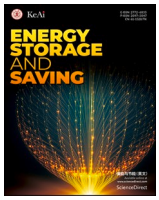
Energy Storage and Saving 3 (2024) 96–105

Contents lists available at ScienceDirect

Energy Storage and Saving

journal homepage: www.keaipublishing.com/en/journals/energy-storage-and-saving/



Research article

A modified transformer and adapter-based transfer learning for fault detection and diagnosis in HVAC systems

Zi-Cheng Wang [a], Dong Li [a,*], Zhan-Wei Cao [a,b], Feng Gao [b], Ming-Jia Li [c]

[a] Key Laboratory of Thermo-Fluid Science and Engineering of Ministry of Education, School of Energy and Power Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] Science and Technology on Space Physics Laboratory, China Academy of Launch Vehicle Technology, Beijing, 100076, China
[c] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China

ARTICLE INFO

Keywords:
Fault detection and diagnosis
Transfer learning
HVAC system
Energy saving
Transformer model

ABSTRACT

Fault detection and diagnosis (FDD) of heating, ventilation, and air conditioning (HVAC) systems can help to improve the energy saving in building energy systems. However, most data-driven trained FDD models have limited generalizability and can only be applied to specific systems. The diversity of HVAC systems and the high cost of data acquisition present challenges for the practical application of FDD. Transfer learning technology can be employed to mitigate this problem by training a model on systems with sufficient data and then transfer it to other systems with limited data. In this study, a novel transfer learning approach for HVAC FDD is proposed. First, the transformer model is modified to incorporate one encoder and two decoders connected, enabling two outputs. This modified transformer model accommodates absent features in the target domain and serves as a robust foundation for transfer learning. It has effective performance in complex systems and achieves an accuracy of 91.38% for a system with 16 faults and multiple fault severity levels. Second, the adapter-based parameter-efficient transfer learning method, facilitating the transfer of trained models simply by inserting small adapter modules, is investigated as the transfer learning strategy. Results demonstrate that this adapter-based transfer learning approach achieves satisfactory performance similar to full fine-tuning with fewer trainable parameters. It works well with limited data amount in target domain. Furthermore, the findings highlight the significance of adapters positioned near the bottom and top layers, emphasizing their critical role in facilitating successful transfer learning.

1. Introduction

The building industry is one of the most energy-intensive sectors globally, accounting for approximately 30% of global energy consumption and 26% of energy emissions [1]. During the operation of a building, heating, ventilation and air conditioning (HVAC) systems consume the largest amount of energy, responsible for around 50% of the total energy consumption [2]. However, these systems are characterized by complex compositions and harsh running conditions, making them susceptible to various failures that can impact their performance. Research indicates that equipment failures and improper controls contribute to a substantial 15% to 30% of energy wastage in commercial buildings [3]. By promptly detecting and diagnosing faults, the long-term healthy operation of building HVAC systems can be ensured. It is possible to effectively enhance the energy efficiency of buildings and achieve energy savings ranging from approximately 10% to 30% [4, 5]. Moreover, in critical applications such as the medical industries and data centers, the operation of HVAC systems with faults may result in severe consequences, including system malfunctions or sudden shutdowns, causing significant losses [6].

Fault detection and diagnosis (FDD) techniques involve monitoring system conditions to detect and identify faults promptly, ensuring safe and efficient system operation [7]. In recent years, the integration of building energy systems with a large number of sensors for HVAC systems has led to the collection of massive operational data. Concurrently, artificial intelligence (AI) technologies have rapidly advanced, offering valuable tools for addressing critical issues. These advances have presented opportunities for employing AI-based methods in FDD of HVAC systems. These methods, called intelligent FDD or data-driven methods, utilize advanced algorithms to intelligently monitor and analyze system

* Corresponding author.
E-mail address: dongli@xjtu.edu.cn (D. Li).

https://doi.org/10.1016/j.enss.2024.02.004
Received 17 January 2024; Received in revised form 22 February 2024; Accepted 22 February 2024
Available online 23 February 2024

operating conditions, effectively ensuring safe and efficient operation [8–11].

In the realm of intelligent FDD methods, the quantity and quality of data play a crucial role. However, obtaining large-scale fault data in experimental settings for each system deployment is costly and even unfeasible. While the FDD model demonstrates a certain level of generalizability, it struggles to maintain universality when confronted with intricate engineering environments [12,13]. Transfer learning, a machine learning approach, facilitates the reuse of a model initially developed for one task on another related task. This approach enhances learning in new tasks by leveraging knowledge obtained from related tasks that may have adequate data [14]. The implementation of transfer learning could potentially address these problems including the lack of data and the limited generalizability of the FDD model, thus facilitating its practical application.

