Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060104
ISSN 2516-5003 Vol. 6 Num. 1

Over-warning of Coal Spontaneous Combustion Risk Based on the Transformer Model

Jidong Yao*,#, Yilong Xiao#, Ruiyuan Su, Wenhao Li, Jianxin Pang, Chao Jiang

College of Artificial Intelligence, North China University of Science and Technology, Tangshan, China
**Corresponding author*
#These authors contributed equally.

Keywords: Transformer Model, Coal Spontaneous Combustion Risk Warning, Time Series Prediction, Coal Mine Safety, Data Preprocessing and Feature Extraction

Abstract: In order to capture the information of different locations in the space more accurately, solve the problem of long-time prediction and parallel computation, and realise the over-warning of coal mine safety, we propose the over-warning method of coal spontaneous combustion risk based on the Transformer model. Firstly, the median filtering method and Lagrange interpolation method are used to detect the outliers, expand the data and fill the missing values. Then the unique self-attention mechanism of Transformer is used for feature extraction and trend prediction of the time series data; finally, the Transformer model is compared with LSTM and RNN by adjusting the size and step size of the sliding window. The experimental results show that, under certain circumstances, using the Transformer model to predict the CO and O2 concentrations can capture the gas changes very well, and the prediction accuracy of the Transformer is improved compared with the Long Short-Term Memory Neural Network (LSTM) and Recurrent Neural Network (RNN), which can be effectively applied to the coal mine safety early warning, reduce the occurrence of spontaneous combustion accidents, and protect the safety of miners and production stability, guaranteeing the safety of miners and production stability.

1. Introduction

China is the world's largest producer and consumer of coal, and also the country most seriously affected by coal mine fires. In recent years, as the phenomenon of the coupling of gas and coal spontaneous combustion in the mining airspace becomes more and more common, serious accidents such as gas explosions caused by coal spontaneous combustion in the mining airspace become more and more prominent, and this kind of disaster poses a serious threat to the safe production of coal mines, and it is necessary to take effective measures for the early identification and prevention of accidents in order to curb the accidents in the nip in the bud [1].

2. Literature Review

Coal spontaneous combustion indicator gases have been widely used as indicators of mine fires.

Early detection of spontaneous coal combustion is mainly carried out by analysing the changes in characteristics of temperature, concentration and ratio of gases in the process of spontaneous coal combustion ignition [2]. Time series forecasting is a method of using historical data to predict data, which is widely used in various industries. Time series forecasting refers to the process of predicting future data based on historical data, i.e., by analysing historical data, mining the internal laws of data, so as to predict the development trend of future data [3].

Common time series prediction algorithms have better prediction results when dealing with data with a single variable, but they often fail to effectively capture the long-term dependencies between data when predicting multivariate data with a large amount of data [4]. The Transformer model, on the other hand, has a very strong parallelisation capability and can better handle sequence data by capturing global information through the self-attention mechanism [5]. Instead of using circular connections to process sequence data, it directly accesses sequence data and the correlation between each time step and other time steps through the self-attention mechanism, thus capturing the long-term dependencies and periodicity between sequences.

Transformer has been widely used in many fields [6-11], but there are fewer studies on coal mine fire prediction. Therefore, this paper proposes a Transformer-based coal spontaneous combustion over-warning. Firstly, the median filtering method and Lagrange interpolation method are used to preprocess the data to improve the quality and availability of the data. Then, a time series prediction model was established based on the Transformer algorithm. Finally, the reliability of the model is verified by comparing with RNN and LSTM.

3. Prediction of Coal Spontaneous Combustion Risk Based on the Transformer Model

3.1 Transformer Architecture

Transformer is a deep learning model architecture for natural language processing and other sequential tasks proposed by Vaswani et al. Transformer architecture is different from traditional neural network models such as CNN, LSTM, and so on, in that it is able to rely on self-attention mechanism to capture the information of different locations in the space, and it can efficiently solve the long-time prediction and parallel computing It can effectively solve the problem of long-time prediction and parallel computation, and has achieved good results in the task of time-series data prediction [12].

