong University of Technology, Guangzhou, 510006 China
[4]School of Applied Mathematics, Guangdong University of Technology, Guangzhou, 510006 China
zhangtianlin172@mails.ucas.ac.cn

Abstract. Traffic congestion seriously affect citizens' life quality. Many researchers have paid much attention to the task of short-term traffic congestion forecasting. However, the performance of the traditional traffic congestion forecasting approaches is not satisfactory. Moreover, most neural network models cannot capture the features at different moments effectively. In this paper, we propose an Attention-based long short-term memory (LSTM) recurrent neural network. We evaluate the prediction architecture on a real-time traffic data from Gray-Chicago-Milwaukee (GCM) Transportation Corridor in Chicagoland. The experimental results demonstrate that our method outperforms the baselines for the task of congestion prediction.

Keywords: Traffic congestion prediction, LSTM, Attention mechanism

1 Introduction

As the population grows and the mobility increase in cities, traffic has received important concern from citizens and urban planners. Traffic congestion is one of the major problems to be solved in traffic management. For this reason, traffic congestion prediction has become a crucial issue in many intelligent transport systems (ITS) applications [1]. Short-Term traffic forecasting have beneficial impact that could increase the effectiveness of modern transportation systems. Therefore, in the past decade, many research activities have been conducted in predicting traffic congestion.

To get better prediction effect, more and more studies use real-time data, which is collected via different devices such as loop detectors, fixed position traffic sensors, or

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

GPS. Compared with loop detectors, fixed position traffic sensors are more cost-effective and equally reliable [2]. Therefore, we use real-time data collected by these sensors to forecast the traffic congestion in our research.

The existing traffic prediction methods can be classified into two groups [3], parametric approach and nonparametric approach. The parametric models are predetermined by some specific theoretical assumptions, such as logistic regression whose parameters can be computed from empirical data. As a commonly used parametric time series method, autoregressive integrated moving average (ARIMA) [4] is suitable for Short-Term traffic congestion prediction. Due to its non-linear complexity characteristic of traffic flow, many researchers tried to employ non-parametric method for prediction. For example, Support Vector Machine (SVM) and Support Vector Regression (SVR) [5] are considered as efficient algorithms. K-nearest neighbors(KNN) [6] is also applied to finding common features in traffic data.

In recent years, as deep learning receiving extensive attention, many neural network-based (NN-based) methods have been proposed. Since the deep learning method has flexible model structure and strong learning ability, it could provide automatic representation learning from high-dimensional data. Huang et al. [7] used Deep Belief Network (DBN) and Lv et al. [8] proposed stacked autoencoder (SAE) method. On this basis, Chen et al. [9] attempted stacked de-noising autoencoder. Due to the dynamic time-serial nature of traffic flow, Recurrent Neural Networks (RNN) that has a chain-like structure may well deal with this sequence data. However, RNN may have the vanishing or blowing up gradient problems during the back-propagation process. In order to overcome this issue, Tian et al. [10] used long short-term memory recurrent neural network (LSTM), which is a type of RNN with gated structure to learn long-term dependencies and automatically determines the time lags. Other researchers have also made some corresponding improvements, like BDLSTM[11], DBLSTM [12]. et. But the current LSTM models are insensitive to time-aware traffic data, which cannot distinguish the importance of different traffic states at different moments.

In order to deal with the issue and improve traffic congestion prediction accuracy, in this paper, we propose a model called Attention-based Long Short-Term Memory Recurrent Neural Network, which can capture the features of different moments more effectively. We evaluated the performance of our proposed Attention-based LSTM model with other basic traffic prediction algorithms. In the experiment, our method is clearly superior to the baselines. The remainder of this paper is organized as follows. Section 2 presents our proposed attention-based LSTM model for traffic congestion predic-tion in detail. Experiments design and results analysis are given in Section 3. Finally, we conclude our work in Section 4.

2 Methodology

To capture the features of traffic flow and take full advantage of time-aware flow data, we propose an Attention-based LSTM method. In Section 2.1, we will present

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

the Attention-based LSTM model. In Section 2.2, the traffic congestion prediction architecture will be explained in detail.

