ISSN 2004-2965 Energy Proceedings, Vol. 17, 2021

Paper ID APEN-MIT-2021_118
Applied Energy Symposium: MIT A+B
August 11-13, 2021 • Cambridge, USA

An Attention-based Seq2Seq Model for Short Term Energy Consumption Prediction

Yuhang Zhang [a, b], Xiangtian Deng [a, b], Yi Zhang [a, b, c], Yi Zhang [a, b, *]

[a] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, 518055, PR China
[b] Future Human Habitats Division, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, PR China
[c] Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, 100084, PR China
*Correspondence: zy1214@sz.tsinghua.edu.cn

Abstract—Building energy consumption prediction is important in energy system management, building operation, and energy supply planning. This study proposes a novel model with attention based seq2seq method, which is a deep learning algorithm, to improve the prediction performance. The developed model is performed with experiment on a real energy profile data of an office building in Shenzhen, China. The prediction performance of the proposed hybrid model is evaluated with indicators of MSE, RMSE, MAE and SMAPE. The results demonstrate that attention mechanism can improve the prediction performance of model whose input are time series. Compared with the metrics of prediction result of other models, the MSE, RMSE, MAE, SMAPE of prediction result of proposed model decrease by more than half percent.

Keywords—energy consumption prediction, seq2seq model, attention mechanism, LSTM

I. INTRODUCTION

The prediction of building energy consumption has been considered not only challenging but also significant. The increasing population extend the demand of energy , and the building energy consumption takes a great percentage of the whole energy consumption in worldwide [1] .The accurate prediction of short term building energy consumption can help to optimize the operation of building energy system(BES) [2] while the accurate prediction of long term building energy consumption is able to give a guide to the policy and strategy making for government and energy providers [3]. For example, the energy supplier can make more reasonable electricity price [4] by the prediction of energy consumption and the government can take measures to allocate the energy more efficient [5] and aid energy saving [6] according to the result of energy consumption prediction, thus benefiting the environment, promoting the economy and supporting the sustainable development of society [7].

There are many methods related to building energy consumption proposed, such as SVM, MLP, random forest. In recent years, attention mechanism attracts many researchers' attention for its great performance in many time series prediction tasks. [8] Additionally, convolutional neural network and long short term memory neural networks always have good performance in prediction tasks in many areas like traffic flow prediction, stock price prediction, emotion prediction, computer vision areas(cv) and natural language processing [9]. There are many approaches related to building energy consumption, but there are still limitations that hamper the prediction performance. Most models need to be iterated step by step to predict energy consumption, and they can not maintain accuracy as the prediction time becomes longer. The data in different time windows are not given different weight of attention while they have different degree of importance to the output. Additionally, hybrid deep learning algorithm has been less considered which can combine the advantages of different algorithms. In order to overcome these shortcomings, we proposed a novel attention-based seq2seq model. The main contributions of this are as follows:

- We combine the attention mechanism and Seq2seq model to improve the performance of building energy consumption prediction.
- We design different time windows to predict the energy consumption of longer time lengths at one forward prediction, and carry out experiments on a real dataset of office building in Shenzhen.
- The proposed model is compared with current prediction models including SVM, MLP, LSTM, CNN-LSTM, attention-based LSTM based on case study.

The remainder of this paper is organized as follows: Section 2 reviews the related work about building energy consumption prediction. Section 3 introduces the specific implementation methodology and the data preparation process. Section 4 describes the development of A-seq2seq model and experimental detail. Section 5 presents the result. Section 6 draws the conclusion.

