Attention-Based Encoder-Decoder Model for Photovoltaic Power Generation Prediction

Xiang Zhu¹, Juntao Hu², Liangcai Song³, Guilong Suo³ and Yong Zhan⁴

¹China Electric Power Research Institute, Nanjing, 210003, China
²School of Computer, Wuhan University, Wuhan, Hubei 430000, China
³Company of Henan Hebi National Grid Power Supply, Hebi, Henan 458000, China
⁴Pannet (Wuhan) New Energy Technology Co. Ltd, Wuhan, Hubei, 430000, China
Email: 476999887@qq.com

Abstract. The weather factors that affect the output of photovoltaic power generation systems have great volatility and discontinuities. Thus how to accurately predict the output of photovoltaic power generation has become a crucial issue. In this paper, we propose an attention-based Encoder-Decoder model for photovoltaic power generation. Filtered data based on maximum information coefficient is used as one of the features to reconstruct the experiment data. Then the attention mechanism is introduced to the Encoder-Decoder model, which constructed by Long Short-Term Memory (LSTM) neurons. We implement this experiment based on actual photovoltaic power plant examples and experimental results confirm the accuracy and applicability of the proposed model in predicting photovoltaic power generation.

1. Introduction

The use of photovoltaic power generation has become the main trend of new energy power generation and as technology developed, photovoltaic power generation gradually shows large-scale and distributed characteristics. In practical applications, the output of photovoltaic power generation systems is mainly affected by meteorological factors. However, the meteorology has great fluctuations, which may cause deviations in the prediction of photovoltaic power generation and eventually bring great challenges to the power system. Therefore, how to construct a prediction model to reduce deviations and improve prediction accuracy is the focus of this paper.

Traditional methods are mainly based on statistical theory to discover the rules between historical data and make related predictions. Which include the time series method, gray theory and regression analysis, etc. [3] based on gray prediction technology and constructed a dynamic prediction model to predict the photovoltaic power generation. [4] proposed an adaptive short-term forecasting model for grid-connected photovoltaic power generation. The adaptive algorithm finally enabled the model to handle sudden weather conditions well. [5] combined the regression analysis with the Markov chain and adopted the state transition probability matrix of the Markov chain to correct the error sequence. Finally improved the prediction accuracy of the algorithm. With the development of machine learning, some scholars have improved the accuracy of predictions by introducing more complex artificial intelligence algorithms. For example, [6] constructed a two-layer neural network model and used a single-step prediction method, which ensured the accuracy of prediction with a small number of samples. [7] reduced the prediction deviation of the model by combining traditional BP neural networks with a genetic algorithm. [8] selected important input features and then used long-term and short-term memory neural network LSTM to learn and model, eventually realized power generation...
prediction. As the photovoltaic power output has a great daily periodicity but the above models didn’t take into account the impact of historical power generation on the current power generation forecast.

In this paper, we first proved that there is a certain correlation between historical power generation data and current power generation based on maximum information coefficient (MIC) method, and next selected a suitable historical power generation data and the original meteorological characteristic observations to form a new input sequence. Then, we build an Encoder-Decoder model through LSTM neurons, and introduce attention mechanism on the model to establish a multi-step and multi-variable photovoltaic power generation prediction model. Finally, we implemented the experiment based on the real power generation data of the photovoltaic power plant in Henan, China. The results show that the Encoder-Decoder model with attention has higher accuracy in photovoltaic power generation prediction compared to RNN, LSTM (stack) and the original Encoder-Decoder model.

2. Encoder-Decoder Framework

The Encoder-Decoder framework is a common structure of LSTM networks [9]. LSTM can processes long-term sequence data by adding cell states and gated structures. We use the Encoder-Decoder model built by LSTM neurons, which encode and decode the constructed input sequence to achieve photovoltaic power generation prediction. As is shown in Fig. 1, the framework consists of two parts, the encoding side and the decoding side.

