Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 99 (2021) 650–655

www.elsevier.com/locate/procedia

14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME '20

A survey on long short-term memory networks for time series prediction

Benjamin Lindemann*, Timo Müller, Hannes Vietz, Nasser Jazdi, Michael Weyrich

Institute of Industrial Automation and Software Engineering, University of Stuttgart, Pfaffenwaldring 47, 70569 Stuttgart, Germany

* Corresponding author. Tel.: +49-711-685-67321; fax: +49-711-685-67302. *E-mail address:* benjamin.lindemann@ias.uni-stuttgart.de

Abstract

Recurrent neural networks and exceedingly Long short-term memory (LSTM) have been investigated intensively in recent years due to their ability to model and predict nonlinear time-variant system dynamics. The present paper delivers a comprehensive overview of existing LSTM cell derivatives and network architectures for time series prediction. A categorization in LSTM with optimized cell state representations and LSTM with interacting cell states is proposed. The investigated approaches are evaluated against defined requirements being relevant for an accurate time series prediction. These include short-term and long-term memory behavior, the ability for multimodal and multi-step ahead predictions and the according error propagation. Sequence-to-sequence networks with partially conditioning outperform the other approaches, such as bidirectional or associative networks, and are best suited to fulfill the requirements.

Peer-review under responsibility of the scientific committee of the 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

Keywords: Recurrent Neural Networks; Long short-term memory; Autoencoder; Sequence-to-Sequence Networks; Time Series Prediction

1. Introduction

Neural networks have been applied in the scope of numerous applications to model and predict complex system dynamics. There is a wide range of network types available but the modeling accuracy is strongly dependent on the fit of network architecture and considered problem. This paper presents an overview on neural networks, with a focus on Long short-term memory (LSTM) networks, that have been used for dynamic system modeling in diverse application areas such as image processing, speech recognition, manufacturing, autonomous systems, communication or energy consumption. The common aim in the scope of all investigated problems is the setup of prediction models based on time series data or data sequences to predict nonlinear time-variant system outputs. This paper analyzes existing approaches regarding the following properties:

- Nonlinear and time-variant prediction ability
- Short-term and long-term memory behavior
- Multidimensional data processing
- Multimodal prediction ability
- Multi-step ahead prediction and error propagation

The rest of the paper is organized as follows: chapter 2 presents a selection of recurrent neural network (RNN) concepts. Chapter 3 introduces two cell architectures that are based on gating mechanisms. Chapter 4 presents an overview of different LSTM architectures. They are divided into LSTM with optimized cell state representations and LSTM with interacting cell states. A detailed description of sequence-to-sequence (Seq2Seq) networks is conducted in chapter 5. The paper is concluded in chapter 6.

2. Recurrent neural networks

RNN are able to capture nonlinear short-term time dependencies. It can be distinguished between fully connected and partially connected RNN. The first RNN was developed by Williams and Zipser in the late 1980s, when an upswing in the development of neural network structures led to numerous fundamental contributions in this area [1]. Partially connected

Peer-review under responsibility of the scientific committee of the 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.
10.1016/j.procir.2021.03.088

RNNs are for instance the Elman net [2] and the Jordan net [3]. They pursue the goal of finding and modeling relations and extract contextual information from time series. Based on this foundation, numerous extensions of RNN have been developed in recent years to tackle a wide range of problems. In [4] a discrete wavelet transformation is integrated into an RNN to replace the regular activation functions. A comparison is drawn to classical activation functions, such as sigmoid as well as tangens hyperbolicus, and it is emphasized that an enhancement in modeling accuracy could be achieved due to an improved control of the information flow. Nevertheless, RNN have the disadvantage of vanishing gradients. Thus, non-stationary dependencies that occur over a long period of time are less well captured by RNN. An approach to tackle the problem concerning long-term dependencies that aims to establish a memory within regular RNNs is described by [5]. RNNs are extended by the usage of dilated recurrent skip connections to capture complex dependencies within time series data and to create a defined memory capacity. The skip connections allow to directly process information from past time steps without its propagation through the entire network. This is illustrated in figure 1.