The self-attention mechanism is one of the core concepts of Transformer, which enables the to consider all positions in the input sequence simultaneously, rather than processing them incrementally as in an RNN or CNN. The multi-head attention mechanism contains multiple independent attention heads, which are spliced together from multiple independent, parallel mechanisms, each with different weights to capture different types of relationships.The Transformer model consists of two main components, the encoder and the decoder, which can be stacked with an equal number of encoders and decoders, and contains a multi-head attention mechanism comprised of multiple self-attention mechanisms that run in parallel. The Transformer model of two parts, encoder and decoder, and can stack multiple encoders and decoders with the same number, and contains multiple self-attention mechanisms running in parallel.

The self-attention mechanism of the Transformer architecture can be formulated as follows:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V$$

Where: Q, K, V computes the matrix of the attention mechanism, the keys and values of the matrix, respectively; d_k is the dimension of the keys.

In the self-attention mechanism, the input matrix X is linearly transformed to obtain the query

(Q), key (K) and value (V). Firstly, Q is dot-producted with K to yield the similarity, which is equivalent to the projection of Q onto K. To avoid too large values, this result is divided by a constant (usually the square root of the dimension of the key) to make the distribution more stable. The result is then softmax normalised to give weight coefficients that sum to 1. These weights reflect the importance of the elements in the input sequence. Finally, these weights are multiplied with V to obtain the weighted values and the output is the sum of these weighted values. In this way, the self-attention mechanism assigns weights to the input sequence and generates a weighted output. As follows Fig.1

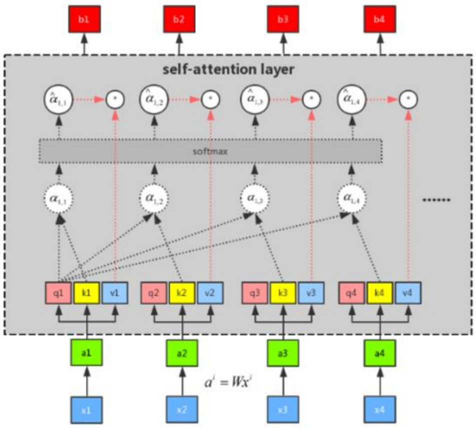

Fig. 1 Schematic diagram of the self-attention mechanism

The multi-attention mechanism can be formulated as follows:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_i)W_0 \quad Head_i = Attention(Q_i, K_i, V_i)$$

Where: *MultiHead* is the multi-attention mechanism; *Concat* is the connection mechanism between the attention; $head_i$ denotes the ith attention mechanism; W_0 is the linear transformation weight matrix after the connection of the multi-attention mechanism; Q_i, K_i, V is the linear transformation weight matrix.

3.2 Analysis of model evaluation

In order to evaluate the performance of the Transformer model for predicting the safety of spontaneous coal combustion, the mean absolute error (MAE) and the coefficient of determination (R2) as well as the root mean square error (RMSE) are used as the indicators to assess the accuracy of the model. The MAE and RMSE can measure the size of the error between the predicted value and the true value, and the smaller the value, the smaller the variability between predicted value and the true value, and thus the higher the accuracy of the model. The smaller the value is, the smaller the discrepancy between the predicted value and the actual value is, and thus the higher the prediction accuracy of the model. The formula of the evaluation index is as follows:

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

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

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (\bar{y}_i - \hat{y}_i)^2 \right)$$

Where: n is the total number of samples; i is the number of samples; y_i is the true value; $\hat{y}_i$ is the predicted value; $\bar{y}_i$ is the mean value.

3.3 Comparative Analysis of Models

Recurrent neural network (RNN) can handle short-term dependency well by processing data through recurrent structure; long short-term memory neural network (LSTM) can handle long-term dependency effectively by adopting input gate, forgetting gate, and output gate to control the flow of information; the prediction results of Transformer are compared and analysed with the two to reflect the prediction accuracy of Transformer model.