II. RELATED WORK

Many researchers study the prediction of energy consumption. The main approaches to predict the building energy consumption can be classified into two categories: physical modelling approaches and data driven approaches [3]. Physical modelling approaches predict energy consumption based on complex physical functions, physical theories and many parameters related to energy system, weather conditions, occupant behavior. These parameters are hard to be available. As the population increases and the

development of technology, there are increasingly huge amount of energy data and the ability of collecting, storing, analyzing of these data has also be improved. Therefore, data driven method is a more practical and easier to perform approach with excellent performance in many studies. Erickson reviewed the applications of data science for different building energy management tasks, like prediction of building energy load, building operation optimization, economic analysis of electric consumption, and fault detection and prevention [10]. Data driven method includes statistical method, machine learning method and deep learning method. The statistical methods used to predict the energy consumption are mainly regression related method, including multiple variable linear regression, ordinary least squares regression, autoregressive (AR) [11], and autoregressive integrated moving average(ARIMA) [12]. However, the statistical method depends on historical data heavily and it is hard to deal with the complex and the trend and fluctuation of energy consumption data [13]. Compared with statistical method, machine learning are more flexible and can deal with the complex nonlinear relationship. Machine learning is widely used in energy consumption prediction currently and have better performance than statistical methods. Support vector machine [14, 15], Back propagate neural network [16-18] , Xgboost are common machine learning methods to predict the energy consumption. Dong et al. used SVM to predict building electrical energy consumption every month in tropical areas [19]. Li et al. proposed SVM have superior accuracy compared with conventional BPNN in predicting the cooling load in office building [20]. In addition, some researchers make some improvements for energy consumption prediction method based on machine learning methods. Some researchers have improved these machine learning methods by combining these models. Li et al. proposed artificial neural network (ANN) and hybrid Genetic Algorithm-Adaptive Network-based Fuzzy Inference System (GA-ANFIS) to predict building energy consumption [21]. The research indicate the result of using GA-ANFIS are more accurate than that of using ANN. Fan et al. Combined the Self Organized Map (SOM) and SVM to predict the short term load [22]. The result shows that the hybrid model perform better than the single model. With the development of deep learning, the prediction accuracy and efficiency of deep learning in some areas have exceeded that of traditional machine learning [23], and there are also some researchers apply it to the area of energy consumption prediction. The deep learning methods used to predict energy consumption are mainly Convolutional neural network(CNN) and Recurrent neural network (RNN), and compared with CNN, RNN considers the interdependence between different hidden states, it can input hidden states of t-1 step and raw features into t step. Kreider et al. used recurrent neural network to predict energy consumption for building heating and cooling. [24] However, RNN has the problems of gradient disappearance and gradient explosion. Additionally, RNN neither capture well periods in time series nor handle well missing values, and many real energy consumption data are periodic and contain missing values [5]. Based on the original RNN, Long short-term memory (LSTM) network [25] and Gated recurrent neural network (GRU) [26] are proposed to alleviate the problem of gradient disappearance by introducing gating mechanism. Jian Qi Wang construct a LSTM to predict the long-term energy consumption, the result shows it perform better than ARMA, ARFIMA, BPNN on the test data [27]. There are also some researches using the hybrid method of deep learning. Tae-Young Kim etal. propose a CNN-LSTM neural network to predict the residential energy consumption which can extract both spatial and temporal features [28]. Fath U Min UIIah etal. combine CNN and multi-layer bi-directional LSTM to predict the short term residential power energy frequency [29]. However, LSTM still have limited memory and it is hard to capture long term input and consider different importance of each time step. With the development of deep learning, Seq2Seq model and attention mechanism are proposed which are used in the area of translation. The combination of Seq2Seq and attention mechanism can better deal with the sequential data, extract the information of sequential data and represent it. However, few researchers have applied it to the energy consumption prediction. To address this gap, this study combine the Seq2Seq model and attention mechanism to improve the building energy consumption prediction.

III. METHODOLOGY

A. LSTM

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

LSTM is a recurrent neural network proposed by Hochreiter and Schmidhuber to solve the gradient disappearance in traditional recurrent neural network. Different from traditional recurrent neural network, it introduces the “gate mechanism”, which control the input and output for each memory cell. The “gate mechanism” include forget gate, input gate and output gate. Figure 1 shows the structure of the memory cell of LSTM. The forget gate determine the information discard from the state of the memory cell at the last moment. The input at time t and the output at time t-1 are used to calculate the forget gate. The formular of forget gate is

$$f_t = \sigma(w_f[h_{t-1}, x_t]) + b_f$$

The input gate determines the part of information that should be added to the cell. The formular of input gate is

