ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Predicting vehicle behaviour using LSTMs and a vector power representation for spatial positions

Florian Mirus[1,2], Peter Blouw [3], Terrence C. Stewart [3] and Jörg Conradt[4]

1- BMW Group - Research, New Technologies, Innovations, Garching, Germany
2- Technical University of Munich - Department of Electrical and Computer Engineering, Munich, Germany
3- University of Waterloo - Centre for Theoretical Neuroscience, Waterloo ON, Canada
4- KTH Royal Institute of Technology - Department of Computational Science and Technology, Stockholm, Sweden

Abstract. Predicting future vehicle behaviour is an essential task to enable safe and situation-aware automated driving. In this paper, we propose to encapsulate spatial information of multiple objects in a semantic vector-representation. Assuming that future vehicle motion is influenced not only by past positions but also by the behaviour of other traffic participants, we use this representation as input for a Long Short-Term Memory (LSTM) network for sequence to sequence prediction of vehicle positions. We train and evaluate our system on real-world driving data collected mainly on highways in southern Germany and compare it to other models for reference.

1 Introduction

Predicting future behaviour and positions of other traffic participants from observations is a key problem intuitively handled by human drivers, that needs to be solved by automated vehicles as well to safely navigate their environment and to reach their desired goal. However, future positions of vehicles not only depend on each vehicle's own past positions and dynamic data (e.g. velocity and acceleration) but also on the behaviour of the other traffic participants in the vehicle's surroundings.
In this paper, we expand our previous work [1] on an automotive environment model based on Vector Symbolic Architectures (VSAs) [2]. Here, our main contribution is an encoding of spatial information for multiple objects in semantic scene vectors of fixed length. We hypothesize that this structured vector representation will be able to capture relations and mutual influence between traffic participants. For prediction the vehicle's future positions, we train a LSTM network using our vector-representation as well as other encoding schemes of the input data and compare their performance against each other as well as against a simple linear model based on a constant velocity assumption.

Related Work: There exist a variety of different approaches for motion prediction in automotive context [3]. Those methods vary in their approach to prediction (data-driven or model-based), the complexity of their motion model and also how they account for interactions between traffic participants or, more generally, agents in the scene. Probabilistic models like costmaps [4] impose physical constraints on the movements, other approaches categorize and represent scenes in a hierarchy [5], model interactions in the learning network's architecture [6] or include distances between other agents and the target directly in the training data [7].

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

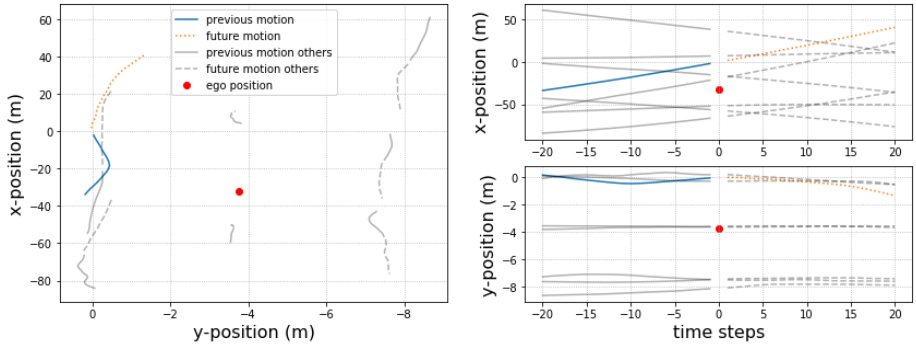


Fig. 1: Data visualization of one driving situation example. The red dot indicates the position of the ego vehicle, blue and orange lines show past and future motion of the target vehicle whereas gray lines depict the other vehicles' motion.

