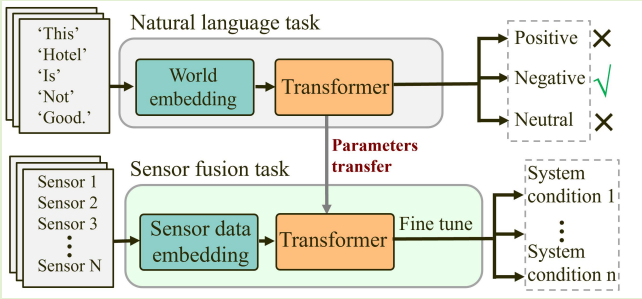
Deep Transfer Learning With Self-Attention for Industry Sensor Fusion Tasks

Ze Zhang, Michael Farnsworth, Boyang Song, *Member, IEEE*, Divya Tiwari, and Ashutosh Tiwari

***Abstract*—Monitoring of complex industrial processes can be achieved by obtaining process data by utilising various sensing modalities. The recent emergence of deep learning provides a new routine for processing multi-sensor information. However, the learning ability of shallow neural networks is insufficient, and the data amount required by deep networks is often too large for industrial scenarios. This paper provides a novel deep transfer learning method as a possible solution that offers an advantage of better learning ability of the deep network without the requirement for a large amount of training data. This paper presents how Transformer with self-attention trained from natural language can be transferred to the sensor fusion task. Our proposed method is tested on 3 datasets: condition monitoring of a hydraulic system, bearing, and gearbox dataset. The results show that the Transformer trained from natural language can effectively reduce the required data amount for using deep learning in industrial sensor fusion with high prediction accuracy. The difficult and uncertain artificial feature engineering which requires a large workload can also be eliminated, as the deep networks are able to extract features automatically. In addition, the self-attention mechanism of Transformer aids in the identification of critical sensors, hence the interpretability of deep learning in industrial sensor fusion can be improved.**



***Index Terms*—Transfer learning, deep learning, natural language processing, sensor fusion, sensor data processing, smart manufacturing.**

I. Introduction

INDUSTRY4.0 or smart manufacturing introduced a clear trend of industrial technology development, which is powered by advanced communication technology and advanced data analytical methods. In such a scenario, the variety and amount of sensor data coming from the production process, products and production machinery may grow exponentially [1]. This often includes state, process, vision, vibration and pressure data etc. Hence, there is a challenge on how to harness sensor data collected from different modalities to extract beneficial information, and that can be used to improve the analysis performance such as useful remaining life (RUL) estimation [1], faults inspection, and diagnosis. Therefore, sensor fusion methods, also referred to as data fusion, can play an important role in solving this challenge. However, in an industrial environment, the nature of data from various sensors is very complex involving different modalities and measuring physical quantities with significantly different sampling rates. This complexity in incoming sensor data introduces difficulties in its utilisation.

Roughly, there are three typical technical routines to employ multi-sensor data to extract useful information. The first type is the model-based method [2]. To use this method, deep domain knowledge is essential, which enable researchers to model the targeting system's behaviour precisely by mathematical and physical formulas [3]. However, building an accurate model for a complex manufacturing process or machine is not an easy task because of the lack of availability of sufficient physical details and the uncertainties involved in some of the processes. The second one is the statistics-based method [4]. This technique makes an inference from sensor data by observing previous system states and analysing the associated sensor measurements, which requires a large amount of good quality observed sensor data and properly estimated noise distribution [5]. Artificial feature extraction is usually needed in these methods to reduce the dimension of input space [6] [7], such as Eigen Decomposition and Principal Component Analysis [8]. The performance could be influenced by the representativeness of the extracted features and how

Manuscript received 1 June 2022; accepted 15 June 2022. Date of publication 1 July 2022; date of current version 1 August 2022. This work was supported in part by the EPSRC through the Future Electrical Machines Manufacturing Hub under Grant EP/S018034/1, in part by the UK Research and Innovation/Engineering and Physical Sciences Research Council (UKRI/EPSRC) through the Made Smarter Innovation-Research Centre for Connected Factories under Grant EP/V062123/1, and in part by the Royal Academy of Engineering (RAEng) and Airbus through the Research Chairs and Senior Research Fellowships Scheme under Grant RCSRF1718/5/41. The associate editor coordinating the review of this article and approving it for publication was Dr. Peng Wang. *(Corresponding author: Ze Zhang.)*

The authors are with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: zzhang99@sheffield.ac.uk; m.j.farnsworth@sheffield.ac.uk; b.song@sheffield.ac.uk; d.tiwari@sheffield.ac.uk; a.tiwari@sheffield.ac.uk).

Digital Object Identifier 10.1109/JSEN.2022.3186505

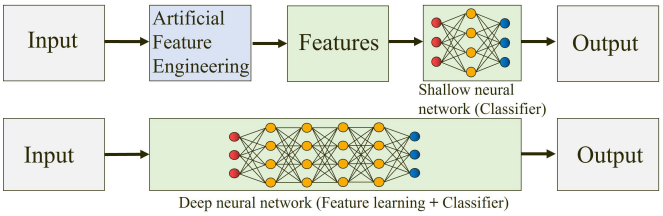


Fig. 1. Comparison between shallow neural network plus artificial feature engineering and deep neural network which can extract features automatically.

good the estimation of noise distribution is. The third method involves the emerging area of Artificial Intelligence, especially Deep Learning [9] [10]. A detailed domain knowledge is not required in this method as it is the case in model-based approaches. The amount of effort on feature engineering which may restrict the performance of statistics-based method can also be reduced [11], as Deep Learning is able to discover inherent features automatically [12]. Hence, Deep Learning may enable more data scientists without sufficient domain knowledge to contribute to manufacturing domains.

However, when dealing with a large and complex input space, such as a large number of sensors with different sampling rates, more complex and deeper networks are often necessary, as deeper networks have stronger feature extraction capabilities [13]. This leads to a significant increase in the depth and width of the model, resulting in high demands on the amount of training data. For industrial scenarios, collecting a large amount of training data means a significant increase in time and expense, which is sometimes not even possible. Alternatively, artificial feature extraction, such as time-domain statistical feature extraction [14], frequency-domain feature extraction [15], plus a shallow neural network [16] [17] can be used to reduce the need for training data. For such a method, the performance is heavily dependent on the quality of manually extracted features, and feature engineering often requires extensive experimentation and an understanding of industrial processes, which can be time-consuming and difficult. Therefore, a method that requires a relatively small amount of data and does not rely on feature engineering is preferred. The hard choices mentioned above can be described by Fig. 1.

In addition, lack of interpretability is another disadvantage of Deep Learning in industrial applications. The mechanism by which input data is mapped to model output of Deep Learning is difficult to obtain. Hence, while using Deep Learning, it is difficult to know which sensor has the decisive influence in a multi-sensor system.

In this paper, we propose a novel transfer learning methodology for sensor fusion that transfers a deep model from the natural language processing (NLP) domain, coined 'Transformer' [18] to industrial applications. NLP is a data-rich domain, deep models in this domain are relatively easy to be adequately trained. Lu *et al.* found that, the feature extraction ability of such a complex deep model have the potential to be transferred to other modalities, since natural language is a modality with a huge amount of data and features [19]. They found that pretrained Tranformer with fine-tuning offers great performance on numerical operations, image classification and protein folding prediction tasks. Therefore, transfer of feature representations identified from NLP to sensor fusion problems, could enable the application scenarios, with high input dimension but insufficient data, to train very deep neural networks from raw to use deep networks. Thus, the proposed method is expected to make use of the strong learning ability of the deep neural network and the advantage of the training ease of the shallow neural network to improve the sensor fusion task for manufacturing. This can be achieved by freezing all the attention and feed forward layers and training the input embedding, layer norm, and final full-connected output layers only. It means that the computation of feature extraction which are inherent in natural language will be transferred to sensor fusion tasks.

In addition, the self-attention mechanism used by Transformer can provide insights into the final decision made by the deep network and obtain the weight information, namely the attention map, of sensor data. This can be used for the identification of critical sensors and hence make up for the lack of interpretability when using the deep learning model in industrial sensor fusion tasks.