![Encoder-Decoder Framework Diagram](image)

Figure 1. The framework structure of Encoder-Decoder

\( h \) denotes the hidden layer state at the encoding end, and \( s \) denotes the hidden layer state at the decoding end. In the encoding phase, the current hidden layer state \( h_t \) is calculated from the current input \( X_t \) and the previous hidden layer state \( h_{t-1} \). The calculation process of \( h_t \) is:

\[
\begin{align*}
    h_t &= f(h_{t-1}, X_t) \\
\end{align*}
\]

Subsequently, the context \( C \) is obtained by summing the hidden layer states at each time and calculated as follows:

\[
C = r(h_1, h_2, h_3...h_n) \\
\]

Where \( R \) denotes some kind of RNN network, which we choose LSTM here.

In the decoding phase, the model predicts the next output \( Y_t \) through the context generated by the encoder and the previously generated \( Y_1, Y_2...Y_{t-1} \). The calculation of \( Y_t \) is:

\[
Y_t = \prod_{t=1}^{T} p(Y_t \mid Y_1, Y_2,...,Y_{t-1}, C) \\
\]

In RNN, \( Y_t \) can be expressed as:
Where $Y_{t-1}$ denotes the output value at the last moment, while $s_{t-1}$ the previous hidden layer state, and $C$ the context vector. Here we use an LSTM neural network followed by a softmax layer as the $g$ function.

Although the Encoder-Decoder framework is widely used in the field of seq2seq, it also has certain limitations. Since the length of the semantic vector is constant, the entire input sequence information needs to be compressed into the context at the encoding end. This will cause two problems. One is that the fixed-length context may not fully represent the entire input sequence information. The second is that during the encoding process, even we choose LSTM neural network as encoder network when faced with a long input sequence, the following input information will also overwrite the previous input information and cause information loss [10], eventually makes predictions less effective. We describe how to make up for this shortcoming of the model by introducing the attention mechanism.

3. Attention Mechanism

The attention mechanism was first proposed by Bahdanau et al. and applied to the field of natural language processing [11]. This mechanism calculates the weight to generate the semantic vector corresponding to each step of the decoder output. This improves information utilization of the Encoder-Decoder model to a certain extent. Fig. 2 displays the model structure after introducing attention mechanism:

![Encoder-Decoder model with attention](image)

As can be seen from Fig. 2, the model firstly encodes the input sequence on the encoder to obtain a vector sequence. Then at the decoder side, each step output is selected to pick up the most important part from the encoded vector sequence to generate semantic vectors $C_1$, $C_2$ ... $C_m$. Therefore, the output of each step can pay attention to useful information through the attention mechanism, thereby improving the adequacy of information utilization.

We use the Luong Attention mechanism [10], and the weight calculation of this mechanism is:

$$a_t(i) = \frac{\exp(\text{score}(s_t, h_i))}{\sum_{i=1}^{n} \exp(\text{score}(s_t, h_i))}$$

(5)

Where $a_t(i)$ denotes the weight value of the encoder hidden layer state $s_t$ and the i-th decoder hidden layer state $h_i$. There are three ways to calculate the score:

$$S = \begin{cases} s^T_t h_i \\ v^T_a \tanh(W_a \cdot \text{concat}(s_t, h_i)) \end{cases}$$

(6) (7) (8)
(6), (7), (8) respectively represent the inner product of Dot, General, and Concat, the S stands for score, \( W_a \) the weight matrix, and superscript T the matrix transpose. After calculating the score and the weight \( a_t(i) \), the weight vector is added to all hidden layer states \( h \) on the encoder side, to obtain the semantic vector \( C_t \) of step t on the decoder side:

\[
C_t = \sum_{i=1}^{n} a_t(i) \cdot h_i
\]

(9)

Then by concatenating \( C_t \) with the hidden layer state \( s_t \) of the original decoder, a new hidden layer state \( s'_t \) is obtained. The calculation process is:

\[
s'_t = \tanh(W_c \cdot \text{concat}(C_t, s_t))
\]

(10)

We multiply the concat stitching result with the fully connected matrix \( W_c \) to recover the shape as the shape will change after stitching. Finally, after calculating the hidden layer state \( s'_t \) of the decoder with "attention", the output \( Y_t \) is obtained. The output can be represented as:

\[
Y_t = W_{ho} s'_t + b_{ho}
\]

(11)

4. Design of Photovoltaic Power Generation Prediction Model

We use the Encoder-Decoder model with attention to realize photovoltaic power generation prediction. This model consists of the following five parts, data preprocessing, input feature construction, network structure design, model training, and model evaluation.