2.1 Attention-based LSTM model

2.1.1 LSTM

Long short-term memory (LSTM) [13] is an effective approach to predict traffic congestion by capturing dependency features. It solves the vanishing gradient problem based on the gate mechanism. The structure is composed of input layer, output layer and recurrent hidden layer that has artificial designed memory cell. This cell can remove and keep the information of the cell state, which consists of three gates, including the input gate, the output gate and the forget gate. The architecture of the LSTM is illustrate in Fig 1.

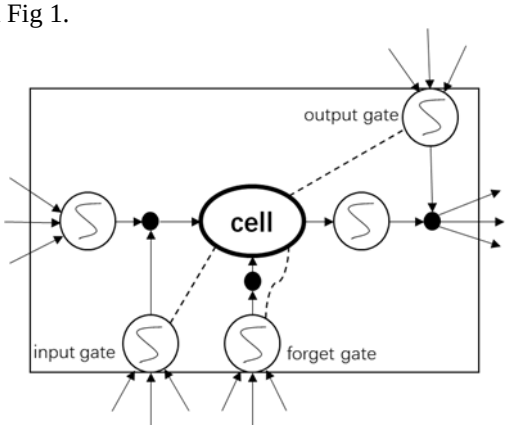

Fig. 1. The architecture of LSTM

In this model, the following equations explain the process and the notations as follow:

$$i_t = \sigma\left(W_j x_t + U_i h_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{3}$$

$$g_t = \tanh\left(W_g x_t + U_g h_{t-1} + b_g\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + f_t \odot g_t \tag{5}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

$$h_t = o_t \odot \tanh(c_{t-1}) \tag{6}$$

Table 1. Notations for LSTM model

Notation	Definition
h_t	hidden state
c_t	memory cell
x_t	the input historical traffic flow
i_t	input gate
f_t	forget gate
o_t	output gate
g_t	the extracted feature
W_* / U_*	weight matrices
b_*	bias vectors
$\odot$	element-wise multiplication

2.1.2 Attention mechanism

The attention mechanism in the neural networks imitates the attention of the human brain. It was proposed in the field of image recognition originally [14]. When people observe images, they often focus on some important information of the image selectively. Recently, many researchers applied the attention mechanism to natural language processing (NLP) [15][16], because the conventional neural networks assume the weight of each word in the input is equal. Thus, they fail to distinguish the importance of different words. Therefore, attention mechanism is added to the basic model to calculate the correlation between the input and output.

Similar to natural language, traffic flow data is sequence data too. The importance of different traffic states in the flow data is not the same either. Nevertheless, since the existing methods did not solve the problem well yet, we propose an attention-based LSTM model.

2.1.3 Attention-based LSTM

As shown in Fig 2, the structure of Attention-based LSTM can be divided into four layers: the input layer, LSTM layer, the attention mechanism layer, and the output layer.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

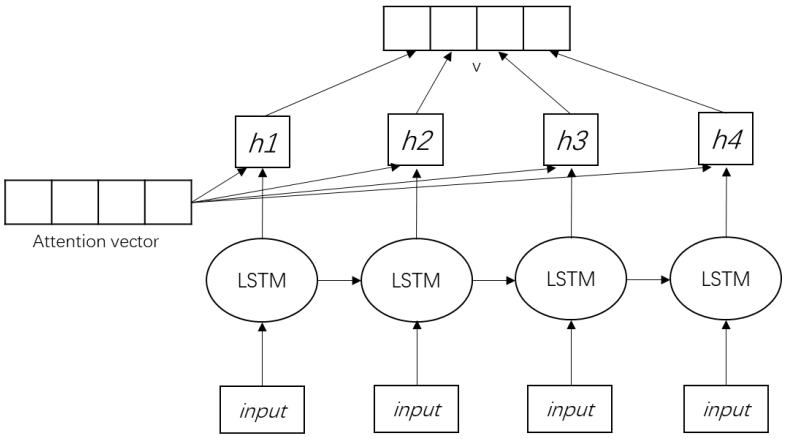

Fig. 2. The structure of the Attention-based LSTM

Attention mechanism layer [17] can highlight the importance of a particular traffic state to the entire traffic flow and consider more contextual association.