Research on the utilization of transfer learning for intelligent FDD of HVAC systems has attracted much attention. Liu et al. [15] conducted a comparative study to assess the effectiveness of transfer learning based FDD in two chillers. This study considered various transfer learning settings. The findings of this study confirmed the efficacy of transfer learning in FDD, especially in cases where experimental data are limited. In the process of transfer learning, features from different fields inevitably have deviation. Data from different sources typically have different formats. To effectively utilize tabular data collected from various building systems, Fan et al. [16] introduced an image-based transfer learning framework. They employed t-SNE to convert the tabular data into image format and evaluated several transfer learning strategies. The study demonstrated that the image-based method was an effective approach for addressing the challenge of compatibility in multi-source data within the building domain. Zhang et al. [17] investigated the deviation of features from different domains. They proposed a transfer learning strategy that involved feature transformation. This method enabled the effective FDD transfer learning for HVAC systems in real-world scenarios, where only limited normal data were available in target domain. Martinez-Viol et al. [18] emphasized the significance of a domain similarity analysis between different domains. They proposed a filtering mechanism called dissimilarity reduction to select training samples prior to transfer, which can mitigate the problem of negative transfer. Consequently, the transfer learning of FDD for HVAC systems was not constrained by the assumption of high similarity between the domains. Some studies explored the competence of diverse transfer learning strategies in HVAC FDD tasks. Zhu et al. [19] introduced a transfer learning strategy for FDD migration between building chillers. The strategy included a heterogeneous data standardization and a domain adversarial neural network (DANN). Their model comprised three parts: a feature extractor, a task predictor, and a domain classifier. Li et al. [20] conducted a study to compare three deep transfer learning strategies for convolutional neural network FDD of HVAC systems. The strategies evaluated were network-based fine-tuning, mapping-based domain adaptive neural network, and adversarial-based domain adversarial neural network. They found that fine-tuning resulted in the best performance among these deep transfer learning techniques. Zhang et al. [21] examined the impact of three critical factors on cross-domain FDD for HVAC systems. These factors included the level of resemblance between the two domains, the availability of labeled data in the target domain, and classifier type employed. They evaluated three strategies: direct prediction-based, feature transformation-based, and pre-training and fine-tuning-based. The study revealed that the selection of classifier had a significant influence on results and the pre-training and fine-tuning method was more robust.

Investigations have demonstrated the substantial potential of transfer learning in FDD. However, achieving satisfactory transfer effects across diverse scenarios remains a challenge. Consequently, there is a need to refine existing transfer learning methodologies or develop novel approaches. Additionally, it is crucial to investigate the extent to which the FDD model can be effectively transferred. A transfer learning strategy that facilitates transfer over a broader range or aligns more closely with practical requirements would be beneficial. Such improvements could significantly enhance the application of FDD in HVAC systems.

Hence, a novel data-driven FDD transfer learning approach is proposed in this study. The approach comprises a basic model, which is a modified transformer with one encoder and two decoders, and an adapter-based parameter-efficient transfer learning strategy. The modified transformer is trained and evaluated on the source domain. Subsequently, the adapter-based transfer learning method is employed to transfer the model to the target domain. Two transfer learning scenarios are designed to evaluate the effectiveness of the proposed method with fine-tuning. Several influencing factors are also examined.

The structure of the presented study is organized as follows: Section 2 provides an introduction of the proposed method, including the modified transformer and the transfer learning methods. Section 3 introduces the setup of data experiments. Section 4 presents the results of direct tests and transfer learning, along with an investigation of influencing factors. Finally, Section 5 summarizes the conclusion.

2. Theoretical background and research methodology

2.1. Modified transformer model

The transformer-based model is a kind of neural networks originally designed for natural language processing (NLP) tasks proposed by Google in 2017 [22]. This model relies on attention mechanisms instead of convolutional layers or recurrent layers. The transformer showed excellent results on various NLP tasks and it soon became the dominant language model architecture [23]. Showing great power of attention mechanism, it has also been widely adapted and used in other tasks including computer vision [24], medical imaging [25], time series application [26], FDD [27,28] and so on.