4. Example Analysis

4.1 Data Pre-processing

As the data collected in the experiment will be interfered by various factors in the actual production, it is easy to appear abnormal values with large deviation from the real value, and due to the influence of experimental equipment and experimental cost, the amount of trained data is far from meeting the requirements, so the median filtering and Lagrangian interpolation are used for the data to remove abnormal values, expand data and fill in the missing values, in order to improve the quality and usability of the data. Median filtering can replace the current value with the median value in the neighbourhood to remove the effect of impulse noise and keep the data smooth.

$$N_{i,j} = \left\{ x_k \,\middle|\, (k \in N) \right\}$$

$$median(N_{i,j}) = \begin{cases} x_{(n+1)/2} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} \end{cases}$$

$$y_{i,j} = median(N_{i,j})$$

Where: x_k value in the neighbourhood, $y_{i,j}$ is the filtered output value, n is the number in the neighbourhood, *median* is the median function

Lagrange interpolation is a method used to construct a polynomial through a given data point. The Lagrange method is used to expand and fill the data, given n+1 known data points $(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)$, the Lagrange interpolation polynomial is:

$$P(x) = \sum_{i=0}^{n} y_i L_i(x)$$

Where $L_i(x)$ is the Lagrange fundamental polynomial defined as:

$$L_i(x) = \prod_{0 \le j \le n, j \ne i} \frac{x - x_j}{x_i - x_j}$$

Where: x is the original observation, y_i is the value of the known data point, x_i is the value of the independent variable of the known data point, and n is the number of known data points minus one.

4.2 Modelling

The Transformer model contains two important components, the encoder and the decoder, each of which consists of multiple identical layers, each of which includes a multi-head attention layer and a feed-forward neural network layer. The encoder receives a sequence of covariates from the past, $x_{t-1}=\{x_1,x_2,x_3,\cdots x_{t-1}\}$, and a sequence of target variables, $y_{t-1}=\{y_1,y_2,y_3,\cdots y_{t-1}\}$, where t represents the starting moment, while the decoder receives a sequence of covariates from the future, $x_{t:T}=\{x_t,x_{t+1},x_{t+2},\cdots x_T\}$, and T is the predicted ending moment. The output of the model is the sequence of future target variables $y_{t:T}=\{y_t,y_{t+1},y_{t+2},\cdots y_T\}$. Due to the characteristics of the sequence data, it needs to be encoded before it can be input into the model, and this process is divided into data encoding and location encoding, and the use of these two forms of encoding can effectively improve the effectiveness of the Transformer model in time series prediction. As follows Fig. 2.

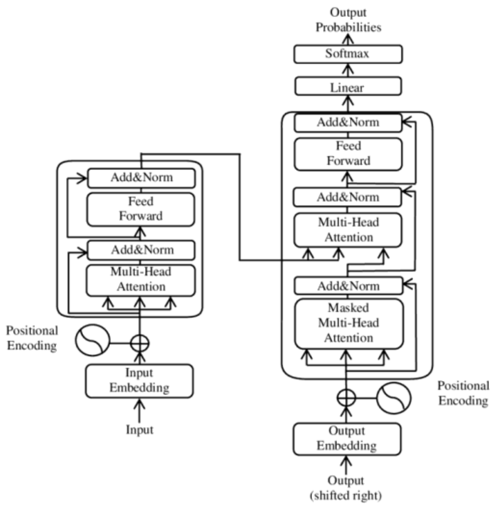

Figure 2 Transformer schematic diagram

Data encoding: Data encoding is the process of converting raw input data into a form that the model can handle, which usually includes input transformation and feature enhancement. We need to input transformation involves converting raw data into a vector representation via the embedding layer, while feature enhancement encompasses normalization, standardization, and other techniques to ensure the data is suitable for model learning.

Position encoding: since the Transformer model cannot directly capture the position information in the sequence data. Therefore, positional encoding is introduced to add positional information to each input element so that the model can sense the relative and absolute positions of the elements in the sequence. Common positional encodings can be classified into three categories: traditional positional encodings, learnable positional encodings and timestamp encodings. For example, Song et al [13] used traditional position coding to develop an attention prediction model when dealing with clinical time-series data; Li et al [8] introduced learnable position embedding in LogSparse Transformer instead of the traditional position coding approach; and Zhou et al [15] incorporated timestamps into the Transformer model that proposed local and global timestamp encoding methods.

