Temporal Attention and Stacked LSTMs for Multivariate Time Series Prediction

Tryambak Gangopadhyay
Department of Mechanical Engineering
Iowa State University
Ames, IA 50011
tryambak@iastate.edu

Sin Yong Tan
Department of Mechanical Engineering
Iowa State University
Ames, IA 50011
tsyong98@iastate.edu

Genyi Huang
Department of Statistics
Iowa State University
Ames, IA 50011
ghuang13@iastate.edu

Soumik Sarkar[*]
Department of Mechanical Engineering
Iowa State University
Ames, IA 50011
soumiks@iastate.edu

Abstract

Temporal attention mechanism has been applied to get state-of-the-art results in neural machine translation. LSTMs can capture the long-term temporal dependencies in a multivariate time series. We use temporal attention mechanism on top of stacked LSTMs demonstrating the performance on a multivariate time-series dataset for predicting pollution. Using attention to soft search for relevant parts of the input, our proposed model outperforms the encoder-decoder model version (using only stacked LSTMs) in most cases. In our approach, the soft alignments highlight the important time-steps that are most relevant in predicting pollution for future time steps.

1 Introduction

The air pollution levels are increasing nowadays in major cities like Beijing with modern urbanization and industrialization. Long-term exposure to fine particulate matters is an environmental risk factor causing cardiopulmonary diseases and lung cancer [1]. Particulate matters (with diameters up to 2.5 microns) is, therefore, one of the important metrics for pollution. Previous works of forecasting air pollution have been performed using deterministic models [2], linear models [3, 4] and support vector regression [5, 6]. LSTM has been used to predict air pollution for single future timestep [7, 8]. LSTM framework demonstrates equivalent accuracy compared to the baseline support vector regression, but LSTM can be used to predict multiple timesteps in the future [7]. We propose a forecasting model to predict air pollution for multiple timestep based on stacked LSTMs and temporal attention mechanism which outperforms the previous results [7] in terms of RMSE values.

LSTM networks use input, output and forget gates to prevent the memory contents being perturbed by irrelevant information. LSTM networks are capable of learning long-range correlations in a sequence and can accurately model complex multivariate sequences [9]. LSTMs have been used effectively for prediction tasks involving multivariate time series data as input [10, 11].

[*] This work was supported in part by U.S. AFOSR YIP grant FA9550-17-1-0220 and Iowa State Plant Science Institute Faculty Scholar award

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

In our encoder-decoder model, we use two layers of stacked LSTMs to predict the pollution in the future using multivariate time series data as input. But, in this model, the entire information in the input sequence is encoded in a vector which is used for decoding the output sequence. As the dimension of the vector is fixed, the performance may degrade when the length of the input sequence increases. The temporal attention mechanism learns the alignments by soft searching the most relevant time steps in the input. Our proposed model jointly learns to align and demonstrates improved performance for different lengths of the input. We can get an enhanced understanding from the attention weights learned by the model which explain the time steps of the multivariate time series input contributing the most for predicting pollution.

2 Experiments

We propose the use of temporal attention mechanism for multivariate time series prediction. The evaluation is performed on the Beijing PM2.5 Data Set. We vary the lengths of the input and output sequences keeping the number of variables in the input sequence fixed. In our approach, we compare the performance of two models for this task using the same dataset.

2.1 Dataset

The Beijing PM2.5 Data Set from UCI Machine Learning Repository is an hourly dataset containing the PM2.5 data of the US Embassy in Beijing and meteorological data from Beijing Capital International Airport [12]. For the problem formulation, we predict the pollution of the upcoming hours with the weather conditions and pollution from the previous hours as input. We have 8 input variables- pollution (PM2.5 concentration), dew point, temperature, pressure, combined wind direction, cumulated wind speed, cumulated hours of snow, cumulated hours of rain. The time period of the data is between Jan 1st, 2010 to Dec 31st, 2014 with 43824 total number of instances. Instead of using different intervals with six months' data as training and two months' data as testing [8], we use first 4 years' data for training and we test our model performance on the last 1 year.

2.2 Models

We use two types of models in this work. The first model is the LSTM Encoder-Decoder Model which doesn't comprise an attention block. The second model is the LSTM Attention Model with the attention block.

2.2.1 LSTM Encoder-Decoder Model

In the Encoder-Decoder Model, the encoder part compresses the information from the entire input sequence into a vector which is generated from the sequence of the LSTM hidden states [13]. The fixed-dimensional representation of the input sequence is given by the last hidden state of the encoding part as shown in Fig. 1. As we observe during experiments that deep LSTMs perform better than shallow LSTMs, we use two layers of stacked LSTMs for encoding the input sequence. The decoding part has one LSTM layer for predicting the output sequence.

