Theory and Practice of Science and Technology
2025, VOL. 6, NO. 3, 175-181
DOI: 10.47297/taposatWSP2633-456933.20250603

Load Prediction and Deployment Optimization of Electric Vehicles Based on CNN-LSTM Model

Yipu Wang

Electrical and Electronic Engineering school, Nanyang Technological University, Zibo, Shandong, China, 255000

ABSTRACT
The development of new energy vehicles in China is currently progressing swiftly. As the number of electric vehicles (EV) increases rapidly, the charging demand has also increased significantly. Numerous electric vehicles linked to the electrical grid at one time can increase power grid's operating pressure, enlarge the variance in power load between peak and valley, making it more difficult for scheduling for the power grid's functioning and upkeep. Therefore, accurate prediction of EV charging is helpful to scientifically plan charging scheme, and is significant to maintain the smooth operation of the grid. This paper use the hybrid neural network model for short-term prediction of charging load, build CNN-LSTM-Attention model, supplemented by attention mechanism and residual connection to optimize the neural network model, introduce the GRU unit to reduce the number of parameters and calculation, make it easier to be deployed on the embedded equipment of charging station.

KEYWORDS
electric vehicle; load prediction; neural network; attention mechanism

1 Introduction

With the widespread adoption of new energy vehicles, electric vehicles gradually occupy the mainstream position in the market. However, with the rise in electric vehicle count, there will be a distinct impact on the power grid. . For example, the increasing charging demand will lead to higher power demand and higher peak load, leading to local grid overload, and increase the pressure of transmission and distribution equipment. In addition, the randomness of EV charging may lead to the fluctuation of voltage and frequency of the power system, increasing the difficulty of operation and maintenance. Therefore, forecasting the charging load for electric vehicles is conducive to the reasonable planning of charging stations, to the stable operation of the power grid, to realize peak load shifting and valley filling, and to reduce the burden of peak electricity consumption.

At present, a variety of algorithms and models have been applied to the charging load forecasting of electric vehicles. Pang Songling, et al., describe the travel characteristics of users based on the travel chain theory, determine the spatial and charging distribution of users, build a regression tree model, trained the model using the LightGBM algorithm and predict the result[1]; Jixiang LU et al combined two effective and popular neural networks, establish the spatiotemporal feature map of the load, use the deep learning method to simultaneously predict the short-time load[2]; Zhang Yanyu et al. used dynamic adaptive maps, introduced the multi-head attention mechanism to express the spatial-temporal relationship of charging station load, predicted the charging load by the convolutional neural network[3]; Lin Xiang et al., improved the LSTM neural network model, and used the three-scale hierarchy method of analysis for determining the significance of charging factors such as working time and temperature, optimized the prediction results [4]. The above literature considers the spatial and temporal relationship of the charging load, but the consideration factors are relatively single, and the accuracy is affected by the data quality, which does not reach the expected level.

To tackle the problem of simple scenario, this paper considers a variety of factors affecting the EV charging load, such as temperature, humidity, wind speed, and month, whether working days, as the characteristics of the charging load. Then, build CNN-LSTM-Attention neural network model for achieving short-term forecasts of charging load of EV, and optimize the model by improving the mechanism of attention, introducing the residual connection and GRU unit to enhance the overall performance of the model, optimize the quantity of training parameters, and make it easier to deploy on the embedded equipment.

2 Model architecture

2.1 Data processing

This paper fully considers the objective reasons such as time, space and weather to achieve the short-term forecast of the charging burden of electric vehicle charging stations with multiple spatial and temporal factors. Charging capacity at the electric vehicle charging station is affected by several factors, which can affect the electric owners travel will, travel journey, thus affect charging power. this paper selects a Nordic countries a community apartment electric vehicle charging pile charging load as training and prediction data set, the community electric vehicle charging infrastructure contains the apartment head in car unit installation of private charging pile and community charging stations within the public charging pile. The data set includes time data including time, week, month, temperature, humidity, air pressure and other factors that may affect the charging load, including a total of 9840 sets of charging pile power data at 1 hour interval, in order to make short-term power load prediction in a step length of 1 hour. The flowchart of the load forecasting is as follows:

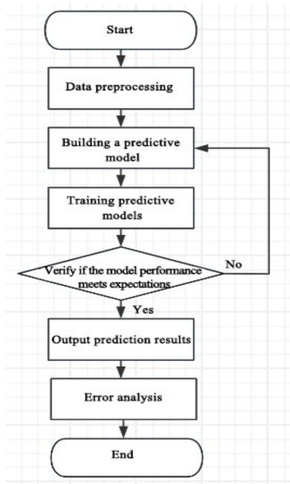

Figure 2-1 Flow chart of load forecasting

The existing datasets are preprocessed before training the model. In this paper, cross-validation is introduced for the purpose of training and evaluating the hybrid model. During training and testing this model, the MinMaxScaler function is introduced to normalize the data, so that the data are scaled between 0 and 1 to enhance the model's stability and generalization ability.

