Short-Term Load Forecasting with Missing Data using Dilated Recurrent Attention Networks

Changkyu Choi[1], Filippo Maria Bianchi[2], Michael Kampffmeyer[1], and Robert Jenssen[1]

[1]UiT The Arctic University of Norway
[2]NORCE Norwegian Research Centre

Abstract

Forecasting the dynamics of time-varying systems is essential to maintaining the sustainability of the systems. Recent studies have discovered that Recurrent Neural Networks (RNN) applied in the forecasting tasks outperform conventional models. However, due to the structural limitation of vanilla RNN which holds unit-length internal connections, learning the representation of time series with *missing data* can be severely biased.

We propose *Dilated Recurrent Attention Networks* (DRAN), a robust architecture against the bias from missing data. This has a stacked structure of multiple RNNs, with each layer leveraging a different length of internal connections to incorporate previous information at different time scales, and updates its output state by a weighted average of the states in the layers. In order to focus more on specific layers that carries reliable information against missing data bias, our model leverages attention mechanism which learns the distribution of attention weights among the layers. The proposed model achieves a higher forecast accuracy than conventional ones from two benchmark time series with missing data that include a real-world electricity load dataset.

1 Introduction

An inaccurate forecast may pay an expensive price for financial and social deterioration which are unanticipated [3, 4]. Since the reliability of the forecast has a strong impact on the economic feasibility of industry [1], Short-Term Load Forecasting (STLF) in time-varying systems has been explored actively. Still, this is a difficult task as it depends on not only the nature of the system but also external influences. In the case of electricity consumption, we initially take distinct time dependencies into account as a nature of the system, namely intra-day, intra-week, and across different seasons [8]. Some external influences, such as calendar effects and rapid change of meteorological conditions, add irregularities on top of it [10].

Complex load patterns driven by the in- and external influences restrict the forecast to a given degree with conventional approaches, as they require strong statistical assumptions. RNN, a member of neural networks known for more flexibility with little prior assumptions, has become a standard framework for STLF tasks after outperforming conventional forecasting models that include AutoRegressive Integrated Moving Average (ARIMA) [4].

Missing data is a classical but critical problem in data analysis. They arise due to imperfect data collection, or various types of censoring [13]. Their possible effect on the results is seldom quantified despite the fact that they are a likely source of bias [15]. RNN can contribute to mitigating the bias from missing data by relying more on the previous information rather than the current missing data, as the internal connections play a role of memory. In addition, this learns rich information from the missing pattern, re-

https://doi.org/10.7557/18.5136

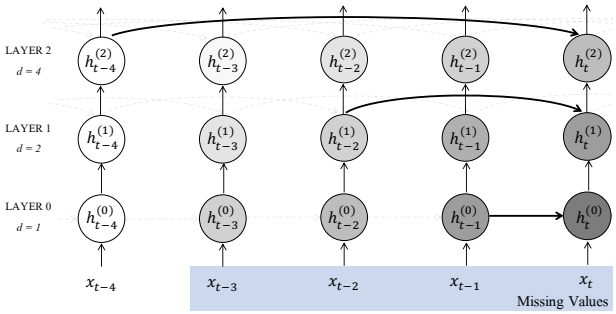


Figure 1: Unfolded graph of the dilated RNN with layer $L = 3$, DRNN(3). Consecutive four values $\{x_{t-3}, x_{t-2}, x_{t-1}, x_t\}$ in the blue window are missing. The gray-scale color of the RNN unit represents the degree of the bias from the missing values.

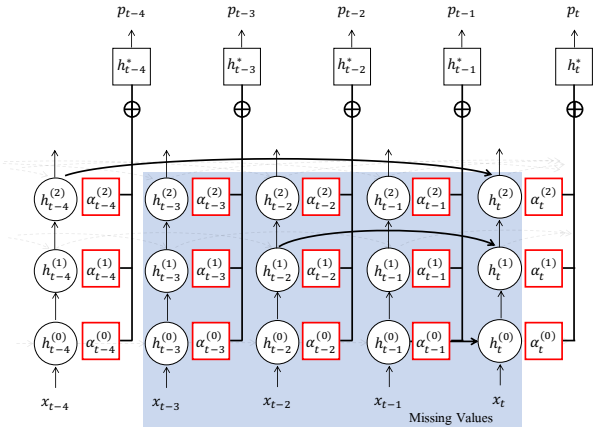


Figure 2: DRAN with layer $L = 3$, DRAN(3), with dilation $d^{(0,1,2)} = \{1, 2, 4\}$.

ferring to *informative missingness* [6]. Several RNN studies successfully attain the classification task with missing data [6, 12], however, there is a room for the study of STLF tasks that focuses on missing data.