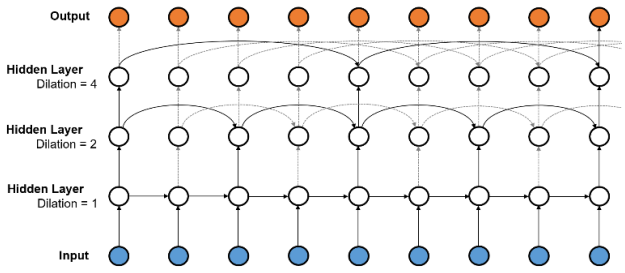

Fig. 1. RNN with dilated recurrent skip connections of different time spans as presented in [5].

Hence, the average length of the skip connections has a major influence on the fact whether short-term or long-term dependencies are primarily incorporated. However, [6] points out that RNN are primarily applicable to predict highly dynamic, time-variant systems subjected to stationary or non-stationary short-term dependencies.

3. Cell architectures

Multiple time dependencies with different characteristics as well as long-term dependencies are not sufficiently captured by RNN due to the vanishing gradient effect. Thus, gating mechanisms are developed to replace the classical activation functions. LSTM cells possess three gates, an input, a forget and an output gate, that allow to make changes on a cell state vector that is propagated iteratively to capture long-term dependencies. This controlled information flow within the cell enables the network to memorize multiple time dependencies with different characteristics. LSTM was introduced by [7] and is mainly used for the modeling of long-term dependencies. Before further LSTM network architectures are presented in this section, the Gated Recurrent Unit (GRU) is introduced as a modification of the LSTM cell. GRU was developed by [8] to model time series with the aim of creating a mechanism that complements the ability to predict long-term dependencies with an improved integration of short-term information. The aim is to enable an adaptive modeling of dependencies over different time horizons. Compared to LSTM, GRU has a simplified cell structure that also operates based on a gating system, but only has an update and reset gate. The main difference to LSTM is the circumstance that the cell state can be completely revised at each iteration and updated with short-term information via the reset gate. LSTM, on the other hand, provides a mechanism that limits the change gradient that can be realized at each iteration. Hence, LSTM does not allow past information to be completely discarded whereas GRU does. Empirical investigations on the cell architectures have been conducted in [9] where cells with gating mechanism achieve significantly better prediction results than the classical RNN approaches. Furthermore, the superiority of LSTM over GRU is determined by [10] in the scope of a large-scale study on variations of different network architectures. Despite a lower number of parameters in GRU cells, no significant advantages with regard to computing time could be substantiated. Furthermore, it could be found that the LSTM gating system contributes to the filtering of irrelevant input information and achieves a higher precision in the modeling of time-variant behavior. For this reason, the present work will further focus on network architectures based on LSTM cells.

4. LSTM network architectures

The following section discusses the state of the art with regard to network architectures that incorporate the LSTM gating mechanism. The architectures are divided into LSTM with optimized cell state representations (4.1 – 4.4), for instance based on attention mechanisms, and LSTM with interacting cell states (4.5 – 4.6), e.g. in the scope of cross-modal predictions.

4.1. Bidirectional LSTM

Bidirectional LSTM networks propagate the state vector introduced in chapter 3 not only in forward but also in reverse direction. This has the advantage that dependencies in both time directions are taken into account. Thus, expected future correlations can be included in current outputs of the network due to the reverse state propagation. Hence, bidirectional LSTM are able to detect and extract more time dependencies than unidirectional LSTM networks and resolve them more precisely. This is tested and evaluated by [11] where bidirectional LSTM networks encapsulate spatially and temporally distributed information and can handle incomplete data by a flexible connection mechanism for the propagation of the cell state vector. For each data sequence, this filter mechanism redefines the connections between cells on the basis of detected data gaps. The architecture is shown in figure 2. The suitability of the bidirectional architecture for the solution of multidimensional problems is shown in [12]. Within this work, features extracted from different dimensions are processed in a parallel architecture and merged using a bidirectional network.

Bidirectional LSTM could also be suitable for the representation of spatially and temporally distributed physics in manufacturing processes due to their properties. However, there is no preliminary work in this area yet.

