Power Load Forecasting Using BiLSTM-Attention

Power Load Forecasting Using BiLSTM-Attention

Jie Du, Yingying Cheng, Quan Zhou, Jiaming Zhang, Xiaoyong Zhang, Gang Li

State Grid Chongqing Electric Power CO. Chongqing Electric Power Research Insititue, YuBei district, Chongqing 401121, China

Abstract. With the development of big data and artificial intelligence, the applications of smart grid have received extensive attention. Specifically, accurate power system load forecasting plays an important role in the safety and stability of the power system production scheduling process. Due to the limitations of traditional load forecasting methods in dealing with large scale nonlinear time series data, in this paper, we proposed an Attention-BiLSTM (Attention based Bidirectional Long Short-Term Memory, Attention-BiLSTM) network to do the accurate short-term power load forecasting. This model is based on BiLSTM recurrent neural network which has high robustness on time series data modeling, and the Attention mechanism which can highlight the key features playing key roles in load forecasting in input data. The verification experiments with real data in a certain area show that the proposed model outperforms other models in terms of prediction accuracy and algorithm robustness.

Keywords: Load Forecasting; Neural Networks; BiLSTM; Attention Mechanism.

1. Introduction

Electrical energy is an instantaneous energy source that cannot be stored in large quantities. The production and consumption of electrical energy need to be carried out simultaneously [1]. Therefore, in order to ensure the economic and reliable power system, it is necessary to develop a reasonable production plan and scheduling plan. Power load forecasting can reflect the trend of future power consumption to a certain extent based on historical data. The power production department can formulate production plans and dispatch plans based on the results of power forecasting, thereby effectively improving the efficiency of the power system.

Power load forecasting can be divided into ultra-short-term forecasts, short-term forecasts, medium- and long-term forecasts, and holiday forecasts by time range [2][3]. Short-term forecasts or ultra-short-term forecasts are mainly used to guide the power sector to arrange power generation plans reasonably. Traditional short-term power load forecasting methods are mainly divided into two categories. The first category is based on traditional methods such as regression analysis, trend extrapolation, expert system methods, and time series methods. The literature [4] uses historical data on population, GDP, and total social electricity consumption to predict power load based on multiple linear regression analysis. The paper [5] proposes a short-term load forecasting method combining wavelet transform and cumulative autoregressive moving average (ARIMA) model. Due to the obvious volatility and randomness in the short-term load data, the traditional method is mainly for the linear model. The linear model is not flexible enough for short-term prediction, and the prediction ability for nonlinear relationships is insufficient, so it cannot be accurately predicted. The other type is a method that includes a support vector machine and an artificial neural network model. The literature [6] applies the BP neural network model to power load forecasting. The literature [2] uses the Support Vector Machine (SVM) to predict the power load. This type of method achieves better results when dealing with nonlinear relationships and improves the problem of limited accuracy. Although the BP neural network has a simple structure, it has poor learning ability and is easy to fall into local optimum [7]. The support vector machine model can be applied to small samples, but it takes a lot of machine memory and runtime when processing large amounts of data [8]. The above two methods mainly consider the relationship between the electrical load and its influencing factors, and ignore the relationship between the serial data of the continuous load samples. Power load data is typical time series data. The data is not only nonlinear, but also correlated and continuous. The load at the current moment is not only related to external factors, but also related to past input characteristics, and future input characteristics can also reflect the current load characteristics to a certain extent. And the short-term power load trend is also cyclical [9], That is, the power load curve is similar in different days, different days of the same day or different statutory holidays. Therefore, the timing of the load data should be considered in the prediction.

With the development of deep learning, big data analysis and high-performance computing technology, the short-term power load forecasting methods commonly used at present are mainly based on the deep cycle neural network model. The most representative one is the LSTM (long short time memory, LSTM), It introduces input gates, forgetting gates and output gates, effectively overcome the problem of the disappearance of the gradient of the traditional cyclic neural network in the long-term sequence. Thereby effectively learning timing information in time series data [11], And it is widely used in machine vision [12], text analysis [13], fault diagnosis [14], and so on. In addition, LSTM's parameter sharing mechanism reduces the amount of computation required for network training, allowing it to be used to process relatively large data. The literature [15][16][17] uses the LSTM model and predicts the electrical load based on historical data of the electrical load, but the LSTM-based approach does not take into account the impact of different time dimensions in the input sequence on the load. The Bi-LSTM model that has emerged in recent years has the advantage of acquiring context information of time series data [18]. And attentionalism (Attention) can assign different weights to different hidden units of the neural network, make the hidden layer pay more attention to the key information in the sequence data [19]. Therefore, this paper combines the Bi-LSTM model with the attention mechanism to mine the context information in time series data and the contribution of different time dimensions to the prediction results based on historical power data, thus further improving the accuracy of power load forecasting [20].