2.2 CNN-LSTM neural network

In this paper, the CNN-LSTM hybrid neural network is introduced as the model for the short-term charing load prediction .

The CNN-LSTM neural network blends the benefits of Convolutional Neural Network (CNN) with Long and Short-term Memory Network (LSTM) for data processing[5]. The following is the fundamental concept behind the CNN-LSTM neural network:

2.2.1 Convolutional Neural Network (CNN)

CNN primarily employed in extracting the spatial features. It captures the local spatial features [6] by using the filter (convolution core) in the sliding window. In time-series data, a CNN can process a segment of a sequence (for example, data within a time window) and extract features within that segment.

CNN increases the profundity of the model by stacking the pooling layers to extract more complex features. The schematic diagram of CNN is as follows:

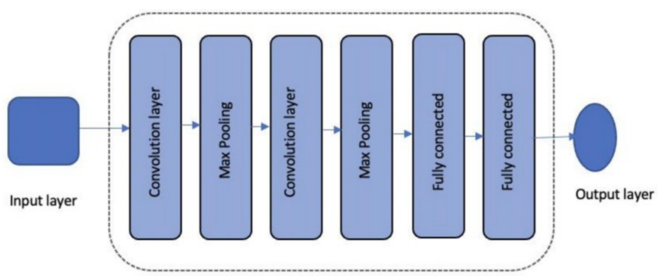

Figure 2-2 A Schematic diagram of the CNN

2.2.2 Long-and Short-term Memory Network (LSTM)

LSTM represents an enhanced version of the Recurrent Neural Network(RNN), which aims to address the issue of gradient explosion or gradient disappearance. LSTM units are able to learn long-term dependencies, permitting the reintegration of previous data subsequently, thus solving the gradient vanishing problem [7]. In LSTM, every neuron functions as a memory cell, and each memory cell contains three types of gates: forgetting gate, input gate, and output gate.

(1) Forget Gate. The function of the forgetting gate is to decide the specific information to be erased from the cell's state. It calculates a gating signal based on the present input along with the hidden state of last time step, which indicates which information should be retained and which should be discarded.

$$f_t=\sigma\left(W_f\cdot\left[h_{t-1},x_t\right]+b_f\right) \tag{2-1}$$

Where f_t is the state of the forgetting gate at time t, h_{t-1} and x_t is the input and output information for round t; σ is the sigmoid function; and the W_f ,b_f is the weight matrix and the bias of the forgetting gate.

(2) Input Gate. The input gate is responsible for deciding the new data to be retained in the cell's state. The structure consists of two distinct parts: one to determine the candidate value (Candidate) of the new information, and the other to determine the proportion of the new information that is actually updated to the cell state. The cell state is at the heart of the LSTM unit, which carries information about the time series and is continuously updated throughout the sequence processing. Information about cell states flows between time steps and is controlled by the forgetting and input gates.

$$i_t=\sigma\left(W_i\cdot\left[h_{t-1},x_t\right]+b_i\right) \tag{2-2}$$

$$\tilde{C}_t=\tanh\left(W_C\cdot\left[h_{t-1},x_t\right]+b_C\right) \tag{2-3}$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \tag{2-4}$$

Where i_t represents the condition of a forgotten gate at the present time step t ; $\tilde{C}_t$ is the temporary state of the memory unit at time step t, C_t is the updated cell state at t; W_i,b_i represents the input gate's weight matrix and its inherent bias, and W_C,b_C is the weight matrix and the state bias component of the cell state.

(3) Output Gate. The output gate calculates the current time step's output values by considering the cell's present state and the concealed state of the previous time step.

$$o_t=\sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right) \tag{2-5}$$

$$h_t=o_t*\tanh\left(C_t\right) \tag{2-6}$$

Where o_t represents the condition of the output gate at time t , and the W_o, b_o represents the input gate's weight matrix and bias component.