4.1. Data Preprocessing

Firstly, we pre-process the original data by empirical checking and completing the missing values. We find out the outliers and correct them by drawing and calculating in the empirical checking. And the completing of the missing value is to find the data at the same time in other records under approximate conditions and take average value to replace the missing data.

Then comes the normalization process. Because the magnitude difference of the original data will introduce errors to the model's prediction and cause neuron saturation [12], we choose to normalize the original data. The normalization formula is:

\[
v = \frac{(v - V_{\text{min}})}{(V_{\text{max}} - V_{\text{min}})}
\]

(12)

\( v \) is the original value of data record, \( V_{\text{max}} \) and \( V_{\text{min}} \) are respectively the maximum and minimum values in the experimental data.

4.2. Input Feature Constructions

The raw input data includes six meteorological observations: wind speed, wind direction, temperature, pressure, humidity, and actual irradiance. As the power generation has a certain time periodicity and trend, so we use the MIC to find the correlation between historical power generation and current power generation, then select the appropriate historical power generation data as one of the input features to add to the input dataset.

4.2.1. Maximum information coefficient

MIC is improved based on the mutual information MI and has universality, fairness, and symmetry [13]. Compared to mutual information, the MIC solves the problem of complicated calculation of continuous variables by mutual information and has higher accuracy in measuring the degree of correlation between two variables.

Here is the formula used by the MIC to calculate the correlation between two variables:

\[
mic(x, y) = \max_{a>b>0} \frac{f(x, y)}{\log_2 \min(a, b)}
\]

(13)
Where the $a$ and $b$ are the number of squares divided by $x$ and $y$ in the two-dimensional space, and $B$ is an empirical variable, which is usually set to about 0.6 power of the data volume.

4.2.2. Correlation analysis between historical generation power and current generation power

Let $P_i = \{p_i(1) \ldots p_i(t)\}$ represent the photovoltaic power generation sequence on the $i$-th day, where $p_i(t)$ denotes the real power generation value of the $t$-th sequence on the $i$-th day. We select historical power generation data through the rate of change and the MIC value.

$d$ denotes the rate of change of photovoltaic power generation between two moments:

$$d = \left(\frac{p_i(t+1) - p_i(t)}{p_i(t)}\right) \times 100\% \tag{14}$$

Through the change rate of $p_i(t-1)$, $p_i(t-2)$, $p_{i-1}(t)$, $p_{i-1}(t-1)$, $p_{i-1}(t-2)$ with $p_i(t)$ at the current moment, which reflects the fluctuation of photovoltaic output at different time intervals. The statistical results are shown in Table 1:

| $p_i$ ($t-1$) | $p_i$ ($t-2$) | $p_{i-1}$ ($t$) | $p_{i-1}$ ($t-1$) | $p_{i-1}$ ($t-2$) |
|--------------|--------------|----------------|----------------|----------------|
| 65.21%       | 58.11%       | 55.69%         | 51.5%          | 48.88%         |

The statistical results identify the proportion of samples with a rate of change within 10%, It can be seen that the fluctuation of the photovoltaic power generation increases over time. And the historical photovoltaic power generation data at the neighboring time has the largest reference value for the current power generation forecast.