The main contributions of this paper include:

- A novel deep transfer learning solution that generalises the feature representation from a data-rich modality, in this case, NLP, to address challenges in sensor fusion. This work demonstrates that when using transfer learning in industrial scenarios, collecting data from similar industrial processes for pre-training may not be necessary. The proposed transfer learning method can effectively reduce the required amount of data when using deep learning for industrial sensor fusion tasks, thus benefiting from the learning ability of deep models and to some extent, eliminating the trade-off between deep and shallow models mentioned above. To the best of my knowledge, it is the first work that uses the model trained from language to solve the industrial sensor data processing problem.
- The problem of poor interpretability when using Deep Learning in sensor fusion tasks is alleviated. Based on the attention mechanism, the decision basis of the deep learning model can be inspected, thus the key sensors that are highly related to the final decisions can be identified. Instead of analysing data from all sensors, it allows for a narrower analysis during diagnostics.
- A novel deep learning solution to automatically establish a unified feature representation and association relationship for the sensor data at significantly different sampling rates from different modalities. For conventional sensor fusion methods, decision-level fusion is a better option if the types of sensors are significantly different [8]. However, the challenge for using decision-level fusion is that, a large amount of features which can be highly correlated have to be created. This could bias the final decision, and these features have to be processed properly [8]. In our proposed method, instead of utilising the raw data directly, the sensor data with different sampling rates will be combined and mapped to an embedding space, and the

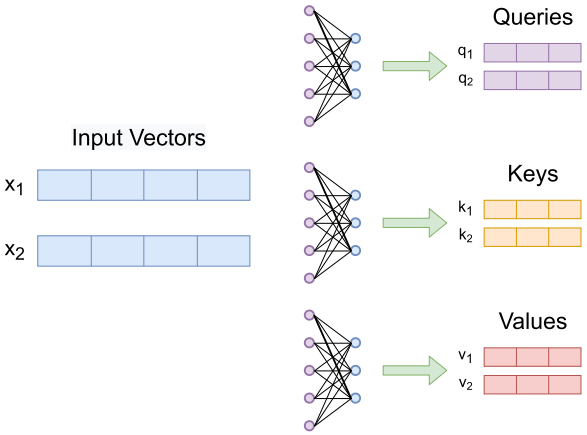


Fig. 2. Queries, keys, and values calculation.

features among different sensors can be extracted from the embedding space by a deep neural network automatically. Hence, the challenge of decision-level fusion can be avoided, the need for artificial feature engineering can be eliminated, and the sensors with different sampling rates can also be combined easily.

The remainder of this paper is organised as follows. In section II, the related research, namely Transformer and its self-attention mechanism will be introduced. In section III, the theoretical basis and detailed steps of the proposed method are explained. In section IV, the proposed method is evaluated by three public datasets from the industrial scenario. The discussion of the results of the experiments is in section V and the conclusion of this paper is in section VI.

II. Related Research - Self-Attention Mechanism and Transformer

Inspired by the fact that humans tend to focus only on the key information in the field of vision, the Attention Mechanism (AM) first appeared in the field of computer vision and gradually became a hot topic [20] [21]. Then, Bahdanau et al applied the AM to Recurrent Neural Network (RNN) to process natural languages. They found that the AM not only visualized deep learning models to some extent but also addressed the fatal flaw of RNN: the forgetting problem when processing long sequences [22]. Later on, researchers found that a deep model built entirely on the principles of the AM improved the performance in machine translation tasks considerably [18]. Here the RNN architecture was abandoned as it was believed that the forgetting problem was rooted in the large number of iterations of RNN, this model was named a Transformer [18].

There are 3 types of AM: self-attention [18], global/soft attention [23], and local/hard attention [24]. This work focused on the self-attention mechanism for industrial sensor fusion tasks. The self-attention could be analogous to retrieval systems: a query vector is used to search information and then the search engine will try to look for the keys in its database and pair the query vector, finally the value vector corresponding to the keys will be the output. In the self-attention mechanism,

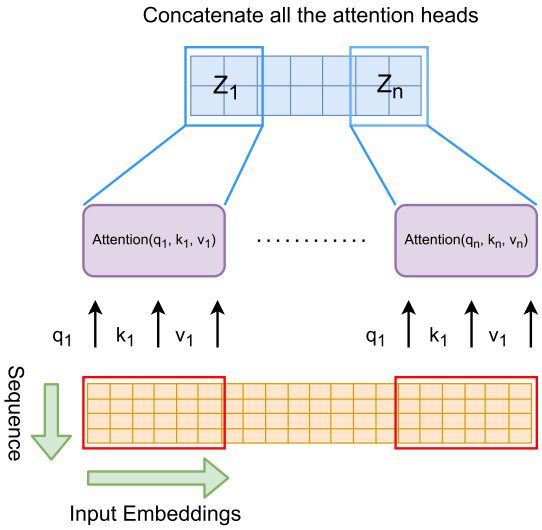


Fig. 3. Multi-heads attention [18].

the input sequences are mapped to the query vector, key vector and value vector by a linear layer as shown in Fig. 2 and try to find the optimized mapping matrix (the weights of these linear layers) in the backpropagation process. Then, the attention weight is calculated by:

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

where Q and K are the query and key vectors respectively, d_k is the dimension of key vector and $1/\sqrt{d_k}$ is the scalar which is used to avoid the dominant term when calculating the softmax function, which may make the gradient difficult to calculate [18]. If Q and K are independent and conform Gaussian distribution: $N(0,1)$, the variance of their dot product, $q \cdot k = \sum_{i=1}^{d_k} q_i k_i$, will be d_k. This effect is not preferable when addressing high dimension data. This AM is called Scaled Dot-Product Attention [18]. Finally, the attention weight is multiplied by the value vector and the final weighted mapping is obtained as shown in equation 2,

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where V is the value vector. It is obvious that the attention weights control the information flow within the network. When the weight of a certain area of input data becomes zero, no information can flow to the next layer of the network. In addition, each element in the attention output sequence is the attention evaluation result of the entire input sequence. Hence, we can know how important each element in the input sequence is against each element in output. As a result, by visualizing the attention weight, to some extent, the basis of network decision-making could be known.

The Transformer method uses multiple isolated attention heads to chunk input data and concatenates the output of each attention head to constitute the attention layers as shown in Fig. 3. After that, multiple attention layers are stacked

to form the Transformer. It consists of an encoder and a decoder. As Transformer are designed to deal with translation tasks, namely the Seq2Seq task, its decoder must be prevented from seeing future information. For example, when outputting the target language translation text verbatim, if the complete input text is obtained when the first few words are output, the translation task will become a memory task [18]. Hence, a masking mechanism is applied to prevent this. For example, if the calculated attention weight matrix W_{Att} and the created mask are:

$$W_{Att} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nn} \end{bmatrix}, \; Mask = \begin{bmatrix} 1 & \cdots & -\infty \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \quad (3)$$

Then, the masked attention can be calculated by the following equation:

$$Attention\_masked = Attention \odot Mask \quad (4)$$

where $\odot$ is the element-wise multiplication. Since the upper right corner of the attention matrix has become negative infinity, it becomes zero after the softmax is calculated.

III. Research Methodology

A. Similarities Between Natural Language Processing and Sensor Fusion

Typically, when deep learning is used to deal with NLP problems, we often map the words of a certain language into an embedding space which could be regarded as the unique identification information of each word. Transformer can automatically learn not only the internal features of each word with the help of multiple attention heads but also the features of the relationships among the input word sequence on a single head of attention as shown in Fig. 3 [18]. This learning mechanism is preferred by sensor fusion tasks, where we want to establish a unified feature representation that includes both inter-sensor and intra-sensor information, where the inter-sensor information means the information contained in the interrelation among different sensors and the intra-sensor information means the information contained in a single sensor.