The state importance vector u_t is calculated by Equation 8. The normalized state weight α_t is obtained through the function (Equation 9). The aggregated of information in the traffic flow v is the weighted sum of each h_t with α_t as the corresponding weights.

$$h_t = LSTM(vec_t) \tag{7}$$

$$u_t = \tan h\left(Wh_t + b\right) \tag{8}$$

$$\alpha_t = \frac{\exp(u_t^T a)}{\Sigma_t \exp(u_t^T a)} \tag{9}$$

$$v = \sum_t a_t h_t \tag{10}$$

Then the vector v is fed to the output layer to perform the final prediction.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

2.2 The Traffic Congestion Prediction Architecture

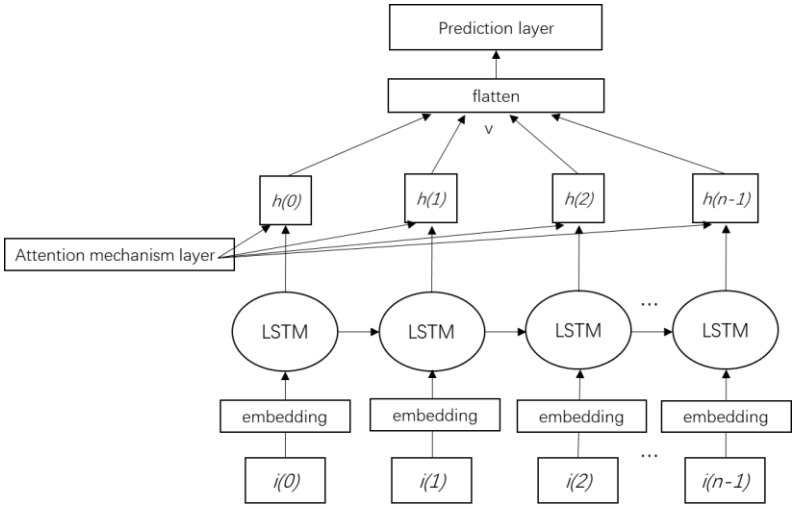

Fig. 3. The traffic prediction architecture

As shown in Fig 3, the prediction architecture mainly consists of four parts: the embedding layer, the LSTM, the attention mechanism layer and the prediction layer. The input is a sequence{i(0), i(1), . . . , i(n - 1)} which represents the traffic flow data, and each i(t) is a piece of data at a time interval encoded by one-hot representation. After the embedding layer, the data is mapped into a same dimensional vector space. Then, the LSTM network will process time-aware embedding vector and produce a hidden sequence {h(0),h(1), , . . . , h(n-1)}. An attention mechanism is used to extract traffic embedding features through the output attention probability matrix that is produced by the process in Section 2.1.3. Then, the prediction layer extracts mean values of the sequence over time intervals and makes the features encoded into a classified vector. Then it is fed into the logistic regression layer at the top of the prediction architecture.

3 Performance Analysis

In this section, to evaluate the effectiveness of our proposed approach, we first introduce our dataset and the experimental settings. Then we present the performances evaluated by different metrics. Finally, we show the comparative results with some baselines.

3.1 Datasets and Experiments settings

1) Dataset Description

In this study, the traffic data is collected by 855 fixed position sensors, located on the highways and roads of Gary-Chicago-Milwaukee (GCM) (consisting of 16 urban-

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

ized counties and covering 2500 miles). Each sensor collects the real-time traffic stream every 5 minutes, which contains attributes like longitude, latitude, length, direction, speed, volume, occupancy, congestion level, etc.

GCM highway system provides congestion levels on the different roads, which is shown in Fig 4.