Besides, we calculate the correlation between $p_i(t-1)$, $p_i(t-2)$, $p_{i-1}(t)$, $p_{i-1}(t-1)$, $p_{i-1}(t-2)$ and the power generation value $p_i(t)$ by the MIC. The results are shown in Table 2.

| MIC ($t-1$) | $p_i$ ($t-1$) | $p_i$ ($t-2$) | $p_{i-1}$ ($t$) | $p_{i-1}$ ($t-1$) | $p_{i-1}$ ($t-2$) |
|-------------|---------------|---------------|----------------|----------------|----------------|
| 0.9174      | 0.8475        | 0.8898        | 0.8540         | 0.8002         |

We can conclude from the table that $p_i(t-1)$ has the highest MIC value of 0.9174, followed by $p_{i-1}(t)$ of 0.8898. This is intuitively because the photovoltaic power output has a certain daily periodicity. Therefore, in addition to the most relevant $p_i(t-1)$, the current photovoltaic power value $p_i(t)$ also has a strong correlation with the power output at the same time of the previous day. From the MIC calculation results, we can find that $p_i(t-1) > p_{i-1}(t-1)$ and $p_i(t-2) > p_{i-1}(t-2)$. Therefore, from a fixed time point of view, $p_i(t-1), p_i(t-2)$ is more valuable than $p_{i-1}(t-1), p_{i-1}(t-2)$ for the current power generation forecast. Based on the results of the photovoltaic power generation rate at each moment, we add $p_i(t-1)$, $p_i(t-2)$ and $p_{i-1}(t)$ as one of the input features to the input data.

4.3. Design of Network Structure

On the encoder and decoder side of the model, we use 100 LSTM neurons for construction and chooses Relu as the activation function.

Let $I_t$ be the input data at time $t$, $O_t$ be the output data at time $t$, which is also the value of photovoltaic power generation. As we record the experimental data 15min / time, so there are a total of 96 records in 24h. On the encoder side of the prediction model, we use the sequence $I_t, I_{t-95}, I_{t-94}, I_{t-93}, \ldots, I_t$ as the input data, the sequence is the data of 24h before the current prediction time $t$. After performing high-level feature learning, encoder reasonably assigns attention
weights through the attention mechanism and passes the generated semantic vector context to the decoder for decoding, finally obtains the predicted output $O_t$.

4.4. Model Training
We use the mean square error (MSE) as the training loss function:

$$MSE(\hat{\theta}) = E(\hat{\theta}, \tilde{\theta})^2$$

(15)

Where $\hat{\theta}'$ denotes the estimated value of the parameter, and $\tilde{\theta}$ denotes the real value.

The model uses Adam algorithm for gradient descent optimization during training, compared with the traditional stochastic gradient descent, the Adam algorithm can dynamically update all the learning parameters through the calculation results, thereby obtaining better results [14].

4.5. Model Evaluations
After obtaining the forecast results, we use the following indicators to test the quality of the forecast results:

$$RMSE(y, p) = \left(\sum_{i=1}^{n} \frac{1}{n} (y_i - p_i)^2\right)^{1/2}$$

(16)

$$MAE(y, p) = \frac{1}{n} \sum_{i=1}^{n} |y_i - p_i|$$

(17)

$$SSE(y, p) = \sum_{i=1}^{n} (y_i - p_i)^2$$

(18)

Where $y_i$ and $p_i$ represent the actual and predicted values of photovoltaic power generation during $t$ period respectively.

5. Example Analysis
We use the weather data and power generation data of the distributed photovoltaic power station in Hebi Power Grid, Henan Province from January 1, 2017, to December 10, 2018, as training samples and use the data from December 11, 2018, to December 17, 2018, as the prediction samples. We compare prediction results by constructing RNN model, LSTM (stack) model, Encoder-Decoder model and Encoder-Decoder model with attention.

5.1. Experimental Data
Fig. 3 shows the real value of photovoltaic power generation for the selected seven-day forecast sample:

![Figure 3. Real values of predicted samples](image)

Figure 3. Real values of predicted samples
It can be seen from the figure that the power generation values in the sample data on December 12 and December 15 fluctuated greatly, which may be caused by the unstable weather. However, the overall power generation trend on December 16 and December 17 was relatively flat, indicating that the actual weather should be relatively stable and no abnormal changes occurred. Due to the limited space, we select two samples with obvious characteristics on December 15 and December 16 to display the results.

