A novel model for the prediction of long-term building energy demand: 
*LSTM with Attention layer*

To cite this article: Yuan Gao et al 2019 *IOP Conf. Ser.: Earth Environ. Sci.* 294 012033

View the [article online](#) for updates and enhancements.
A novel model for the prediction of long-term building energy demand: *LSTM with Attention layer*

Yuan Gao¹, Chengkuan Fang¹ and Yingjun Ruan²

¹ Postgraduate Student, Tongji University, China
² Associate Professor, Tongji University, China
E-mail: ruanyj@tongji.edu.cn

Abstract. In modern building management systems, the prediction of building energy demand is of great importance for energy saving in buildings. Nowadays, the long-term building load forecasting usually depends on the meteorological data. Therefore, most researches in this field focus on traditional machine learning methods to resolve intricate linear relations between energy demand and meteorological features, such as temperature or humidity. On the contrary, those traditional methods cannot attain ideal accuracy in long-term prediction tasks since those methods ignore the correlation of results in each timestep. Hence, people may use recurrent neural network (RNN) to memory the information of time series to improve prediction performance such as Long short-term memory (LSTM). However, RNN have the problem of long-term time dependency, which means this model may not run well in long-term prediction. In this paper, we presented a new deep learning method with LSTM and attention layer to solve the problems mentioned above. A set of measured data from an office building in Qingdao were used to verify and test our model network. Based on the mean absolute percentage error (MAPE), the results indicate that our model can get the 6.4% result for the whole year prediction, which is 2.9% less than the result of using only LSTM.

1. Introduction

Owing to the advancement of intelligent online meters and computer calculating ability, the administrators of buildings currently rely on measured data and precise prediction of energy demand to cope with their tasks. Therefore, it is crucial for administrators to get accurate prediction of energy demand in a long-term time period in order to manage building’s equipment reasonable and save building’s energy use. Thus, the problem of prediction of long-term energy demand has been considered by other authors and has been approached in different ways.

A multitude of prediction methods based on traditional machine learning methods such as Support Vector Regression (SVR)[1], Neural Networks[2], and autoregressive integrated moving average (ARIMA) models[3] were used to solve this problem. Some instances follow. C. Deb et al.[4] proposed a data-driven method to forecast diurnal cooling and electrical demand for public buildings. They studied three public buildings within two years in detail. The energy consumption data are divided into five categories to improve the accuracy of prediction. W. Yu et al.[5] developed a residential energy consumption prediction model based on the Back Propagation (BP) neural network. They introduced 16 kinds of energy consumption indexes as the characteristics of residential buildings in Chongqing. The results of their prediction were good. Paudel et al.[6] used Support Vector Machines (SVM) to carry out energy demand forecasting for low-energy buildings. They adopted two training methods, ‘all data’ and ‘relevant data’, to improve accuracy of prediction. The results showed that using partial relevant data can enhance the generalization of their models.
Be that as it may, during the process of traditional machine learning algorithms, these algorithms concentrate on matching a kind of intricate linear equation which demonstrate the relations between features such as humidity and temperature and building energy demand. Indeed, the results showed that traditional machine learning algorithms run well in short-term prediction\cite{7}. Contrarily, these algorithms cannot forecast long-term energy demand in a relatively sufficient accuracy. That could be owed to the fact that conventional machine learning tools such as Artificial Neural Network neglect the inner logical relationships and sequences among the long-term energy demand time series which are significant factors for long-term load prediction. In other words, the sequence of training data will not influence the final forecasting results in traditional machine learning algorithms, that is why traditional machine learning cannot obtain sufficient accuracy in long-term prediction.

To overcome the shortage of machine learning, this paper proposed Recurrent Neural Network (RNN) to emphasize the inner logical relations among data set during the process of prediction. This special Neural Network will consider both the relations between ‘label to features’ and ‘label to label’. In this network, the training cells combine the input features with previous outputs. To improve the accuracy a step further, this paper also proposed attention layer with RNN to resolve the problem of long-term dependencies which is crucial for long-term prediction.

The paper is organized as follows. In section 2, we introduced a special kind of RNN named Long Short-Term Memory (LSTM) and attention layer. In section 3, we introduced the data we used and the methods to normalize the data. The results of prediction were shown in section 4. Finally, in section 5, we gave summaries and conclusions.

2. Long Short-Term Memory and Attention layer

Traditional neural networks and machine learning tools could hardly deal with the persistence problems, which means those methods could not use prevent events to inform later one. However, Recurrent neural networks can address this issue. There are loops in them, which allow information to pass and persist. In recent few years, there have been numerous success applying RNNs to a variety of problems such as language modelling, translation and image captioning. Figure 1 shows the basic topology of a simple RNN.