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i$$

The candidate memory cell is

$$\hat{C}_t = \tanh\,(w_c[h_{t-1}, x_t] + b_c)$$

Then, the memory cell of time t constructed by the state of memory cell t-1 and the candidate memory cell is calculated as follows

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t$$

The output gate determines the part of information that be conveyed to the next memory cell. The output gate is calculated as follows

$$O_t = \sigma(W_o[h_{t-1}, x_t]) + b_o$$

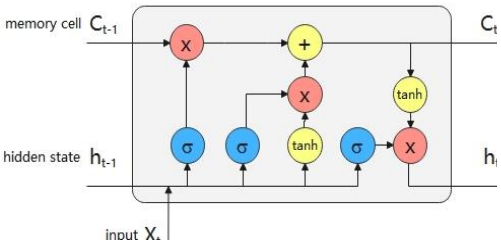

Fig. 1. Memory cell of LSTM

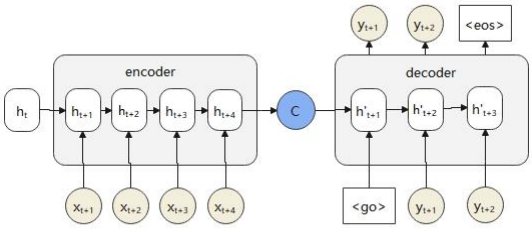

Fig. 2. Seq2seq model

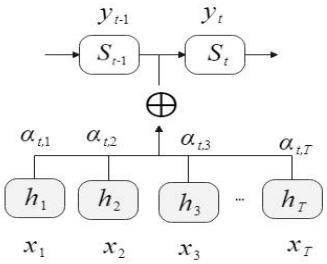

Fig. 3. Attention mechanism

The information of hidden layer at time t that will be input to the next memory cell is

$$h_t = O_t \cdot \tanh(C_t)$$

B. Seq2seq Model

Seq2seq model consists of encoder and decoder. Both encoder and decoder consist of a series of cells, usually LSTM cell or GRU (Gated Recurrent Unit) cell. The encoder extracts the features of input and encode it into a fixed length vector and the decoder generate the final output. Figure 2 shows the structure of Seq2seq model. Seq2seq model is usually used in the translation task. In this model, encoder encodes all input sequences into a unified semantic vector context, and then decodes by decoder. In the decoding process, the output of the previous time is continuously taken as the input of the next time, and the decoder circularly decodes until the stop character is output. In our study, we take the feature sequence of one week (168 hours) as input, and then decode the energy consumption sequence of a day (24 hours).

C. Attention mechanism

Attention mechanism is a mechanism to calculate the weighted sum of values of vector set according to vector query. Attention mechanism was first proposed in the field of visual image. Google Mind team used attention mechanism in RNN model for image classification [30]. Then, Bahdanau et al. used attention mechanism to translate and align simultaneously in machine translation task [31]. Currently, attention mechanism is widely used in various NLP tasks based on RNN / CNN and other neural network models. Figure 3 shows the structure of attention mechanism. The attention mechanism considers the importance of different time steps. e_{ti} is the attention score that shows the similarity of S_{t-1} (the hidden layer output at time t) and h_i (the hidden unit).

$$e_{ti} = f(S_{t-1}, h_i)$$

Then, softmax function is used to activate the attention score to obtain the weight α_{ti} of each hidden cell.

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_k \exp(e_{tk})}$$

The context vector is the weighted sum of h_i, which is calculated as follows

$$C_t = \sum_i \alpha_{ti} h_i$$

S_t is the hidden layer output at time t, Y_t is decoder output.

$$Y_t = f(C_t, S_t, Y_{t-1})$$

These formulars are used to build our model.