5.2. RNN Model

Fig. 4 shows the comparison between the prediction results of the RNN model and the true values:

![Figure 4. Results of RNN model on December 15](image1)

![Figure 5. Results of RNN model on December 16](image2)

It can be seen that the predicted power generation value of the RNN model is consistent with the basic trend of the real power generation value and the actual power generation value on the 15th fluctuated strongly between 12:30 and 17:00. On the 16th, except that the power generation value fluctuated between 10:30-11:00, the remaining parts were relatively smooth. The RNN model not only has a large prediction deviation in a period of high volatility but under the stable condition on the 16th, there is still a certain gap between the predicted value and the true value. This is because when the input time series accepted by the RNN model is too long, the model will lose the information learned from the previous input sequence, which ultimately leads to poor prediction results.
5.3. Encoder-Decoder Model

Fig. 6 shows the comparison of the prediction results of the Encoder-Decoder model and the RNN model on the 15th and 16th days:

![Figure 6. Results of Encoder-Decoder model on December 15](image)

![Figure 7. Results of Encoder-Decoder model on December 16](image)

It can be seen from the prediction results that the prediction results of the Encoder-Decoder model are better than the RNN model, especially when the weather fluctuation is small. However, in the face of large weather fluctuations, it still has a gap with the actual value although improved prediction compared to the RNN model. This is due to the limitations of the model itself, which makes it impossible to include all the information in the input sequence when encoding to generate the semantic vector.

5.4. Encoder-Decoder With Attention

Fig. 8 and Fig. 9 show the comparison between power generation forecast of the Encoder-Decoder model with attention and the traditional Encoder-Decoder model on the 15th and 16th:
Figure 8. Results of Encoder-Decoder model combined with attention on December 15

Figure 9. Results of Encoder-Decoder model combined with attention on December 16

It can be seen that the prediction effect of Encoder-Decoder with attention is improved compared to the traditional Encoder-Decoder model whether the power generation is with large fluctuations or a relatively stable. From the results on December 15, it can be seen that except for the three periods of 11:30-12:00, 12:30-13:00, and 15:30-15:45, which are in the early period of fluctuations, the model’s prediction value has a certain error with the true value. In other fluctuation stages, due to the addition of the attention mechanism, the model can generate a certain amount of "attention" for the current prediction when making predictions, to make full use of the existing information and make the prediction effect closer to the actual power generation value.

5.5. Quantitative Analyses of Prediction Results

We also use the RMSE, MAE and SSE indicators determined in the previous article for further quantitative analysis. The recorded results of each indicator are shown in Table 3:
Table 3. Statics of forecasting errors for three forecasting models

| Time          | Evaluation indicators | RNN  | LSTM (stack) | Encoder-Decoder | E-R (with attention) |
|---------------|------------------------|------|--------------|-----------------|----------------------|
| Total seven days | RMSE                   | 0.734| 0.576        | 0.545           | 0.379                |
|               | MAE                    | 0.409| 0.304        | 0.304           | 0.22                 |
|               | SSE                    | 361.57| 222.58      | 199.28          | 96.47                |
|               | RMSE                   | 1.069| 1.120        | 0.966           | 0.814                |
| 15th          | RMSE                   | 0.926| 0.795        | 0.824           | 0.532                |
|               | SSE                    | 41.12| 45.15        | 33.60           | 24.51                |
|               | RMSE                   | 1.401| 0.793        | 0.941           | 0.606                |
| 16th          | RMSE                   | 1.267| 0.692        | 0.856           | 0.412                |
|               | SSE                    | 70.66| 22.61        | 31.85           | 13.59                |

It can be seen that the Encoder-Decoder model with attention is superior to the Encoder-Decoder model without the attention, the RNN model and the LSTM (stack) model in both the overall prediction effect of the seven days and the specific date prediction effect selected for the weather fluctuations. Because the Encoder-Decoder model is built based on LSTM, compared with LSTM (stack), the overall power generation prediction results have a little difference. However, due to the structural differences of LSTM (stack), the Encoder-Decoder prediction effect is better in the case of large weather fluctuations. From the perspective of the RMSE, MAE, and SSE, the prediction performance of the Encoder-Decoder with attention is improved by 30.4%, 27.6%, and 51.6% compared with the traditional Encoder-Decoder model. It can be seen that by increasing the attention mechanism, the shortcomings of the original Encoder-Decoder model have been compensated to a certain extent, so that the attention model can better deal with the problem of photovoltaic power generation prediction.