In NLP tasks, the model needs to recognize the meaning of a single word, and it also needs to infer the information contained by a sentence or a paragraph. Similarly, in a multi-sensor system, the information of each sensor could be compared to a single word in NLP, and the information of all sensors can be regarded as a sentence or paragraph. If the multiple sensor data from different modalities can be mapped to a mapping space with a unified format, the extraction of multi-sensor information could also be analogized to the understanding of paragraph semantics as shown in Fig. 4. This gives the possibility to use NLP models on sensor fusion tasks. However, finding an optimised embedding space is not an easy task. In this paper, we use a linear layer to learn the mapping rules automatically instead of designing mapping rules artificially.

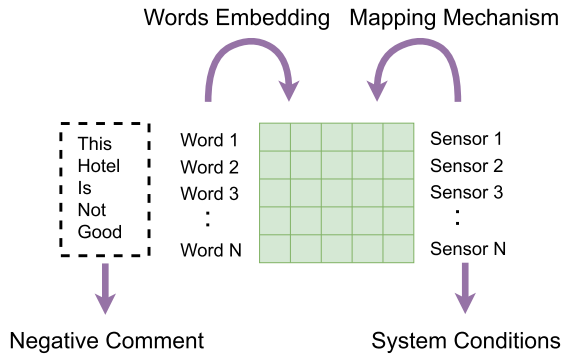


Fig. 4. Sensor data mapping.

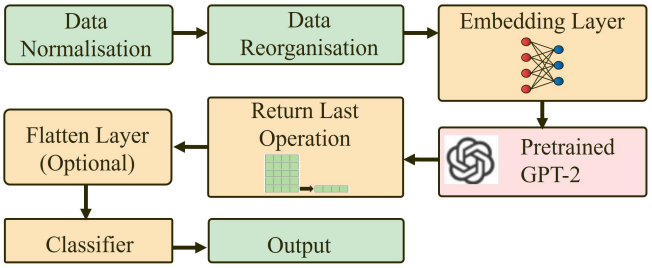


Fig. 5. Proposed model architecture.

B. Proposed Model Architecture

The architecture of the proposed model can be described as shown in Fig. 5. Each part of this model will be explained in the following items.

1) Data Normalisation: As the value range of different sensors will vary by orders of magnitude, in order to facilitate training, the data from each sensor will be normalized using the following formula:

$$X_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (5)$$

2) Data Reorganisation: The input of the Transformer is a two-dimensional matrix. Each row is a word embedding vector and the two-dimensional input matrix is formed by stacking the embedding vectors of each word. In this research, word embedding is replaced with sensor data, hence different rows of input are different sensor data. Since different sensors usually have different sampling rates corresponding to the nature of the object being measured, in order to maintain the same embedding vector length, data with a high sampling rate need to occupy multiple embedding vectors, as shown in Fig. 6. To keep the length of the embedding vector the same, the sampling rates of the sensors need to be in multiples, and to achieve that some sensor values may be discarded or data padding may be used. The reorganised input space can be described by:

$$\mathcal{X} \in \mathbb{R}^{d \times L} \quad (6)$$

where d is the sequence length, and L is the dimension of each element in the input sequence.

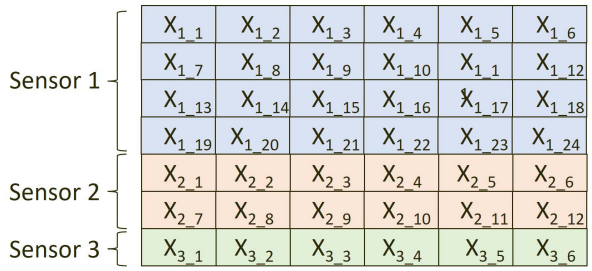


Fig. 6. Example of data reorganisation using 1 second sensor data as a input data point (Sampling rates: sensor 1: 24Hz, sensor 2: 12Hz, sensor 3: 6Hz).

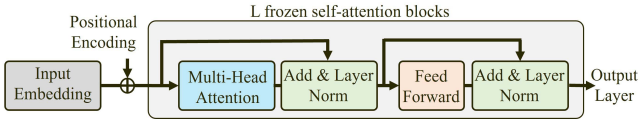


Fig. 7. GPT-2 architecture.

3) Embedding Layer: The mapping of sensor data to the embedding space is done by a linear layer with a dropout rate of 0.1 and orthogonally initialised weights. The mapping rules will be automatically learned as the parameters are updated during the training process. At the same time, this layer also has the functions of preliminary feature extraction and dimension adjustment for the upcoming layers. The embedding layer can be described by the following equation:

$$\text{Embedding}(X) = ReLU\left(XW_{embedding} + b_{embedding}\right) \quad (7)$$

where $W_{embedding}, b_{embedding}$ are the weight matrix and bias matrix. ReLU(*) is the activation function. $X \in \mathbb{R}^{d\times L}$, $W_{embedding} \in \mathbb{R}^{L\times L_{embedding}}$, and $b_{embedding} \in \mathbb{R}^{d\times L_{embedding}}$. $L_{embedding}$ is the embedding dimension which is used for matching the downstream layers.

4) Generative Pre-Trained Transformer 2 (GPT-2:) In this research, a Transformer is expected to be the feature extraction engine, therefore only its decoder part is required. OpenAI provides with a pre-trained decoder of Transformer called GPT-2 with 1.5 billion parameters [25]. It is trained from a 40GB non task-specific training dataset which was crawled from 8 million web pages. In sensor fusion tasks for industrial scenarios, it is difficult to train a deep model with a large amount of data, however, this is not very difficult to achieve in NLP scenarios. The architecture of GPT-2 is shown in Fig. 7 [26]. It is composed of 12 attention layers and each attention layer consists of a 12 heads attention which is used to generate an attention map, two shortcut connections with layer normalisation, and a fully connected feed-forward network. The input dimension of GPT-2 is (N, 768), where N is sequence number, namely the number of rows of the input matrix. For example in Fig. 4, there are 5 rows, so the sequence number is 5. The second number 768 is the size of each row. Similar to the embedding layer described earlier, each row in this sequence will share the same feed-forward network.

In this proposed method, similar to [19], 99.9% of the GPT2 model parameters which were trained from the natural language dataset were frozen. Only the normalisation layer, linear input and output layer were fine-tuned with a learning rate of 0.001 as this learning rate can balance the training speed and stability based on practice, and Adam was used as its optimiser.

5) Return Last Operation: Because of the mask mechanism of GPT-2, only the output result at the last position of the output sequence contains all the input sequence information, which could be compared to the last hidden state in RNN. Therefore, since our work is not a Seq2Seq task, only the last position in the output sequence will be used as the input of the classifier.

6) Flatten Layer (Optional): Depending on the sampling rate of the sensor, if the data of the last sensor occupies multiple rows in the sequence, we need to return all the rows of this sensor, because the expected output result is the inter-sensor information. Thus, in this case, a flatten layer is needed.

7) Classification Layer: After the above calculations, we have obtained a unified feature representation of the fused sensor data. The last layer of this model is a feed-forward neural network for classification. The last position in the output sequence of GPT2 attention block is the input of the classifier, and the number of output neurons is the number of classes. This work only used a single-layer feed-forward network, and the cross-entropy loss was used as its loss function.

IV. Experiments and Results

Based on the description in the last section, the proposed model consists of an embedding layer, GPT-2, and a classifier. As we transferred the parameters of the pre-trained GPT-2, the GPT-2 has to be kept to its original structure. Since the input and output dimensions of GPT-2 are (N, 768), the output dimension of the embedding layer and the input dimension of the classifier were set to (N, 768) to match the dimension of GPT-2. The input dimension of the embedding layer depends on the dimension of the specific dataset and the output dimension of the classifier can be determined by the target categories of specific tasks.

The hyperparameters and model configuration can be found in Table I. The feed-forward layer and multi-head attention layer were frozen, as these two parts contain all the features learned from natural language. The layer normalisation was set to trainable since the data distribution is different for different datasets. The learning rate was set to 0.001 because it was found in practice that smaller learning rates lead to slow training and larger learning rates lead to increased instability. The optimiser and initialisation remained the same as those used in training the original GPT-2 model [25].