IV. Case Study

A. Data Preprocessing

The building consumption data is from an office building located in Shenzhen. The interval of data collection is 15 minutes. The energy consumption includes lighting socket power, air conditioning power, dynamic power and special power. We aggregate the subitem of energy consumption by hour and extract the index of hour in a day which is a kind of time feature. According to some studies, weather also affect the energy consumption. Therefore, we match the weather condition in Shenzhen and the energy consumption according to the time. Finally, total 9 features are selected to be variables and energy consumption is the output variables. Additionally, we processed the raw data as follows:

- There are some missing values of energy consumption in some periods and we fill them with the mean energy consumption in that day.
- In order to reduce the prediction errors, we use the MinMaxScaler to standardize the variables. The formular is as follows

$$x_{it} = \frac{x_{it} - \min(x_{it})}{\max(x_{it}) - \min(x_{it})}$$

- For the categorical variable, we encode it to the numerical variable. The wind direction is encoded in our study.

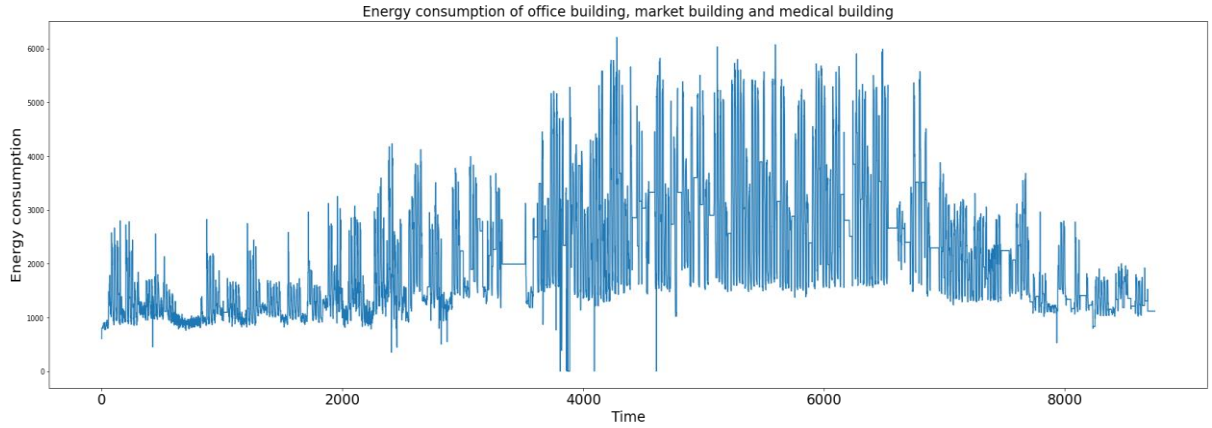

Fig. 4. The original data of energy consumption

B. *Trend Analysis and Correlation Analysis*

The data used in this study were collected for the whole year, Figure 4 shows the trend of energy consumption after data cleaning. It can be found that the energy consumption follows the periodic change in some periods. It's a time sequence with periodicity. The energy consumption in summer is generally higher than that in other seasons. In order to detect the relationship between variables, we do the correlation analysis using the Pearson correlation coefficient. Figure 5 shows the correlation of variables. Temperature and pressure have great correlation with energy consumption while the humidity has little correlation with energy consumption which is dropped.

C. *The Attention-based Seq2seq Model*

The core idea of attention-based seq2seq model is to combine the attention mechanism and seq2seq model to obtain higher prediction accuracy. In our basic seq2seq model, the encoder consists of two BiLSTM layers and the decoder consist of two BiLSTM layers and fully connected layer. Compared with LSTM, BiLSTM can capture both near and far position information. The encoder state and the information of the t-1 time step decoder output is the t-step input of the decoder. Then we add the attention mechanism between the encoder output and decoder output to get the final output. In first step, we calculate the similarity between the decoder output and encoder output to get the attention score of the encoder output. In second step, we activate the attention score using the softmax function to obtain the weight of encoder output.

$x^1, x^2, \dots x^T$ represents the encoder output. $\alpha_t^1, \alpha_t^2, \dots \alpha_t^n$ represents the attention weight of nth dimension of encoder output at time t. The importance of encoder output can be expressed as follows:

$$X_t = (x_t^1 * \alpha_t^1, x_t^2 * \alpha_t^2, \dots, x_t^n * \alpha_t^n)$$

Then, X_t is concatenated with the decoder output Y_t, the concatenated vector is

$$[X_t, Y_t]$$

The final output is obtained through a fully connected neural network.

$$Y_t' = f([X_t, Y_t])$$

Figure 6 shows the structure of the A-seq2seq model in our study.