6. Conclusions
In this paper, we have presented the attention-based Encoder-Decoder model for photovoltaic power generation prediction. Firstly, we use MIC to prove that there is a certain correlation between historical power generation data and photovoltaic power generation at the current moment. Next the appropriate historical power generation data selected based on MIC is added to the input datasets as the input features. Then we use LSTM neurons to build the Encoder-Decoder model and introduce the attention mechanism to overcome the information loss problem of the traditional model on the long input sequence. Finally, we perform the experiment based on the real power generation data. The results show that the Encoder-Decoder model with attention can improve the accuracy in the prediction of photovoltaic power generation compared to the traditional RNN neural network, LSTM (stack) model and the Encoder-Decoder model.

7. Acknowledgments
This work is supported by the Science and Technology Project of Henan Power Company of State Grid (SGHAHB00XTJ51900147).

8. Reference
[1] Jia Ke, Xuan Zhenwen, Lin Yaoyi, et al. An islanding detection method for grid-connected photovoltaic power system based on Adaboost algorithm [J]. Transactions of China Electrotechnical Society, 2018, 33(5): 1106-1113.
[2] Li Yanbin, Zhang Jiuju, Xiao Junming. Photovoltaic power short-term forecasting based on Gray-Markov chain statistical combination model [J]. Electrical Measurement & Instrumentation, 2015, 52 (23): 111-116.
[3] Hou Wei, Xiao Jian, Niu Liyong. Analysis of Power Genera-ti-on Capacity of Photovoltaic Power Generation Sys-te-m in Electric Ve-hicle Charging Station [J]. Electrical Technology, 2016, 32 (04): 53-58.
[4] Yang Zhichao, Zhu Feng, Zhang Chenglong, Ge Le, et al. Photovoltaic Power Generation Short-Term Power Forecasting Based on Adaptive Fuzzy Time Sequence Method [J]. Journal of Nanjing Institute of Technology (Natural Science Edition), 2014, 12 (01): 6-13.

[5] Wang Jituo, Wang Wancheng, Chen Hongwei. Photovoltaic power short-term forecasting based on gray-markov chain statistical combination model [J]. Electrical Measurement & Instrumentation, 2019, 56 (01): 76-81.

[6] Zhang Chengyi, Tang Yajie, Li Yongjie, Gao Qiang, Jiang Quanyuan. Photovoltaic power forecast based on neural network with a small number of samples [J]. Power Automation Equipment, 2017, 37 (01): 101-106+111.

[7] Ma Yiming. Study on Short-term Prediction of Photovoltaic Output Power Based on Optimized Neural Network [D]. Shenyang Agricultural University, 2017

[8] Huang Zhenling, Chi Xuebin, Xu Ke, Wang Tieqiang, Shi Ming, et al. Prediction of Photovoltaic Power Generation Based on Long Short-Term Memory Network [J]. e-Science Technology & Application, 2019, 10 (02): 31-41

[9] Kang Yunyun, Peng Dunlu, Chen Zhang, Liu Cong. ED-GAN: Judicial Document Generating Model Based on Improved Generative Adversarial Networks [J]. Journal of Chinese Computer Systems, 2019, 40 (05): 1020-1025.

[10] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation [J]. Computer Science, 2015.

[11] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. Computer Science, 2014.

[12] Song Renjie, Liu Fusheng, Ma Dongmei, Wang Lin. A very short-term prediction model for photovoltaic power based on similar days and wavelet neural network [J]. Electrical Measurement & Instrumentation, 2017, 54 (07): 75-80.

[13] Justin B K, Gurinder S A. Equitability, mutual information, and the maximal information coefficient [J]. PNAS, 2014, 111(9): 3354-3359.

[14] Li Zhaoyu, Ai Qian, Zhang Yufan, Xiao Fei. A LSTM Neural Network Method Based on Attention Mechanism for Ultra Short-term Load Forecasting [J]. Distribution & Utilization, 2019, 36 (01): 17-22.