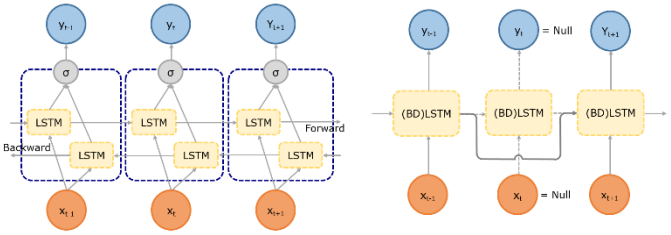

Fig. 2. Bidirectional LSTM (left) and filter mechanism for processing incomplete data sets (right) according to [11].

4.2. Hierarchical and attention-based LSTM

Hierarchical LSTM networks solve multidimensional problems by dividing the overall problem into sub-problems and organizing them in a hierarchical structure. This has the advantage that the focus can be placed on one specific or multiple connected sub-problems. This is done by shifting weights within the network which thereby acquires the ability to generate a certain amount of attention. Thus, hierarchical LSTM networks can be viewed as attention-based networks according to the description in [13]. In this work, an attention-based network is introduced that hierarchically structures images and aims to conduct a parallel prediction of the trajectories of different objects. Further approaches that deal with the topic of attention are presented, for instance, in [14] where separate attention layers are introduced and incorporated into the network. These layers can be enriched by a priori knowledge to effectively control the focus of data processing and prediction. Thus, attention can be created concerning output prediction but also with regard to the efficient processing of input sequences. In [15] hierarchical LSTM networks are used to predict long-term dependencies under the consideration of a weighting-based attention mechanism that processes and filters input sequences.

4.3. Convolutional LSTM

Input data that has been collected over a longer time horizon can be filtered and reduced based on convolution operations that are incorporated into LSTM networks or directly into the LSTM cell structure. These approaches aim to improve the prediction accuracy of long-term dependencies based on a more efficient processing of input sequences by projecting the data into a lower-dimensional feature space. In [16] the regular LSTM cell is extended by convolution operations that are directly integrated in the cell. Current input sequences, recurrent output sequences as well as weight matrices are convolved and correlations are extracted. The gates receive the generated features as new inputs. They are a reduced representation solely capturing the most relevant information so that the efficiency of the cell state update mechanism is enhanced. The method is depicted in figure 3.

Approaches for the incorporation of convolution operations into LSTM networks are presented by [17] or [18]. In the former case, the network is suitable for modeling locally distributed relations and for extracting corresponding representative features. LSTM cells, on the other hand, are used to learn temporal feature dependencies so that a composition of both network types in form of a stacked architecture shows decent prediction results. An advantage of convolutional LSTM is the fact that features can capture a long time horizon so that a larger amount of past information can be incorporated in the predictions.

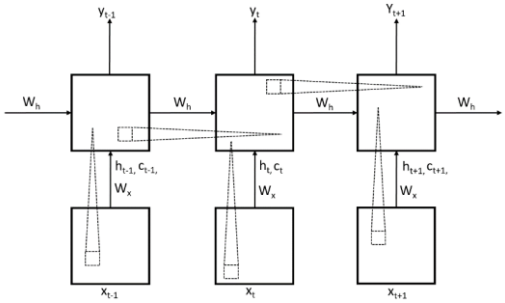

Fig. 3. Convolution operations within LSTM cells according to the approach of [16].

Furthermore, convolutional LSTM networks are also suitable for modeling multiple quantities e.g. spatially and temporally distributed relations due to their characteristic properties as demonstrated in the second work. However, multiple quantities can solely be predicted collectively in terms of a reduced feature representation. To predict multiple output quantities not as features but based on their original units, decoding or deconvolving layers are necessary.

4.4. LSTM autoencoder

The decoding and encoding of information is often realized in an autoencoder structure. In [19] a stacked LSTM autoencoder solves the problem of high dimensional input sequences and the prediction of high dimensional parameter spaces by a reducing and an expanding network. The autoencoder structure is trained separately with the aim of an exact reconstruction of input data as described by [20]. During test and operation, solely the encoder is applied for the extraction of low-dimensional features that are fed into the LSTM. An approach to directly integrate an autoencoder in the LSTM cell structure is introduced by [21]. The approach is developed to extend LSTM for multimodal prediction. Encoder and decoder are directly integrated in the LSTM cell structure to compress input data as well as cell states. This combined reduction optimizes the information flow in the cell and leads to an improved update mechanism of the cell state with respect to both short-term and long-term dependencies. Nevertheless, the encoding and decoding procedure is always connected to a finite loss of information. The extraction of features, on the other hand, significantly increases the density of prediction-relevant information so that the information loss effect can mostly be outreached with regard to the prediction accuracy.