The input sequence is 168 timesteps (one week) and the output sequence is 24 timesteps. As encoder inputs, temperature, pressure, wind direction, wind speed, horizontal

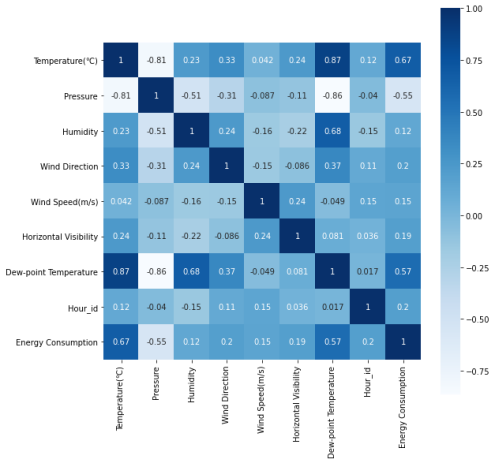

Fig. 5. Correlation analysis

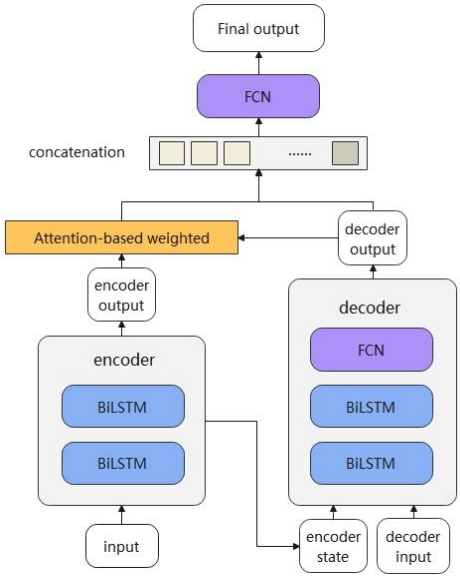

Fig. 6. Attention based seq2seq model architecture

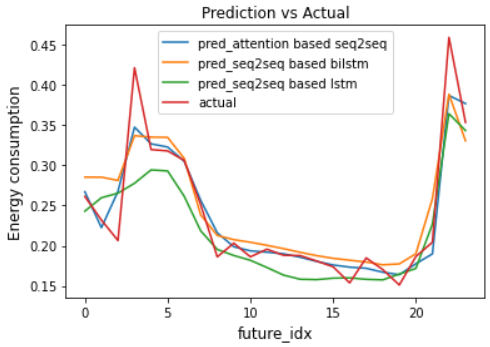

Fig. 7. Prediction result for the next 24 hours

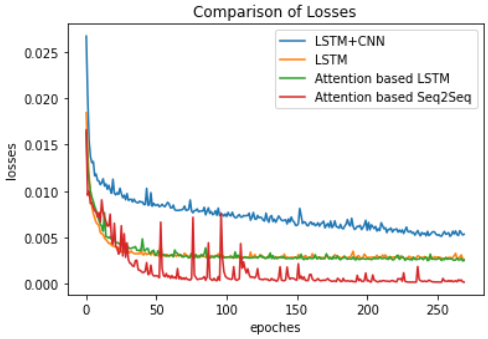

Fig. 8. Comparison of losses

visibility, dew-point temperature, hour, previous energy consumption are used. As decoder inputs, 1-step ahead energy consumption data are used.

We compare the performance of different model structure when developing our model. Figure 7 shows the prediction of energy consumption in the next 24 hours. The result shows attention based seq2seq using BiLSTM layers perform better.

Furthermore, we set the parameters, the number of hidden units in each layer is 60, batch size is 32, epochs is 300. We adopt the mean squared error as the loss function.