The model generally follows an encoder-decoder architecture. The encoder component facilitates the transformation of an input sequence into a latent representation. Subsequently, the decoder produces final output sequence based on the encoder output. In adapting the transformer model for the FDD tasks of HVAC systems, we innovated beyond the traditional encoder-decoder arrangement. As shown in Fig. 1, the modified transfer model has one encoder and two decoders connected. The encoder processes time series data from multiple sensors, creating a latent representation. The first decoder uses this representation to determine the HVAC system's fault type or confirm normal operation, and also utilizes the same representation as the key and value in the cross-attention process. Subsequently, the second decoder, taking the output of the first decoder as its input, evaluates the fault's severity level using the encoder's latent representation for cross-attention. This two-stage decoding process ensures a precise and thorough FDD, harnessing the foundational transformer mechanisms, but it is tailored to the specificity of HVAC systems.

The processing of input vectors initiates with a positional encoding layer. In the transformer architecture, which solely relies on the attention mechanism and lacks convolution or recurrence, there is no inherent understanding of the order or position of input vectors. To provide the model with a sense of sequence order, positional encoding is incorporated before inputting the vectors into the model encoder. The positional encoding takes the form of vectors that are added to the input embeddings. The values of these positional encoding vectors are determined by the position of each vector in the input sequence, as shown in the Eqs. (1) and (2).

$$PE(pos, 2i) = \sin\left(pos \,/\, 10000^{2i/d_{\text{model}}}\right) \tag{1}$$

$$PE(pos, 2i+1) = \cos\left(pos \,/\, 10000^{2i/d_{\text{model}}}\right) \tag{2}$$

where, pos denotes the location of the vector in the sequence; i denotes

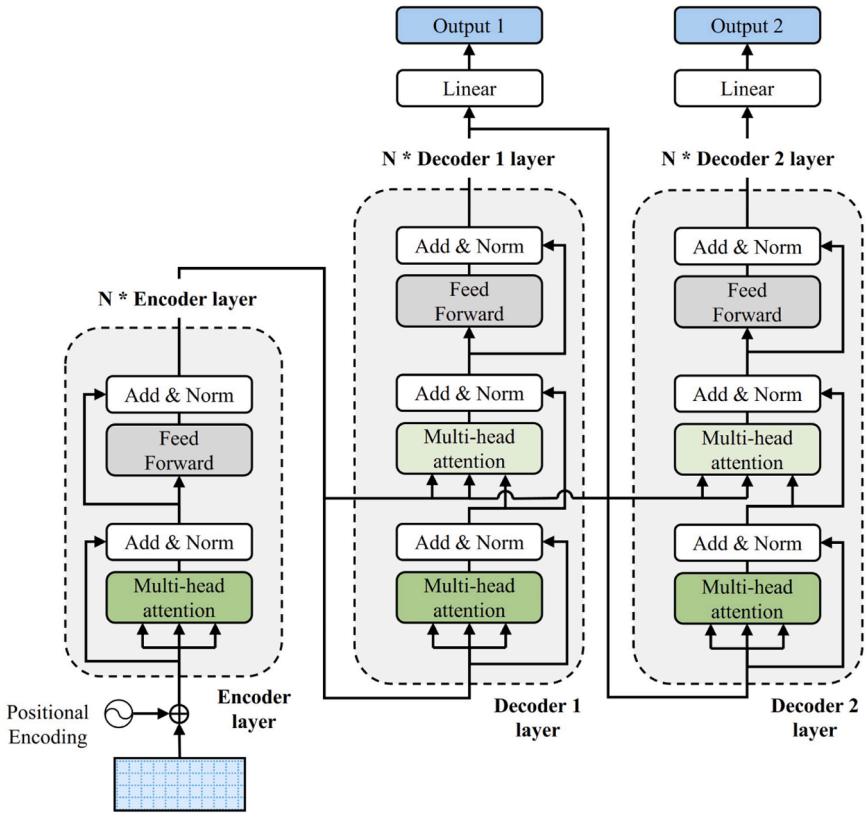


Fig. 1. Modified transformer model with one encoder and two decoders.

the dimension index of the positional encoding; d_{model} denotes the dimension of the model.

Following that, the positional encoding is incorporated into the input. This enables the model to learn the relative positions of the input vectors in the sequence. By incorporating positional information, the model can understand relationships between vectors according to where they appear in the sequence.

After incorporating positional encoding, the transformer model consists of both encoder and decoder components. These components are composed of multiple encoder layers and decoder layers, respectively. In both the encoder and decoder layers, several crucial elements are shared, including the multi-head attention module, the feedforward network, residual connections, and layer normalization. The primary difference is that the encoder layer only has a self-attention module, while the decoder layer includes both a self-attention module and a cross-attention module. In summary, the transformer encoder and decoder utilize multi-head attention and feedforward networks in a stacked architecture. This design allows them to effectively capture sequential dependencies and intrinsic patterns.