The main task of the encoder is to process the input sequence and transform it into a high-dimensional vector representation. Among them, the multi-head attention mechanism captures the complex relationships in the sequence by processing different subspaces in parallel, while residual connectivity and layer normalisation help to improve the stability and training efficiency of the model. The encoder consists of a stack of N identical layers, each layer includes a multi-head attention layer and a feed-forward neural network layer, and the output of the multi-head attention layer in the encoder after residual normalisation is:

$$Z^1 = LayerNorm(Z + Connat(A_1, A_2, A_3 \cdots A_n)W_0)$$

Where: Z is the input, Z^1 is the output and *LayerNorm* is the normalisation function.

The role of the feed forward neural network is to introduce an activation function to transform the representation of each position with nonlinear features in order to process each position independently and to improve the expressive power of the model. The output of the feed forward neural network layer in the encoder is:

$$output = W_2 \cdot f\left(W_1 \cdot x + b_1\right) + b_2$$

Where: W_1, W_2 are the first and second linear layer weights, f is the activation function ReLU, b_1,b_2 are the first and second learning biases.

The decoder's task is to predict the next target sequence based on the encoder's output and the generated target sequence. The decoder also consists of a stack of N identical layers, containing three parts: the multi-head self-attention layer, the encoder-decoder attention layer and the feed-forward neural network layer. In the decoder, the ungenerated positions in the multi-head self-attention layer need to be masked out by a mask, and the self-attention with mask is:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \cdot M$$

The output after residual normalisation is:

$$Z^1 = LayerNorm\left(Z + Connat(A_{M1}, A_{M2}, A_{M3} \cdots A_{Mn})W_0\right)$$

Where: M is the mask operation.

In this paper, the beam pipe data from the fire warning system at the 5103 return corner of the Hongshuliang coal mine serves as the experimental data. Eighty percent of this data is utilized for training, while the remaining twenty percent is used for testing. Based on the dimensions and size of the processed data, the initial parameters for the encoder and decoder in the Transformer model are determined as follows: the number of heads in the multi-head attention mechanism is set to 6, the size of the hidden layer is 128, the size of the internal layer is 64, the dimension of the feed-forward neural network is set to 6, and the activation function employed is ReLU. The size of the hidden layer is 128, the size of the internal layer is 64, the dimension of the feed-forward neural network is 6, and the activation function adopts ReLU.The time series prediction is carried out after analysing the concentration data of CO and O2 in one day and one week, respectively, so as to provide early warning of spontaneous combustion of coal in the mining area, and to take preventive measures accordingly.

5. Discussion of Results

5.1 Analysis of Model Prediction Results

The predictive analyses of O_2 and CO concentrations were selected for the time periods of the day as follows Fig.3-6. The blue line represents the true value, and the red line represents the predicted value.