The proposed framework was tested on three different datasets based on condition monitoring of a hydraulic system, bearing condition of an electric motor, and gear and bearing working conditions of a gearbox. The experimental results from the three mentioned datasets are provided in this section.

TABLE I
MODEL PRAMETERS AND TRAINING DETAILS

	Experiment	Condition monitoring of a hydraulic system	Bearing condition	Gearbox condition
Pretrained GPT2 model	Return last operation	True	True	True
	Position embedding	None	None	None
	Layer normalisation	Trainable	Trainable	Trainable
	Multi-head attention	Frozen	Frozen	Frozen
	Feed forward layer	Frozen	Frozen	Frozen
	Flatten layer	False	True	False
Training	Parameter initialisation	Orthogonal initialisation (Gain = 1.41)	Orthogonal initialisation (Gain = 1.41)	Orthogonal initialisation (Gain = 1.41)
	Learning rate	0.001	0.001	0.001
	Optimiser	Adam	Adam	Adam
	Batch size	8	64	16
	Loss function	Cross entropy	Cross entropy	Cross entropy

Algorithm 1 Self-Attention-Based Deep Transfer Learning Method for Sensor Fusion

Input: Multi-sensor data samples and the corresponding labels

Step 1 (Data preprocessing):
Normalising the multi-sensor data:

$$X_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Constructing multi-sensor inputs to $\mathcal{X} \in \mathbb{R}^{d \times L}$ for Embedding.

Step 2 (Sensor data embedding):
Constructing the embedding space:

$$\text{Embedding}(X) = ReLU\left(XW_{embedding} + b_{embedding}\right)$$

Step 3 (Transfer learning and feature extraction):
Construction the feature representation by GPT-2 pretrained from language dataset as shown in Fig. 7.

Step 4 (Return Last Operation):
Return the last element of the output sequence.
if *The last sensor occupies more than 1 row of embedding space* **then**
  Do flatten operation and return this vector
else if then
  Return the last row of output sequence

Step 5 (Fault classification):
Constructing the fault classifier, which is a single fully connected layer:

$$\text{Classifier}(X) = XW_{classifier} + b_{classifier}$$

Training the model with cross-entropy loss.

Output: The target categories.

A. Experiment 1: Condition Monitoring of a Hydraulic System

1) Task Description: This classification task requires determining the operating conditions of a complex hydraulic system based on 17 sensors with different sampling rates (7 x 100Hz sensors, 2 x 10Hz sensors, 8 x 1Hz sensors). The dataset was created by Helwig *et al.* [27] on a test rig that was able to simulate a reversible degradation of system performance. This hydraulic system was composed of a primary working circuit and a secondary cooling circuit. Different loads were cyclically applied to a pre-defined work cycle. The data was recorded in every one minute snapshot, therefore the total number of attributes for one data snapshot will be $8(sensors) \times 60(s) \times 1(Hz) + 2(sensors) \times 60(s) \times 10(Hz) + 7(sensors) \times 60(s) \times 100(Hz) = 43680$. This experiment was repeated 2204 times, hence the dataset had 2204 snapshots included.

TABLE II
EXPERIMENT 1-OPERATING CONDITIONS

Task	System conditions
1 Cooler condition	(1) Close to failure (2) Reduced efficiency (3) Full efficiency
2 Valve condition	(1) Optimal switching behaviour (2) Small lag (3) Sever lag (4) Close to total failure
3 Internal pump	(1) No leakage (2) Weak leakage (3) Sever leakage
4 Hydraulic accumulator	(1) Optimal pressure (2) Slightly reduced pressure (3) Sever reduced pressure (4) Close to failure
5 Stable flag	(1) Condition were stable (2) Non-static conditions

This experiment has a total of five tasks to determine the operating conditions of the five different parts of the system as shown in Table II. This system may have multiple faults at the same time.

2) Data Organisation: The sensors have three different sampling rates, 1Hz, 10Hz, and 100Hz respectively. The one-minute snapshot data of the 1Hz sensor is used as the length of the embedding vector (60 columns), hence the 10 Hz and the 100 Hz sensors occupy 10 and 100 embedding vectors respectively. Therefore, one snapshot of input data from these 17 sensors, total 43680 attributes, will be reshaped as a matrix with 728 rows and 60 columns.

3) Results of Prediction Accuracy: The experimental results of model prediction accuracy and comparison with other research are shown in the Table III. In this table, each

TABLE III
EXPERIMENT 1-ACCURACY COMPARISON

Research group	Method	Cooler	Valve	Pump	Accumulator	Stable flag	Mean (without stable flag)
Helwig et al. [27] (2015)	LDA	100%	100%	98.0%	90.4%	N/A	97.1%
Berghout et al. [28] (2021)	Auto-NAHL	100%	100%	100%	96.4%	N/A	99.1%
Wu et al. [29] (2020)	EGMSVMs	100%	100%	100%	76.5%	N/A	94.1%
Gupta et al. [30] (2021)	SECM	N/A	N/A	N/A	N/A	N/A	92.3%
Lei et al. [31] (2019)	PCA+XGBoost	N/A	96.58%	N/A	N/A	N/A	N/A
Huang et al. [16] (2021)	Deep CNN	100%	100%	99.0%	99.4%	N/A	99.6%
Prakash et al. [32] (2020)	ANN+XGBoost	99.54%	N/A	N/A	N/A	N/A	N/A
Our results	**Our method**	**100%**	**100%**	**98.2%**	**91.4% (96.4%)**	**94.4%**	**98.7%**

percentage means accuracy which can be described by the following equation:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{8}$$

The reason for using accuracy to assess model performance is that there is no order of magnitude difference in the number of samples in each category in this dataset. The accuracy in brackets in this table means the accuracy after sensor selection has been applied, and this will be explained in details in the next section. Based on Table III, it can be found that our proposed method achieves an accuracy of 98.7%, ranking in the top three of published work in recent years based on this dataset.

4) Interpretability: Key Sensor Identification Based on Attention Mechanism: In a multi-sensor system, a large number of sensors can be employed, however, not all of them will have an impact on the final decision. When we are dealing with a complex industrial system, it can be difficult to identify critical sensors. Using too many redundant sensors may increase the complexity of the system, waste communication bandwidth and increase the computational burden. Hence, deep learning methods that are interpretable to some extent are preferred by the industry.

The concept of interpretability of deep learning models can be divided into the following two aspects:

- Models are transparent to humans. The corresponding model parameters and the model decisions can be predicted before the model is trained for a specific task [33].
- Decision interpretability. After a model makes a decision, humans can understand the reasons for that decision [34].

This subsection shows the decision interpretability of the proposed method.

In general, AMs are used to control the information flow of deep networks. In the process of backpropagation, the part of the data that has less impact on the results will be masked gradually, and only the information that is decisive for the final decision will be retained. The importance of input sensor data is reflected by the attention weights [35]. Hence, compared with the conventional black-box deep learning models that give no information on which sensors it relies, the key sensors can be identified by visualising the attention weights. As shown in Table III, the classification accuracy of the fourth task is only 91.4%. This indicates that there may still be room for improvement. Therefore, we chose this task to show the key sensor identification capability of AM.

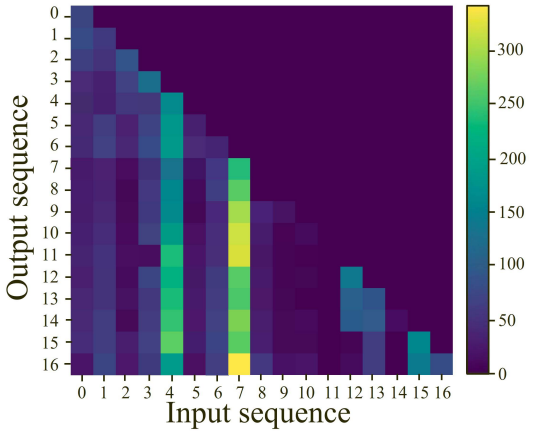


Fig. 8. Attention heat map for accumulator task of experiment 1.