Fig. 4. Gary-Chicago-Milwaukee (GCM) Corridor Transportation System

By analyzing the correlation matrix of attributes shown in Fig 5, dark colors represent high correlation between two attributes. We select attributes (speed, travel time, volume) that are more correlated to the congestion level.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

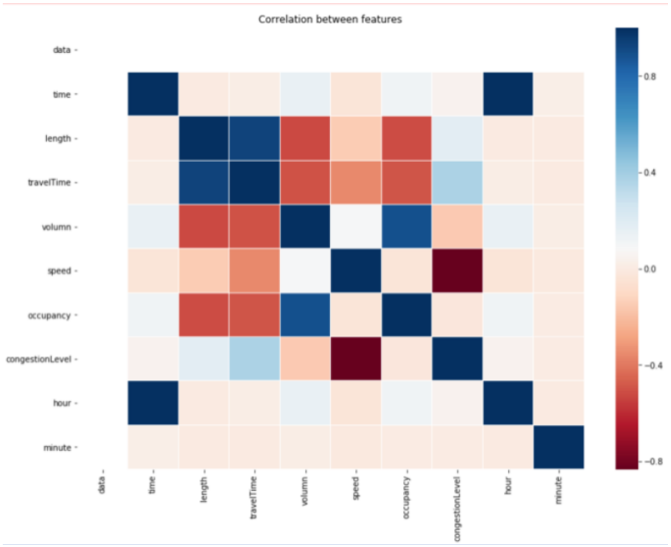

Fig. 5. Correlation between attributes

2) Experimental settings

Our method is implemented in Keras framework. The embedding dimension is 10. We take the traffic congestion values of the first 20 days as the training set, and the next 5 days as the validation set for the purpose of tuning parameters. The number of hidden units of LSTM is 64. We then use the stochastic gradient descent (SGD) method with the RMSprop [18] is set at 0.001 to minimize the square errors between our predictions and the actual congestion levels. Moreover, the mini-batch size is set at 64.

To improve the generalization capability of our model and alleviate the overfitting problem [19], we adopted the dropout method proposed in [20][21], which randomly drops units (along with their connections) from the network. The dropout rate of the output layer is set at 0.7.

3.2 Measures

To evaluate the effectiveness of the congestion prediction, we use two performance metrics, Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE), which are defined as:

$$MAPE\left(f,\hat{f}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|f_i-\hat{f}_i\right|}{f_i} \tag{11}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

$$RMSE\left(f,\hat{f}\right)=\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left|f_i-\hat{f}_i\right|\right)^2\right]^{\frac{1}{2}} \tag{12}$$

Where f is the real value of traffic congestion, and $\hat{f}$ is the predicted value.

3.3 Experimental Results and Discussion

We compare our proposed Attention-based LSTM with several methods in predicting the short-term traffic congestion levels. We use the same dataset and measures to ensure a fair comparison.

XGBOOST: extreme gradient boosting[22]
ARIMA: autoregressive integrated moving average
KNN: K-nearest neighbors[23]
LSTM: long short-term memory network

Table 2. Performance of different methods

method	MAPE(%)	RMSE
XGBOOST	10.34	67.25
ARIMA	9.13	61.86
KNN	8.96	59.32
LSTM	6.21	50.32
Attention-based LSTM	6.01	48.12

The congestion prediction performance of the five models is listed in Table 2. Both MAPE and RMSE of attention-based LSTM are lowest among the prediction models. Therefore, our proposed method is superior to the baselines.

Fig.6 presents the traffic congestion prediction vs. the observed congestion values collected from the data of No.IL-54 in one day. It is evident that the prediction results are satisfactory, and most of the fluctuations are captured by our Attention-based LSTM. Table 3 shows some results of congestion prediction of some sensors at 12:00am when we set the size of time slot at 30 minutes. We can see the congestion trends are almost the same (0 means normal, 1 means light, 2 means medium, 3 means heavy).

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

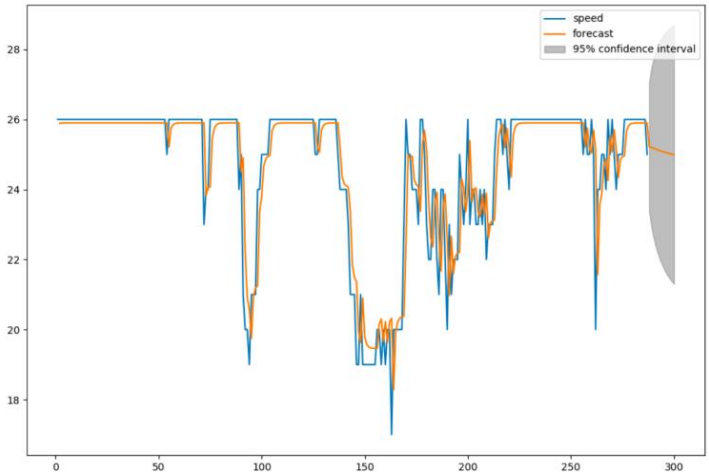

Fig.6. Traffic congestion prediction vs. observation

Table 3.