Specifically, in one encoder layer, the input data, which consists of time series sequences of vector representations, first passes through the multi-head self-attention module. The attention mechanism is a category of neural network techniques utilized in the deep learning models to selectively prioritize critical components within the input data [22, 29]. It endows the model with the ability to weigh the significance of various portions of the inputs. Attention mechanisms have proven to be particularly efficacious when processing sequential data. The attention mechanism is composed of three parts: the query, key, and value components. Self-attention, as shown in Fig. 2, is a specific form of attention mechanism in which the query, key, and value are derived directly from the input data. Each element in the sequence concurrently serves as a query, key, and value. This enables the model to concurrently allocate the attention to diverse segments of the sequence, thus facilitating the learning of complex patterns. Multi-head attention represents a variant

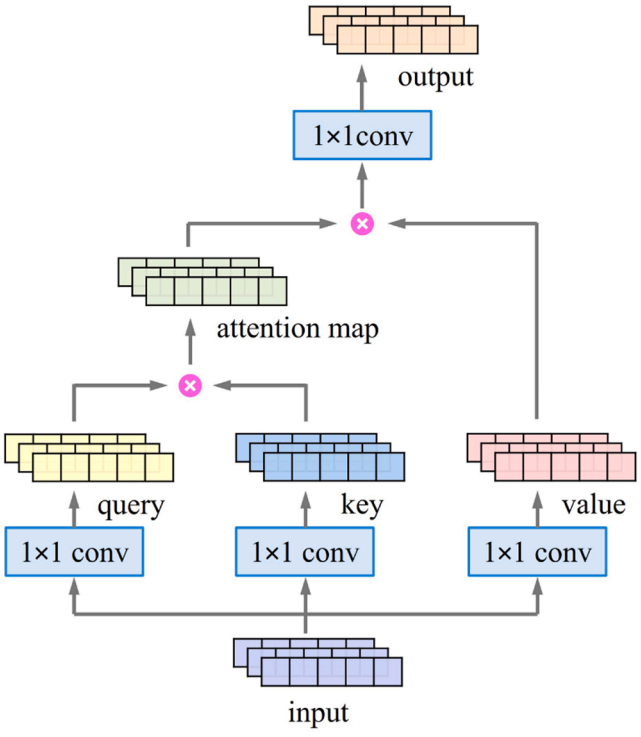


Fig. 2. Self-attention mechanism.

of the attention mechanism, as shown in Fig. 3. Unlike typical attention mechanism which operates on a single set of query, key, and value, it partitions the query, key, and value into several distinct sets or heads, which are subsequently processed in parallel [30]. Attention scores are

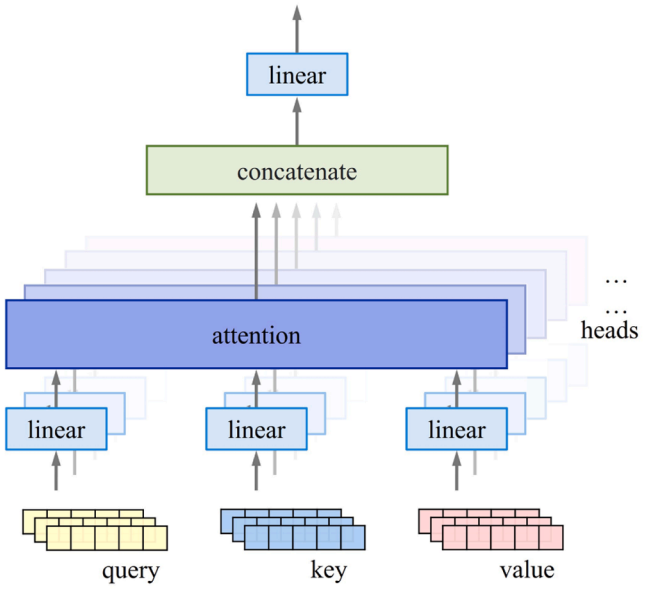


Fig. 3. Multi-head attention mechanism.

computed independently for each head, then the outputs from the heads are concatenated and passed through a fully connected layer to generate the final output. Each individual head directs its attention towards a different facet of the input sequence, enabling the model to find diverse representations and perspectives of the input. This allows the model to integrate information from diverse sources and learn more intricate relationships between the inputs and outputs. The multi-head mechanism provides increased modeling power and flexibility compared to regular single-head attention.