We propose DRAN, a novel framework tailored for STLF tasks with missing data. This inherits the properties of Dilated RNN (DRNN) [5], featured by a multi-layer and cell-independent architecture, where each layer has a different internal connection, referred to *dilation*. To the best of our knowledge, this is the first STLF paper that applies RNN on the missing data problem. The model we suggest is readily applicable to other types of tasks but we limit ourselves to STLF tasks in this paper.

2 Dilated Recurrent Neural Networks

DRNN [5] is featured by *dilation* $d^{(l)}$, which is defined by initial length d_0, and base M. It is specified in Equation (1), where layer $l = 0, 1, \cdots, L-1$, state $\mathbf{h}_t^{(l)}$, and input x_t corresponding to layer $l = -1$.

$$\begin{aligned} \mathbf{h}_t^{(l)} &= f(\mathbf{h}_{t-d^{(l)}}^{(l)}, \mathbf{h}_t^{(l-1)}) \\ d^{(l)} &= d_0 M^l \end{aligned} \tag{1}$$

This enables the capture of multiple time dependencies and aggregate multi-scale temporal context into output. This provides more flexibility and capability in learning representation of the time series. The literature suggests to let $d^{(l)}$ have exponentially increasing length, as introduced in WaveNet [14] and Dilated CNN [16].

Role of Dilation towards Missing Data

Figure 1 represents how dilations operate in a missing window that consists of consecutive missing values $\{x_{t-3}, x_{t-2}, x_{t-1}, x_t\}$ represented by a blue box in the figure. As input values within the missing window are biased, it is reasonable to argue that a less number of the state update will protect the networks from the bias. Dilation is closely linked with the update frequency of the state $\mathbf{h}_t^{(l)}$. By comparing two dilations in LAYER 0 and LAYER 2 in Figure 1, it is evident that exploiting layers with longer dilation more in the missing window will reduce the update frequency of the state.

3 Dilated Recurrent Attention Networks

Figure 2 illustrates DRAN with layer $L = 3$ that improves DRNN(3) in Figure 1. The idea of DRAN is to leverage the attention mechanism [2] in regulating the exploitation of the layers when dealing with missing data. The attention mechanism is to make specific internal states contribute more to the output state, where a weighted average is the general form

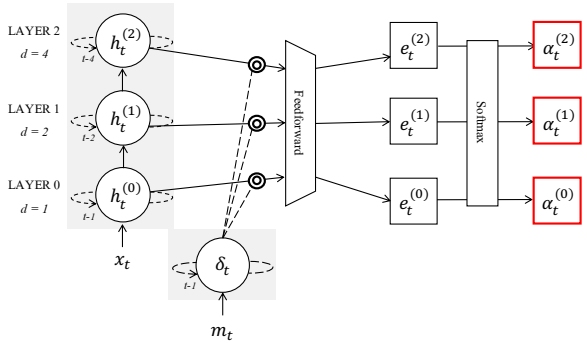
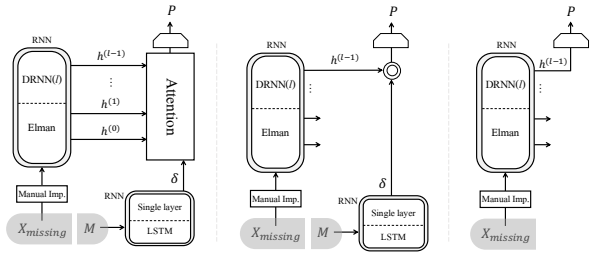


Figure 3: Schema of constructing attention for DRAN(3). The attention $\alpha_t^{(l)}$ is obtained by the score $e_t^{(l)}$ applied by a softmax function. The score is derived by the concatenation of missing history δ_t and the state $h_t^{(l)}$ processed by feedforward neural networks.



