Short-term Load Forecasting for Distribution Substations Based on Residual Neural Networks and Long Short-Term Memory Neural Networks with Attention Mechanism

Hailong Li1*, Pan Zhang2 and Chunhui Li1
1 State Grid Tianjin Binhai Electric Power Supply Company, Tianjin, 300450, China
2 Electric Power Research Institute of State Grid Tianjin Electric Power Supply Company, Tianjin, 300201, China
*Corresponding author’s e-mail: hailong.li@tj.sgcc.com.cn

Abstract. Electric load at distribution substation level has strong volatility due to various customer characteristics, making it challenging for short-term load forecasting. This paper proposes a novel short-term load forecasting method for distribution substations based on combined residual neutral networks (ResNet) and Long Short-Term Memory (LSTM) neutral networks with attention mechanism (ResNet-LSTM-Attention) where the advantages of ResNet and LSTM are combined together. A 34-layer ResNet is built to extract the latent features of data. Then, a two-layer LSTM is added to learn the time series characteristics of the extracted features, followed by an attention mechanism to selectively pay attention to the state of the hidden layer. The K-fold cross-validation is further adopted to make full use of data and improve the generalization ability of the model. Finally, the data set from North Carolina and the smart meter energy consumption data from the Low Carbon London project are employed to verify the validity of the method. Compared with traditional LSTM model, the proposed ResNet-LSTM-Attention method shows smaller mean absolute percentage error and better performance on load forecasting.

1. Introduction
Short-term load forecasting refers to forecasting the load in the next few hours or several days. Accurate load forecasting results can be used to arrange day-ahead scheduling, equipment maintenance, monitor system operation status, and prevent accidents. It is of great significance to improve resource utilization and economic benefits, and to ensure the normal production of society and people's daily life.

At present, the methods applied to short-term load forecasting can be divided into three categories: classic methods, traditional methods and intelligent methods. Among the classical methods, the regression analysis method has a simple structure, fast calculation speed, and good extrapolation performance. However, linear equations are used to express complex problems and cannot accurately predict the impact of various factors on the results[1]; the time series method requires less data and can reflect the continuous characteristics of the load in a short period of time, but the uncertain factors that have a greater impact on the load such as holidays are not considered enough[2]. Among the traditional methods, the Kalman filter method performs better: the load is divided into random components and deterministic components. The random components are represented by state variables, and the deterministic components are described by a first-order linear model[3]. A state space model is...
established to achieve prediction. The best estimate of the state of the moment, combined with the future state of the system, makes the model prediction results more accurate, but in actual scenarios, it is difficult to obtain the statistical characteristics of noise. With the development of data collection and storage technology, historical load data has increased exponentially[4], and various intelligent methods have been applied. Among them, neural network can deal with massive data[5], and because of the characteristics of fast convergence and strong adaptive ability, it performs best.

Nowadays, two types of deep learning models, convolutional neural networks (CNN) and recurrent neural networks (RNN), are widely used in load forecasting. CNN is a good machine learning method that can learn the relationship of input data. Using the existing model to train the convolutional neural network, it has the ability to express the relationship between input and output data. CNN has been widely used in pattern recognition tasks such as speech recognition, image recognition, and behaviour cognition, and has shown good performance. The special feature of RNN is that it uses the relationship between the current output in the network and the previous information, and can selectively use the information of the sequence to process the current input, and therefore it is more suitable to deal with problems with correlation between the samples[6]. As variants of CNN and RNN, ResNet and LSTM have advantages. Deep residual networks can effectively overcome the problem of overfitting[7] while LSTM can solve the problem of gradient vanishing and gradient explosion during RNN training[8].

In this paper, we propose a novel short-term load forecasting method for distribution substations based on ResNet and LSTM with attention mechanism (ResNet-LSTM-Attention), where the advantages of ResNet and LSTM are combined together. The attention mechanism is further added to take the different contributions of historical data at different times into account.

The rest of this paper is organized as follows. The proposed load forecasting methodology via ResNet-LSTM-Attention is described in Section II. In Section III, simulation results are demonstrated. Finally, conclusion will be drawn in Section IV.