2 Methods

Data and Preprocessing: In this work, we use real-world data gathered during test drives mainly on highways in southern Germany. The data contains object-lists with a variety of features obtained from different sensor sources. Apart from features about motion and behaviour of the dynamic objects in the scene like position, velocity and acceleration, which are estimated from Light Detection and Ranging (LIDAR) sensors, there is also visual information like object type probabilities or lane information, which is acquired from additional camera sensors.
We aim to predict future positions of dynamic objects 5 s into the future based on their positions 5 s prior to their current location. To improve consistency and applicability, we interpolate the available data over 20 equidistant steps to achieve intervals of 0.25 s (see fig. 1). Finally, to improve suitability of the data as input for neural networks, we divide all x-positions by a factor of 10 such that x-/y-values are scaled to a similar order of magnitude. We created two data-sets, D_1 (102 vehicles) and D_2 (3891 vehicles), containing roughly 20 min and 10 h of driving data respectively for rapid and more in-depth training and evaluation with $D_1 \subsetneq D_2$. We split both data-sets into training $T_i \subset D_i$ and validation data $V_i \subset D_i$ containing 90 % and 10 % of the objects respectively with $T_i \cap V_i = \varnothing$ to avoid testing the system on vehicles it has been trained with.

Convolutive vector-power: The Semantic Pointer Architecture (SPA) [8] is based on Plate's Holographic Reduced Representations (HRRs) [9], which is one special case of a Vector Symbolic Architecture [2]. Here, atomic vectors are picked from the real-valued unit sphere, the dot product serves as a measure of similarity and the algebraic operations are component-wise vector addition $\oplus$ and circular convolution $\otimes$. In this work, we make use of the fact that for any two vectors $\mathbf{v}, \mathbf{w}$, we can write

$$\mathbf{v} \otimes \mathbf{w} = IDFT\left(DFT(\mathbf{v}) \odot DFT(\mathbf{w})\right), \tag{1}$$

where $\odot$ denotes element-wise multiplication, DFT and IDFT denote the Discrete Fourier Transform and Inverse Discrete Fourier Transform respectively.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

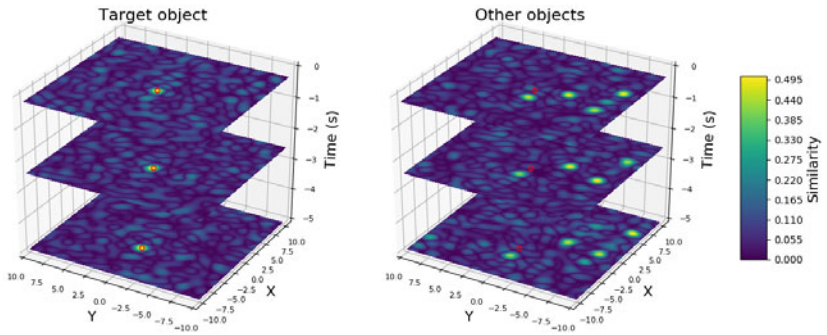


Fig. 2: Visualization of the convolutive vector-power representation over time as a heat map. The red circles indicate the measured position of the target vehicle.

Using eq. 1, we define the *convolutive power* of a vector $\mathbf{v}$ by an exponent $p \in \mathbb{R}$ as

$$\mathbf{v}^p := \Re\left(IDFT\left((DFT_j(\mathbf{v})^p)_{j=0}^{D-1}\right)\right), \tag{2}$$

where $\Re$ denotes the real part of a complex number. Furthermore, we call a vector $\mathbf{u}$ *unitary*, if $\|\mathbf{v}\| = \|\mathbf{v} \otimes \mathbf{u}\|$ for any other $\mathbf{v}$ (see [9, Sec. 3.6.3 and 3.6.5] for more details on the convolutive power and unitary vectors).