As the data from 17 sensors are reorganised into 728 vectors of 60 columns, the original attention heat map is a matrix of 728 rows and 728 columns. The x-axis represents the input sequences and the y-axis represents the output sequences, for example, the first row is the attention scores of the first element in the output sequence corresponding to all inputs. The dark blue colour means these part of data is masked and no information can flow to the next layers. As mentioned in previous section, the input sequence has 728 vectors of size 60. Since the sensors with different sampling rates occupy different number of input vectors, attention scores for input vectors from the same sensor need to be added together. The processed attention heat map is shown in Fig. 8. As mentioned above, the last row of the processed attention heat map, in this case, the 17th row, is the output attention scores based on all of these 17 sensors.

The attention weights shown in the 17th row can be visualised as Fig. 9. Based on this information, the sensors are divided into two groups, highest attention weights group and lowest attention weights group. 8 of the sensors are contained in the highest attentions weights group, whereas the another 8 sensors are contained in the lowest attention weights group. Then, instead of using all these 17 sensors, the model is trained by these two groups separately. The training history is shown in Fig. 10 and the accuracy of using the top 8 attention sensors and the last 8 attention sensors are 96.4% and 83.0%.

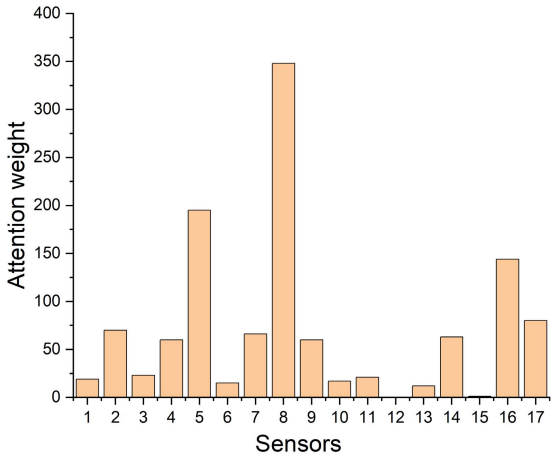


Fig. 9. 17 sensors attention weights.

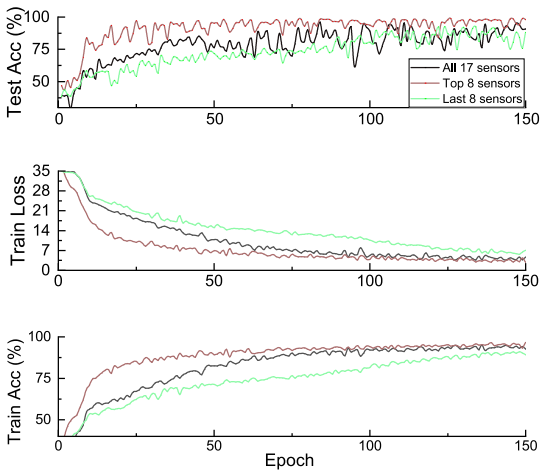


Fig. 10. Training history of 2 groups of sensors vs. Using all sensors.

It can be found that using the highest attention weights group only to train the model can obtain better results than using all the sensors, which means faster convergence, higher accuracy and less fluctuation. The reason for this phenomenon may be that selecting only the more important sensors can effectively reduce the dimension of the input space and thus reduce the difficulty of feature learning of the model. This result illustrates the key sensors identified by the proposed method contain sufficient information for decision making. As the proposed method provides access to the decision basis of the model, it has a higher degree of interpretability than traditional black-box models.

5) Model Performance Tests Under Different Amount of Training Data: This section compares the performance difference under different amount of training data between the pretrained model and scratch model to evaluate whether the parameters trained from natural languages can help to reduce the necessary amount of training data. In this experiment, the training dataset was trimmed to 6 subsets with different size, 100%, 80%, 60%, 40%, 20%, and 5% of the original size respectively. The testing dataset for model evaluation was kept the same as in the Experiment 1: Condition monitoring of a hydraulic system subsection 3) for all tests.

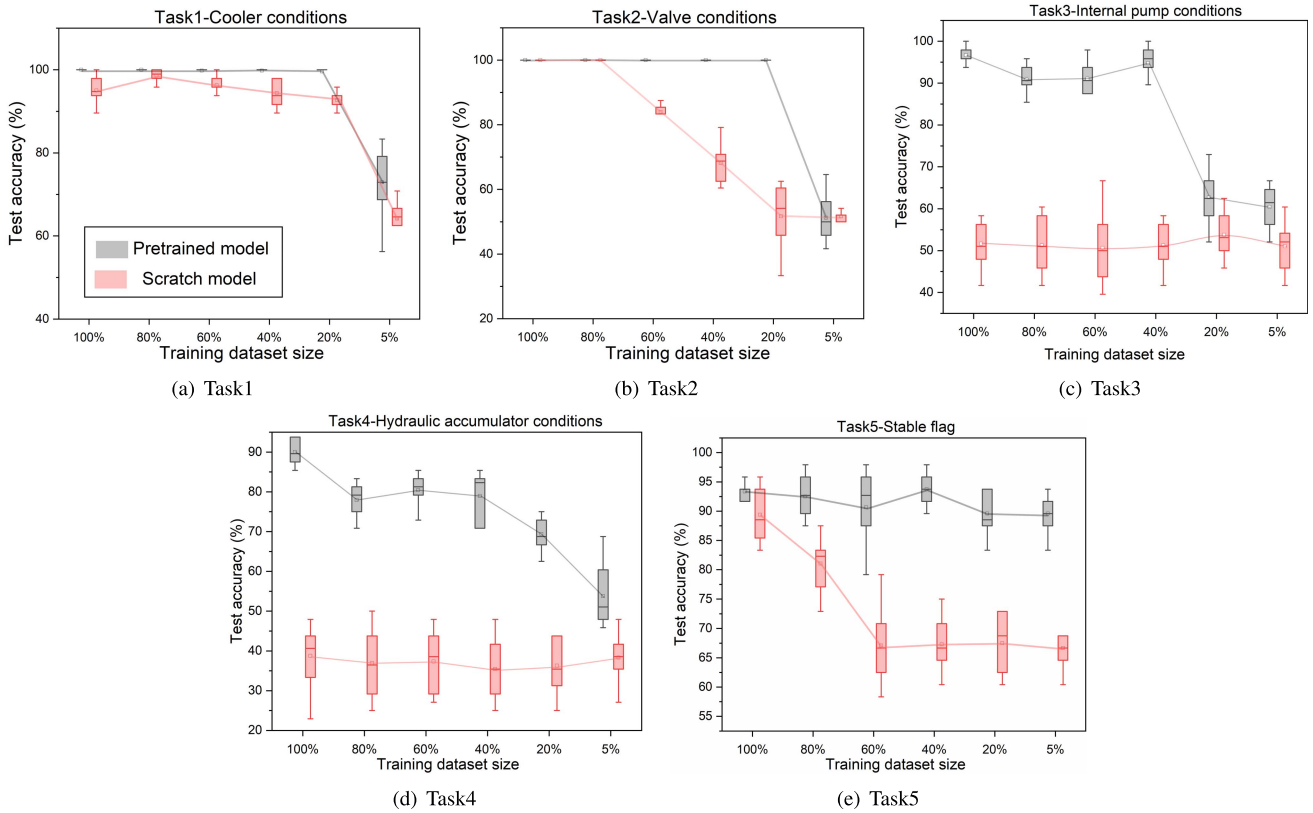
TABLE IV
EXPERIMENT 1-METHOD COMPARISON

Research group	Feature engineering	Dimension reduction	Transfer learning
Helwig et al. [27]	Yes	Yes	No
Berghout et al. [28]	Yes	Yes	No
Wu et al. [29]	Yes	Yes	No
Lei et al. [31]	No	Yes	No
Huang et al. [16]	No	Yes	No
Prakash et al. [32]	Yes	Yes	No
Our method	No	No	Yes

The test results are shown in the box plot shown in Fig. 11. It can be found that regarding to the prediction accuracy and its stability, the pre-trianed model outperforms the scratch model for all 5 tasks under all different training data amount. In task 1 as shown in Fig. 11 (a), although there is a significant performance drop when the size of training data is reduced from 20% to 5% for both models, the lower accuracy and larger fluctuation of the scratch model can be observed. In task 2 as shown in Fig. 11 (b), when the amount of training data is greater than or equal to 80% of its original size, these two models perform almost the same. However, the performance of scratch model decreases obviously when the train data size shrinks to 60%, while the pre-trained model keeps stable until 20% of training data. As for task 3 and 4, the scratch model fails to capture enough information to predict system conditions as shown in Fig. 11(c) and (d). In terms of task 5 as shown in Fig. 11(e), compared with the fact that the performance of the pre-trained model is still relatively high even at 5% of training data, the scratch model degrades remarkably after shrinking the size of training data to 80%.