![Figure 1](image.png)

Figure 1. The structure of RNN

Where $X$ and $o$ indicate the input data and output data; $h$ denotes the output of hidden layer; $U$ and $V$ indicates weight matrixes delineating the relationship between $X$ and $h$ and $h$ and $o$; $W$ is a square matrix which demonstrates the relationship between two hidden layers. However, the hidden layer $h$ only connects directly to the hidden layer output one timestep before. Thus, it is sensitive to short-term information. But there are also cases where we need more context. Considering predicting next whole year’s energy demand, it’s entirely possible for a large gap between the relevant information.
Unfortunately, as that gap grows, RNNs become unable to learn to connect the information. Consequently, we select a special RNN named Long Short-Term Memory (LSTM) to predict the next whole year energy demand.

2.1. LSTM

LSTM is a special RNN which is effective for time series data. The traditional RNN only applies the transformation result of the active function to the current input data and the hidden node output of the previous timestep as the output of the current hidden node. Compared with simple RNN, each module in LSTM use special LSTM cell and has a different structure called gates, including Forget gate, input gate and output gate. Figure 2 shows the structure of the LSTM cell.

Figure 2. The structure of LSTM cell

The three gates are determined by the current input and the output of the hidden layer at the previous one timestep.

The purpose of the input gate in_t,i is to provide a weight to receive the current normal input information. The input state at timestep t and node i is provided as:

\[ \text{in}_t,i = \text{sigmoid}(b_{in}(i) + \sum_j U_{in}(i, j)x_t,j + \sum_j W_{in}(i, j)h_{t-1,j}) \]  

where \( x_t \) indicates the current input data; \( h_{t-1} \) denotes the output of all the nodes in previous one timestep; \( b_{in}, U_{in}, W_{in} \) indicates respectively the Threshold vector, the Input weight matrix, and the Output weight matrix of the input gate.

The second gate is called forget gate \( f_t,i \), which determine a weight to forget current state data. The formulation is presented as:

\[ f_t,i = \text{sigmoid}(b_f(i) + \sum_j U_f(i, j)x_t,j + \sum_j W_f(i, j)h_{t-1,j}) \]  

where \( b_f, U_f, W_f \) indicates respectively the Threshold vector, the Input weight matrix, and the Output weight matrix of the forget gate.

LSTM cell must generate new memory \( S_{t+1,i} \), which provided as:

\[ S_{t+1,i} = f_t,iS_t,i + \text{in}_t,iC_{t,i} \]  

\[ C_{t,i} = \tanh(b(i) + \sum_j U(i, j)x_t,j + \sum_j W(i, j)h_{t-1,j}) \]  

where \( C_{t,i} \) demonstrated the new memory calculated by the new input data. The tanh is a kind of active function.

Above all, the last gate is obviously output gate. It can provide a weight to evaluate the recognition level of current output. Its value depends on the current input and the output of the previous hidden layer. The formulation is provided as:
4

\( o_{t,i} = \text{sigmoid}(b_o(i) + \sum_j U_o(i,j)x_{t,j} + \sum_j W_o(i,j)h_{t-1,j}) \)  

\( \text{The final LSTM output is defined as:} \quad h_{t,i} = o_{t,i} \cdot \text{tanh}(S_{t,i}) \)

2.2. Attention layer

According to the characteristics of the RNN and LSTM, they seem to be important tools in time series sequence research and could be used to predict the long-term building energy demand. But one of the obvious shortcomings of RNN is that it can't be paralleled, so it's slower, which is a natural defect of recursion. In addition, RNN can't learn the global structural information well because it is essentially a Markov decision process. Besides, even the LSTM cannot avoid the flaw of long-term forgetting, which means the network could not remember long-term information or state and convey it to the current LSTM cell. Therefore, using LSTM alone to deal with long-term prediction will not attain ideal and sufficient accuracy.