Vector representation: In this paper, we adopt and improve the vector representation for automotive scenes introduced in earlier work [1]. Here, we focus on investigating the expressive power of encoding spatial positions using the convolutive vector-power introduced in eq. 2. We assign a random ID-vector to each category of dynamic objects (e.g. car, motorcycle, truck) as well as random unitary vectors $\mathbf{X}$ and $\mathbf{Y}$ to encode spatial positions. Given a situation as shown in fig. 1 with a sequence of prior positions (x_t, y_t) for the target vehicle at time step $t \in \{t_0, \ldots, t_N\}$ and equivalent sequences $(x_{obj,t}, y_{obj,t})$ for all other visible objects closer than 40 m to the target, we encapsulate this information in a scene vector

$$\mathbf{S}_t = \mathbf{THIS} \otimes \mathbf{TYPE}_{target} \otimes \mathbf{X}^{x_t} \otimes \mathbf{Y}^{y_t} \oplus \sum_{obj} \mathbf{TYPE}_{obj} \otimes \mathbf{X}^{x_{obj,t}} \otimes \mathbf{Y}^{y_{obj,t}}, \tag{3}$$

where **THIS** denotes an additional ID-vector chosen at random to indicate the target object to be predicted. We use the 40 m threshold to avoid accumulation of noise in the vectors while focusing on the objects most relevant for prediction. This yields a sequence of scene vectors $\mathbf{S}_t$ for $t \in \{t_0, \ldots, t_N\}$ encoding the past spatial development of objects of interest in the current driving situation. Fig. 2 depicts the aforementioned scene vector representation: the left plot shows similarities (depicted as heat map) between the vector $\mathbf{S}_t$ encoding the scene from fig. 1 and $v = \mathbf{THIS} \otimes \mathbf{TYPE}_{target} \otimes \mathbf{X}^{\bar{x}_i} \otimes \mathbf{Y}^{\bar{y}_t}$, with a set of discrete position samples $\bar{x}_i, \bar{y}_i$. Similarly, the right plot shows similarities between $\mathbf{S}_t$ and $\mathbf{CAR} \otimes \mathbf{X}^{\bar{x}_i} \otimes \mathbf{Y}^{\bar{y}_t}$ visualizing all other objects in the scene of type *car*. Thus, we can encode spatial information of several different objects in a sequence of semantic vectors and reliably decode it back out. This allows us to encode automotive scenes with varying number of dynamic objects in a vector representation of fixed dimensionality.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

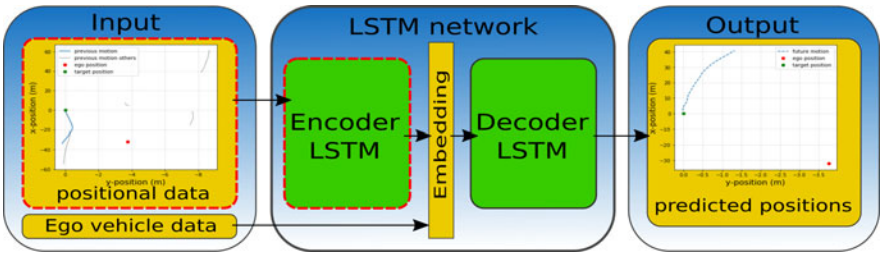


Fig. 3: Visualization of our learning architecture. Modules that change with varying encoding of the input data are highlighted through dashed red borders.

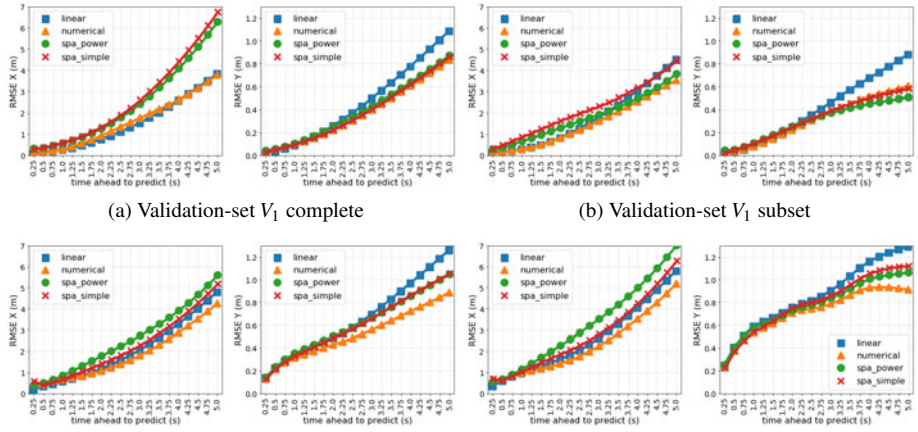
Network architecture and training: In this work, we use a Long Short-Term Memory (LSTM) [10] network-architecture consisting of one LSTM encoder and decoder cell with 150 hidden states each, for sequence to sequence prediction of vehicle positions. We use a similar network for all encoding schemes, whereas only the input dimensionality of the encoder cell changes when varying the representation of the input data. Fig. 3 visualizes this architecture indicating changing modules by a dashed red border. We use a batch-size of 150 and train the networks for 10 epochs on T_1 and for 20 epochs on the T_2 training data set.