However, a direct but cooperative prediction of multiple quantities can have a high potential for certain applications.

4.5. Grid LSTM

In addition to the afore-mentioned LSTM networks, numerous other network architectures have been developed in recent years. In [22] an LSTM cell on the basis of a matrix structure is proposed (Grid LSTM). In addition to regular connections between the layers, the Grid LSTM possesses connections regarding e.g. spatial or temporal dimensions of the input sequences. Hence, connections in multiple dimensions within the cells extend the regular information flow. The Grid LSTM is therefore suitable for a parallel prediction of multiple output quantities that can be either independent or linearly or nonlinearly dependent. In contrast to afore-mentioned approaches, a direct prediction of considered quantities without a projection onto abstract features is feasible. Each quantity is modeled within a separate dimension whereas the dependencies between the quantities are modeled by the newly created connections. The design of network and dimensions depend on the input data structure and the prediction goal [23]. Figure 4 illustrates a Grid LSTM with two dimensions is comparison to a simple stacked LSTM network.

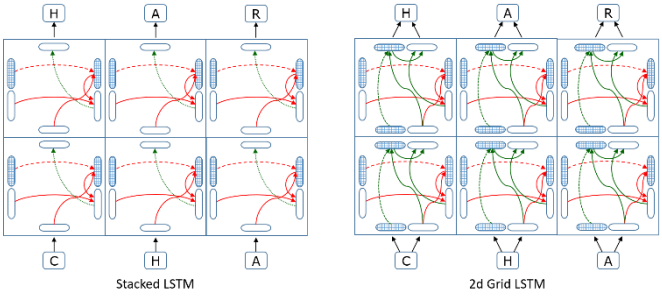

Fig.4. LSTM cells evolved by one dimension within a 2d Grid LSTM (right) compared to regular LSTM cells within a stacked LSTM network (left) [22].

The green information flow within the cells shows that the grid uses LSTM outputs of one dimension as additional inputs to other dimensions. By linking the dimensions, the prediction accuracy of multidimensional problems can be improved in contrast to one-dimensional stacked architectures. The latter are characterized by a stacking of several functional network layers as described in [19].

4.6. Cross-modal and associative LSTM

A novel approach is provided by [24] to cooperatively predict multiple quantities. It combines multiple regular LSTM that are used to separately model the individual quantities. These LSTM streams interact via recurrent connections to account for the dependencies of the quantities. Outputs of defined layers are utilized as additional inputs of previous and subsequent layers in other streams. Thus, a multimodal prediction can be realized (cross-modal LSTM). The concept is visualized in figure 5. Additionally, there exist further approaches that aim to conduct a multimodal prediction of multiple output variables. A dual holographic LSTM was developed by [25] that can detect correlations in data streams. It consists of two or more LSTM networks that process the data streams. The generated representations are examined with regard to dependencies by the usage of a circular cross-correlation. The method has the advantage that the extracted correlation vectors have identical dimensions as the input vectors. Thus, there is neither an increase concerning the parameter space and nor of computing time. An LSTM concept with external memory is presented by [26] that augments the network without an increase in parameters. The approach is also based on holographic reduced representations to realize a key-value based memory and storage of data representations. The key and the associated content are stored in a distributed manner and without predefined location.

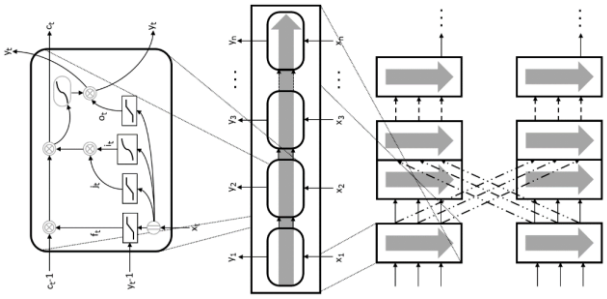

Fig. 5. Cross modal LSTM with recurrent connections to exchange information between different prediction streams [24].