2. Establishment of BiLSTM- Attention Neural Network Model

2.1. LSTM Neural Network Model

The LSTM neural network evolved from the RNN neural network [21]. Compared with the traditional neural network, the RNN neural network not only interconnects the hidden layers, but also the currently accepted input is related to the previous timing input, which makes the neural network better in processing timing-related inputs.

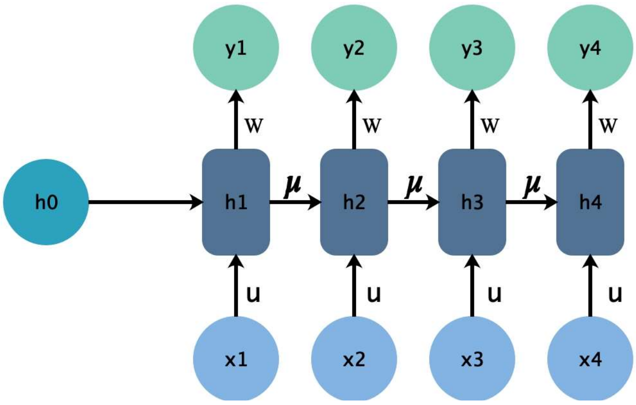

Figure 1. RRNN neural network structure

As shown in FIG. 1, U, V, and W are the input layer to the hidden layer, the hidden layer to the hidden layer, and the weight of the hidden layer to the output layer, respectively. X, O are the input and output of the neural network, respectively, and S is the current state of the hidden layer. Unlike traditional neural networks, the U, V, and W parameters of each layer of the RNN are shared, which reduces the parameters that the network needs to learn and shortens the training time. However, when the time interval is large, the gradient descent method will be reduced [22], The influence of the "memory" of the distant moment on the input will be attenuated very little, causing the problem of gradient demise. The LSTM proposed by Hochreiter is an improvement on RNN [23]. LSTM adds a cell state to solve the problem of gradient demise, and its structure is shown in Figure 2.

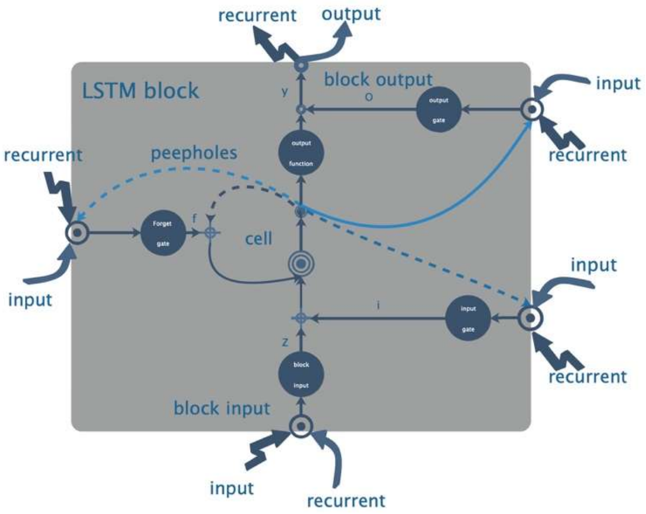

Figure 2. LSTM cyclic neural network structure

The LSTM uses gates to control access to cellular status. Each LSTM memory unit contains three control gates, namely Input Gate, Forget Gate, and Output Gate, which are responsible for controlling information input, controlling cell history state information retention and control information output. The forgotten gate calculation formula is:

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

The input gate formula is:

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

The current state of the cell is the formula:

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

The memory unit status value is:

$$C_t = f_t C_{(t-1)} + i_t * \widetilde{C}_t$$

The output gate output is:

$$o_t = \sigma(w_o * [h_{(t-1)}, x_t] + b_o)$$

The hidden layer output is:

$$h_t = o_t * tanh\ (c_t)$$

The w and b in the equations are the corresponding weight coefficient matrix and the offset term, respectively. σ is the sigmoid activation function, $tanh$ is a hyperbolic tangent activation function. In one calculation, the input to the LSTM unit consists of three parts: Last moment state memory $C_{(t-1)}$, Current input information x_t and the last moment hidden layer output $h_{(t-1)}$. The output is divided into 2 parts, Current time hidden layer output h_t And memory at the moment C_t.