3 Experiments

Reference encodings: For evaluation purposes, we also use different, simpler encoding schemes of our input data to compare our approach against. For a simple vector representation, we add the positional vectors $\mathbf{X}$ and $\mathbf{Y}$ scaled with the target vehicle's prior positions (x_t, y_t) at each time step t, yielding the sequence $\tilde{\mathbf{S}}_t = x_t \cdot \mathbf{X} + y_t \cdot \mathbf{Y}$. Finally, we also use the numerical position values $p_t = (x_t, y_t)$ as input data. Therefore, we have three different instantiations of the input data $(\mathbf{S}_t)_{t_0}^{t_N}$, $(\tilde{\mathbf{S}}_t)_{t_0}^{t_N}$ and $(p_t)_{t_0}^{t_N}$, which we refer to as "SPA-power", "SPA-simple" and "numerical". Both "SPA"-representations use 512-dimensional vectors. Note, that only the SPA-power representation $(\mathbf{S}_t)_{t_0}^{t_N}$ contains positional information about vehicles other than the target.

Results: Fig. 4 visualizes the Root-Mean-Square Error (RMSE) of all approaches on both validation-sets V_1 and V_2 for each dimension. Fig. 4a and 4c show the performance on the complete validation-set, whereas fig. 4b and 4d show only situations with at least 3 other vehicles present, the distance between the target and the ego vehicle being lower than 20 m and the distance between the target and the closest other vehicle being less than 10 m. We observe that all approaches yield comparable results with notable differences in certain situations. Although the SPA-power encoding scheme tends to perform worse in x-direction (longitudinal), we observe that it performs better in y-direction (lateral) in crowded situations with closely driving vehicles. Remarkably, the SPA-power representation performs best in y-direction in such situations when trained on the smaller data set T_1 (fig. 4b) and second best on validation set V_2 (fig. 4d).
To further investigate these results, we evaluated certain metrics, chosen to identify crowded and potentially dangerous situations, for items in both validation-sets, where the SPA-power approach outperforms all other approaches with respect to the RMSE in y-direction (see fig. 5). We observe that the ratio of situations, where the distance

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.



(a) Validation-set V_1 complete

(b) Validation-set V_1 subset

(c) Validation-set V_2 complete

(d) Validation-set V_2 subset

Fig. 4: RMSE evaluation in x and y-direction (left and right plot in all sub-figures) for the complete validation-sets (4a, 4c) and situations with at least 3 other vehicles present and distance between the target and ego vehicle lower than 20 m and between target and closest other vehicle lower than 10 m (4b, 4d).

