Photovoltaic output prediction of regional energy Internet based on LSTM algorithm

Geping Weng¹, Chuanxun Pei¹, Jiaorong Ren¹, Chen Ye¹, Qinyue Cui¹, Hua Qing¹, Yuan Liu², Xinyu Guan²*
¹Ningbo Supply Company, State Grid Zhejiang Electric Power Co., Ltd, Ningbo, China
²Beijing Inshine Technology Co., Ltd., Beijing, China

*Corresponding author e-mail: hd_mrg@163.com

Abstract. By the huge development of large scale and modular photovoltaic power generation, accurate photovoltaic (PV) output prediction can help PV power station, scheduling department and power system operate safely and economically. In the process of PV output prediction, the data density is large, and the output data is relatively regular. Therefore, this paper considers the use of long-term and short-term memory neural network algorithm to optimize the problem of algorithm gradient vanishment in recurrent neural network, and complete the output prediction of PV power in the regional energy Internet on the basis of historical output data. In this paper, LSTM algorithm is used to analyze the historical output data of PV stations in an industrial zone of a certain city. It can be found that LSTM algorithm has good adaptability for short-term PV output prediction, which can meet the needs of application.

1. Introduction

In the recent period, the utilization and huge development of solar energy have become an important field in the global energy transition. Photovoltaic (PV) power generation has fully entered the stage of large-scale development, showing good development prospects. In 2017, the global installed capacity of PV power generation reached 403.47GW, an increase of 33.8% compared with 2016, and Chinese domestic installed capacity of PV power generation reached 137.62GW, an increase of 76.3% compared with 2016 [1]. The output of PV power generation is greatly influenced by meteorological ingredients, with strong volatility and intermittent. These characteristics make the high proportion of PV access to the power system caused a huge impact and challenge [2]. If we can accurately predict output of PV power generation, not just can improve the operational efficiency of PV power plants, but can help adjust the operating mode of scheduling department, to ensure security and stability and economic operation after the access of high proportion of PV power systems.
Multi-energy system analysis builds its foundation on district energy network. It is also a firm manifestation of the characteristics of multi-energy system. From a functional point of view, a systematic energy system can organically and smoothly integrate various forms of energy. It can also allocates and adjusts according to factors such as price and environmental impact. From an energy service point of view, the user’s multiple needs are statistically considered, and through reasonable dispatch to achieve the purpose of peak-shaving and valley-filling and rational use of energy; from the perspective of energy networks, through collaborative analysis of electric nets, crude oil and gas nets, heating and vapor nets and other networks, the development of multiple energy technologies is promoted. The area can be as small as an industrial park, big company, and new building, as big as a city, town, and community. On the whole, it includes integrated energy systems such as electricity supply, natural gas supply, thermal energy supply, hydrogen energy supply, automated transportation, and related communication and information foundations. The basic feature of the facility is to have power generation, power transmission, voltage transformation, power storage, load and other links. In this kind of regional network where multiple energy sources are integrated, the information involves "electricity flow", "gas and liquid flow", "information flow", "transportation flow", etc.

Compared with the inter-regional backbone energy Internet, the regional energy Internet uses various industrial enterprises and residents in a local area as user groups. It collects information data such as energy production, consumption, transmission, and storage, and relies on data analysis, energy coordination and optimization. The scheduling mechanism meets the load demand of users in the domain. Internet energy region on the basis of the formation of some energy characteristics different from the inter-regional Internet, for example, multiple kinds of energy source mutually complementary, consisting of a two-way controllable energy flow, etc.

Accurately predicting the generation power, cooperating with the power dispatching department to reasonably dispatch power and arrange the reserve capacity, not only can improve the power quality and enhance the stability of the grid, but also can lessen the operation and maintenance cost of the power system from some aspect. In this environment, it is more and more important to predict the power of PV power station. Researchers at home and abroad use different methods to establish prediction models with different prediction time and resolution to meet different requirements, and are committed to enhancing the prediction performance of the models.