Number of sensors	12:30 observation	12:30 prediction
No.IL-239	(1,1,1,1,1,1)	(1,1,1,1,1,1)
No.IL-161	(3,3,3,3,3,3)	(3,3,3,3,3,3)
No.WI-7022	(1,1,1,2,2,2)	(1,1,1,1,2,2)
No.WI-9019	(2,3,3,3,3,3)	(2,3,3,3,3,3)
No.WI-33018	(2,2,2,2,2,2)	(2,2,2,2,2,2)

4 Conclusion

In this paper, we propose an Attention-based LSTM model to predict short-term traffic congestion, which is able to capture more features at different moments and take full advantage of the time-aware traffic data. In the experimental results, both the MAPE and RMSE of our model are the lowest when compared with XGBOOST, ARIMA, KNN, LSTM models in the real traffic data from Gray-Chicago-Milwaukee (GCM) Transportation Corridor in Chicagoland. It is demonstrated that the proposed method outperforms baselines significantly.

5 Acknowledgements

This project was partially supported by Guangdong Provincial Science and Technology Project 2016B010127004 and Grants from Natural Science Foundation of

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

China #71671178/ #91546201/ #61202321, and the open project of the Key Lab of Big Data Mining and Knowledge Management.

Reference

1. Vlahogianni E I , Karlaftis M G , Golias J C . Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach[J]. Transportation Research Part C Emerging Technologies, 2005, 13(3).
2. Barros J , Araujo M , Rossetti R J F . Short-term real-time traffic prediction methods: A survey[C]// 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, 2015.
3. Tian Y , Pan L . Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network[C]// 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity). IEEE, 2016.
4. Levin M, Tsao Y D. On forecasting freeway occupancies and volumes (abridgment)[J]. Transportation Research Record, 1980 (773).
5. Castro-Neto M, Jeong Y S, Jeong M K, et al. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions[J]. Expert systems with applications, 2009, 36(3): 6164-6173.
6. Xia D, Wang B, Li H, et al. A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting[J]. Neurocomputing, 2016, 179: 246-263.
7. Huang W, Song G, Hong H, et al. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning[J]. IEEE Trans. Intelligent Transportation Systems, 2014, 15(5): 2191-2201.
8. Lv Y, Duan Y, Kang W, et al. Traffic flow prediction with big data: A deep learning approach[J]. IEEE Trans. Intelligent Transportation Systems, 2015, 16(2): 865-873.
9. Chen Q, Song X, Yamada H, et al. Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference[C]//AAAI. 2016: 338-344.
10. Tian Y, Pan L. Predicting short-term traffic flow by long short-term memory recurrent neural network[C]//Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on. IEEE, 2015: 153-158.
11. Cui Z, Ke R, Wang Y. Deep Stacked Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction[C]//6th International Workshop on Urban Computing (UrbComp 2017). 2016.
12. Wang J, Hu F, Li L. Deep Bi-directional Long Short-Term Memory Model for Short-Term Traffic Flow Prediction[C]//International Conference on Neural Information Processing. Springer, Cham, 2017: 306-316.
13. Kawakami K. Supervised Sequence Labelling with Recurrent Neural Networks[D]. PhD thesis. Ph. D. thesis, Technical University of Munich, 2008.
14. Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in neural information processing systems. 2014: 2204-2212.
15. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
16. Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016, 2: 207-212.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24

17. Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.
18. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26-31.
19. Hawkins D M. The problem of overfitting[J]. Journal of chemical information and computer sciences, 2004, 44(1): 1-12.
20. Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.
21. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
22. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.
23. Habtemichael F G, Cetin M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns[J]. Transportation Research Part C: Emerging Technologies, 2016, 66: 61-78.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_24