V. Results and Discussion

We use the data of the first half of the year as the training set, and use the data of the last two months to test the model. In the training process of attention based seq2seq model, the convergence condition curve of loss with the number of iterations is shown in Figure 8. Figure 8 shows the convergence process of different models on the verification set. The red line represents the curve of the attention based seq2seq model on the test set. With the increase of iteration times, loss of proposed model decreases rapidly. After 100 iterations, attention based seq2seq model proposed by us tends to be stable, and the loss of it is less than that of other models (CNN-LSTM, LSTM, attention-based LSTM).

We compare the prediction performance of LSTM, attention-based LSTM and attention based seq2seq for forecasting energy consumption in the next month. Figure 9 shows the prediction of the models. It can be found that the attention based seq2seq model perform best in the prediction, although there will be a small amount of deviation over time

We also evaluate and compare the performance of attention based seq2seq and SVM, MLP, LSTM, CNN-LSTM, attention-based LSTM, attention based seq2seq from four evaluation indexes: root mean square error (RMSE), mean square error (MSE), mean absolute percentage error (MAE) and symmetric mean absolute percentage error (SMAPE). The table shows the evaluation results of the prediction of each model.

As shown in the table 1, each model has good prediction results. The MSE (0.0011), RMSE (0.0332), MAE (0.0177) and smape (0.4107) of attention based seq2seq model were the lowest, and the prediction effect was the highest. In the short term, the prediction performance of attention based seq2seq model is significantly better than LSTM model and some traditional machine learning models, and the prediction effect of LSTM model with attention mechanism is better than that of traditional LSTM model, which indicates that introducing attention mechanism to adjust weight in LSTM model can improve the prediction performance. Attention

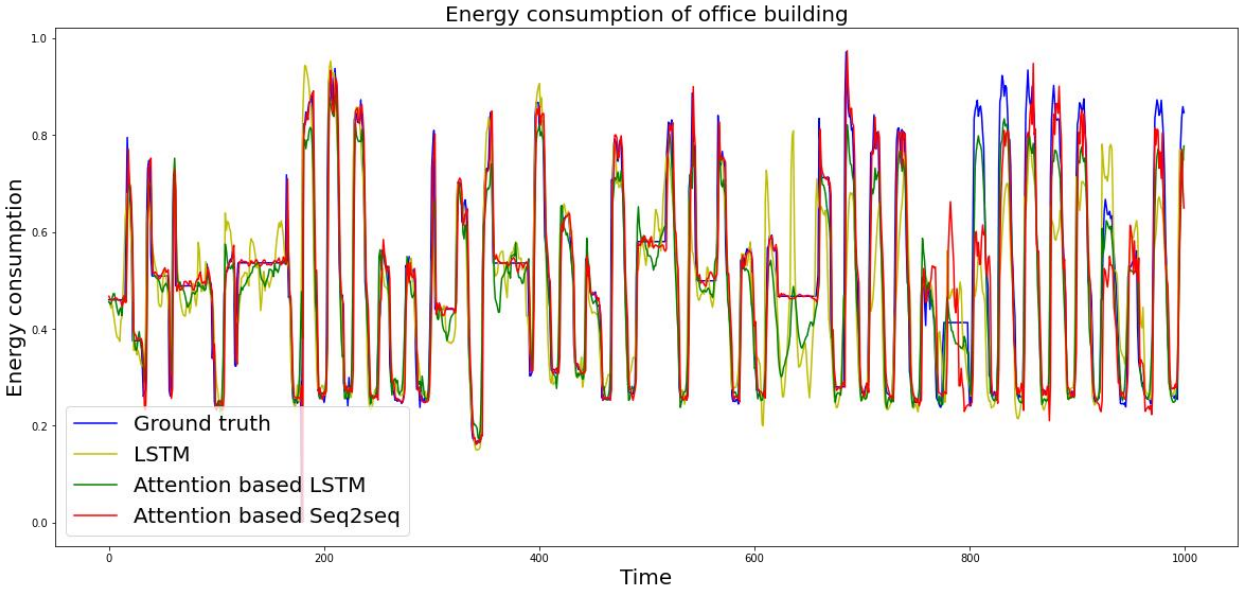

Fig. 9. Prediction result of LSTM, attention based LSTM, attention based seq2seq

TABLE I. PERFORMANCE OF MODELS FOR BUILDING ENERGY CONSUMPTION PREDICTION

Model	MSE	RMSE	MAE	SMAPE %
SVM	0.0262	0.1619	0.1261	4.0302
MLP	0.0236	0.1537	0.1186	3.8621
LSTM	0.0192	0.1386	0.1096	2.3969
CNN-LSTM	0.0203	0.1426	0.1123	3.8936
attention+LSTM	0.0163	0.1278	0.0964	3.6419
attention+seq2seq	**0.0011**	**0.0332**	**0.0177**	**0.4107**