In power predictions of different durations, the pros and cons of the short-term power prediction effects will directly affect the arrangement and dispatch of the power system, as well as the safety and stability of the power grid. In this regard, relevant practitioners at home and abroad will focus on PV power prediction for short-term power prediction.
1.1. Indirect prediction

In the indirect prediction method, accurate prediction of illumination amplitude is the primary factor affecting the overall prediction effect of the model. According to the actual situation of different regions, the relevant literature selects different input parameters and puts forward a variety of light amplitude prediction models. The indirect prediction method usually uses a physical model to calculate the power generation after obtaining the predicted value of the surface illumination amplitude. The physical model is a combination of the working principle of PV panels actual installation, the power generation is derived by experimental measurements, theoretical calculation or the like corresponding to the physical parameters related to the relationship, which is also known empirical formula square method.

In reference [3], the power generation of each time of the next day was predicted by using the least square method with the input of the external irradiance and the amplitude of the surface illumination at each time of the day. Using numerical weather forecast data can improve the adaptability of weather prediction model and improve the prediction accuracy. In reference [4], the recursive neural network (RNN) and the feedforward neural network (FNN) were used to predict the future 1 hour to 3 hours by using the historical light amplitude, meteorological data combined with the data provided by the numerical weather forecast of extraterrestrial irradiance and hourly temperature. The prediction results of Japanese RNN model with different durations are 18.53% to 20.3%, which is better than that of fuzzy neural network. Reference [5] used NWP data to measure PV power stations from four locations, combined with the data of local weather stations, realized the dimensionality reduction of NWP data, and used the sunshine index as the output parameter, and used BP neural network to predict the optical amplitude. The MSE of the prediction results in different periods of the next day are twenty to fifty, and similar NWP data can be obtained without using local NWP data to improve the prediction performance of the model.

1.2. Direct prediction

Compared with indirect forecasting, direct forecasting methods are more widely used in short-term power forecasting. Linear regression models based on time series such as ARMA and ARIMA are more common in early direct power prediction. The literature [6] adds seasonal factors to the ARIMA model, and proposes the SARIMA model to predict hourly power generation. RMSE of the prediction result is 9.56%. Literature [7] uses a partial functional linear regression model (PFLRM) on the basis of multiple linear regression to predict the next day's power generation. While using historical data, it takes into account the characteristics of the change in power generation within a day. Compared to traditional linear regression models, PFLRM's ability to deal with nonlinear problems has also been improved.

Artificial neural network is developed from imitating the working principle of neurons in the human brain. Through the error feedback algorithm, the complex relationship between the input and output of the model is processed, which shows great advantages in nonlinear modeling.

The literature [8] models the historical data in a year according to the seasons, uses the power generation at each time of the day and the maximum temperature and minimum temperature forecasted on the next day, and uses RNN to predict the hourly power of the day.

The literature [9] used a new neural network (Nonlinear Auto - Regressive Exogenous the Inputs Models with the Model, NARX), using the predicted power and the PV plant history NWP same time, the historical power as the adjacent PV plant NARX input. The MAPE of this method on sunny and rainy days is 10.06% and 18.89%, respectively.

In addition to weather factors, the effectiveness of historical data used in model training, the selection of model input parameters and the length of the training set will also have a greater impact on the prediction effect.

2. Methodology

Long-Short Term Memory (LSTM) is an improved Recurrent Neural Network (RNN), promoted by Hochreiter et al. [11] and improved and proposed by Alex Graves. LSTM can learn the long- and short-termed dependence information of the time series. Because the long-term memory neural
network contains internal memory and storage units, it can be used to process and predict time interval and time delay event data under time series distribution.