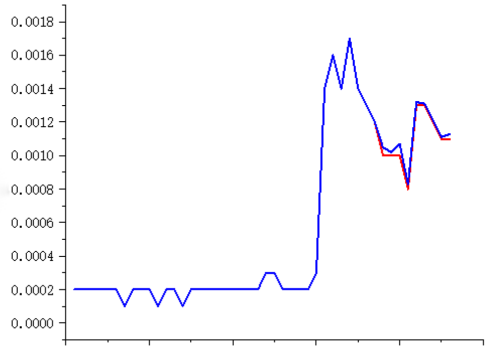

Figure 3 Predicted 24-day CO concentration

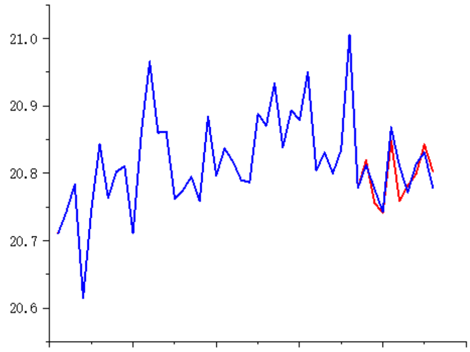

Figure 4 Predicted 24-day O_2 concentration

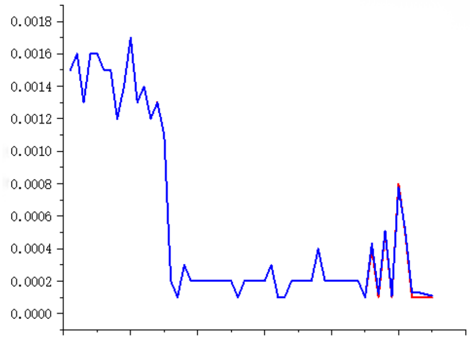

Figure 5 Predicted 25-day CO concentration

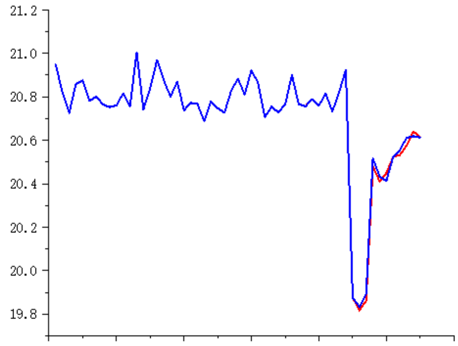

Figure 6 Predicted 25-day O_2 concentration

A one-week time period was chosen to analyse the predicted concentrations of O2 and CO as follows Fig7-8:

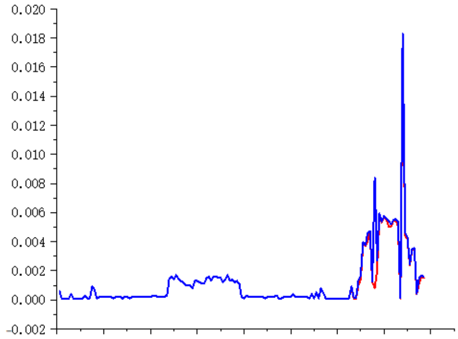

Figure 7 CO concentration prediction

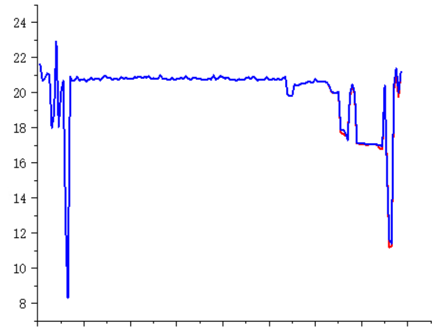

Figure 8 O2 concentration prediction

From the figure, it can be seen that the error between the real CO concentration value and the predicted value of the model is small, so it can be seen that the Transformer model is more effective in predicting the CO, O_2 concentration.

According to the Coal Mine Safety Regulations, the CO concentration should not exceed 0.0024% and the O2 concentration should not exceed 20.5%, and this is used as the threshold value, and then timely precautions are taken according to the prediction results to reduce the occurrence of accidents.

5.2 Comparative Analysis with RNN, LSTM

The prediction performance of Transformer is compared with that of LSTM, RNN by setting different prediction lengths at the same time step, and the results are shown in the Table 1.

Table 1 Errors at 15 time steps

modelling	lengths	MAE	RMSE	R2
Transformer	7	0.000075	0.00000015	99.17
	15	0.000071	0.00000013	99.24
LSTM	7	0.000078	0.00000017	98.47
	15	0.000074	0.00000016	98.52
RNN	7	0.000081	0.00000020	97.64
	15	0.000079	0.00000019	97.81

In terms of MAE and RMSE, the prediction accuracy of the Transformer model is higher than that of LSTM and RNN; and the coefficient of determination R2 of the Transformer model is greater than that of LSTM and RNN; it is evident that the use of the Transformer model for early warning of the risk of spontaneous combustion in coal mines is a highly efficient and reliable prediction model at present. The Transformer model can be effectively applied to coal mine safety.