2.2. Bi-LSTM Neutral Network

There is only a backward-propagating LSTM in the LSTM, which allows it to obtain only the previous information in the sequence data while processing the data. Graves [24] proposed a Bi-LSTM based on LSTM. Different from one-way LSTM, Bi-LSTM adds a layer of reverse LSTM. The reverse LSTM reverses the data and the hidden layer synthesizes the forward and reverse information so that cells in the network can simultaneously obtain context information. The structure of the bidirectional LSTM is shown in Figure 3:

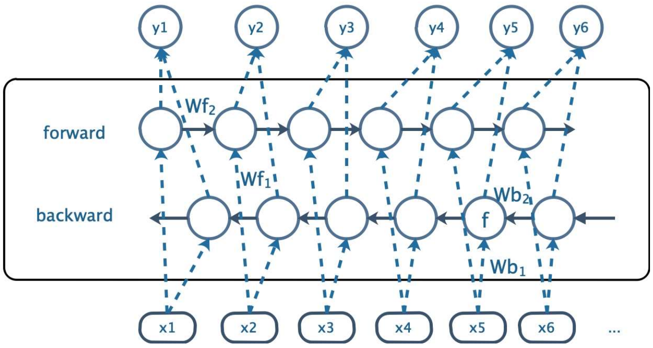

Figure 3. Bi-LSTM structure diagram

The reverse layer LSTM is calculated in a similar manner to the forward LSTM, except that the direction is reversed to obtain the subsequent time information. The Bi-LSTM network calculation formula is as follows:

$$h_f = f(w_{f1}x_t + w_{f2}h_{t-1})$$

$$h_b = f(w_{b1}x_t + w_{b2}h_{t+1})$$

h_f is the forward LSTM network output, h_b is the reverse LSTM network output. The final output of the hidden layer is:

$$y_i = g(w_{o1} * h_f + w_{o2} * h_b)$$

2.3. Attention mechanism

The Attention mechanism is a probability weighting mechanism that mimics the attention of the human brain [19], When the human brain observes things, it will focus on specific places and ignore other places. The Attention mechanism increases the accuracy of the model by highlighting more important factors by assigning different probability weights to the inputs. Therefore, the introduction of the Attention mechanism into the Bi-LSTM model can predict the load. Attention structure as shown:

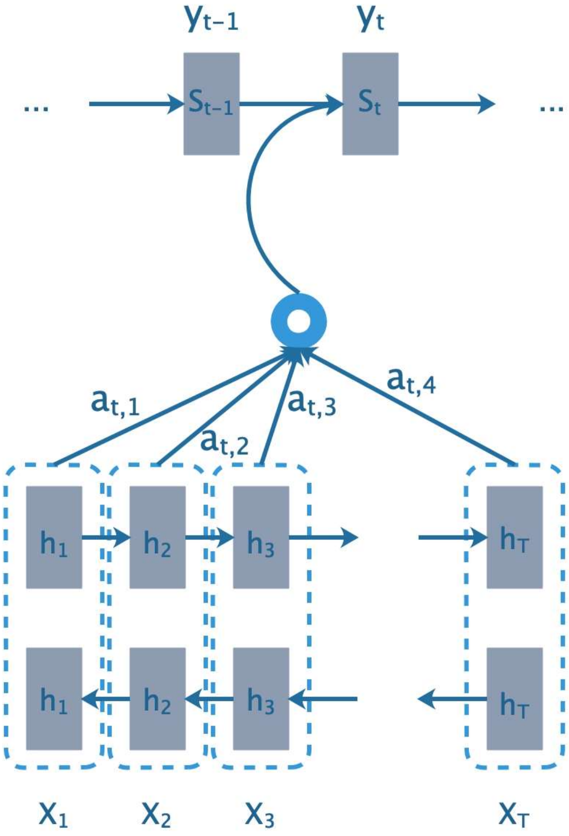

Figure 4. Attention mechanism structure diagram

In Figure 4, the input sequence values are x_1 to x_k, The hidden layer status values are h_1 to h_k, a_{ki} is the attention weight of other hidden layers for the current input, The calculation method is:

$$a_{ki} = \frac{\exp(e_{ki})}{\sum_{j=i}^{Tx} \exp(e_{kj})}$$

$$e_{ki} = vtanh(Wh_k + Uh_i + b)$$

$$C = \sum_{i=1}^{Tx} a_{ki}\, h_i$$

h'_k is the final hidden layer state value of the final output, calculated as:

$$h'_k = H(C, h_k, x_k)$$

2.4. BiLSTM-Attention Model

BiLSTM-Attention Model includes an input vector, a forward LSTM hidden layer, a reverse LSTM hidden layer, an Attention layer, a fully connected layer, and an output layer, and the structure is as shown in FIG 5:

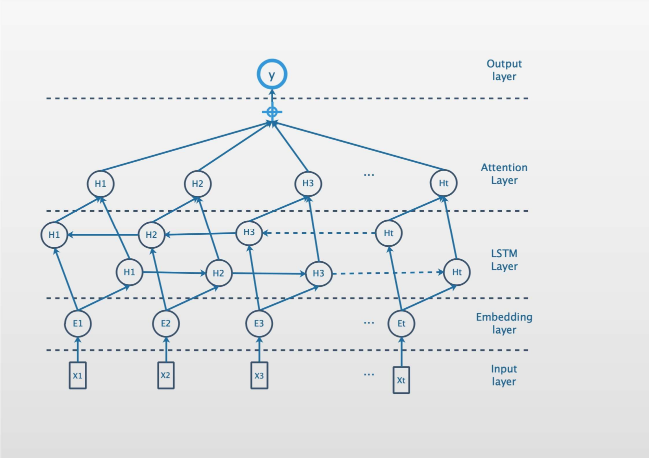

Figure 5. BiLSTM-Attention Model structure diagram

After receiving the input, the model passes the sequence data to the forward LSTM hidden layer and the reverse LSTM hidden layer, and the two combine to output the processed vector. The Attention layer takes the data processed by the LSTM layer as input, calculates the weight vector, and then combines the weight vector with the shallow output to obtain a new vector input into the fully connected layer. Finally, the fully connected layer calculates the predicted value.

The LSTM layer can remember important information and forget important information. In theory, the more layers, the better the model fits the nonlinear data. However, too many layers can lead to overfitting and consume a lot of time, so this model sets up a 2-layer LSTM. Generally, the number of neurons is 2, and the number of the first layer is too small, which makes the model unable to learn the regularity. Therefore, the number of neurons in the first layer is finally set to 128. Too many neurons will cause the fully connected layer to train a large number of parameters, thus compressing the number of second layer LSTM neurons, the number of second layer neurons is 64. The purpose of the Attention layer is to highlight the influence of key features on the sequence by assigning the feature weights learned by the model to the input vector of the next time step. The final data is input to the fully connected layer, and after being processed by the virtual function of the fully connected layer, the predicted load value is obtained.

3. Case Analysis and Experimental Results

3.1. Data Preprocessing

Data standardization (normalization) processing is a basic work of data mining. Different evaluation indicators often have different dimensions and dimension units. Such situations will affect the results of data analysis, in order to eliminate the dimension between indicators. Impact, data standardization needs to be done to resolve the comparability of data metrics. After the original data is processed by data standardization, each indicator is in the same order of magnitude, which is suitable for comprehensive comparative evaluation. Therefore, before training and verification, the data is processed first by the method of maximum and minimum normalization.

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

The data is mapped to the [-1, 1] interval after normalization, to reduce the impact of the dimension on the results.

3.2. Periodic Verification Of Data

This paper combines the actual power data to verify the proposed prediction method. The data comes from the power load data of two companies in a certain day. The day in the data is divided into 48 time points. In order to make the experiment more rigorous, the periodicity of the data is first verified. Figures 6 and 7 show the charge trend plots of the two companies in a week.