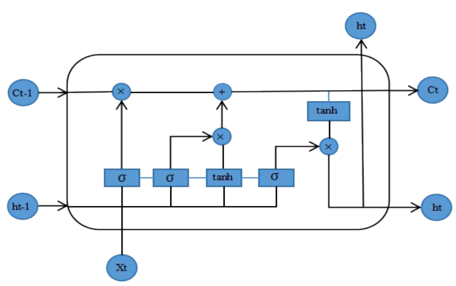

Figure 2-3: A Schematic diagram of the LSTM

2.2.3 CNN-LSTM

In the CNN-LSTM network, the input data is first processed using the CNN to extract the spatial features.

The output of the CNN (which can be the activation map of the last convolution layer) is then flattened into a sequence that serves as an input to the LSTM. LSTM processes these feature sequences to further extract temporal features and learn the time-dependent [8] between features. Finally, the output of the LSTM can be used for classification, prediction, or other tasks.

2.3 Multi-head attention mechanism

The mechanism of attention is inspired by the properties of the animal's visual attention, which helps us to get the desired results faster by focusing on the key information. From the 1990s, this mechanism was also introduced into computer vision and was widely used. In recent years, specialists in machine learning have studied deeply into the algorithm and application of the attention mechanism, integrating it into diverse neural networks to enhance their efficiency[9]. In this paper, the Multi-Head Attention Mechanism is inserted behind the LSTM layer of the model, enhancing the model's efficiency and its capacity for broad application [10].

2.4 Residual connection

Residual Connection, also known as jump connectivity, is a technique used in deep learning to train very deep neural networks. Residual connection was initially proposed in ResNet, and has accomplished remarkable achievements in the domain of image recognition. Thereafter, residual connectivity has been widely used in a variety of deep learning models. The core concept of this technology is to allow the signal (or residuals) in the network to skip one or more layer of direct transmission, solving issues related to the disappearance and explosion of gradients[11] in deep network training. The introduction of residual connectivity not only simplifies the network training, but also improves the model performance, making it possible to build and train deeper networks.

The core idea of the residual link is to add the input directly to the output. Specifically, presuming x represents a layer's input and F(x) denotes its learning function, then the output of the residual connection is:

$$y=F(x)+x \tag{2-7}$$

Where x represents the input, known as the jump connection; F(x) denotes the transformation required for the network's learning, usually expressed as a combination of several layers of neural networks (such as convolution, activation function, etc.); the y is the output of the residual connection.

In this way, the neural network actually learns the residual F (x), rather than directly learning the complete mapping of the input to the output.

2.5 Neural network structure

This paper uses the CNN-LSTM model with multi-head attention for the purpose of forecasting the immediate charging capacity of electric vehicle charging points. The primary framework of this model is: the first part is the convolutional neural network, primarily made up of two layers of CNN network, each layer of which has corresponding pooling layer and regularization to reduce the chance of overfitting; the second part is the LSTM neural network, which is also composed of two layers of LSTM network.The residual connection is placed between the second layer of CNN and the first layer of LSTM; the mechanism for multiple attention is integrated following the second LSTM, and the output result is obtained after the full connection layer.

The basic parameters for each layer of the neural network are shown below:

Table 2-1 Structure table of the CNN-LSTM-Attention model

network layer	network topology	main parameter	network layer	network topology	main parameter
1	input layer		8	residual	
2	Convolutional layer 1	in_channels=12, out_channels=16, kernel_size=3, stride=1, padding=1	9	LSTM1	input_size=32, hidden_size=16
3	pooling layer	kernel_size=2	10	LSTM2	input_size=16, hidden_size=8
4	Dropout	p=0.35	11	Attention layer	input_dim=8,num_heads=4
5	Convolutional layer 2	in_channels=16, out_channels=32, kernel_size=3, stride=1, padding=1	12	fully connected layer 1	In features=8, Out features=4
6	pooling layer	kernel_size=2	13	fully connected layer 2	In features=4, Out features=1
7	Dropout	p=0.25			

2.6 AI chip deployment optimization

The Gated Recurrent Unit (GRU) is a Recurrent Neural Network (RNN) variant employed in the handling of sequential data. The GRU is designed to make it fewer in number of parameters than LSTM, which makes GRU training faster in some cases and easier to deploy [12] on embedded devices. Two primary elements of the GRU are the Update Gate and the Reset Gate. The basic principles of GRU are as follows:

(1) Update Gate:

The update gate determines the quantity of data regarding the present time step to be transmitted to the subsequent time step. It calculates a gating signal based on the current input and the final concealed state ranges between 0 and 1, determining the mixing ratio of new and old information.

$$z_t = \sigma\left(W_z \cdot x_t + U_z \cdot h_{t-1} + b_z\right) \tag{2-8}$$

Where σ represents the sigmoid function, W_z serves as the weight matrix for the update gate, h_{t-1} represents the concealed condition of the final time step, and x_t is the input to the present time step.