6. Conclusion

Early warning of coal spontaneous combustion risk is an important means to ensure mine safety, and advanced and effective early warning methods are adopted to provide reliable support for mine safety management. Through the early warning system and timely preventive measures to reduce the occurrence of spontaneous combustion accidents, to protect the safety of miners and production stability, so proposed based on the Transformer model of the coal spontaneous combustion risk early warning method, the main research content is described below.

In this paper, the Transformer model is used to process sequence data to predict the concentration of CO, O2, and through its unique attention mechanism adaptively focuses on the information that has more influence on the results and ignores the unimportant information in order to improve the accuracy and stability of the model prediction.

In terms of data processing, the median filtering method is used for impulse noise reduction of the acquired data, and then the Lagrangian interpolation method is used to expand and fill in the data to improve the quality and availability of the data, which is finally inputted into the Transformer model for time-series prediction.

The experimental results show that, under certain circumstances, the prediction of CO and O2 concentration using the Transformer model can capture the gas changes well, so as to take early warning measures even though, through the example verification, the prediction accuracy of the Transformer algorithm is improved compared with the LSTM algorithm and the RNN algorithm, and it can be effectively used in the coal mine safety early warning.

Acknowledgment

Jidong Yao, Yilong Xiao are co-first Authors.

This article is funded by North China University of Science and Technology 2024 College Student Innovation and Entrepreneurship Training Program Project (No. S202410081056).

References

[1] Luo Zhenmin, Zhang Lidong, Song Zeyang. Realisation of multi-step prediction of CO in mining area based on fully connected long and short-term memory network[J]. Journal of Tsinghua University (Natural Science Edition), 2024, 64(06):940-952.DOI:10.16511/j.cnki.qhdxxb.2024.22.011.
[2] Li Yiman. Research on time series prediction of multiple indicators of coal spontaneous combustion in the mining hollow area[D]. North China Institute of Science and Technology, 2023.DOI:10.27861/d.cnki.gnckj.2023.000002.
[3] Mao Yuanhong, Sun Chenchen, Xu Luyu, et al. A review of time series forecasting methods based on deep learning[J]. Microelectronics and Computers, 2023, 40(04):8-17. DOI:10.19304/J.ISSN1000-7180.2022.0725.
[4] Liang Hongtao, Liu Shuo, Du Junwei, et al. A review of research on deep learning applied to timing prediction[J]. Computer Science and Exploration, 2023, 17(06):1285-1300.
[5] Meng Xiangfu, Shi Haoyuan. A review of time-series data prediction methods based on Transformer model[J/OL]. Computer Science and Exploration, 1-24[2024-11-03].
[6] Ye Mi. Transformer-based multi-feature futures price trend prediction [D]. Nanjing: Nanjing University of Finance and Economics, 2023.
[7] Zhang Shuai, Liu Wenxia, Tang Haoyang, et al. A short-term load forecasting method based on Transformer multi-feature fusion[J/OL][2023-06-07]. Journal of North China Electric Power University (Natural Science Edition):1-9
[8] Zhou H Y, Zhang S H, Pen G J Q, et al. Informer: Beyond efficient transformerforlong sequencetime-seriesforecasting[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 1 1 106-1 1 1 1 5.
[9] Geng Xinyue, Hu Changhua, Zheng Jianfei, et al. Transformer-based residual life prediction of lithium batteries in dual time scales[J]. Space Control Technology and Applications, 2023, 49(4): 119-126.
[10] Wang Yiwen, Lai Jianjun, Qu Zaipeng. Research on short-term blood glucose prediction method based on Transformer[J]. Journal of China University of Weights and Measures, 2023, 34(3): 372-378.
[11] Tian Sheng, Hu Xiao. Vehicle trajectory prediction based on Transformer model[J]. Journal of Guangxi Normal University (Natural Science Edition).2024 , 42 (03)
[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information pro-cessing systems, 2017, 30:5998-6008.
[13] Li S Y, Jin X Y, Xuan Y, et al. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting [C]. NeurIPS 2019: 5244-5254.