(a) DRAN(l) (b) DRNN(l) with missing mask (c) DRNN(l)

Figure 4: Model comparison: (a) DRAN(l); (b) DRNN(l) with missing history binary mask; (c) DRNN(l). Elman RNN refers to the vanilla RNN. Every model has input with missing values $\mathbf{X}_{missing}$. The effect of attention is compared by the model (a) and (b), where model (b) concatenates the output states of two RNNs. Model (c) are suggested to see the effect of missing mask by comparing with model (b). $\mathbf{M}$ and $\mathbf{P}$ represent binary mask and forecast respectively.

of the contribution. We define the trainable weights $\{\alpha_t^{(l)}\}$ as attention parameters.

We argue that DRAN simultaneously learns the representation of the states $\{\mathbf{h}_t^{(l)}\}$ and the distribution of the attention weights $\{\alpha_t^{(l)}\}$ over the layers in order to determine the exploitation of the layers with different dilations.

Depicted in Figure 3 and Equation (2), the construction of attention parameters that DRAN utilizes is unique and is inspired by two different methods, the attention mechanism [2] and missing history setting from GRU-D [6].

$$\begin{aligned} \alpha_t^{(l)} &= \frac{exp(e_t^{(l)})}{\sum_{k=0}^{L-1} exp(e_t^{(k)})} \qquad : \textit{softmax} \\ e_t^{(l)} &= g(\boldsymbol{h}_t^{(l)}; \boldsymbol{\delta}_t) \qquad\quad g\textit{: FFNN} \end{aligned} \tag{2}$$

The attention parameters $\{\alpha_t^{(l)}\}$ are derived from the scores $\{e_t^{(l)}\}$, processed by the softmax function so that they have values within the interval $[0, 1]$ and the sum over the layers is one. The scores $e_t^{(l)}$ play a role in incorporating current $\boldsymbol{h}_t^{(l)}$ and $\boldsymbol{\delta}_t$, representing the state at each layer and the missing history of input x_t respectively. Scores are derived by the concatenation of these two vectors, processed by a feedforward neural networks(FFNN).

$$\begin{aligned} \boldsymbol{\delta}_t &= f(\boldsymbol{\delta}_{t-1}, m_t) \qquad f\textit{: external RNN} \\ m_t &= \begin{cases} 1, & \text{if } x_t \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

Missing history $\boldsymbol{\delta}_t$ is the state of an external/small RNN. It is derived from binary mask time series m_t in Equation (3), processed by other RNN which are trained jointly, such as LSTM.

4 Experiments

The experiments are designed to compare DRAN(l) in Figure 4(a) with two reduced models, reduction of the attention unit in Figure 4(b), referring to DRNN(l) with *missing mask*, and reduction of the external RNN(LSTM) in Figure 4(c), referring to DRNN(l). Two baseline models, Gated Recurrent Unit(GRU) [7] and ARIMA(p, d, q), are chosen and compared with the three models mentioned above. The order of ARIMA(p, d, q) is carefully selected by following commonly used practices for the design of the coefficients[1].

[1] `https://people.duke.edu/~rnau/arimrule.htm`

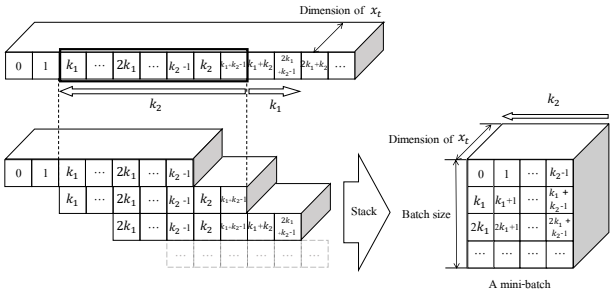


Figure 5: Formulation of a mini-batch for tBPTT(k_2, k_1).

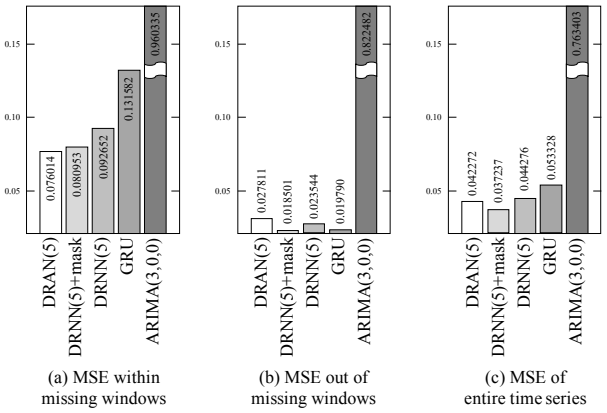


Figure 6: MSE comparison among DRAN(5) and others with MG set.