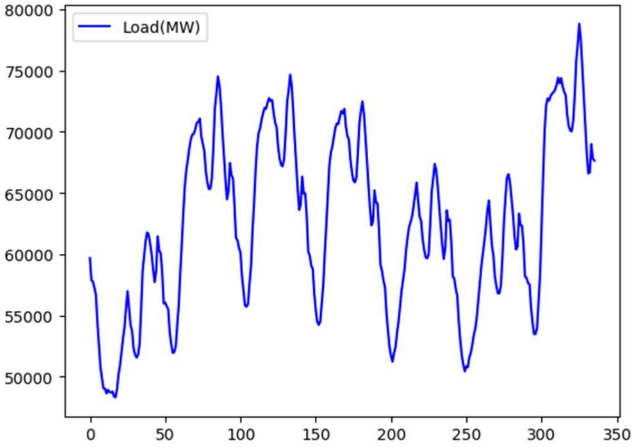

Figure 6. Dataset 1 - power trend graph within one week

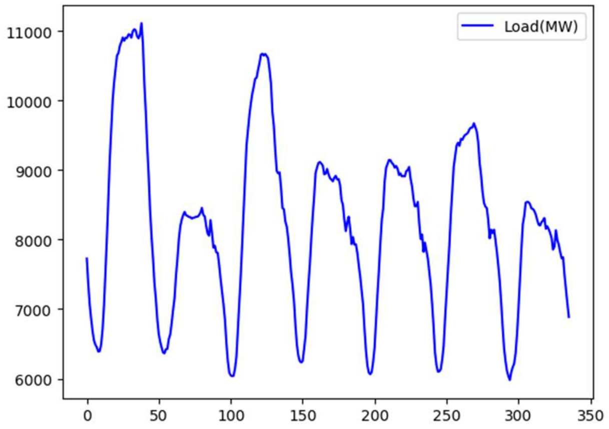

Figure 7. Dataset 2 - power trend graph within one week

It can be seen from the figure that the data of the two data sets fluctuate according to a certain frequency, and the whole has periodicity, and it is reasonable to use LSTM.

3.3. Error indicator

The error indicators used in this paper are: MAPE (mean absolute percentage error) The formula is as follows:

$$e_{MAPE} = \frac{\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}\,|}{y_i}}{n}$$

The real value is y_i, the predicted value is $widehaty_i$, and the predicted total is n. The smaller the value of MAPE, the better the accuracy of the prediction model.

3.4. Comparison of different model effects

The experimental environment of this experiment is a GPU server, the CPU is Intel Xeon E5-2680 v4, the memory is 32GB, the GPU is Nvida Titan XP, and the software platform is Google Tensorflow. The following is the experimental comparison of the LSTM, Bi-LSTM, Attention-LSTM, and Attention-BiLSTM prediction models. All the network structures used in this paper are two-layer structure, the number of neurons in each layer is set to 128, 64, the activation function selects $tanh$, the fully connected layer activation function is $Relu$, and the steps of LSTM are all 5. Both dataset inputs are historical load and time. One day is divided into 48 moments, and all models output one time prediction value each time. This article divides 80% of the total data in each dataset into training sets. Tables 1 and 2 list the e_{MAPE} values for each of Data Set 1 and Data Set 2 models, respectively.

Table 1. Error comparison of different models

Model	A-BiLSTM	A-LSTM	BiLSTM	LSTM
$e_{MAPE}\%$	1.1526	1.4708	1.4087	1.7795

Table 2. Error comparison of different models

Model	A-BiLSTM	A-LSTM	BiLSTM	LSTM
$e_{MAPE}\%$	1.1496	1.1812	1.3156	1.5677

As can be seen from Table 1, the total value of e_{MAPE} of the Bi-LSTM model based on the Attention mechanism is 1.1526%, which is reduced by 0.3182%, 0.2561%, 0.6269% compared to Attention-LSTM, Bi-LSTM and LSTM, respectively. From the results, the results of the Attention-Bi-LSTM model are significantly better than other models. The two-way LSTM model is better than the one-way LSTM model. The model with the Attention mechanism is better than the model without the Attention mechanism. Bidirectional LSTM is better able to discover features in a sequence. And the Attention mechanism further improves the prediction effect. The Bi-LSTM model based on the attention mechanism in Table 2 predicts the same results.