In summary, the model transferred from natural language domain can effectively reduce the necessary amount of training data when using deep learning in industrial sensor fusion domain. This means the workload and the time consumption for collecting industrial data can be effectively saved. Current transfer learning solution for industrial sensor fusion requires similar industrial process data to pre-train the deep learning model [36], [37]. Our proposed method proves that similar industrial process data may not the only option to perform transfer learning for industrial sensor fusion. Hence, the limitations of using deep learning in industrial scenarios can be reduced significantly.

6) Discussion of Results and Comparison of Performance With the Alternative Methods: The accuracy comparison of the works published in recent years can be found in Table III, and the comparison of the characteristics of these different methods can be found in Table IV. As the method proposed by Gupta *et al.* was focused on detecting system degradation earlier, and it is different from the condition monitoring focused on in this paper, their work is not included in the comparison. As shown in Table IV, manual feature engineering is required for all methods except the methods proposed in [31], [16], and our proposed method. In contrast to other methods, our approach also does not require dimension reduction. This demonstrates the end-to-end nature of our proposed approach.



(a) Task1 (b) Task2 (c) Task3

(d) Task4 (e) Task5

Fig. 11. Classification accuracy under different amount of training data.

In [27], Helwig *et al.* extracted a large number of features in the time and frequency domain for all sensors, and for each extracted feature they analysed its correlation with system faults by Pearson's correlation coefficient and Spearman's rank correlation coefficient. Then, only the highly relevant features are used as the input of their classifier called Linear Discriminant Analysis (LDA). Their accuracies for different tasks shown in Table III were achieved by different combinations of features obtained by different correlation analyses. This means that when working on a new task, a large number of feature combinations have to be traversed to find the best model. Similar feature engineering can also be found in [28] and [29]. In [32], Prakash *et al.* used XGBoost technology to select features before feeding data to a shallow neural network. The studies mentioned above all rely heavily on the use of artificial features to represent information from multiple sensors, thus combining the sensors with different sampling rates and reducing the dimension of the input space to ensure that the complexity of the input space does not exceed the capacity limit of the classifier. In contrast, in our proposed method, manual feature extraction is not required. On the one hand, the sensor data will be mapped to a unified embedding space, thus combining sensors with different sampling rates. On the other hand, the deep learning model trained from natural language, as feature extractors, can extract and select the features from the embedding space automatically. Hence, it can significantly reduce the workload and avoid the difficulty of artificial feature engineering.

In terms of the research of Huang *et al.* in [16], they used multiple independent parallel convolutional neural networks to extract features for each sensor. The output of each of the convolutional neural networks was kept the same, thus, the sensors with different sampling rates can be combined and automatic feature extraction was also achieved. They have achieved excellent accuracy on this dataset. However, such a network structure widens significantly as the number of sensors increases. Cohen *et al.* pointed out that it is necessary to increase the network depth if the width increases [13]. As the size industrial dataset is usually limited, the depth of a network that can be trained is also limited. Therefore, the number of input sensors has to be reduced to prevent the network from being too wide. Huang *et al.* noted that the large input dimension of their model was unacceptable as this will lead to training failure [16]. Hence, they reduced the input dimension from 43680 to 6000 based on artificial sensor selection. In contrast, our proposed method is relatively more insensitive to high input dimensions. Since the amount of data in natural language is very large, this allows for training deeper models and therefore it can handle higher input dimensions compared to shallow networks. This is a preferred advantage when dealing with large sensor numbers and a lack of a priori knowledge of these sensors. In our experiment, we used the original size of the input (43680) without any artificial dimension reduction.

Moreover, in our proposed method, the model decision basis provided by the self-attention mechanism is also not available

TABLE V
EXPERIMENT 2-SUBDATASET

Subdataset	1	2	3	4	5	6
Training data	0 hp	1 hp	2 hp	3 hp	0-3 hp	0-2 hp
Testing data	0 hp	1 hp	2 hp	3 hp	0-3 hp	3 hp

TABLE VI
EXPERIMENT 2-RESULTS COMPARISON

Method	Subdataset					
	1	2	3	4	5	6
[41]	88.9%	-	-	-	-	-
[42]	-	-	-	92.5%	-	-
[39]	98.8%	98.8%	99.4%	99.4%	99.8%	96.8%
[40]	100%	100%	100%	99.96%	99.95%	98.80%
Our method	99.12%	99.61%	99.49%	99.9%	99.84%	81.52%

in other methods, and this alleviates the black-box nature of deep learning models.

B. Experiment 2: Bearing Dataset

1) Task Description: This dataset, created by the Case Western Reserve University Bearing Data Center [38], is based on a task that identifies the bearing conditions of a electric motor. Two vibration sensors are mounted on the drive end and the fan end respectively. The faults occurred in three locations, bearing rolling element, inner raceway, and outer raceway. Each location has three different levels of faults severity, small, medium, and large. Thus, these nine different faults categories plus the healthy condition make a total of 10 different categories. It is a classification problem of 10 categories, and the data of the 10 categories were collected for four different working loads (0-3hp). Several subdatasets are created based on the methods provided in [39], [40] as shown in Table V, and the proposed method is evaluated on these 6 subdatasets. The sixth subdataset was slightly different from the others, it was based on predicting the bearing condition under high working load (3hp) based on the data collected from low working load (0-2hp).

2) Data Organisation: As the vibration signal from the 2 sensors was recorded during a certain time period for each type of condition, and the length of each file was more than 120,000 data points, they need to be chunked into small portions. A 2×120 was chosen as the window size of one portion (120 data points for each vibration sensor) and the data was reorganised to a matrix with 4 rows and 60 columns as the input of our deep network.

3) Experiment Results: The experiment results of the proposed method and the results obtained in previous research are shown in Table VI. It was observed that the best results of previous research have been achieved in [40]. They used a 16 layer Visual Geometry Group (VGG-16) as a feature extractor to extract features in time-frequency images of vibration signals. Compared with their results, the proposed method has similar accuracy in tasks 1-5 (nearly 100%). However, this method shows a worse accuracy than the accuracy in [40] in task 6. This may be because this research used the raw data in the time domain as compared to the frequency domain analysis used in [40]. As mentioned in the task description, the requirement of this task is to use low working load data to predict failures under heavy working load and the speed of motor remains constant. In frequency domain bearing vibration analysis, the rotation speed of the shaft is a main factor affecting the frequency feature distribution of vibration signal [43], and the severity of defects, rotation speed, and working load mainly affect the amplitude of each frequency element [15]. Therefore, as the rotation speed remains constant in this experiment, the frequency domain analysis is less sensitive than time domain analysis.

TABLE VII
EXPERIMENT 3-RESULTS COMPARISON

Classification method		Bearing		Gear	
		Low load	High load	Low load	High load
[44]	SAE-DNN	87.5%	92.1%	92.7%	91.9%
	GRU	91.2%	92.4%	93.8%	90.5%
	BiGRU	93.0%	93.6%	93.8%	90.7%
	LFGRU	93.2%	94.0%	94.8%	95.8%
[40]	Pre-trained CNN	99.94%	99.42%	99.64%	99.02%
Our proposed method		99.10%	99.85%	99.90%	99.92%

C. Experiment 3: Gearbox Dataset

1) Task Description: This dataset was created by Shao et al on a dynamic simulator which comprised of a motor, a shaft, a gearbox, and a brake [40]. This task required the identification of gear and bearing working conditions of the gearbox based on the vibration signal. Gear and bearing have four faulty working conditions and one healthy working condition.