We analyze both a synthetically generated time series; Mackey-Glass (MG) system, and a time series from real-world load data from a public dataset; GEFCom 2012 competition [11], in order to provide controlled and easily replicable results for the architectures under analysis. MG dataset is given without missing values, hence, we assign missing values in the time series. To observe the performance when values are missing consecutively, we set the missing lasts to the next 50 time points once it happens. We refer the 50 consecutive missing values to a missing window with length 50. Missing windows are randomly assigned without overlap to make 30 % of the whole time series are missing. GEFCom dataset is given with consecutive missing values. Each missing window consists of length 168 and 4 windows are included in the time series.

The forecast accuracy is represented by the Mean Squared Error (MSE) obtained on the unseen values of the test set. The lower MSE implies the higher forecast accuracy. In order to obtain a forecasting problem that is not too trivial, it is reasonable to select forecast time interval that guarantees to become linearly decorrelated. Hence, we consider the first zero of the autocorrelation function of the time series [4], 12 time steps ahead for Mackey-Glass (MG) system [9] and 24 time steps ahead for GEFCom 2012 dataset [11].

All RNNs are trained by truncated backpropagation though time, tBPTT(k_2, k_1) [4] with its tailored mini-batch formulation illustrated in Figure 5. Note that a chunk of tBPTT(k_2, k_1) have overlapped information of length $k_2 - k_1$ with neighboring chunks. This redundancy, obtained from the overlapped information, alleviates the impact that occurs in the drawback where the gradient is not fully backpropagated.

5 Results

Mackey-Glass Dataset

Figure 6 reports the forecast accuracy of MG test set with respect to MSE obtained from each model. To show the difference between the prediction performance of the different models with or without missing values in the input, the MSE presented in each subplot is computed on (a) within the missing windows; (b) out of the missing windows; and (c) entire time series.

In Figure 6(a), DRAN(5) outperforms other models with the lowest MSE(0.076), meanwhile, in Figure 6(b), DRNN(5) with missing mask outperforms other models with the lowest MSE(0.018). In Figure 6(c), DRNN(5) with binary mask outperforms other models with the lowest MSE(0.037) and DRAN(5) follows by 0.042.

An important sanity check for DRAN consists of observing the change of each attention weights $\{\alpha_t^{(l)}\}$ between when the input data is missing or not. We keep track of each weight and compare the change

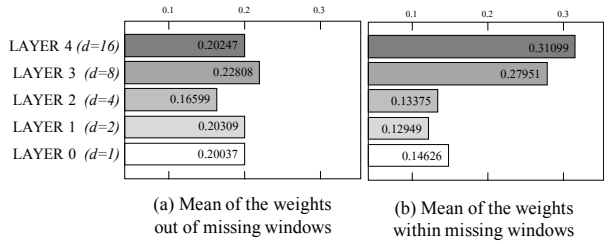


Figure 7: Comparison of the attention weights $\{\alpha_t^{(l)}\}$ of DRAN(5) depending on input missingness with MG set.

of mean values as attention weights play an indicating role revealing the layer that RNNs exploit. We argue that the change in the performance when the input data is missing or not supports the hypothesis that DRAN exploits the layer with the longer dilation more by redistributing finite attention resources when input value is consecutively missing. Figure 7(a) and (b) reveal that the average of attention weights of layer 3 ($d = 8$) and 4 ($d = 16$) strikingly increase while the weights of layer 1, 2 and 3 decrease within the missing windows, that supports the argument.

GEFCom Dataset

Figure 8 reports the forecast accuracy of GEFCom test set with respect to MSE in the same manner

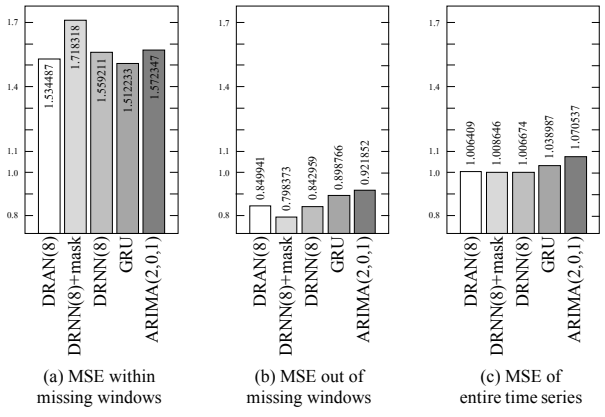


Figure 8: MSE comparison among DRAN(8) and others with GEFCom 2012 set.