mechanism changes the efficiency of model using information. By adding attention mechanism into the seq2seq model, we can selectively pay attention to the vector information of encoder output which changes dynamically with time, and combine it with the output of decoder to map to the energy consumption of the final output. In the first mock exam, this model improves the performance and efficiency of short-term building energy consumption. Seq2seq model is a sequence to sequence prediction. When using fixed time step input, this model can be used to predict the energy consumption of future multiple time steps at one time, not just the energy consumption of a certain step after the input time step. It enriches the application of artificial intelligence in building energy consumption prediction. In the future, we will further consider how to improve the stability of this model applied in a longer period, and further explore its application in long-term building energy consumption prediction.

The data used in this study is the building energy consumption of 15 minutes a year gathered to 1 hour in office buildings, and we will experiment on more types of building data in the future.

VI. CONCLUSION

Building energy consumption prediction plays an important role in the areas of energy system management, energy supply. This study proposes an attention based seq2seq model to forecasting the building energy consumption prediction. The results show that the performance of proposed model is better than traditional models in predicting short term energy consumption. The proposed model can better fit the nonlinearity of energy consumption.

In addition, the proposed model is more efficient. Compared with LSTM, it can predict energy consumption of multiple time steps more accurately at one time while the input has equal time steps.

We introduce the mechanism to the model which can dynamically deal with the input data by assigning different weights to the encoder output, thus improving the performance of the model. Compared with SVM, MLP, LSTM, CNN-LSTM, attention-based LSTM, the proposed model have less loss and lower MSE, RMSE, MAE, MAPE.

Although the prediction effect of the proposed model is better in the short term, we expect it to have a stable and excellent performance in the long term. We will study and improve the model in the future, and carry out experiments on more types of buildings to improve the generalization ability of the model.

REFERENCES

[1] N. Somu, R. Gauthama, K. J. R. Ramamritham, and S. E. Reviews, "A deep learning framework for building energy consumption forecast," vol. 137, p. 110591, 2021.

[2] F. A. Xi, A. Gg, B. Gl, C. A. Liang, A. Wl, and P. A. J. E. Pei, "A hybrid deep transfer learning strategy for short term cross-building energy prediction," vol. 215, 2020.

[3] K. Amasyali, N. M. J. R. El-Gohary, and S. E. Reviews, "A review of data-driven building energy consumption prediction studies," vol. 81, no. pt.1, pp. 1192-1205, 2018.

[4] H. X. ~Zhao, F. M. J. Renewable, and S. E. Reviews, "A review on the prediction of building energy consumption," vol. 16, no. 6, pp. 3586-3592, 2012.

[5] J. Liu, Y. Chen, J. Zhan, and F. S. J. I. T. o. V. Technology, "An On-line Energy Management Strategy based on Trip Condition Prediction for Commuter Plug-in Hybrid Electric Vehicles," pp. 1-1, 2018.

[6] B. Han, D. Zhang, and Y. Tao, "Energy consumption analysis and energy management strategy for sensor node," in *2008 International Conference on Information and Automation*, 2008.

[7] M. Molina-Solana, M. Ros, M. Dolores Ruiz, J. Gomez-Romero, M. J. J. R. Martin-Bautista, and S. E. Reviews, "Data science for building energy management: A review," vol. 70, no. APR., pp. 598-609, 2017.

[8] Q. Yao, D. Song, H. Chen, C. Wei, and G. W. Cottrell, "A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[9] J. F. Torres, A. Fernández, A. Troncoso, and F. M.-Á. J. L. N. i. C. Science, "Deep Learning-Based Approach for Time Series Forecasting with Application to Electricity Load," vol. In press, no. 2, pp. 1-10, 2017.