2.2.2 LSTM Attention Model

The use of a fixed-length vector for encoding the input sequence can be a bottleneck in improving the performance of the LSTM Encoder-Decoder Model [14]. The extension of this approach is to learn the alignments by soft searching the set of time steps in the input having the most relevant information. Similar to the Encoder-Decoder Model, we use two stacked LSTM layers for the encoding part and one LSTM layer for decoding as shown in Fig. 2. The input sequence is encoded into a sequence of vectors instead of encoding into a fixed-length vector. Each annotation $a^{<t>}$ focuses on information surrounding the time step $<t>$ in the sequence. The context vector is generated by taking a weighted sum of the annotations.

$$context^{<t>} = \sum_{j=1}^{T_x} \alpha^{<t,j>} a^{<j>}$$

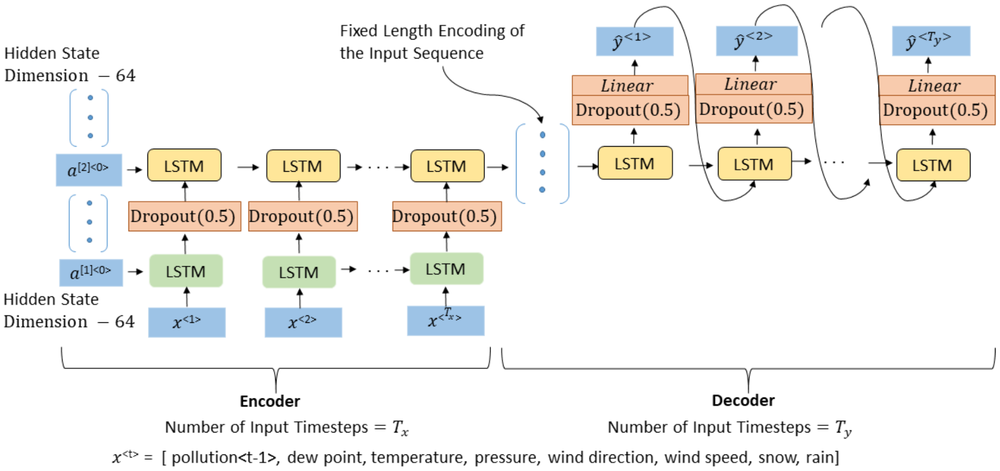

Figure 1: Encoder-decoder model using stacked LSTMs for encoding and one LSTM layer for decoding.

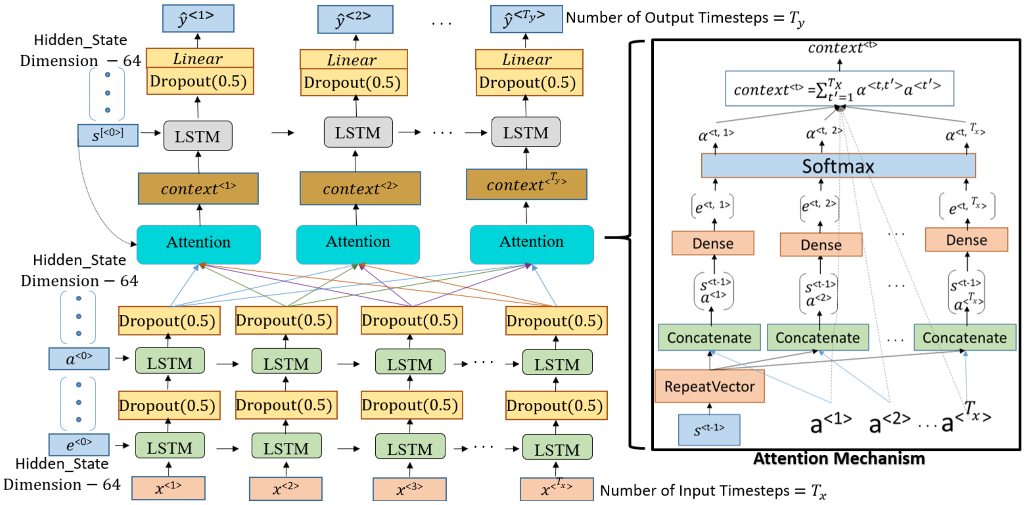

Figure 2: Proposed model predicts multiple output timesteps using temporal attention. The alignments are learned by the attention mechanism.

We use softmax to compute the attention weights for each annotation $a^{<t>}$ as:

$$\alpha^{<t,j>} = \frac{exp(e^{<t,j>})}{\sum_{j=1}^{T_x} exp(e^{<t,j>})}$$

The alignment model (dense layer,d) is parametrized as a feedforward neural network and is jointly trained with the entire network. The alignment model which scores the inputs around timestep t is given as:

$$e^{<t,j>} = d(s^{<t-1>}, a^{<j>})$$

3 Results

For the Encoder-Decoder model, we get test set RMSE values (for $T_y = 5$) ranging from 40.08 to 41.56 as we increase the number of input timesteps from 20 to 100. When we are predicting pollution for 10 future timesteps ($T_y = 10$), due to fixed vector length encoding the RMSE values increase