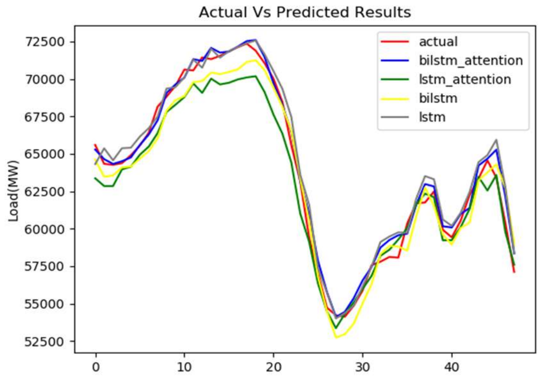

Figure 8. Data set 1 load prediction curve

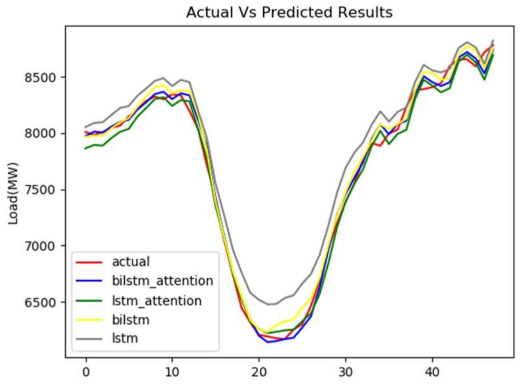

Figure 9. Data set 2 load prediction curve.

Table 3. Data set 1 MAPE value of each model in one day

	A-BiLSTM	**A-LSTM**	**BiLSTM**	**LSTM**
0	0.4641	3.5077	1.5222	1.9832
1	0.4791	2.3599	1.3561	1.5851
2	0.0950	2.2519	1.1063	0.4673
3	0.1971	0.6493	0.4469	1.5218
4	0.2269	1.1982	1.1773	0.7675
5	0.0585	0.9400	1.3407	0.9017
6	0.1728	1.4073	1.8775	0.4617
7	1.3562	2.6607	3.2000	0.9831
8	0.3562	1.5021	1.3918	0.7696
9	0.1277	1.8543	1.3912	0.1288
10	0.7670	2.6667	2.5783	0.7272
11	1.0530	1.2443	1.1216	1.0103
12	0.3324	3.4267	2.2683	0.9674
13	1.0410	1.8596	1.2688	0.9564
14	0.3000	2.7422	1.7391	0.1673
15	0.0196	3.0077	1.9575	0.0131
16	0.0535	3.0793	2.1032	0.0246
17	0.2255	3.2268	1.7327	0.0421
18	0.9776	2.4280	0.8918	0.9361
19	0.5678	2.7983	0.6591	0.8231
20	0.4186	3.5408	0.8599	0.7206
21	0.3011	3.1956	0.4613	1.2530
22	1.3368	1.7839	1.4496	2.7935
23	0.0270	3.5153	0.2632	0.7014
24	2.5336	0.4697	1.8752	3.5527
25	1.4449	1.1986	0.3731	0.6183
26	1.8513	0.4528	0.4473	1.7732
27	0.2141	1.6190	2.8324	0.4018
28	0.5925	0.3580	2.2017	0.4503
29	0.9017	0.4070	2.2168	0.0034
30	1.1886	0.1815	1.5635	0.4849
31	0.2160	1.2388	2.2455	0.1744
32	1.6053	0.6688	0.8650	2.2239
33	1.9109	0.8219	1.2556	2.2890
34	2.5104	2.0402	1.2594	2.8242
35	1.1261	0.5034	3.0848	1.0254
36	0.1896	0.1166	1.4709	0.4691
37	1.9558	0.9480	1.6099	2.7798
38	0.5361	0.6136	1.5601	1.2853
39	0.3970	1.1879	0.5437	1.1315
40	1.0929	0.3087	0.8119	1.2686
41	0.7327	0.7213	0.8261	0.8545
42	1.4145	1.5129	3.0541	0.3176
43	1.3442	0.1770	0.2071	1.6779
44	0.1801	3.2128	1.2527	0.5495
45	2.8649	0.2809	1.3977	3.8372
46	2.8650	1.0003	3.6762	3.8224
47	2.1229	0.8289	3.1314	2.1366
AVG	0.8905	1.6191	1.5401	1.1804