2) Data Organisation: The time window size for the vibration signal was kept at 4000, this implies that the input data had 4000 data points, organised as a 40×100 matrix. This model treats each 40×100 input as 40 different data sources and attempts to capture features inter- and intra- data sources.

3) Experiment Results: The experiment results are shown in Table VII. As can be seen from the table, the accuracy obtained by our proposed method is slightly higher than that obtained by the best results in previous studies [40] [44]. Compared with the method proposed that used wavelet transformation of vibration signal as input data [40], the proposed method uses the raw vibration signal and does not require any artificial feature extraction or data transformation.

V. DISCUSSION

The experiment results demonstrate that it is feasible to use the deep model transferred from NLP, in this case, a Transformer-based model called GPT-2, to handle the task of multi-sensor fusion for industrial applications. The proposed method offers the advantages in the following aspects:

- Interpretability: The attention heat map indicates the decision basis of the deep learning model which can be used for key sensor identification, thereby assisting engineers to conduct system diagnosing or maintenance or reducing the redundancy of the system. The results from experiment 1 demonstrate that using the data from

important sensors to classify system conditions is more accurate than using data from all the sensors. This suggests that the AM has a positive effect on identifying key sensors. Such a feature is not available in most other deep learning models. However, it is worth noting that no attention allocated to a sensor does not necessarily mean that the sensor does not have enough information. It can only indicate that the sensor data with strong attention are more easy to harness during the back propagation process. In other tasks with accuracy above 95%, there was little difference between the results obtained by using the sensors with high attention and all the sensors.

- Enable the use of deep learning models with limited industrial data. Normally, transfer learning is a hot topic to solve the data shortage in industrial scenarios requiring that the two datasets have some kind of similarity [37] [45]. However, even collecting enough data on similar industrial processes is costly and finding similar industrial processes also has limitations. Our proposed method demonstrates that a deep model pre-trained in natural language can be transferred to industrial sensor fusion tasks. As natural language is a data-rich modality, deep learning models can be sufficiently trained from it. Therefore, the use of deep learning for sensor fusion tasks is less likely to be limited by the limited industrial data and the lack of similar industrial processes.
- Eliminate the need of artificial feature engineering: Manual feature extraction and selection is labour intensive, difficult and has high uncertainty. However, in order to combine different sensors and reduce the complexity of the input space, it is usually necessary. In the proposed method, the sensor data with different sampling rates were combined and mapped to a unified embedding space, and a deep learning model was used to extract features from the embedding space automatically. Therefore no manual feature engineering was required.

However, while the model performs well on classification problems as shown in this paper, it shows less capacity on the regression tasks, such as remaining useful life prediction tasks. The reason for this may be that stacked attention layers are better at searching for key information rather than mapping features to a specific value. As mentioned in section II, the output of AM are calculated based on Q, K, and V vectors, where Q and K vectors are used to search and match information and only V vector is used to extract features. As the V vector is obtained only using a linear layer, therefore AM may be less capable of extracting abstract features from the data. In order to adapt the AM to the regression task, the computational mechanism of the attention layer may need to be adapted, which might be a possible future research direction. In addition, the large model size and high memory usage are other limitations of this approach. The computational complexity and memory usage of this method grow by the square of the length of the input sequence [46]. As a result, it may cause high occupancy of computing resources in industrial applications. How to minimise the model and find a balance between model performance and resource consumption maybe another future research direction.

VI. Conclusion

In conclusion, this work proposed a new deep transfer learning method to deal with industrial sensor fusion tasks. The results of condition monitoring of a hydraulic system show that the proposed method has achieved high accuracy without feature engineering in an extremely large input space. This proposed method allows industrial scenarios to use deep models with a relatively small amount of data. In its accumulator conditions classification task, the accuracy can be further improved from 91.4% to 96.4% if only the sensor data with high attention scores are used as input. This phenomenon suggests that AM has a positive effect on improving the interpretability of deep learning model. As can be seen from the results of bearing condition classification, the proposed method achieves similar accuracy to the best results of previous studies in Tasks 1 to 5. However, it shows an unsatisfactory result in Task 6. The reason for the result in task 6 may be due to the fact that the frequency domain analysis used in [40] has an advantage in predicting high workloads conditions using low workloads data when compared to the time domain analysis used in this study. As for gearbox condition classification, the proposed method is slightly more accurate in comparison with the accuracy obtained in [40].

The results of this research show that the pretrained NLP model GPT-2 based on the architecture of the Transformer also has the potential to handle multi-sensor fusion tasks. The GPT-2 acts as a feature extraction engine that could replace manual feature extraction thus eliminating the difficult choice of using a deep model or a shallow model with artificial feature extraction for industrial sensor fusion tasks. Moreover, the experimental results show that deep learning model, in this case, GPT-2, trained from natural language can be transferred to industrial sensor data, which means it may not be necessary to collect data for specific kinds of machine or processes before using deep transfer learning. Hence, the cost and time required for industrial data collection can be significantly reduced. In addition, the AM allows the model to provide not only the prediction information but also the basis for the model's decision-making which might be beneficial for industrial applications. Currently, the limitations of this approach are that the current model does not perform very well on the regression task and the high computational resource usage. These may be two valuable directions for future research.

References

[1] T. Zonta, C. A. da Costa, F. A. Zeiser, G. de Oliveira Ramos, R. Kunst, and R. da Rosa Righi, "A predictive maintenance model for optimizing production schedule using deep neural networks," *J. Manuf. Syst.*, vol. 62, pp. 450–462, Jan. 2022.

[2] F. Sell-Le Blanc, J. Hofmann, R. Simmler, and J. Fleischer, "Coil winding process modelling with deformation based wire tension analysis," *CIRP Ann.*, vol. 65, no. 1, pp. 65–68, 2016.

[3] M. Ferdowsi, A. Benigni, A. Monti, and F. Ponci, "Measurement selection for data-driven monitoring of distribution systems," *IEEE Syst. J.*, vol. 13, no. 4, pp. 4260–4268, Dec. 2019.

[4] N. Li, N. Gebraeel, Y. Lei, X. Fang, X. Cai, and T. Yan, "Remaining useful life prediction based on a multi-sensor data fusion model," *Rel. Eng. Syst. Saf.*, vol. 208, Apr. 2021, Art. no. 107249.

[5] R.-T. Wu and M. R. Jahanshahi, "Data fusion approaches for structural health monitoring and system identification: Past, present, and future," *Structural Health Monitor.*, vol. 19, no. 2, pp. 552–586, Mar. 2020.

[6] A. J. Torabi, M. J. Er, X. Li, B. S. Lim, and G. O. Peen, "Application of clustering methods for online tool condition monitoring and fault diagnosis in high-speed milling processes," *IEEE Syst. J.*, vol. 10, no. 2, pp. 721–732, Jun. 2016.

[7] B. Sun and X. F. Liu, "Significance support vector machine for high-speed train bearing fault diagnosis," *IEEE Sensors J.*, early access, Dec. 20, 2021, doi: 10.1109/JSEN.2021.3136675.

[8] A. Stief, J. R. Ottewill, J. Baranowski, and M. Orkisz, "A PCA and two-stage Bayesian sensor fusion approach for diagnosing electrical and mechanical faults in induction motors," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9510–9520, Dec. 2019.

[9] A. Essien and C. Giannetti, "A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6069–6078, Sep. 2020.

[10] D. Tiwari, M. Farnsworth, Z. Zhang, G. W. Jewell, and A. Tiwari, "In-process monitoring in electrical machine manufacturing: A review of state of the art and future directions," *Proc. Inst. Mech. Eng. B, J. Eng. Manuf.*, vol. 235, no. 13, pp. 2035–2051, 2021.