The attention mechanism functions by initially computing weight coefficients through the utilization of the query and key. Subsequently, these weight coefficients are employed to aggregate the corresponding values. To implement self-attention calculation, the initial step involves computing the query, key, and value shown in Eqs. (3)–(5) for each element of a sequence of input vectors $X = [x_1, x_2, \ldots, x_n]$, where, x_i represents the input vector for the ith element.

$$Q_i = W_q x_i \tag{3}$$

$$K_i = W_k x_i \tag{4}$$

$$V_i = W_v x_i \tag{5}$$

where W_q, W_k, and W_v denoted the weights updated during training.

The calculation of the attention map is then performed by considering the query and key, as shown in Eq. (6).

$$w_{ij} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) \tag{6}$$

where, d_k is the dimensionality.

The output vector for the ith element is calculated by obtaining the weighted summation of the value vectors shown in Eq. (7), utilizing the weights that were computed before.

$$O_i = \sum_{j=1}^{n} w_{ij} V_j \tag{7}$$

The feedforward module is a crucial component in transformer architectures, following the attention module. Typically, this module comprises of two fully-connected linear transformations, separated by a ReLU activation function. The feedforward module is used to further process the output of the attention module. This module helps the transformer-based architecture to model complex non-linear interactions between the input sequence elements and to learn more expressive representations.

The key distinction between the encoder and decoder lies in the fact that the decoder incorporates two attention modules: a self-attention module and a cross-attention module. In the cross-attention modules of the decoder, the output of the encoder serves as the key and value, while the output of the first self-attention module acts as the query. This design enables the decoder to attend not only to the first output but also to the encoded latent representations of the input sequence, providing a crucial link between the encoder and decoder. The cross-attention module, by allowing the decoder to access source context and exchange information with the encoder, serves as a bridge between the two components. This enriches the decoder's representations by incorporating information from the encoder, resulting in more contextually relevant and accurate output generation.

In this study, a modified transformer architecture was proposed for the FDD of HVAC systems. The encoder layers are structured in a stacked manner to constitute the encoder, which generates a latent representation of the input data. The first decoder module utilizes this latent representation from the encoder to classify fault types or determine the normal state of the system. Building upon the outputs of both the encoder and the first decoder, the second decoder assesses the severity level associated with any identified fault. This model architecture facilitates a hierarchical FDD approach, where fault types are initially identified, followed by an assessment of the relative severity of faulty conditions through the second decoder. This architecture serves as the basic model for both direct FDD and transfer learning.

Furthermore, in the context of transfer learning, the self-attention structure of the modified transformer model proves beneficial for handling absent features within the target domain in comparison to the source domain. Unlike most transfer learning methods that assume an identical feature list in both domains, which may not align with reality, the proposed modified transformer allows the direct padding of missing feature locations in the target domain with zeros. The self-attention mechanism in transformers calculates attention weights based on the similarity between positions, and zero values at certain positions effectively result in those positions having no influence on the attention weights and the final output. This characteristic is particularly valuable in transfer learning scenarios, as it facilitates the process by mitigating the impact of absent features.

2.2. *Transfer learning strategies*

Transfer learning is a machine learning technology which can transfer learned expertise and experience from a task or domain (source domain) with abundant data to a task or domain (target domain) with limited data [31]. By reusing existing data, rules, or models, learning efficiency and performance on new tasks can be effectively improved. There are many methods to transfer learned experience or knowledge. They can be categorized as instance-based approach, feature-based approach, model-based approach, and relation-based approach [32]. Some popular techniques used in FDD of HVAC systems include feature transformation methods [16,17,21], domain-adaptive neural network [20], domain adversarial neural network [19,20], pre-training and fine-tuning method [20,21], and some others. Among them, pre-training and fine-tuning method is the most popular method. It provides a simple yet effective way to transfer learning. Additionally, parameter-efficient transfer learning represents a novel transfer learning technique [33]. One of the parameter-efficient transfer learning methods enables transfer simply through adapter modules inserted into the model architecture. But its potential for FDD of HVAC systems has not been investigated. In this study, the performance of parameter-efficient transfer learning method will be evaluated with pre-training and fine-tuning method as the transfer strategies for FDD of HVAC systems. The descriptions of these two methods will be given in this section.

