Supplementary Information

Gaussian-Linearized Transformer with Tranquilized Time-series Decomposition Methods for Fault Diagnosis and Forecasting of Methane Gas Sensor Arrays

Kai Zhang, Wangze Ning, Yudi Zhu, Zhuoheng Li, Tao Wang, Wenkai Jiang, Min Zeng* and Zhi Yang*

1. Different Kinds of Faults Models

The experimental system of the methane sensor array is composed of six parts: gas path part, gas chamber part, environmental stress simulator, fault trigger, communication module, and upper computer. The gas path part includes a standard air bag and program-controlled gas injection mechanism; The air chamber part includes four parts: sensor array, environmental sensor, humidifier, and heater. The environmental stress simulator includes a vibration generator and swing generator; The fault trigger includes a program-controlled relay, program-controlled potentiometer, and program-controlled voltage device. The communication module realizes bidirectional communication and data transmission with the host computer. As the main platform of the test system, the upper computer is responsible for sending down the control command, collecting the output signal of the sensor array, and processing the fault data. The simulation of real-world sensor faults requires the collaboration of functional components of the experimental system. Typical failure modes of methane sensors include welding, cracking, falling off, open circuit, voltage fluctuation, etc.

In the experiment system, the on-off of the programmed relay can simulate the open lead and open heating wire faults of the sensor in the real world. The program control relay instantaneous on-break on-break on can simulate the sensor lead solder spot welding, and heating wire solder spot welding failure; The cracking and falling off of the sensitive film of the sensor are simulated by the sudden change of resistance value of the programmable potentiometer. The superposition fault of an open heating wire and open lead can be simulated by the combination of a programmed relay and a programmed potentiometer. Programmable voltage controller simulates sensor heating voltage fluctuations, operating voltage fluctuations, etc.

TABLE S1

Different Kinds of Fault Models

Models	1	2	3	4	5	6
Fault description	a. HWD b. ESB c. HWDI d. ESBI e. PESB f. OF	a. ESBI b. HWDI c. ESB d. HWD e. PESB f. Normal	a. HWDI b. ESBI c. HWD d. ESB e. Normal f. OF	a. ESB b. HWD c. ESBI d. HWDI e. OF f. Normal	a. HWDI+HWD b. ESB+HWD c. HWD d. ESB e. OF f. OF	a. ESBI+ESB b. ESBI+ESB c. ESBI+ESB d. HWD e. PESB f. HWD
Training samples	37.2k	37.2k	37.2k	37.2k	37.2k	37.2k
Testing samples	37.2k	37.2k	37.2k	37.2k	37.2k	37.2k

2. Equivalence between self-attention and Gaussian-linearized attention

The output of dot-product attention is，

$$R(Q,K,V) = \delta(QK^T)V \tag{S1}$$

The normalization function of the two choices is,

$$Scaling\colon \delta(X) = \frac{X}{n}$$

$$softmax\colon \delta(X) = \omega_{row}\,(X) \tag{S2}$$

where ω_{row} denotes applying the softmax function along each row of matrix X. Substituting the scaling normalization formula in Equation (11) into Equation (10) gives

$$R(Q,K,V) = \left(\frac{QK^T}{n}\right)V \tag{S3}$$

Similarly, plugging the scaling normalization formula in Equation (8) into Equation (7) results in

$$E''''(Q,K,V) = \frac{\emptyset(Q)}{\sqrt{n}}\left(\frac{(\varphi(K))^T}{\sqrt{n}}V\right) \tag{S4}$$