Input Sequence Length (T_x)	Model	RMSE($T_y = 5$)	RMSE($T_y = 10$)
20	Attention	39.55	51.65
	Encoder-Decoder	40.08	55.54
30	Attention	39.64	53.16
	Encoder-Decoder	42.03	54.79
40	Attention	40.06	53.93
	Encoder-Decoder	41.57	55.03
50	Attention	41.32	56.82
	Encoder-Decoder	41.45	57.99
100	Attention	40.14	57.03
	Encoder-Decoder	41.56	56.58

Table 1: Comparison of performance of the two models using test set rmse values for different input and output sequence lengths

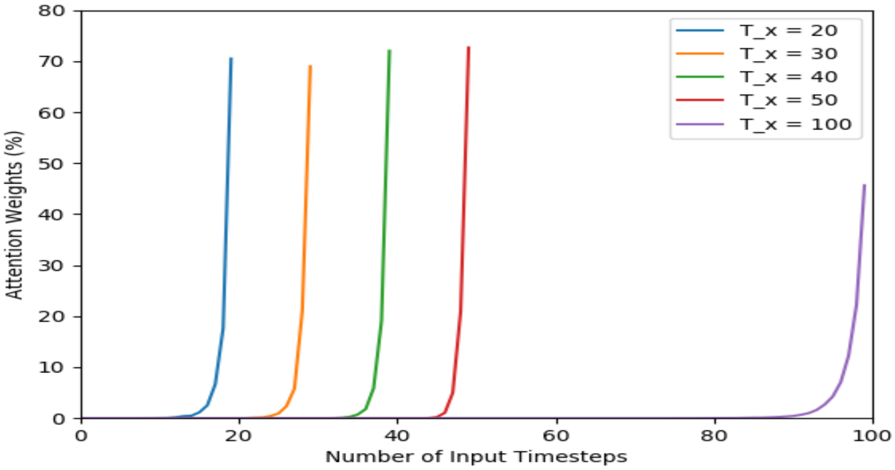

Figure 3: Distribution of attention weights for different lengths of input sequence.

55.54 to 57.99 with increasing T_x. With increasing T_x our proposed model (with Attention) performs better than the Encoder-Decoder model in most cases except for some cases with comparable results as shown in Table 1. We show the temporal attention weights for varying number of input timesteps in Fig. 3. The explanations provided by our model demonstrate that the model learns to focus only on the last few timesteps for predicting the output, even when we increase T_x from 20 to 100. For predicting pollution, Fig. 3. suggests that using less number of input timesteps is sufficient and that excess input information may degrade the performance.

4 Conclusions

Predicting pollution for future timesteps is an important task and accurate predictions can lower the risk with adequate warnings and implementing beneficial policies. In this work, we propose a model based on stacked LSTMs and temporal attention mechanism which focuses on relevant input timesteps for prediction. Our approach performs better than the Encoder Decoder model without using attention. We outperform other previous results on the same dataset with both of our models. The model also shows that regardless of the length of the input sequence it focuses mostly on the recent timesteps to predict the pollution. The input timesteps with the highest weights contribute the most to predict the output sequence. The alignments learned by the model may provide a better understanding of different types of multivariate time series prediction problems.

References

[1] C Arden Pope III, Richard T Burnett, Michael J Thun, Eugenia E Calle, Daniel Krewski, Kazuhiko Ito, and George D Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, 287(9):1132–1141, 2002.

[2] Carlie J Coats Jr. High-performance algorithms in the sparse matrix operator kernel emissions (smoke) modeling system. In *Proc. Ninth AMS Joint Conference on Applications of Air Pollution Meteorology with A&WMA, Amer. Meteor. Soc., Atlanta, GA*, pages 584–588. Citeseer, 1996.

[3] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] Can Li, N Christina Hsu, and Si-Chee Tsay. A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmospheric Environment*, 45(22):3663–3675, 2011.

[5] Bing-Chun Liu, Arihant Binaykia, Pei-Chann Chang, Manoj Kumar Tiwari, and Cheng-Chin Tsao. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *PloS one*, 12(7):e0179763, 2017.

[6] Wei-Zhen Lu and Dong Wang. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Science of the total environment*, 395(2-3):109–116, 2008.

[7] Vikram Reddy, Pavan Yedavalli, Shrestha Mohanty, and Udit Nakhat. Deep air: Forecasting air pollution in beijing, china.

[8] Chiou-Jye Huang and Ping-Huan Kuo. A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. *Sensors*, 18(7):2220, 2018.

[9] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings*, page 89. Presses universitaires de Louvain, 2015.

[10] Johnathon M Shook, Linjiang Wu, Tryambak Gangopadhyay, Baskar Ganapathysubramanian, Soumik Sarkar, and Asheesh K Singh. Integrating genotype and weather variables for soybean yield prediction using deep learning. *bioRxiv*, page 331561, 2018.

[11] Zehui Jiang, Chao Liu, Nathan P Hendricks, Baskar Ganapathysubramanian, Dermot J Hayes, and Soumik Sarkar. Predicting county level corn yields using deep long short term memory models. *arXiv preprint arXiv:1805.12044*, 2018.

[12] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proc. R. Soc. A*, 471(2182):20150257, 2015.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.