2. Methodology

2.1. ResNet model

Theoretically, deeper CNN is more complex, and the prediction effect should be better. However, the learning ability of the network begins to degenerate, although the network can be converged through regularization. Because when the number of network layers increases, some values will be infinitely close to zero, i.e. the phenomenon of gradient vanishing. The residual structure breaks through the inherent idea that the neural network can only propagate sequentially when it propagates forward. That is, the output of the current layer of the network can only be the inertial thinking of the next layer of network as the input, so that the output of the ith layer of network can directly pass the adjacent. The jth layer network is used as the input of the (i+j)th neural network. The neural network is no longer degraded by factors such as gradient disappearance, and can reach hundreds of layers as needed.

By superimposing $y=x$ on a shallow convolutional neural network, it can be ensured that the prediction result of the neural network will not degenerate compared to the shallow network[9]. It is customary to call the residual structure of $y=x$ an identity connection, as shown in the Figure 1 below.

![Figure 1 ResNet basic block](image-url)
Let the first weight layer be $w_1$ and the second weight layer be $w_2$. Therefore, the final result $y$ should be:

$$y = F(x, \{w_1\}) + x$$

(1)

Therefore, what the residual network learns is the residual function $F(x)$, $F(x) = H(x) - x$. When the number of convolutional layers is deep, even if some parameters in $F(x)$ tend to zero, due to the existence of the $y=x$ layer, it can still be guaranteed that the learning ability will not degenerate.

In the residual basic learning unit, by adding an identity connection layer, the input $x_i$ is directly passed to the output as the final result, so the output $H(x) = F(x) + x$. When the number of layers increases, $F(x)$ is approximately equal to 0, then the output $H(x)$ of the residual result is approximately equal to $x$, which is the so-called identity connection. When $F(x)$ appears in the back propagation process when the gradient disappears, it can directly pass the identity connection part as the output. At this time, the neural network no longer learns the sum $H(x)$ of $F(x)$ and $x$, but learns the difference $F(x)$ between the output function $H(x)$ and $x$, which is the so-called residual: $F(x) = H(x) - x$, therefore, if the learning ability of $F(x)$ is poor, the input will be output directly, so that the accuracy of the entire network will not decrease as the number of layers deepens.

2.2. LSTM model

LSTM can consider the long-term and short-term dependencies in the time series. The structure of a single LSTM neuron is shown in the following figure. On the basis of the RNN model structure, three gate controllers are added: input gate $i$, forget gate $f$ and output gate $o$.

The forget gate determines how much information of the last moment in the memory unit will be transferred to the current moment for learning. Realized by the parameter $\sigma$, the value range of $\sigma$ is $(0, 1)$, and the forgetting gate function uses the sigmoid function to control the ratio of output.

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

(2)

where $W_f$ is the intermediate output of the hidden layer, $h_t$ and input data $x_t$ are the weight parameters of the gate operation; $b_f$ is the bias of the gate operation; $\sigma$ is the non-linear activation function sigmoid.

The input gate determines how much new information is added to the unit. To achieve this, two functions are needed as:

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

(3)

$$g_t = \sigma(W_g x_t + W_g h_{t-1} + b_g)$$

(4)

The output gate determines the proportion of the memory stored in the memory unit that will be output:

$$o_t = \sigma(W_o x_t + W_o h_{t-1} + b_o)$$

(5)

$$S_t = g_t \odot i_t + S_{t-1} \odot f_t$$

(6)

$$h_t = o_t \odot S_t$$

(7)
2.3. Attention mechanism

Attention mechanism is essentially a resource allocation mechanism that can highlight the impact of important information. This method is added to LSTM[10].

\[
\begin{align*}
    e_i &= u \tanh(w h_t + b) \\
    \alpha_i &= \frac{\exp(e_i)}{\sum_{j=1}^{T} \exp(e_j)} \\
    s_t &= \sum_{i=1}^{T} \alpha_i h_i
\end{align*}
\]

where \(e_i\) is the attention probability distribution value determined by the output \(h_t\) of the LSTM layer at time \(t\); \(u\) and \(w\) are the weight coefficients and \(b\) is the bias; \(s_t\) is the output value of the attention layer at time \(t\).