In contrast to the afore-mentioned attention mechanism, the associative LSTM delivers a novel addressing mechanism of the distributed memory system. It is possible to realize multi-step ahead predictions based on the introduced LSTM networks by iterating the corresponding number of time steps over the network. Predicted outputs are fed back to predict additional outputs. A problem with this procedure for predicting an output sequence is the propagation of errors. The prediction on the basis of outputs that have already been predicted and are subjected to prediction errors further increases the propagated error. To be able to predict multiple time steps ahead at any point in time, sequence-to-sequence (Seq2Seq) networks are necessary.

5. Sequence-to-Sequence networks

5.1. Regular and multivariate Seq2Seq

The Seq2Seq architecture was introduced in [27] for the prediction of output sequences based on input sequences. Thus, a multi-step ahead prediction at every point in time is feasible. RNN and regular LSTM are solely able to predict output sequences by a simultaneous error propagation [8]. The above-mentioned approaches do not propose any concept to convert sequences of different lengths due to the fact that for each prediction a new input is required. Seq2Seq solves these problems by means of an encoder-decoder structure that incorporates a copying mechanism between the two network parts.

The input sequence is transformed and projected onto a vector of fixed dimensions. This abstract representation of the input is copied as initial cell state into the decoder and transformed into an output sequence of variable length. A comparison of Seq2Seq LSTM and regular LSTM is given by [28] based on different benchmark data sets. Seq2Seq was able to predict both short-term dependencies over a horizon of seconds to minutes as well as long-term dependencies over days whereas regular LSTM showed higher inaccuracies when modeling short-term dependencies.

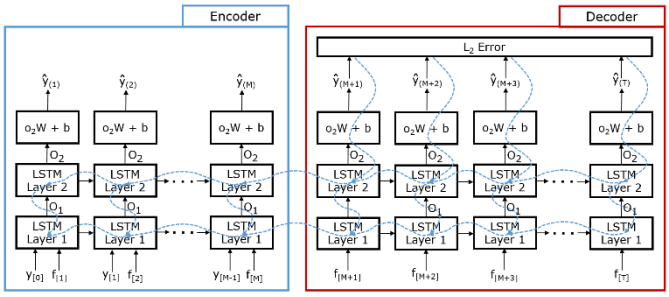

Fig. 6. Seq2Seq network with LSTM cells for the prediction of output sequences of variable time horizons according to [28].

The network architecture used is shown in figure 6. The blue arrows visualize the error feedback during training by the backpropagation through time (BPTT) algorithm. Seq2Seq networks can be realized in terms of different architecture variants depending on the number of inputs and outputs. Examples that have been applied in a hybrid approach presented in [29] or [30] are many-to-one, one-to-many or many-to-many architectures. The prediction of multiple quantities as discussed in previous sections is projected onto Seq2Seq LSTM in the scope of the work introduced in [31]. The presented model is able to perform a multivariate multi-step ahead prediction. A sliding window techniques is applied to construct the input vectors that encapsulates representations of all quantities of the multivariate time series.

5.2. Partially conditioned Seq2Seq

A disadvantage of regular Seq2Seq networks is the copy mechanism between encoder and decoder that considers a vector of defined length. A novel mechanism to realize a more flexible copying process is introduced by [32]. For this purpose, distribution functions that depend on the current context are applied to compress and copy the input representations to defined locations in the decoder. This means that not every value of the input sequence has to be included equally in the summation of the copy vector. This method is extended in [33] by introducing a transducer that slides over the encoder and generates a probabilistic output based on the first value of the input sequence. The probabilistic output is refined with each further input value. Thus, an expected output can be determined at any point in time during the processing of input sequences. This approach is visualized in figure 7.

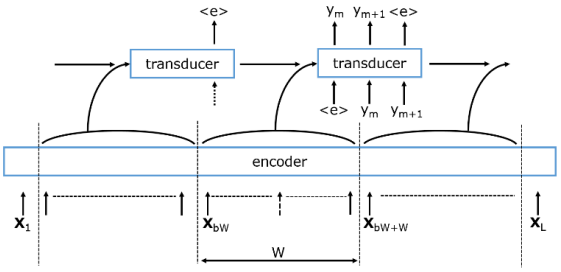

Fig. 7. Seq2Seq network based on partially conditioning [33].