In 2017, Google[8] promoted a new and special network called attention. The original of attention is to stimulate the brain activity of human being. When human study a time series, they can easily capture the global trend and can also focus on small periods of whole sequence. The frame and definition of Attention layer is provided as Figure 3[8] and equation 7[8]:

\[
\text{Figure 3. The structure Attention layer}
\]

\[
\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_i}})V
\]
where \( Q \in \mathbb{R}^{n \times d_k}, K \in \mathbb{R}^{n \times d_k}, V \in \mathbb{R}^{n \times d_v} \), if we neglect the active function \( \text{softmax} \), equation 7 actually indicate a process of Multiply three matrices. This process encodes a sequence whose shape is \( n \times d_k \) to a \( n \times d_v \) sequence. Equation 8 demonstrate a particular calculation of a single vector in equation 7:

\[
\text{Attention}(q_k, K, V) = \sum_{i=1}^{m} \frac{1}{Z} \exp\left(\frac{(q_k, k_i)}{\sqrt{d_k}}\right)v_s
\]

(8)

where \( Z \) is the normalization factor, \( q_k, K, V \) are abbreviations for query, key, value, respectively, \( K, V \) are one-to-one correspondence, they are like a key-value relationship; \( \sqrt{d_k} \) is an adjustment factor, in order to prevent the inner product from being too large. The purpose of the formula is to obtain the similarity between \( q_k \) and each \( v_s \).

However, equation 7 and 8 describe a generalized attention layer which always applied in Natural Language Processing (NLP). In this paper, the model uses Self-Attention which will more suitable to our task of long-term energy demand prediction. The definition of Self-Attention is as follow:

\[
Y = \text{Attention}(X, X, X)
\]

(9)

which means we just use attention to input sequence itself, in order to discover the inner relationships between input sequence.

In this paper, we used a layer LSTM and a layer Self-Attention to build our model. This model considered global trend in our training data, thus significantly improved our accuracy of prediction.

3. Case study: an office building in China

3.1. Datasets

This case study aimed to use LSTM network with attention layer to improve the accuracy of an entire year energy demand prediction in an office building in China. The goal is to attain a higher accuracy than using LSTM alone and using simple Dense Back Propagation (BP) Neural Network.

The datasets contain two years of daily data. The time period is from January 1, 2016, to December 31, 2017. The datasets contain 12 attributes, including two temporal attributes: (1) datetime, (2) holiday, 10 external weather attributes: (3) minimum outdoor-temperature, (4) maximum outdoor-temperature, (5) mean outdoor-temperature, (6) minimum wind speed, (7) maximum wind speed, (8) mean wind speed, (9) dew point, (10) minimum humidity, (11) maximum humidity, and (12) mean humidity; and finally the label (13) total electrical energy consumption, which is measured and collected by the local government. The holiday is a category-binary feature which ‘0’ indicates a day of business day, ‘1’ refers to a day of public holiday according to Chinese official schedule.

Figure 4 and 5 demonstrates the energy of this office building in the year 2016. Since this office building is located in North of China, the local government will support the heat demand during the winter. Consequently, the electrical consumption in winter is higher than transition seasons but lower than summer, which can be attribute to the fact that the whole building relies on split air conditioner to meet all cooling demand in the summer but opening the air conditioner in winter just as an extra assist to public heating system.

Figure 5 depicts each month electrical consumption in the whole year, the trend is similar in some months since it is an office building which contain regular running schedule. This characteristic is crucial in training model. Since model cannot accept all the training data at one batch, therefore, in the training process of the simple Dense Neural Network, we select 30 as a mini-batch size to train our model. In this case, the process of training time will significantly decrease without any punishment in accuracy. Figure 6 is a heatmap of 2016 electrical consumption in target building which delineates that the summer consumption is much higher than other seasons. Besides, the working periods of this building were also depicted precisely in this figure, which are normally a five-days working period and a two day off.

3.2. Datasets pre-processing
As for machine learning and deep learning, the pre-processing of datasets is a significant and non-substitutable step since it can dramatically decrease the training time. In this paper, we used z-score, a traditional normalization method. It is presented by the equation 10:

$$z = \frac{x - \mu}{\sigma}$$  \hspace{1cm} (10)

**Figure 4.** The consumption of 2016

**Figure 5.** The consumption of 2016 in each month
The numeric attributes will follow a normal distribution after applying z-score normalization. This method can be allocated to all the attributes in the datasets except datetime and holiday. Since the datetime will set to be the index of the datasets to introduce the order in LSTM, the holiday still needs another pre-processing method owing to the fact that its’ value doesn’t indicate its’ weight in the model. The model will recognize holiday more important than business day since 1 is greater than 0 in mathematic field.

Therefore, this paper introduced an encoding method named One-Hot Encoding to manage category attribute holiday. One-Hot Encoding is effective to deal with category attributes or discrete attribute, since this method can also expand the dimension of training data which will be benefit for model. Consequently, this paper proposed One-Hot Encoding to handle holiday feature and after that the dimension of holiday is from 1 to 2. As a result, the whole attributes of datasets are 13 after pre-processing.

