Deep Learning-Based Sensor Data Imputation System – A Case Study In Water Quality

Yi-Fan Zhang
Department of Agriculture and Food
CSIRO
Brisbane, Australia
yi-fan.zhang@csiro.au

Peter Thorburn
Department of Agriculture and Food
CSIRO
Brisbane, Australia
peter.thorburn@csiro.au

***Abstract*—Water quality high-frequency monitoring offers a comprehensive and improved insight into the temporal and spatial variability of the target ecosystem. However, most monitoring system lacks the consideration of sensor data quality control. The sensor data missing, background noises and signal interference have long been a huge obstacle for the users in understanding and analysing the sensor data, therefore makes the utilisation of sensor data much inefficient. We proposed an deep learning-based data imputation system for water quality sensor data. The model is based on the state-of-the-art sequence-to-sequence deep learning architecture and deployed on the Amazon Sagemaker service. It was tested in the 1622™WQ real-time water quality monitoring platform and can be a promising approach for data quality control in wireless sensor networks.**

***Index Terms*—Time Series, Missing Data, Data Imputation**

I. Introduction

The widespread use of in-situ high-frequency monitoring instrumentation enables a better characterisation of water quality processes, leading to more meaningful decision making. The large amount of data collected by the high-frequency sensors creates new opportunities for machine learning methods to better understand data-intensive processes in aquatic ecosystems and improve data streams coming from sensors. However, the issue of missing data is relatively common in wireless sensor networks and can have a negative effect on the conclusions drawn from the data.

II. Missing Data In Water Quality

Missing data are unavoidable in real-time monitoring networks. Although multiple methods have been proposed for filling gaps in the data, most methods give poor estimates when multiple data points are missing. The greater number of missing data points, the more difficult the gap to fill.

Fig. 1 illustrates a common scenario for missing time series sensor data. In this case, all the sensor data during the same period of time are missing. This can be caused by a variety of factors including unstable sensor power supply, data transmission errors or regular device maintenance.

Fig. 1. Time series with continuous missing data. Gray blocks highlight the available time series data (red solid lines), while the white blank spaces represent the missing data sequences for different time series.

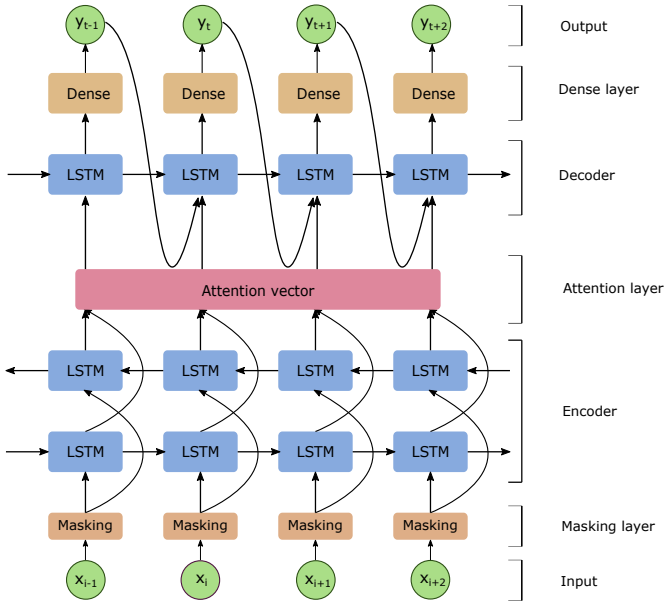

Fig. 2. SSIM architecture from . The encoder in the SSIM is a BiLSTM, which is comprised of a forward LSTM and a backward LSTM. The decoder in the SSIM is an unidirectional LSTM. The dropout Layers are also included in both the encoder and decoder. A masking layer is added to remove the zero-padded vectors in the input sequences

III. Deep Learning-Based Data Imputation System

A. Imputation Model

We proposed a new sequence-to-sequence imputation model (SSIM) for recovering missing data in sensor networks [1].

This work was supported by the CSIRO Digiscape Future Science Platform.

The SSIM uses the state-of-the-art sequence-to-sequence deep learning architecture. In conjunction with Long Short Term Memory Network (LSTM), the memory and attention mechanisms utilize both the past and future information for infilling data points in a period of time.

The SSIM utilizes the sequence-to-sequence architecture with the attention mechanism as depicted in Fig. 2, where the encoder and decoder are two key functional components. The encoder processes an input time series and maps it to a high-dimensional vector. The decoder takes input from the vector and yields target data sequences. Also, the attention mechanism enables the decoder to learn how to focus on a specific range of the input sequence for the differing outputs.