Since scalar multiplication is commutative with matrix multiplication and matrix multiplication is associative, we have

$$E''''(Q,K,V) = \frac{\emptyset(Q)}{\sqrt{n}}\left(\frac{(\varphi(K))^T}{\sqrt{n}}V\right) \quad (S5)$$

$$= \frac{1}{n}\emptyset(Q)\left((\varphi(K))^T V\right)$$

$$= \frac{\emptyset(Q)(\varphi(K))^T}{n}\mathrm{V}$$

Due to the Gelu function property, if data are non-negative, then Gelu(x) $\approx$ x > 0, and our fault diagnosis and fault forecast are non-negative, so

$$E''''(Q,K,V) = R(Q,K,V) \quad (S6)$$

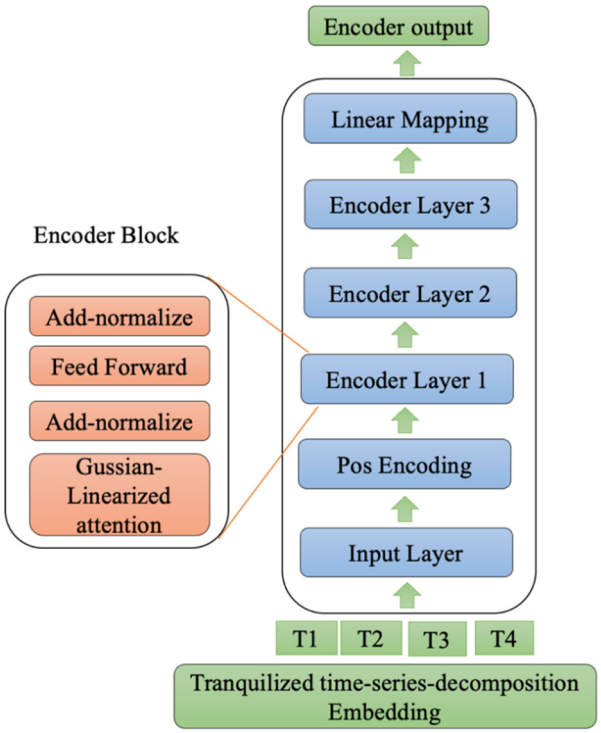

Fig. S1 Encoder block of fault forecast task

3. Sequence length and Necessity of linearized attention

If n is the length of the sequence, d is the "head-size" (base 64) and H is the number of heads (base 12); then, *hd* is "hidden-size" (base 768). For self-attention, the projection transformation of *Q*, *K*, *V*, *n*hd* matrix times the *hd*hd* matrix was done three times, so the calculation is $3n*(hd)^2$. Then, we calculated the *h* attention heads. Each head's $Q(n*d)$ is multiplied by $K^T(d*n)$, and the N^2 attention matrix is then finished (ignore softmax and normalization). Next, the n*n matrix is multiplied by $V(n*d)$ to obtain the $n*d$ matrix. Both of these steps are n^2d, so the total amount of computation is $h(2n^2d)$. The final output has a projection transformation, which is also an *n*hd* matrix times an *hd*hd* matrix, and the computation is $n(hd)^2$, so the total computation for self-attention (SA) is

$$3\mathrm{n}(hd^2) + \mathrm{h}(n^2\mathrm{d} + n^2\mathrm{d}) + \mathrm{n}(hd)^2 = 4\mathrm{n}h^2d^2 + 2n^2\mathrm{hd} \quad \text{(S7)}$$

Regarding a feed-forward network (FFN), two fully connected layers are involved; that is, it is a two-matrix transformation (the activation-function calculation is also ignored), and the general parameter setting is as follows: The first layer is an *n*hd* matrix multiplied by an *hd*4hd* matrix and the second layer is an *n*4hd* matrix multiplied by a 4hd matrix, so the total amount of computation is

$$n \times hd \times 4hd + n \times 4hd \times hd = 8nh^2d^2 \quad \text{(S8)}$$

Thus, if the SA is larger than the FFN,

$$4nh^2d^2 + 2n^2hd > 8nh^2d^2 \quad <=> \quad n > 2hd \quad \text{(S9)}$$

For the base version of the model, this means that *n*>1,536; in other words, only when the sequence length exceeds 1,536 is the calculation amount of SA greater than that of the FFN. Before this, the FFN with linear complexity is dominant.

From the above results, we can obtain the total computation amount of the Transformer layer as

$$4nh^2d^2 + 2n^2hd + 8nh^2d^2 = 12nh^2d^2 + 2n^2hd \quad \text{(S10)}$$

This is the sum of the first and second terms of n, and when n is large enough, it is $O(n^2)$, but only if

$$2n^2hd > 12nh^2d^2 \quad <=> \quad n > 6hd \quad \text{(S11)}$$

For the base model, this means that *n*>4,608; that is, when the sequence length approaches 4,608, the Transformer's complexity is truly quadratic.