2.4. Output

The input of the output layer is the output of attention layer, which is fed to a fully connected layer, and the activation function is the Relu function. The calculation formula is:

\[
y_t = \text{Relu}(w_s s_t + b_s)
\]

Finally, the predicted data is de-normalized to obtain the actual value:

\[
X_{\text{scaled}} = X_{\text{old}} \times (\text{max} - \text{min}) + \text{min}
\]

2.5. ResNet-LSTM network structure

Combine several basic block residual basic modules with LSTM to build a ResNet-LSTM hybrid network model, whose composition is shown in the figure 4. The ResNet network consists of 34 layers of convolution. In order to prevent the model from overfitting, in the two-layer LSTM, a part of neurons is randomly discarded using the dropout method.

![Figure 4: Structure of the proposed ResNet-LSTM-Attention network](image)

3. Simulation results

In this section, the 2017 global energy forecasting (GEF2017) competition dataset from Charlotte, North Carolina[11] and the smart meter energy consumption data from the low carbon London (LCL) project[12] are employed to verify the validity of the method. The sampling interval of the GEF2017 dataset is one hour, including date, load, temperature, humidity, and whether it is a holiday. The LCL
project records kWh of end customers every half hour. We convert the kWh data to kW data and randomly aggregate data of some customers in year 2013 to simulate the total load curve of a distribution substation. The performance of the proposed model is compared with LSTM which is commonly used in short-term load forecasting.

3.1. Data preprocessing
Firstly, load the original data, and encode the hour, month, day of the week, whether it is weekend, whether it is holiday or not, into one-hot coding format. Second, discard the features that the neural network model cannot directly use in the original data. Thirdly, add the interactive influence of temperature, humidity information and month to enrich the characteristic dimension of the data. Then, take each consecutive 24 moments as a sequence. In order to facilitate calculations, the feature dimensions are split so that the length and width are approximately equal. The data is normalized using the maximum and minimum normalization method to eliminate the adverse effects of singular samples. The normalization method formula is as follows:

\[
X_{\text{norm}} = \frac{X - X_{\text{min}}(\text{axis} = 0)}{X_{\text{max}}(\text{axis} = 0) - X_{\text{min}}(\text{axis} = 0)}
\]

3.2. Evaluation Index
The mean absolute percentage error (MAPE) is used as the loss function for each iteration. Its expression is as follows:

\[
Y_{\text{MAPE}} = \frac{1}{n} \sum_{t} \left| \frac{x_{\text{true}}(t) - x_{\text{pred}}(t)}{x_{\text{true}}(t)} \right|
\]

where \(x_{\text{true}}(t), x_{\text{pred}}(t)\) are the true load and predicted load at time \(t\).

The adaptive moment estimation (Adam) is used as the optimization algorithm\[13\]. This algorithm can be regarded as a combination of momentum method and RMSprop algorithm. Not only does it use momentum as the parameter update direction, but it can also adjust the learning rate adaptively.

3.3. K-fold cross-validation method
This method uses the K-fold cross-validation method to improve the data set to obtain stronger randomness. First, the data set is randomly shuffled and divided into K sub-data sets evenly. When predicting, randomly select K-1 sub-data sets as the training set of the model, and use the remaining sub-data set as the test set of the model. The above process is repeated K times, and finally the K results obtained are averaged as the final output\[14\]. In this way, every piece of data can participate in the two processes of forward propagation and backward propagation. K-fold cross-validation not only prevents the model from overfitting, but also prevents the inability to make full use of the data.

3.4. Simulation results
Training loss curves of the LSTM model and the proposed model using the GEF2017 dataset are shown in Figure 5. It can be seen that compared with traditional LSTM model, the proposed ResNet-LSTM-Attention method shows smaller MAPE on the training set.
Figure 5 Training loss of the LSTM model and the proposed model using the GEF2017 dataset

Figure 6 shows the comparison of the 24 hours (one day) load forecasting results using the GEF2017 dataset and the LCL dataset. The blue, red, green curves correspond to the actual value, the prediction result of the LSTM model, the prediction result of the proposed model, respectively. Compared with the actual value, the time of the maximum point and the minimum point of the predicted result of the proposed model completely coincide, and the trends of the curves are almost the same. Compared with the prediction result of the LSTM model, the prediction results of the proposed model are closer to the true value, especially during the low load period at night, when the predicted values are almost completely consistent with the actual values.