[10] V. L. Erickson, M. A. Carreira-Perpinan, and A. E. J. A. T. o. S. N. Cerpa, "Occupancy Modeling and Prediction for Building Energy Management," vol. 10, no. 3, pp. 1-28, 2014.

[11] W. Xu, H. Hu, and W. J. I. A. Yang, "Energy Time Series Forecasting Based on Empirical Mode Decomposition and FRBF-AR Model," vol. PP, pp. 1-1, 2019.

[12] C. Yuan, S. Liu, and Z. J. E. Fang, "Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model," vol. 100, no. apr.1, pp. 384-390, 2016.

[13] M. J. I. J. o. E. R. Mohandes, "Support vector machines for short‐term electrical load forecasting," vol. 26, no. 4, pp. 335-345, 2002.

[14] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. The Top Ten Algorithms in Data Mining, 2009.

[15] T. Wang and L. J. I. Qin, "Application of SVM based on rough set in smart grid energy-saving prediction," 2010.

[16] S. Karatasou, M. Santamouris, V. J. E. Geros, and Buildings, "Modeling and predicting building's energy use with artificial neural networks: Methods and results," vol. 38, no. 8, pp. 949-958, 2006.

[17] S. E. Reviews, "Energy models for demand forecasting—A review," 2012.

[18] C. W. Yan and J. J. I. Yao, "Application of ANN for the prediction of building energy consumption at different climate zones with HDD and CDD," 2010.

[19] None, "06/00462 Applying support vector machines to predict building energy consumption in tropical region: Dong, B. et al. Energy and Buildings, 2005, 37, (5), 545–553," vol. 47, no. 1, pp. 63-0, 2006.

[20] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. J. A. E. Mochida, "Applying support vector machine to predict hourly cooling load in the building," vol. 86, no. 10, pp. 2249-2256, 2009.

[21] K. Li, H. Su, C. J. E. Jian, and Buildings, "Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study," vol. 43, no. 10, pp. 2893-2899, 2011.

[22] S. Fan and L. J. I. T. o. P. S. Chen, "Short-Term Load Forecasting Based on an Adaptive Hybrid Method," vol. 21, no. 1, pp. 392-401, 2006.

[23] C. Fan, Y. Sun, Y. Zhao, M. Song, and J. J. A. E. Wang, "Deep learning-based feature engineering methods for improved building energy prediction," vol. 240, no. APR.15, pp. 35-45, 2019.

[24] J. F. Kreider, P. Curtiss, R. Dodier, M. Krarti, D. E. Claridge, and J. S. Haberl, "Recurrent neural networks for building energy use prediction and system identification -- A progress report," 1995.

[25] S. Hochreiter and J. J. N. C. Schmidhuber, "Long Short-Term Memory," vol. 9, no. 8, pp. 1735-1780, 1997.

[26] B. Wang and J. J. E. Wang, "Energy futures price prediction and evaluation model with deep bidirectional gated recurrent unit neural network and RIF-based algorithm," vol. 216, 2021.

[27] J. Q. Wang, Y. Du, and J. J. E. Wang, "LSTM based long-term energy consumption prediction with periodicity," vol. 197, pp. 117197-, 2020.

[28] T. Y. Kim and S. B. J. E. Cho, "Predicting Residential Energy Consumption using CNN-LSTM Neural Networks," vol. 182, no. SEP.1, pp. 72-81, 2019.

[29] F. Ullah, A. Ullah, I. U. Haq, S. Rho, and S. W. J. I. A. Baik, "Short-Term Prediction of Residential Power Energy Consumption via CNN and Multi-Layer Bi-Directional LSTM Networks," vol. 8, pp. 123369-123380, 2020.

[30] V. Mnih, N. Heess, A. Graves, and K. J. A. i. N. I. P. S. Kavukcuoglu, "Recurrent Models of Visual Attention," vol. 3, 2014

[31] D. Bahdanau, K. Cho, and Y. J. C. S. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2014.