2.2.1. Pretraining and fine-tuning method

The pre-training and fine-tuning method is an extensively used transfer learning approach that has demonstrated success [31]. This approach involves initially pre-training a neural network model on a large source dataset and subsequently fine-tuning the pre-trained model on a smaller target dataset for a specific purpose. During the fine-tuning process, the weights of individual layers in the neural network can be treated in various ways: they can be fixed, used as initialization (tuned), or re-trained based on experimental results or prior knowledge. The strategy chosen for each layer depends on its specific role within the network. Additionally, factors such as the size and similarity between the two datasets need to be taken into account. Typically, earlier layers in the neural network, which learn more general features, are often fixed, while later layers, responsible for learning more specialized features, are fine-tuned. However, these are general guidelines, and the optimal choice of fixing or fine-tuning layers may vary depending on the specific task and dataset.

The pre-training and fine-tuning method will be employed to compare with the adapter-based transfer learning method described below. In this study, the pre-training and fine-tuning methods will be performed starting from tuning the top layers.

2.2.2. Parameter-efficient transfer learning

More recently, parameter-efficient transfer learning methods have been proposed to promote the use of transfer learning. One popular approach involves the incorporation of small adapter modules into the pre-trained model architecture [33]. On the target domain, only the adapter parameters are updated, while the parameters of the original pre-trained model remain unchanged. This strategy facilitates knowledge transfer with minimal re-training of parameters.

In this study, the adapter module includes a down projection, a ReLU activation function, an up projection, and a residual connection at the end. This is the most common setting method for adapter module. The locations of the adapters are shown in Fig. 4. There are two adapters in each encoder block. Specifically, the first adapter is situated after the self-attention module and before the add and norm operations, while the second adapter is positioned after the feed forward module and before the add and norm operations. Similarly, each decoder block features three adapters, placed after the self-attention module, cross-attention module, and feed-forward module, preceding the add and norm operations.

3. Data experiments

To investigate the presented transfer learning methods for FDD of HVAC systems, a series of data experiments have been designed. The data utilized in this study are sourced from the LBNL fault detection and diagnostics data sets [34,35]. This database comprises labeled time series data from several typical HVAC systems operating under both normal and various fault states. The experimental setup is outlined in Fig. 5.

Two typical scenarios are designed to assess the transfer performance of the presented transfer learning methods for FDD. The first scenario involves a cross-system transfer, which simulates the condition in which a basic model is trained on one system and subsequently transferred to a similar system. This scenario is intended to explore the extent to which transfer learning can facilitate the generalization of knowledge across similar but not identical systems. The second scenario involves a simulation-to-reality transfer, which simulates the condition in which a basic model is trained on simulation data and subsequently transferred to real-world data. This scenario aims to investigate the effectiveness of transfer learning in bridging the gap between simulation and reality, and thereby, improving the performance of FDD for real-world applications.

Specifically, the comparison between data in the source domains and target domains of these two scenarios is shown in Table 1. In the first cross-system scenario, the source domain is a dual-duct air handling unit (AHU), while the target domain is a single-duct AHU. The dual duct system comprises two separate parallel duct systems that carry hot and