4. Conclusion

In this paper, a Bi-LSTM short-term power load forecasting model based on Attention mechanism is designed. Before the model is verified, it is first analyzed that the data has periodicity, and the current time data is affected by past and future data. The data was then normalized and the criteria for evaluating the validity of the model were developed. Finally, the data is brought into the model calculation verification. It is verified that the two-way network and Attention mechanism have a positive effect on the accuracy of power load forecasting.

References

[1] Powell, Lynn. Power system load flow analysis[J]. McGraw-Hill, 2005(1).

[2] Niu, Wang D A, Wu Y A, etal. Power load forecasting using support vector machine and ant colony optimization[J]. Expert Systems with Applications, 2p1p, 37(3):2531-2539.

[3] Almeshaiei, Soltan E A, Hassan. A methodology for electric power load forecasting[J]. Alexandria Engineering Journal, 2011, 50(2):137-144.

[4] Bakirtzis, Petridis A A, Kiartzis V A, etal. A neural network short term load forecasting model for the Greek power system[J]. IEEE Transactions on power systems, 1996, 11(2):858-863.

[5] Zhang, Peter G. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003(50):159-175.

[6] Xiao, Ye Z A, Zhong S A, etal. BP neural network with rough set for short term load forecasting[J]. Expert Systems with Applications, 2009, 36(1):273-279.

[7] Hsu, Chen C A, ChiaYon. Regional load forecasting in Taiwan----applications of artificial neural networks[J]. Energy conversion and Management, 2003, 44(12):1941-1949.

[8] Hsu, Chang C A, Lin C A, etal. A practical guide to support vector classification[J]. -, 2003(-).

[9] Li, Guo H A, Li S A, etal. A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm[J]. Knowledge-Based Systems, 2013(37):378-387.

[10] Concordia, Ihara C A, Susumu. Load representation in power system stability studies[J]. IEEE transactions on power apparatus and systems, 1982(4):969-977.

[11] Gers, Eck F A A, Schmidhuber D A, etal. Applying LSTM to time series predictable through time-window approaches[J]. Neural Nets WIRN Vietri-01, 2002(1).

[12] Byeon, Breuel W A, Raue T M A, etal. Scene labeling with lstm recurrent neural networks[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015(-):3547-3555.

[13] Wang, Yu J A, Lai L A, etal. Dimensional sentiment analysis using a regional CNN-LSTM model[J]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016(-):225-230.

[14] Yuan, Wu M A, Lin Y A, etal. ault diagnosis and remaining useful life estimation of aero engine using LSTM neural network[J]. 2016 IEEE International Conference on Aircraft Utility Systems (AUS), 2016(-):135-140.

[15] Kong, Dong W A, Jia Z Y A, etal. Short-term residential load forecasting based on LSTM recurrent neural network[J]. IEEE Transactions on Smart Grid, 2017, 10(1):841-851.

[16] Mauch, Yang L A, Bin. A new approach for supervised power disaggregation by using a deep recurrent LSTM network[J]. 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015(-):63-67.

[17] Zheng, Xu J A, Zhang C A, etal. Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network[J]. 2017 51st Annual Conference on Information Sciences and Systems (CISS), 2017(-):1-6.

[18] Zeng, Yang Y A, Feng H A, etal. A convolution BiLSTM neural network model for Chinese event extraction[J]. Natural Language Understanding and Intelligent Applications, 2016(-):275-287.

[19] Wang, Huang Y A, Zhao M A, etal. Attention-based LSTM for aspect-level sentiment classification[J]. Proceedings of the 2016 conference on empirical methods in natural language

processing, 2016(-):606-615.

[20] Bin, Yang Y A, Shen Y A, etal. Describing video with attention-based bidirectional LSTM[J]. IEEE transactions on cybernetics, 2018, 7(49):2631-2641.

[21] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Eleventh annual conference of the international speech communication association. 2010.

[22] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(02): 107-116.

[23] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[24] Wang, Y., Huang, M., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. *In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 606-615).*