2.1. Introduction to RNN Recurrent Neural Network

Recurrent neural network (RNN) is an improved multilayer perceptron network (as shown in the figure), which includes an input layer, a hidden layer and an output layer. In the hidden layer, there is a transport tool chain to input the current state to the next hidden layer. After extending the time series, the hidden input layer matching with it can be found. There are two processing times, that is, the moment before the input of the hidden layer and the current time of the input layer.

![Figure 2. Basic structure of Recurrent Neural Network.](image)

2.2. Long and short-term memory neural network

In view of the disappearance of the gradient of the recurrent neural network[10], the long and short-term memory neural network redesigned the computing node on the basis of maintaining the RNN network structure.

The LSTM network structure adopts a gate control mechanism, which is composed of memory cells, input doors, output doors, and forget doors. The specific structure is shown above. In the structure diagram, $X_t$ represents the input at time $t$, and $h_t$ represents the state value of the cell at time $t$. The three different big boxes in the figure represent the states of the cells in different timings. The small boxes with $\sigma$ in the cells represent the feedforward network layer with the activation function sigmoid. Similarly, the activation function with tanh is a feedforward network layer with tanh. Among them, the number of hidden neurons in the feedforward network layer is continuously trained and debugged, and an optimal value is determined after comparing and measuring the prediction accuracy of each model.

![Figure 3. Cell of LSTM](image)

The computing node is composed of input doors, output doors, forget doors, and cells. Basic cell is the foundation to record the current state. The calculation formula is:
In the formula, \( \sum_{i} x_i(t)w_{ic} \) is the input gate input at time \( t \), and \( \sum_{h} b_h(t-1)w_{hc} \) is the forgetting gate input at time \( t-1 \).

At the same time:

\[
s_c(t) = b_o(t)s_c(t-1) + b_t(t)g[a_c(t)]
\]

In the formula, \( b_o(t)s_c(t-1) \) is the product of the forget gate \( a_c(t) \) mapping at time \( t \), \( b_t(t) \) is the product of the forgetting gate at time \( t \) and the basic cell output at time \( t-1 \), \( g() \) is the mapping function, and \( s_c(t) \) is the state output of the cell at time \( t \).

The following is the calculation principle of each control gate. First, calculate the input gate \( i_t \) values and the candidate cell state value \( \hat{C}_i \) which is inputted at time \( t \), the formula is like this form:

\[
i_t = \delta(W_i \ast (X_t, h_{t-1}) + b_i)
\]

\[
\hat{C}_i = \tanh(Wc \ast (X_t, h_{t-1}) + b_c)
\]

Next, calculate the activation value of forgetting gate at time \( t \), the following formula:

\[
f_t = \delta(W_f \ast (X_t, h_{t-1}) + b_f)
\]

From the above two calculations, the cell state update value \( C_t \) at time \( t \) can be calculated, the formula is as follows:

\[
C_t = i_t \ast \hat{C}_i + f_t \ast C_{t-1}
\]

After the cell state update value is calculated, the value of the output gate can be calculated at last, and the calculation formula is as follows:

\[
O_t = \delta(W_o \ast (X_t, h_{t-1}) + b_o)
\]

\[
h_t = O_t \ast \tanh(C_t)
\]

2.3. Proposed framework

a) The first step is to transform the original data and generate a model training data set:

The historical output data output plan has been explained in the previous chapter. After the historical output data is processed, the output data is divided into 3 groups, namely training data set (80%), verification data set (10%) and test data set (10%). The grouping is arranged according to time, and comprehensive consideration of data sufficiency and annual distribution of output. Set and test set validation data requires more typical and representative data.

b) The second step is to determine and adjust the structure of the LSTM network model:

Use the LSTM module in the Keras data tool library under the Python language to construct the LSTM network, and use the training set data to train it. The training process of the network involves the adjustment of many parameters, such as: Determining the activation function of the LSTM module (the parameter is \( \tanh \) by default in Keras); Determining the full connection activation function of LSTM output data (Keras default is linear); In order to prevent over-fitting, it is necessary to determine the rejection rate of each layer of network nodes, and the default value is set to 0.2; Determining an error and the validity of the calculated analysis indicators will be discussed in the following sections; Determining the iterative update method of weight parameters;

c) The third step is to import the training set data, training the LSTM network.