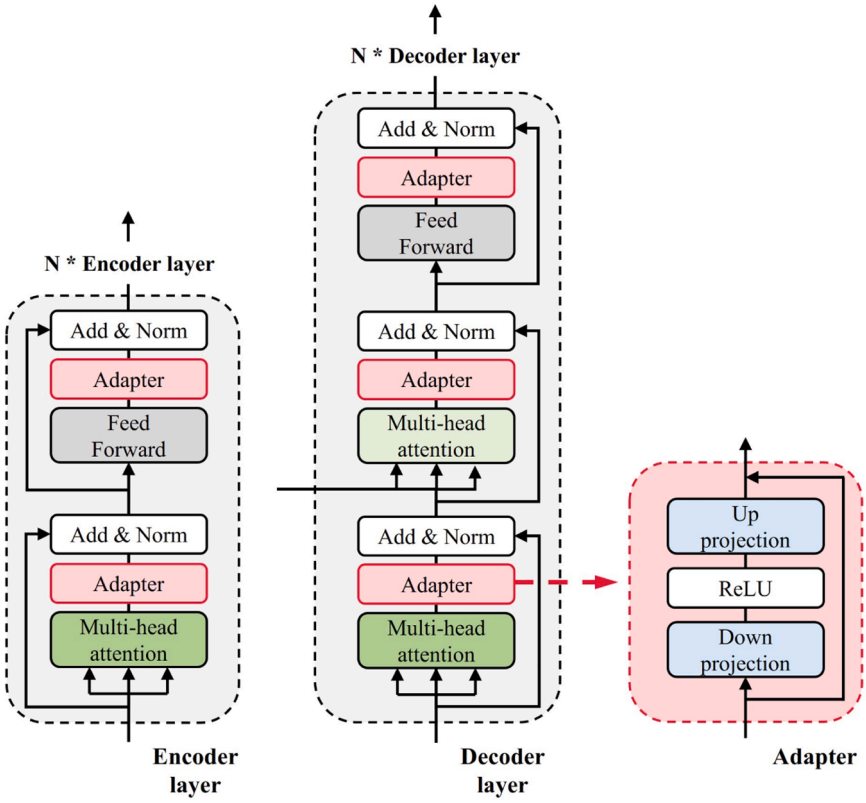


Fig. 4. Inserted adapters and their locations.

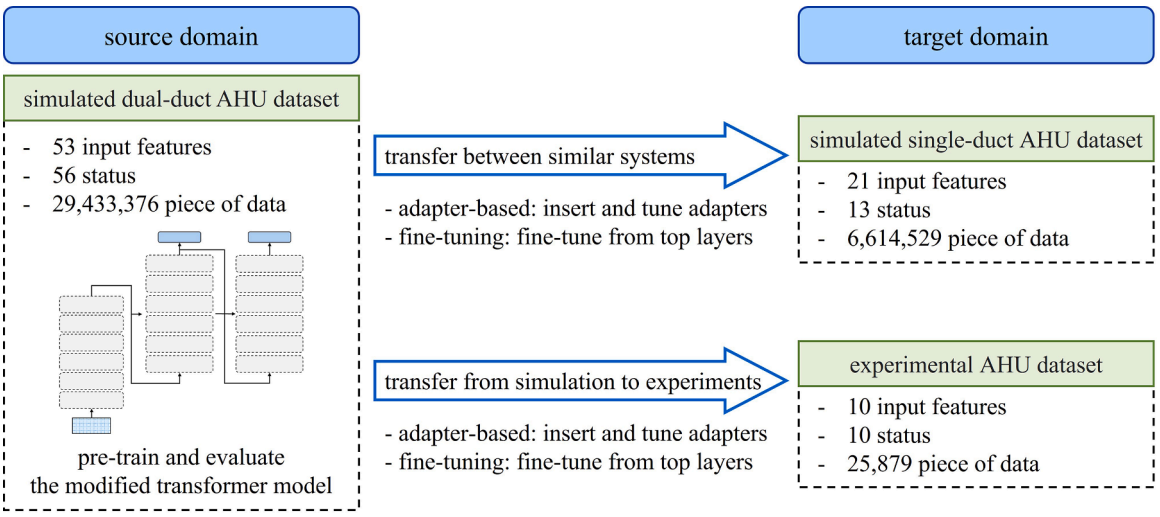


Fig. 5. The outline of data experiments.

Table 1
Comparison of data for 2 scenarios.

Scenario	Domain	System	Features	Status	Data amount
Cross-system	Source domain	Dual-duct AHU	53	56	29,433,376
	Target domain	Single-duct AHU	21	13	6,614,529
Simulation-to-reality	Source domain	Simulated AHU	53	56	29,433,376
	Target domain	Experimental AHU	10	10	25,879

cold air flows individually. This is a bigger data set. In contrast, the single-duct system only consists of one duct and it is a smaller data set. All the data of the source domain (dual-duct AHU), including 56 fault and fault-free states with different severity levels and 53 features, are incorporated to build a powerful basic model using the proposed modified transformer model. In target domain (single-duct AHU), only the corresponding features that are contained in the source domain are selected (21 in total) for the transfer learning of the FDD model. There are 13 fault and fault-free states with different severity levels in target domain, which are also less than source domain. In the second simulation-to-reality scenario, the source domain is also the dual-duct AHU in the first scenario, which is the simulation data. The target domain is a single zone constant air volume or variable air volume (fixed supply air fan speed or not) AHU. The experiments were conducted by Lawrence Berkeley National Laboratory in the LBNL's FLEXLAB test facility. Similarly, all the data in source domain are used to build a basic FDD model using the proposed modified transformer model. There are 10 corresponding similar features in the target domain that are also contained in the source domain, while there are 10 fault and fault-free states with different severity levels in target domain. All the data are read by applying a sliding window of size 5 directly on the transient data. More detailed description about the data and tested systems can be found in the inventory of data sets for AFDD evaluation of the used LBNL fault detection and diagnostics data sets.