An approach for the integration of attention to optimize the copying procedure is presented by [34]. The attention mechanism is utilized to focus and extract relevant correlations within multivariate input sequences. The approach also applies a probabilistic processing of inputs that provides a weighted mapping of input sequences to each individual output. The weightings are determined based on the extracted correlations.

Table 1 shows the evaluation of the LSTM architectures with regard to the investigated properties. In contrast to the architectures described in the previous sections, a prediction of multiple quantities can be realized by Seq2Seq not only in terms of a multi-step ahead prediction with error propagation but also as a sequence at once over an arbitrarily variable prediction horizon. In general, Seq2Seq LSTM approaches are predestined for multi-step ahead prediction with minimized error propagation. Grid and associative LSTM networks can cooperatively predict multiple quantities with high precision. Hierarchical as well as autoencoder-based LSTM concepts deliver improved update mechanisms of the cell state to process multidimensional data. All approaches are able to accurately predict nonlinear time-variant behavior but partially conditioned Seq2Seq LSTM show the best suitability to model both short-term and long-term dependencies.

Table 1. Investigated properties concerning all LSTM architectures.

Prediction properties	Bidirectional LSTM	Hier. & atte. LSTM	Convolut. LSTM	LSTM Autoencoder	Grid LSTM	Cro. & asso. LSTM	Reg. & mult. Seq2Seq	Part. cond. Seq2Seq
Nonlinear and time-variant prediction ability	+	+	+	+	++	++	++	++
Short-term and long-term memory behavior	0	+	0	+	+	++	+	++
Multidimensional data processing	+	++	+	++	+	+	+	+
Multimodal prediction ability	-	0	+	+	++	++	+	+
Multi-step ahead prediction and error propagation	-	-	--	-	0	-	++	++

6. Conclusion

The paper presents an overview of LSTM architectures that are developed to predict nonlinear time series behavior. There is a wide range of architectures available that have been used in different application areas such as image processing, manufacturing or autonomous systems. In the scope of this paper, the approaches are categorized and evaluated with regard to defined properties. The key findings are summarized as follows:

- LSTM with optimized cell state representations, such as hierarchical and attention-based LSTM, show an improved ability to process multidimensional data
- LSTM with interacting cell states, such as Grid and cross-modal LSTM, are able to cooperatively predict multiple quantities with high precision
- Seq2Seq LSTM can predict multiple quantities in terms of a multi-step ahead prediction with minimized error propagation
- Moreover, partially conditioned Seq2Seq LSTM show the best suitability to model both short-term and long-term dependencies

References

[1] Williams, R. J. and Zipser, R. A., A learning algorithm for continually training neural networks. Neural computation, vol. 1, 1989, pp. 270–280.

[2] Elman, J. L., Finding structure in time. Cognitive science, vol. 14, no. 2, 1990, pp. 179–211.

[3] Jordan, M. I., Attractor dynamics and parallelism in a connectionist sequential machine. in Artificial neural networks: concept learning, 1990, pp. 112–127.

[4] Motazedian, Z. and Safavi, A. A., Nonlinear and Time Varying System Identification Using a Novel Adaptive Fully Connected Recurrent Wavelet Network. in 2019 27th Iranian Conference on Electrical Engineering (ICEE), 2019, pp. 1181–1187.

[5] Chang, S. et al., Dilated recurrent neural networks. in Advances in Neural Information Processing Systems, 2017, pp. 77–87.

[6] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., and Cottrell, G., A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971, 2017.

[7] Hochreiter, S. and Schmidhuber, J., Long short-term memory. Neural computation, vol. 9, no. 8, 1997, pp. 1735–1780.

[8] Cho, K. et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[9] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y., Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[10] Britz, D., Goldie, A., Luong, M.-T., and Le, Q., Massive exploration of neural machine translation architectures. arXiv preprint arXiv:1703.03906, 2017.

[11] Cui, Z., Ke, R., Pu, Z., and Wang, Y., Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:1801.02143, 2018.

[12] Ma, X., Zhang, J., Du, B., Ding, C., and Sun, L., Parallel architecture of convolutional bi-directional lstm neural networks for network-wide metro ridership prediction. IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 6, 2018, pp. 2278–2288.