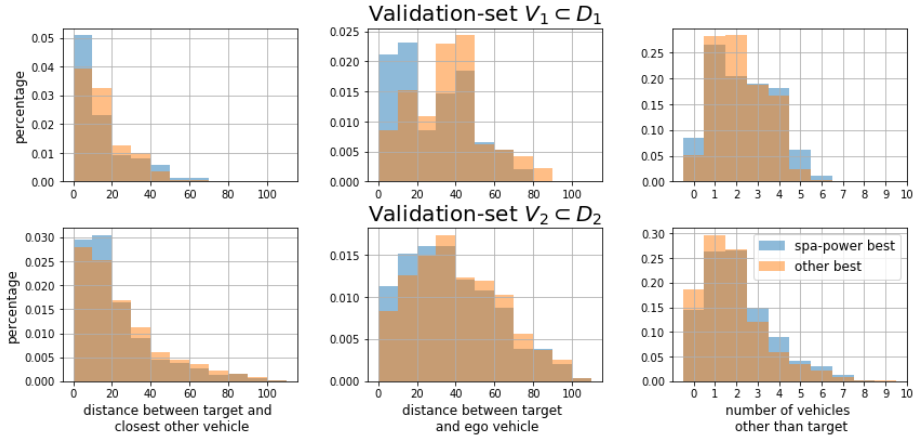


Fig. 5: Metric evaluation based on the RMSE in y-direction regarding situations where the model using SPA-power representation outperforms all other approaches.

between the target and the ego vehicle and/or the closest other object being small is significantly higher when the SPA-power representation outperforms all other approaches. Furthermore, the ratio of situations with at least 3 other vehicles present is also higher.

4 Discussion

Conclusion: In this paper, we showed a novel approach to encapsulate spatial information of multiple objects in a sequence of semantic pointers of fixed vector length.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

We used a LSTM network, to predict future vehicle positions from this representation. Our system clearly outperforms the simple linear model in y-direction and shows an comparable performance in y-direction and slightly worse performance in x-direction to benchmark-LSTMs using simpler input encodings. However, the results indicate that our approach gives improvements in crowded driving situations, where the target is close to the ego and/or other vehicles. We consider reliable predictions in such situations to be of crucial importance for collision avoidance and thus safe motion planning. Although this hints that the proposed SPA-power representation is actually able to capture mutual influence between vehicles, the results still demand further investigation.

Future work: We aim to analyze the results achieved in this paper in more detail through e.g. other partitions or subsets of the data-sets, tuning of the LSTM's hyper-parameters as well as potential improvements of the vector representation itself. Apart from that, the performance of our approach, especially in potentially dangerous driving situations, encourages us to extend and improve our system. We envision a future online-learning system, which chooses at runtime between several available predictors depending on the current driving situation to achieve the best possible forecast.

References

[1] F. Mirus, T. C. Stewart, and J. Conradt. "Towards cognitive automotive environment modelling: reasoning based on vector representations". In: *26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium*. 2018-04-25, pp. 55–60.

[2] R. Gayler. "Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience". In: *ICCS/ASCS International Conference on Cognitive Science*. Ed. by P. Slezak. University of New South Wales. CogPrints, 2003, pp. 133–138.

[3] S. Lefèvre, D. Vasquez, and C. Laugier. "A survey on motion prediction and risk assessment for intelligent vehicles". In: *ROBOMECH Journal* 1.1 (2014-07-23), p. 1. ISSN: 2197-4225. DOI: 10.1186/s40648-014-0001-z.

[4] M. Bahram, C. Hubmann, A. Lawitzky, M. Aeberhard, and D. Wollherr. "A Combined Model- and Learning-Based Framework for Interaction-Aware Maneuver Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 17.6 (2016-06), pp. 1538–1550. ISSN: 1524-9050. DOI: 10.1109/TITS.2015.2506642.

[5] S. Bonnin, F. Kummert, and J. Schmüdderich. "A Generic Concept of a System for Predicting Driving Behaviors". In: *2012 15th International IEEE Conference on Intelligent Transportation Systems*. 2012-09, pp. 1803–1808. DOI: 10.1109/ITSC.2012.6338695.

[6] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. "Social LSTM: Human Trajectory Prediction in Crowded Spaces". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016-06, pp. 961–971. DOI: 10.1109/CVPR.2016.110.

[7] F. Altché and A. de La Fortelle. "An LSTM Network for Highway Trajectory Prediction". In: *ArXiv e-prints* (2018-01). arXiv: 1801.07962.

[8] C. Eliasmith. *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press, 2013.

[9] T. Plate. "Distributed Representations and Nested Compositional Structure". PhD thesis. University of Toronto, 1994.

[10] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997-11), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.