[11] W. Zhang, D. Yang, and H. Wang, "Data-driven methods for predictive maintenance of industrial equipment: A survey," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019.

[12] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[13] N. Cohen, O. Sharir, and A. Shashua, "On the expressive power of deep learning: A tensor analysis," in *Proc. Conf. Learn. Theory*, 2016, pp. 698–728.

[14] J. Liu, Y. Hu, Y. Wang, B. Wu, J. Fan, and Z. Hu, "An integrated multi-sensor fusion-based deep feature learning approach for rotating machinery diagnosis," *Meas. Sci. Technol.*, vol. 29, no. 5, May 2018, Art. no. 055103.

[15] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *J. Sound Vib.*, vol. 289, nos. 4–5, pp. 1066–1090, 2006.

[16] K. Huang, S. Wu, F. Li, C. Yang, and W. Gui, "Fault diagnosis of hydraulic systems based on deep learning model with multirate data samples," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 10, 2021, doi: 10.1109/TNNLS.2021.3083401.

[17] M. Cheng *et al.*, "Intelligent tool wear monitoring and multi-step prediction based on deep learning model," *J. Manuf. Syst.*, vol. 62, pp. 286–300, Jan. 2022.

[18] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.

[19] K. Lu, A. Grover, P. Abbeel, and I. Mordatch, "Pretrained transformers as universal computation engines," 2021, *arXiv:2103.05247*.

[20] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," 2014, *arXiv:1406.6247*.

[21] Z. Cheng, Y. Zhang, and C. Tang, "Swin-Depth: Using transformers and multi-scale fusion for monocular-based depth estimation," *IEEE Sensors J.*, vol. 21, no. 23, pp. 26912–26920, Dec. 2021.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[23] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[24] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[25] A. Radford, W. Jeffrey, D. Amodei, J. Clark, M. Brundage, and I. Sutskever. (2019). *Better Language Models and Their Implications*. Accessed: Nov. 2021. [Online]. Available: https://openai.com/blog/better-language-models/

[26] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[27] N. Helwig, E. Pignanelli, and A. Schütze, "Condition monitoring of a complex hydraulic system using multivariate statistics," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I MTC)*, May 2015, pp. 210–215.

[28] T. Berghout, M. Benbouzid, S. M. Muyeen, T. Bentrcia, and L.-H. Mouss, "Auto-NAHL: A neural network approach for condition-based maintenance of complex industrial systems," *IEEE Access*, vol. 9, pp. 152829–152840, 2021.

[29] J. Wu, P. Guo, Y. Cheng, H. Zhu, X.-B. Wang, and X. Shao, "Ensemble generalized multiclass support-vector-machine-based health evaluation of complex degradation systems," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 5, pp. 2230–2240, Oct. 2020.

[30] A. Gupta, H. P. Gupta, B. Biswas, and T. Dutta, "An unseen fault classification approach for smart appliances using ongoing multivariate time series," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 3731–3738, Jun. 2020.

[31] Y. Lei, W. Jiang, A. Jiang, Y. Zhu, H. Niu, and S. Zhang, "Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost," *Processes*, vol. 7, no. 9, p. 589, Sep. 2019.

[32] J. Prakash and P. Kankar, "Health prediction of hydraulic cooling circuit using deep neural network with ensemble feature ranking technique," *Measurement*, vol. 151, Feb. 2020, Art. no. 107225.

[33] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[34] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.

[35] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," 2019, *arXiv:1908.04626*.

[36] Y. Deng, D. Huang, S. Du, G. Li, C. Zhao, and J. Lv, "A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis," *Comput. Ind.*, vol. 127, May 2021, Art. no. 103399.

[37] X. Pei, X. Zheng, and J. Wu, "Rotating machinery fault diagnosis through a transformer convolution network subjected to transfer learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[38] *Case Western Reserve University Bearing Data Center*. Accessed: Sep. 2021. [Online]. Available: https://engineering.case.edu/bearingdatacenter

[39] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017.

[40] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2018.

[41] W. Du, J. Tao, Y. Li, and C. Liu, "Wavelet leaders multifractal features based fault diagnosis of rotating mechanism," *Mech. Syst. Signal Process.*, vol. 43, nos. 1–2, pp. 57–75, Feb. 2014.

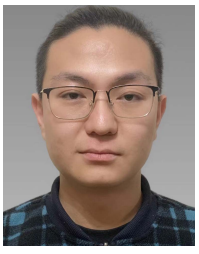
[42] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Trans. Ind. Electron.*, vol. 61, no. 5, pp. 2441–2451, May 2013.

[43] P. D. McFadden and J. D. Smith, "Model for the vibration produced by a single point defect in a rolling element bearing," *J. Sound Vib.*, vol. 96, no. 1, pp. 69–82, Sep. 1984.

[44] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2017.

[45] C.-G. Huang, J. Zhu, Y. Han, and W. Peng, "A novel Bayesian deep dual network with unsupervised domain adaptation for transfer fault prognosis across different machines," *IEEE Sensors J.*, vol. 22, no. 8, pp. 7855–7867, Apr. 2022.

[46] J. Xu *et al.*, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22419–22430.



Ze Zhang received the B.Sc. degree from Jiangnan University in 2012, and the M.Sc. degree in electronic and electrical engineering from the University of Sheffield in 2018, where he is pursuing the Ph.D. degree. Before his research career, he worked in the electric power industry after graduating from his B.Sc. degree. His research interests include artificial intelligence and sensor fusion for digital manufacturing.

Michael Farnsworth received the degree in biochemistry with molecular biology from Cardiff University in 2004, the M.Sc. degree in computer science from the University of the West of England in 2007, and the Ph.D. degree in computer science from Cranfield University in 2013.

He is currently a Research Associate with the Department of Automatic Control and Systems Engineering, the University of Sheffield, and a Research Lead with the EPSRC Future Electrical Machines Manufacturing Hub. He has published over 30 peer-reviewed articles. His research interests include digital manufacturing and the application of machine learning and bioinspired artificial intelligence. His interest also includes understanding how evolutionary processes can be used to develop systems of general intelligence that are able to tackle hard problems within the field of manufacturing and robotics. He is a Chartered Scientist, an Associate Fellow of the Higher Education Authority, and a member of the IET.

Boyang Song (Member, IEEE) received the B.Sc. degree in 2010, and the M.Sc. and Ph.D. degrees in manufacturing operation management and manufacturing informatics from Cranfield University in 2014 and 2019, respectively. Before his research career, he worked in the automotive manufacturing industry after graduating from his B.Sc. degree. He is a Research Associate at the University of Sheffield. His research interests include converging the operation technology, information technology and artificial intelligence for manufacturing and supply chain.

Divya Tiwari received the B.Eng. degree in electronics and communication in 2002, and the Ph.D. degree from Cranfield University in 2010.

Before joining academia, she worked in the electronics industry on aerospace and automotive applications in the U.K. She is a Research Fellow working with the Digitisation Laboratory for Manufacturing, Department of Automatic Control and Systems Engineering, the University of Sheffield. Her work focuses on sensors and simulation for high-value manufacturing processes. She has written over 21 peer-reviewed articles.

Dr. Tiwari was awarded the Daphne Jackson and Royal Academy of Engineering Fellowship in 2013 for the Development of Novel Photonic Sensors for Industrial Applications. She previously worked in the area of development of photonic sensors for manufacturing and automotive applications.

Ashutosh Tiwari holds the prestigious RAEng/Airbus Research Chair in Digital Manufacturing at the University of Sheffield. He is internationally renowned for research in digital manufacturing. He is the Deputy Director of the £20mn EPSRC Future Electrical Machines Manufacturing Hub, the Executive Committee of £7mn Made Smarter Research Centre for Connected Factories, and serves on the EPSRC Strategic Advisory Team (SAT) for Manufacturing. With over 20 years of experience, he has a track record of leading cross-TRL projects worth over £15mn, produced 343 publications (155 journals and 133 conference papers), graduated 36 Ph.D.s, and was awarded an EPSRC HVM Catapult Fellowship. He is a C.Eng. He is a Fellow of IMechE and IET.