(2) Reset Gate:

The reset gate assesses the extent to which past data should affect the concealed condition at the present moment. It similarly calculates a gating signal grounded on the current input and the previously concealed state, determining the importance of the old information.

$$r_t = \sigma\left(W_r \cdot x_t + U_r \cdot h_{t-1} + b_r\right) \tag{2-9}$$

Where W_r represents the weight matrix of the reset gate.

(3) Candidate Hidden status:

The GRU also calculates a candidate hidden state, which is an intermediate state, representing the new state that can be generated based on the current input if the old information is completely ignored only.

Candidate hidden status $\tilde{h}_t$:

$$\tilde{h}_t = \tanh\left(W_h \cdot x_t + U_h \cdot \left(r_t \odot h_{t-1}\right) + b_h\right) \tag{2-10}$$

Where W_h represents the weight matrix of the candidate hidden states, ⊙ represents element-by-element multiplication.

(4) Final hidden status:

The Final hidden status represents a cumulative weighted total of the existing hidden and potential hidden states, with the update gate governing the weights. Meanwhile, the reset gate affects the computation of candidate hidden states, determining how much past information should be considered.

New hidden state: h_t

$$h_t = z_t \odot h_{t-1} + \left(1 - z_t\right) \odot \tilde{h}_t \tag{2-11}$$

Where z_t controls the mixture of the former hidden state h_{t-1} and the candidate hidden state $\tilde{h}_t$.

Compared to LSTM network, the GRU unit merges the forgotten and output gates into a single gate(update gate) and introduce the reset gate. The two gates together determine which information should be retained and which should be forgotten. For the purpose of reducing the number of parameters, improving the operation speed and deployment, in this paper, the second layer of LSTM in the prediction model is replaced with GRU layer. The new model is re-trained in the same way to compare the effect on the prediction performance of the model after the replacement.

3 Outcomes of the experiments

3.1 Training and forecasting the models

During the model's training and evaluation phases, the MSELoss function is utilized to assess the discrepancy between the actual and forecasted values; the optimizer selects Adam optimizer and combines the learning rate regulator to dynamically modify the rate of learning in line with the established validation set performance to improve the convergence efficiency.

The time step of the loaded dataset is 128, and the training times are epochs=500. In each training round, the batch size of forward propagation and back propagation is 32. After training, the model that has been trained is preserved, the test dataset divided before is used to make load prediction on the trained model, and multiple evaluation indexes are calculated. The error evaluation indexes used in this paper are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), which are used to detect the actual effect of the model in the load prediction task. The formula for the three evaluation indexes are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3-1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3-2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3-3}$$

Regarding the implementation and enhancement of the model, the neural network model incorporating GRU undergoes identical training to evaluate its predictive efficacy, and compare its accuracy to verify whether the deployment optimization is feasible.

3.2 Prediction results and performance evaluation

A total of three neural network models are mentioned, the tradition CNN-LSTM model with attention mechanism, the model which upgrades attention mechanism from single-head to multi-head, and the model that replaces the second LSTM layer with the GRU layer. The same validation set is used in these three models to confirm their predictive precision, along with the predictive curve of EV charging load is shown in the following figures:

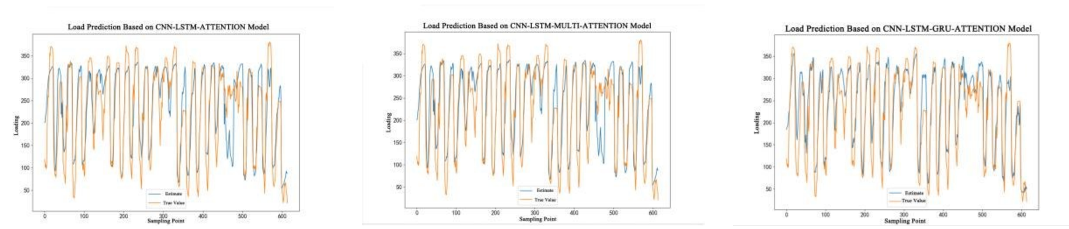

Figure 3-1 Load prediction plots of the three models

As can be observed from the graph of EV charging load in this region, the charging load of EVs is periodic and fluctuates in a daily cycle (the time interval of each sampling point is one hour). In addition, due to the subjectivity and randomness of ev charging, there are certain differences between the daily charging peak and trough, which increases some difficulty to the load prediction, especially the daily peak and valley value prediction. According to the figure, although there will be some differences between the prediction results of the model and the actual situation, the overall accuracy is relatively good. Furthermore, the above error indicators were used to evaluate the model performance. The evaluation index values of each model are shown in the table:

Table 3-1 Evaluation indicators of each model

	MAE(kW)	RMSE(kW)	MAPE(%)	No.of Params
CNN-LSTM-ATTENTION	36.527	47.709	24.0727	24738
CNN-LSTM-MULTI-ATTENTION	37.749	46.429	23.7482	25017
CNN-LSTM-GRU-ATTENTION	33.444	43.223	22.4869	24809

From the table above shows, when the attention mechanism upgrade to multiple attention mechanism, the average absolute percentage error before slightly reduced, performance has a small improvement, and when the second LSTM layer is replaced with GRU layer, MAPE decreases from 23.7482% to 22.4869%, whose performance has improved, and the model of the peak prediction is significantly better than the other two models. Therefore, this optimized deployment scheme can not only reduce the quantity of computing parameters, enhance the speed of computing, but also improve the model performance in certain situations.

4 Summary and outlook

The rapid advancement of the electric vehicle sector has led to a substantial number of electric cars connected to power grids for recharging, greatly impacting the grid's stability. For that reason, accurately predicting the daily charging load of EVs contributes to the reasonable distribution of power, improving the dispatching capacity and stability of power systems. This paper details the construction of a neural network model designed to predict the charging load of electric vehicles. Based on the currently popular and efficient CNN-LSTM neural network, the model utilized the multi-head attention mechanism to enrich the structure of CNN-LSTM-Attention model. In addition, for the purpose of enhancing the model's prediction performance, optimization methods such as residual connection and regularization are integrated into the model for enhance its operation efficiency and precision. Finally, in order to better deploy in embedded devices, the model deployment is optimized, and the GRU layer is introduced instead of the LSTM layer for enhancing the comprehensive efficacy of the model. Through the performance comparison of each model, the conclusion can be drawn that the optimized CNN-LSTM-GRU-Attention model is the most suitable model selection based on the principle of guaranteeing precise accuracy.

References

[1] Pang Songling, Fan Kaidi, Chen Chao, et al. Multi-time-scale prediction model of EV charging load based on LightGBM algorithm and travel chain theory [J]. Automotive Technology, 2024,(06):9-16.DOI:10.19620/j.cnki.1000-3703.20230993.

[2] A hybrid model based on convolutional neural network and long short-term memory for short-term load forecasting

[3] Zhang Yanyu, Zhang Zhiming, Liu Chunyang, et al. Charging load prediction of EVs based on a dynamic adaptive graph neural network [J]. Automation of Electric Power System, 2024,48 (07): 86-93.

[4] Lin Xiang, Zhang Hao, Ma Yuli, et al. Charging load prediction of electric vehicles based on the improved LSTM neural network [J]. Modern Electronic Technology, 2024,47 (6): 97-101.

[5] Rui Yan, Jiaqiang Liao, Jie Yang, Wei Sun, Mingyue Nong, Feipeng Li,
Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering,Expert Systems with Applications,Volume 169,2021,114513,ISSN 0957-4174.

[6] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021).

[7] Alex Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, Physica D: Nonlinear Phenomena, Volume 404, 2020, 132306, ISSN 0167-2789, https://doi.org/10.1016/j.physd.2019.132306.

[8] M. Alhussein, K. Aurangzeb and S. I. Haider, "Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting," in IEEE Access, vol. 8, pp. 180544-180557, 2020, doi: 10.1109/ACCESS.2020.3028281.

[9] HUANGFU Xiao-ying, QIAN Hui-min, HUANG Min. A Review of Deep Neural Networks Combined with Attention Mechanism[J]. Computer and Modernization, 2023, 0(02): 40-49.

[10] Zheng Zheng, et al. " Power load prediction based on multi-head attention convolution network. "Journal of Nanjing University of Information Science & Technology (Natural Science Edition)/Nanjing Xinxi Gongcheng Daxue Xuebao (ziran kexue ban) 14.5 (2022).

[11] Wang Jin, et al. " named entity identification based on hierarchical residue connection LSTM. "Journal of Jiangsu University (Natural Science Edition)/Jiangsu Daxue Xuebao (Ziran Kexue Ban) 43.4 (2022).

[12] Shewalkar, Apeksha, Deepika Nyavanandi, and Simone A. Ludwig. "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU." Journal of Artificial Intelligence and Soft Computing Research 9.4 (2019): 235-245.