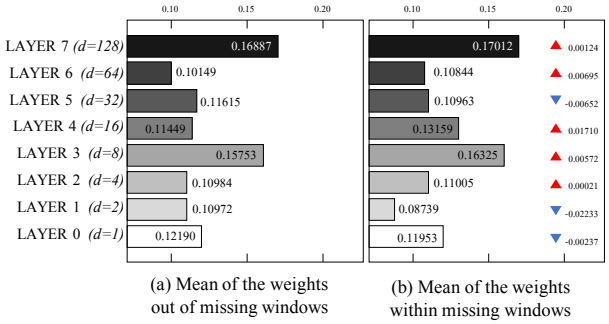


Figure 9: Comparison of the attention weights $\{\alpha_t^{(l)}\}$ of DRAN(8) depending on input missingness with GEFCom set.

shown in Figure 6. In Figure 8(a), DRAN(8) results in the lowest MSE(1.534) among the dilated RNNs class, and second lowest MSE, followed by GRU(1.512) with small difference.

Figure 8(b), DRNN(8) with missing mask achieves the lowest MSE(0.798) and other DRNN-based models are followed by, DRNN(8)(0.843) and DRAN(8) (0.850). For the MSE of the entire time series shown in Figure 8 (c), DRNN-based models indicate similar MSE, achieving a lower MSE than two baselines.

The change between Figure 9(a) and (b) follows similar phenomenon in Figure 7 between two classes. The attention weights with dilation $d = \{64, 128\}$ increase, while others turn to decrease. It implies that DRAN(8) uses attention to find more reliable information on its own, although the attention mechanism has not shown a definite improvement in the forecasting performance.

6 Conclusion

In the paper, we propose a novel model DRAN(l) tailored for STLF tasks with missing data. The consistent results from the different datasets support that DRAN(l) learns how to capture the missingness and utilize multiple dilations to improve forecasting accuracy.

References

[1] E. Almeshaiei and H. Soltan. A methodology for electric power load forecasting. *Alexandria Engineering Journal*, 50(2):137–144, 2011.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[3] F. M. Bianchi, E. De Santis, A. Rizzi, and A. Sadeghian. Short-term electric load forecasting using echo state networks and PCA decomposition. *IEEE Access*, 3:1931–1943, 2015.

[4] F. M. Bianchi, E. Maiorino, M. Kampffmeyer, A. Rizzi, and R. Jenssen. *Recurrent neural networks for short-term load forecasting: an overview and comparative analysis*. Springer, 2017.

[5] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. Hasegawa-Johnson, and T. Huang. Dilated recurrent neural networks. *NeurIPS*, 30:77–87, 2017.

[6] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, 2018.

[7] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Conference on Empirical Methods in Natural Language Processing*, 1:1724–1734, 2014.

[8] T.-H. Dang-Ha, F. M. Bianchi, and R. Olsson. Local short term electricity load forecasting: Automatic approaches. *International Joint Conference on Neural Networks*, 7:4267–4274, 2017.

[9] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Physical Review Letter*, 59(8):845–848, 1987.

[10] T. Hong and M. Shahidehpour. Load forecasting case study. *EISPC, US Department of Energy*, 2015.

[11] Kaggle. GEFCom global energy forecasting competition, 2012.

[12] Z. C. Lipton, D. Kale, and R. Wetzel. Modeling missing data in clinical time series with RNNs. *Machine Learning for Healthcare Conference*, 56:253–270, 2016.

[13] I. Shpitser, K. Mohan, and J. Pearl. Missing data as a causal and probabilistic problem. *Conference on Uncertainty in Artificial Intelligence*, 31:802–811, 2015.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *Arxiv*, 2016.

[15] M. Woodward, W. Smith, and H. Tunstall-pedoe. Bias from missing values: sex differences in implication of failed venepuncture for the scottish heart health study. *International journal of epidemiology*, 20(2):379–383, 1991.

[16] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016.