[13] Xue, H., Du Huynh, Q., and Reynolds, M., SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1186–1194.

[14] Su, Z. and Jiang, J., Hierarchical Gated Recurrent Unit with Semantic Attention for Event Prediction. Future Internet, vol. 12, no. 2, 2020, p. 39.

[15] Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., and Lee, H., Learning to generate long-term future via hierarchical prediction. in Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017, pp. 3560–3569.

[16] Chu, K.-F., Lam, A. Y. S., and Li, V. O. K., Deep multi-scale convolutional LSTM network for travel demand and origin-destination predictions. IEEE Transactions on Intelligent Transportation Systems, 2019.

[17] Huang, C.-J. and Kuo, P.-H., A deep cnn-lstm model for particulate matter (PM2. 5) forecasting in smart cities. Sensors, vol. 18, no. 7, 2018, p. 2220.

[18] Kim, T.-Y. and Cho, S.-B., Predicting residential energy consumption using CNN-LSTM neural networks. Energy, vol. 182, 2019, pp. 72–81.

[19] Gensler, A., Henze, J., Sick, B., and Raabe, N., Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. in 2016 IEEE international conference on systems, man, and cybernetics (SMC), 2016, pp. 2858–2865.

[20] Lindemann, B., Fesenmayr, F., Jazdi, N., and Weyrich, M., Anomaly detection in discrete manufacturing using self-learning approaches. Procedia CIRP, vol. 79, 2019, pp. 313–318.

[21] Hsu, D., Multi-period time series modeling with sparsity via Bayesian variational inference. arXiv preprint arXiv:1707.00666, 2017.

[22] Kalchbrenner, N., Danihelka, I., and Graves, A., Grid long short-term memory. arXiv preprint arXiv:1507.01526, 2015.

[23] Cheng, B., Xu, X., Zeng, Y., Ren, J., and Jung, S., Pedestrian trajectory prediction via the Social-Grid LSTM model. The Journal of Engineering, vol. 2018, no. 16, 2018, pp. 1468–1474.

[24] Veličković, P. et al., Cross-modal recurrent models for weight objective prediction from multimodal time-series data. in Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2018, pp. 178–186.

[25] Tay, Y., Phan, M. C., Tuan, L. A., and Hui, S. C., Learning to rank question answer pairs with holographic dual lstm architecture. in Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, 2017, pp. 695–704.

[26] Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., and Graves, A., Associative long short-term memory. arXiv preprint arXiv:1602.03032, 2016.

[27] Sutskever, I., Vinyals, O., and Le, Q. V., Sequence to sequence learning with neural networks. in Advances in neural information processing systems, 2014, pp. 3104–3112.

[28] Marino, D. L., Amarasinghe, K., and Manic, M., Building energy load forecasting using deep neural networks. in IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, 2016, pp. 7046–7051.

[29] Cherrier, N., Castaings, T., and Boulch, A., Deep sequence-to-sequence neural networks for ionospheric activity map prediction. in International Conference on Neural Information Processing, 2017, pp. 545–555.

[30] Lindemann, B., Jazdi, N., and Weyrich, M., Detektion von Anomalien zur Qualitätssicherung basierend auf Sequence-to-Sequence LSTM Netzen. at-Automatisierungstechnik, vol. 67, no. 12, 2019, pp. 1058–1068.

[31] Du, S., Li, T., and Horng, S.-J., Time Series Forecasting Using Sequence-to-Sequence Deep Learning Framework. in 2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), 2018, pp. 171–176.

[32] Gu, J., Lu, Z., Li, H., and Li, V. O. K., Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393, 2016.

[33] Jaitly, N., Le, Q. V., Vinyals, O., Sutskever, I., Sussillo, D., and Bengio, S., An online sequence-to-sequence model using partial conditioning. in Advances in Neural Information Processing Systems, 2016, pp. 5067–5075.

[34] Cinar, Y. G., Mirisaee, H., Goswami, P., Gaussier, E., Aït-Bachir, A., and Strijov, V., Position-based content attention for time series forecasting with sequence-to-sequence rnns. in International Conference on Neural Information Processing, 2017, pp. 533–544.