Figure 7 shows the comparison of the 168 hours (one week) load forecasting results using the GEF2017 dataset and the LCL dataset. Similar conclusion can be drawn that the proposed model gives lower MAPE and better forecasting performance than the LSTM model. The results show that the method can also be used to predict the load of even a week.
4. Conclusions
This paper proposes a novel load forecasting method for distribution substations based on combined residual neural networks and long short-term memory neutral networks with attention mechanism. A 34-layer ResNet is built to extract the latent features of data. Then, a two-layer LSTM is added to learn the time series characteristics of the extracted features, followed by an attention mechanism to selectively pay attention to the state of the hidden layer. The K-fold cross-validation is adopted to make full use of data and improve the generalization ability of the model. Finally, the data set from North Carolina and the smart meter energy consumption data from the low carbon London project are employed to verify the validity of the method. Compared with traditional LSTM model, the proposed ResNet-LSTM-Attention method shows smaller mean absolute percentage error and better performance on load forecasting.

Therefore, the proposed method is very effective, which can help the power grid company adjust the dispatching plan and work arrangement, predict the load change in advance, and take targeted measures to ensure the stable and efficient operation of the power grid, so as to provide better power supply services.

Future work will be focused on integrating multiple models and improving the generalization capabilities of models.

Acknowledgments
This work is supported by the State Grid Tianjin Electric Power Company (No.KJ21-1-17).

References
[1] Amral, N., Ozveren, C.S., King, D. (2007) Short term load forecasting using Multiple Linear Regression. In: International Universities Power Engineering Conference. Brighton. pp. 1192-1198.
[2] Wang, Y., Xia, Q., Kang, C. (2011) Secondary Forecasting Based on Deviation Analysis for Short-Term Load Forecasting. IEEE Transactions on Power Systems., 2: 500–507.

[3] Guan, C., Luh, P. B., Michel, L. D., Chi, Z. (2013) Hybrid Kalman Filters for Very Short-Term Load Forecasting and Prediction Interval Estimation. IEEE Transactions on Power Systems., 4: 3806–3817.

[4] Quilumba, F.L., Lee, W.J., Huang, H., Wang, D.Y., Szabados, R.L. (2015) Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities. IEEE Transactions on Smart Grid., 2: 911-918.

[5] Wang, Y., Chen, Q., Gan, D., Yang, J., Kirschen, D.S., Kang, C. (2015) Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities. IEEE Transactions on Smart Grid., 2: 911-918.

[6] Bedi, J., Toshniwal, D. (2019) Deep learning framework to forecast electricity demand. Applied Energy., 238: 1312-1326.

[7] He, K., Zhang, X., Ren, S., Sun, J. (2016) Deep Residual Learning for Image Recognition. In: Computer Vision and Pattern Recognition. Las Vegas. pp. 770–778.

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

[9] Chen, K., Chen, K., Wang, Q., He, Z., Hu, J., He, J. (2018) Short-term load forecasting with deep residual networks. IEEE Transactions on Smart Grid., 4: 3943-3952.

[10] Luong, M.-T., Pham, H., Manning, C.D. (2015) Effective Approaches to Attention-based Neural Machine Translation. In: Empirical Methods in Natural Language Processing. Lisbon. pp. 1412–1421.

[11] Global Energy Forecasting Competition (2017) http://www.drhongtao.com/gefcom/2017.

[12] Smart Meter Energy Consumption Data from the Low Carbon London project https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

[13] Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on Artificial intelligence. Montréal. pp. 1137–1143.

[14] Chen, X., Liu, S., Sun, R., Hong, M. (2018) On the convergence of a class of Adam-type algorithms for non-convex optimization. In: International Conference on Learning Representations. New Orleans.