d) The fourth step is to use the trained LSTM model to perform regression analysis on the test data set and compare it to the results.

e) The fifth step is to analyze the effectiveness of the forecast.
3. Case Study

3.1. Proposed framework

The collected original historical output data cannot be directly used for modeling calculations. It needs to be pre-processed first to meet the input parameter requirements of different prediction models. Data preprocessing can be divided into four steps: overall monitor abnormal value, determine the valid time interval, fill loss of data and normalize data.

In the actual data acquisition process, sensor failure, staff maloperation, transmission error and the impact of individual extreme weather will lead to extreme values, these extreme points are also known as abnormal points, which either usually deviates greatly from the adjacent data points or does not conform to the basic laws of physics. So, it is necessary to remove the abnormal points from the original data. The power abnormal point removal process is as follows:

a) Eliminate power points that are less than zero or greater than the rated input of the inverter’s DC terminal;

b) When the extra-terrestrial irradiance and the surface illumination range are both positive values, the power points with a value of zero are eliminated, which is mainly used to eliminate the data which was collected when PV power station was shut down.

The missing data in the collection transmission process will destroy the continuity of the time series, and cause a great interference to the prediction results of the time series method for different prediction models. Therefore, the completeness of data in each interval needs to be considered before the data is smoothed. Divide the original data set into several time intervals at a fixed time interval, and check the integrity of the different attribute data in each time interval. When the missing data of any attribute reaches more than 30%, the entire time interval is judged to be an invalid time interval and no longer participate in subsequent smoothing calculations.
3.2. Evaluation indicators

Evaluating the prediction error of PV power prediction scientifically can help analyze the characteristics of a variety of forecasting techniques, and tap their underlying causes and laws, in order to attempt to further enhance the predictive accuracy of the output of PV power generation.

There are many evaluation indicators for calculating the prediction results of PV power generation output points, the actual output variation range and the predicted output variation error are small, and the prediction results are more accurate.

From previous studies situation, there are relatively many indicators for evaluating when using point prediction methods, wherein the mean absolute error (MAE), mean absolute percentage error, root mean square (Root on Mean Square, the RMS) and root mean square error are more commonly adopted.

Different indicators have different evaluation angles, and their values have certain differences. Therefore, it is not scientific to select a single evaluation indicator. Usually, multiple indicators are used for comprehensive evaluation.

The average absolute error MAE is:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |P_i - \bar{P}_i|$$

The root means square error RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{P_i - \bar{P}_i}{\bar{P}_i} \right)^2}$$

In the formula, \(N\) is the number of predicted points of PV power generation; \(P_i\), \(\bar{P}\) is the predicted value and the actual value.

3.3. Results and discussion

According to the above method, we select several distributed PV power sources in an industrial park of City N in the first half of 2020, and conduct model analysis on the PV power output data sampled by these power sources every minute, so as to predict the PV output distribution of the energy Internet in the region, and compare it with the actual PV output data.

Part of the data is presented as below table:

Table 1. Part of original PV output data from city n.

| \(h_{\text{min}}\) | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 7                 | 0.275 | 0.277 | 0.275 | 0.277 | 0.278 | 0.277 | 0.279 | 0.279 | 0.278 |
| 8                 | 0.531 | 0.531 | 0.531 | 0.533 | 0.531 | 0.535 | 0.534 | 0.536 | 0.539 |
| 9                 | 1.023 | 1.026 | 1.026 | 1.028 | 1.029 | 1.031 | 1.032 | 1.034 | 1.034 |
| 10                | 1.814 | 1.815 | 1.811 | 1.814 | 1.810 | 1.816 | 1.834 | 1.820 | 1.824 |

Adopt the trained LSTM model to process and analysis regression analysis on the test data set, and the fitting results are shown in the figure.
Figure 5. Results of LSTM prediction

Where in thick solid lines as the test data, the imaginary line fitting results. Calculating the MAE is 0.1373, the RMSE is 0.3195. It can be seen that the error between actual output distribution and predicted output distribution is small and the prediction results are almost accurate.