B. Imputation Workflow

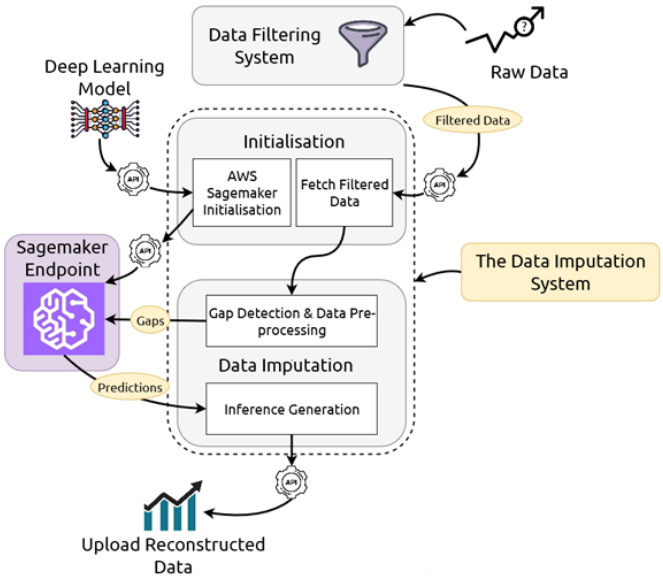

Fig. 3. Workflow of the data imputation system.

Fig. 3 illustrates the workflow of the data imputation system. At first, raw water quality measurements are passed through a data filtering system to remove the obvious outliers. After that, the configured imputation models are fetched from bitbucket, and a Sagemaker endpoint is created. Once the endpoint is active, every gap will be progressively filled, assigned a quality code, and used in future gap predictions to ensure there is a sufficient amount of input for the deep learning model. Finally, the data is uploaded back to the data storage system and displayed in the 1622™water quality monitoring platform [2]. If any of the uploads failed, the data will be stored in Amazon S3 for inference at a later scheduled time.

IV. System Evaluation

An example of the application of the SSIM is shown in Fig. 4. The missing data points are predicted by the SSIM one by one from 17/8/2017 to 23/8/2017. Each time the model yields one output, it will combine this output with the previous inputs to generate the next new output. The SSIM utilises

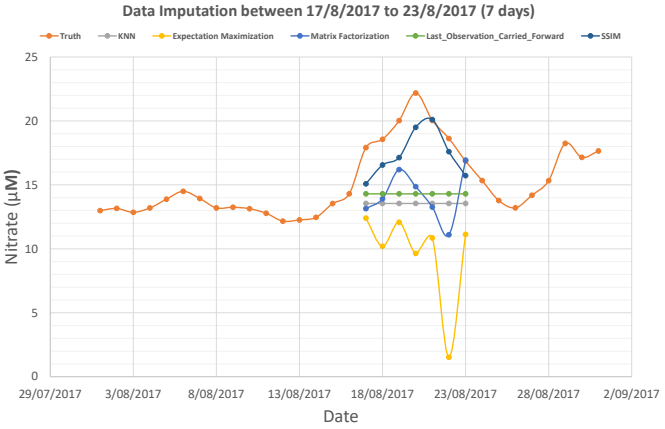

Fig. 4. Example data imputation from the SSIM method and four traditional imputation methods.

the available information both from the past and future time steps, which enhances model's ability to capture the trend through a period. Processing information from two directions can efficiently reduce accumulated predictive error.

Fig. 5. Imputation running performance.

Fig. 5 shows the running performance of the imputation system. On average, each Sagemaker instance executes around 550 seconds for the imputation task of one water quality variable. Benefit from AWS cloud's automation and scale-up capacity, the imputation model is secluded to run every 4 hours for the water quality monitoring data.

References

[1] Y. Zhang, P. J. Thorburn, X. Wei, and P. Fitch, "SSIM -a deep learning approach for recovering missing time series sensor data," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6618–6628, 2019.

[2] M. P. Vilas, P. J. Thorburn, S. Fielke, T. Webster, M. Mooij, J. S. Biggs, Y.-F. Zhang, A. Adham, A. Davis, B. Dungan *et al.*, "1622wq: A web-based application to increase farmer awareness of the impact of agriculture on water quality," *Environmental Modelling & Software*, p. 104816, 2020.