Combining the above results, we can conclude that the Transformer's complexity is nearly linear for base versions when the sequence length does not exceed 1,536; when the sequence length exceeds 1,536, the Transformer's computation becomes more attention oriented and its complexity gradually becomes quadratic until the sequence length exceeds 4,608.

For the Gaussian-linearized Transformer model, in the fault-diagnosis task, $d = 16$ and $h = 4$, so when the sequence length is greater than 384 the Gaussian-linearized Transformer model must be linearized. In the fault-forecasting task, $d = 3$ and $h = 3$. Therefore, when the sequence length is greater than 54, the Gaussian-linearized Transformer model must be linearized. The sequences herein are much larger than these, so it is necessary to linearize the model.

4. Correlation between Sensors of Fault Diagnosis task

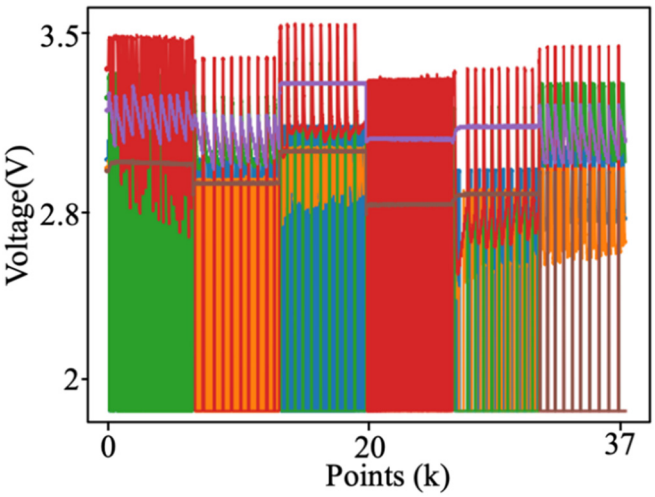

Fig. S2. Sensors fault diagnosis task data

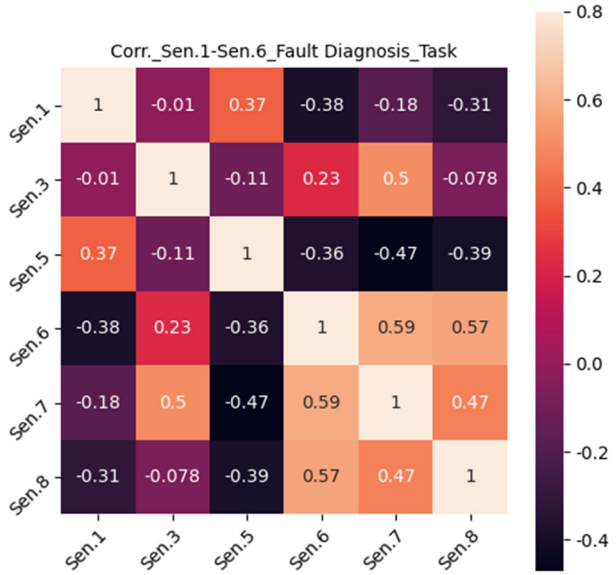

Fig. S3. Correlation between sensors of fault diagnosis task

5. Parameters of Fault Diagnosis Model and Fault Forecast Model

TABLE S2

Parameters of Fault Diagnosis Models

Layer Number	layer type	Layer description	dropout	Mask Number
1	d_{ff} ,d_k ,d_v	128,32,32	no	-
2	d_{model}	16	no	-
3	Multi-attention	Head=4	yes	2
4	Encoder	N=3	yes	2

TABLE S3

Parameters of Fault Forecast models

Layer	layer type	Layer description	dropout	Mask Number
1	d_{ff} , d_k , d_v	128, 3, 3	no	-
2	d_{model}	3	no	-
3	Multi-attention	Head=3	yes	2
4	Encoder	N=2	yes	2