3.3. Model Building and Evaluation

To verify the impact of attention layer, this paper implemented three models to the same datasets:
• L: normal LSTM network with a year of data from the target school, used as a benchmark to compare with AL.
• AL: LSTM model combined with Attention layer.
• D: Simple Dense Deep Learning network, used only as a benchmark.

The main goal of these three models were to verify the hypothesis that the accuracy of prediction in AL is higher than L, thus, proving the effect of attention layer in long-term accuracy. Besides, the accuracy of L should be higher than D. consequently, it is meaningful to use more sophisticated model LSTM.

4. Prediction Results

In this paper, we selected Mean Absolute Percentage Error (MAPE) for evaluation of each model. MAPE expresses average absolute error as a percentage. MAPE is calculated as follows:

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| 
\]  

(11)
where $y_i$ is the actual energy consumption, $\hat{y}_i$ is the predicting value, and $N$ is the number of the testing data. Table 1 shows the results of the three models. The results also been presented in the Figure 7 and Figure 8, where Figure 8 is a timeline chart to depict the actual energy consumption and the prediction results of each model.

![Mean Absolute Percentage Errors](chart1)

**Table 1** shows the results of the three models.

| Model   | MAPE       |
|---------|------------|
| D model | 0.129      |
| L model | 0.093      |
| AL model| 0.064      |

![Comparison of models with their MAPE](chart2)

**Figure 7.** Comparison of models with their MAPE

**Figure 8.** The first month prediction results

It was obvious that D model attains the worst forecasting accuracy since this model just added some simple Dense layer together, as a result, neglected the relationship in timeseries. However, this simple model still got relatively low MAPE, which is 12.9%, due to the similarity between train data and test data. We set the results of D model as benchmarks. The MAPE of L model was lower than the benchmarks, which demonstrated the importance of sequence during the process of training. This model has already considered the relation between previous timesteps and current inputs, therefore, improved the accuracy about 3.6%. Given to the fact that LSTM cannot convey and remember long-term
information. This model would not achieve the lowest MAPE as AL model in the field of long-term prediction. The AL model got the lowest MAPE 6.4%, demonstrating the importance and effectiveness of Attention layer in long-term prediction.

5. Conclusion
This paper has proposed a new deep learning model which contains LSTM cells and Attention layer. The purpose of this model is to solve the problem of ignoring the relationship between each result in the perspective of time sequence and the problem of long-term memory. Therefore, we use LSTM to train the data as a time sequence and consider the relationship in the time series. The Attention layer is introduced in order to improve the ability of the model to convey and remember the long-term information, thus, to increase the accuracy of long-term prediction.

To verify Attention layer, a case study on long-term energy consumption prediction for an office building in China was presented. In this paper we use mean absolute error as the evaluation of our results. The result of LSTM is 9.3% in average of the whole year results, which is 3.6% lower than the traditional BP neural network. Furthermore, the result of LSTM with attention layer is 2.9% lower than the simple LSTM and 6.5% lower than the traditional BP neural network, which denotes that using LSTM and attention layer will improve the prediction result in the field of long-term energy demand prediction.

Future work will focus on transfer the fresh methods from NLP to this field, such as Sequence to Sequence, in order to improve the accuracy of prediction and enhance the research field to other type of buildings.

References
[1] F. Zhang, C. Deb, S. E. Lee, J. Yang and K. W. Shah 2016 Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique Energy and Buildings 126 94-103
[2] R. Kumar, R. K. Aggarwal and J. D. Sharma 2013 Energy analysis of a building using artificial neural network: A review Energy and Buildings 65 352-358
[3] M. A. Mat Daut, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah and F. Hussin 2017 Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review Renewable and Sustainable Energy Reviews 70 1108-1118
[4] C. Deb, L. S. Eang, J. Yang and M. Santamouris 2016 Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks Energy and Buildings 121 284-297
[5] W. Yu, B. Li, Y. Lei and M. Liu 2011 Analysis of a Residential Building Energy Consumption Demand Model Energies 4 475-487
[6] S. Paudel et al 2017 A relevant data selection method for energy consumption prediction of low energy building based on support vector machine Energy and Buildings 138 240-256
[7] J. Massana, C. Pous, L. Burgas, J. Melendez and J. Colomer 2016 Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes Energy and Buildings 130 519-531
[8] A. Vaswani et al 2017 Attention Is All You Need Energy arXiv:1706.03762 [cs]

Acknowledgments
This research has been supported by the national key R & D project (No.2016YF0700104-05)