Table 2. Evaluation indicators.

| Result No. | MAE   | RMSE   |
|------------|-------|--------|
| 1          | 0.1373| 0.3195 |
| 2          | 0.0858| 0.2633 |
| 3          | 0.1592| 0.5174 |
| 4          | 0.0627| 0.2045 |

4. Conclusions

This paper analyzes the importance of PV power generation output forecasting of regional energy Internet. Under the core characteristics of multi energy complementary, aiming at the features of PV output forecasting in regional energy Internet, a PV power generation output forecasting model based on LSTM algorithm is proposed. The main conclusions are as follows:

a) For PV power generation in regional energy Internet, the collected data density is relatively large and the distribution law obviously follows the rise and fall of the sun. There is no effective output data at night. The historical data required smoothed interpolation processing to meet the requirements of the algorithm.

b) The output model of PV power generation based on LSTM algorithm improves the general recurrent neural network regression prediction model, and makes use of the unique memory of LSTM algorithm to better match the data type of PV output. The results of the example analysis show that the method can achieve the short-termed PV output prediction in regional energy Internet, and can meet the needs of practical application.

There are still the following shortcomings, which can be improved in future research:

a) Due to the limitation of data collection time, it is not enough to analyze the change characteristics of power output in different seasons of the year;
b) The related research on the selection method of the trained set and the influence of the trained set length on the prediction results is still relatively rough. In the future, the prediction accuracy can be enhanced by analyzing the data characteristics of the training set to establish a typical sample set.

5. Acknowledgments
This article is sponsored by the State Grid Zhejiang Electric Power Co., Ltd. Science and Technology Project (No. 5211NB1900VT).

References
[1] China new energy. Global new energy development report 2016, [R], 2016.
[2] G Gulin, M., Vasak, M., & Peric, N. (2013). Dynamical optimal positioning of a photovoltaic panel in all weather conditions. Applied Energy, 108(aug.), 429-438.
[3] Zhijia Wang, Liang He, Xu Cheng, & Jieqiong He. (2014). Method for short-term photovoltaic generation power prediction base on weather patterns. China International Conference on Electricity Distribution. IEEE.
[4] Atsushi Yona, Tomonobu Senjyu, Toshihisa Funabashi,... & Chul-Hwan Kim. (2013). Decision Technique of Solar Radiation Prediction Applying Recurrent Neural Network for Short-Term Ahead Power Output of Photovoltaic System. Smart Grid and Renewable Energy (6).
[5] Amrouche, B., & Pivert, X. L. (2014). Artificial neural network based daily local forecasting for global solar radiation. Applied Energy, 130(oct.1), 333-341.
[6] Bouzerdoum, M., Mellit, A., & Pavan, A. M. (2013). A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. Solar Energy, 98(pt.C), 226–235.
[7] Wang, G., Su, Y., & Shu, L. (2016). One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. Renewable Energy, 96(pt.a), 469-478.
[8] Zhang, Y. X., & Zhao, J. (2011). Application of recurrent neural networks to generated power forecasting for photovoltaic system. Power System Protection and Control.
[9] Vaz, A. G. R., Elsinga, B., van Sark, W. G. J. H. M, & Brito, M. C. (2016). An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in utrecht, the netherlands. Renewable Energy, 85(JAN.), 631-641.
[10] Hochreiter S, Bengio Y, Frasconi P, et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies m. Wiley-IEEE Press, 237-243.
[11] Hochreiter S, Schmidhuber J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.