4. Results and discussion

4.1. Evaluation metrics

In this study, the diagnostic process follows a hierarchical order including two outputs. The involved faults are relatively complex, taking into account the types of faults and their severity levels. Therefore, the evaluation process and metrics need to be clarified. For a multi-class classification problem, the confusion matrix is utilized. A binary class confusion matrix as an example is presented in Table 2. The model makes the right prediction if the predicted label matches with the true class. Accuracy, Precision, Recall and F1 can be defined as Eqs. (8)–(11).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{11}$$

The overall accuracy is expressed in Eq. (12), representing the proportion of the number of samples with both outputs being correct to the total number of samples. For each output, there are accuracy 1 (for fault types) and accuracy 2 (for fault severity levels), given in Eqs. (13) and (14), respectively.

$$\text{overall accuracy} = \frac{N_{\text{correct samples for both outputs}}}{N_{\text{all samples}}} \tag{12}$$

$$\text{accuracy 1} = \frac{N_{\text{correct samples for the first output}}}{N_{\text{all samples}}} \tag{13}$$

$$\text{accuracy 2} = \frac{N_{\text{correct samples for both outputs}}}{N_{\text{correct samples for the first output}}} \tag{14}$$

4.2. Direct FDD results in source domain

To evaluate the proposed modified transformer model, it is employed for the direct detection and diagnosis of faults in the source domain. The model hyper-parameters are optimized with the Python toolkit Optuna.

Table 2
A binary class confusion matrix.

	True class	
Predicted label	Class 1	Class 2
Label 1	True positive (TP)	False positive (FP)
Label 2	False negative (FN)	True negative (TN)

Specifically, the number of encoder/decoder layers is selected within the range of 1 to 10 in integer. The encoder/decoder heads are tuned within the range of 8 to 32, with a step of 8. The dimensionality of the encoders/decoders is calibrated to be a multiple of the number of heads (with the rate ranging from 8 to 64, in steps of 8) to satisfy architectural constraints. The size of the feedforward dimension is considered among options of 256, 512, 1024, 2048, and 3072. Some parameters are also further tuned by experience. The details of hyper-parameters are shown in Table 3. This optimized modified transformer model has 157.85 million trainable parameters, indicating a powerful large model in the context of HVAC FDD problem.

The FDD results show an overall accuracy of 91.38%, with accuracy 1 (for fault types) of 95.86% and accuracy 2 (for severity level) of 95.33%. The recalls for the first output and the second output are shown in Fig. 6. This modified transformer exhibits a satisfying FDD performance, considering the complexity of the system and the multiple faults.

4.3. Transfer learning FDD in cross-system scenario

The proposed modified transformer and the transfer learning strategies are tested in this scenario. The dual-duct AHU dataset is used as the source domain and the target domain is the single-duct AHU dataset. In this scenario, the pre-trained model is tested directly on the target domain at first. The overall accuracy is 12.69%, with accuracy 1 of 36.33% and accuracy 2 of 34.92%. This unsatisfactory result indicates that a transfer learning is necessary. Adapter-based transfer learning, as well as fine-tuning, is employed with multiple parameter settings. In adapter-based transfer learning, the adapter size ranges from 2, 4, 12, 48, 96 to 192. The fine-tuning starts from the top layers, from the output layer, to the first/second/third/forth layers in the decoders, and to all layers. After transfer learning with different settings. The outcomes of the testing are shown in Figs. 7 and 8.

As can be seen, the adapter-based method demonstrates a comparable result to full fine-tuning but with significantly fewer parameters. It achieves high test accuracy even with limited parameters, and the accuracy increases slightly with the increase of the trainable parameters. On the other hand, fine-tuning exhibits less satisfactory performance with a constrained number of trainable parameters. When incrementally adding tuning layers from the top down, introducing the first layer in the decoder leads to a substantial increase in accuracy. However, subsequent layers (the second, third, and fourth layers) do not contribute significantly to the transfer learning performance. Finally, by tuning all parameters, fine-tuning has the potential to match or even surpass the effectiveness of the adapter-based method.

The trend of accuracy variation with trainable parameters is basically consistent for both outputs. The first output is more accurate than the second one. One of the interesting things is that the first accuracy even decreases with the second, third and fourth layers tuned. This means tuning more than one layers in the decoder even harms the transfer learning process. One layer in the decoder provides enough flexibility for this task. Accuracy does not increase with the increase of trainable parameters from the top down, which indicates fine-tuning is

Table 3
Hyper-parameters of the modified transformer.

Hyper-parameters	Value	Hyper-parameters	Value
Encoder layers	6	Encoder heads	16
Decoder 1 layers	6	Decoder 1 heads	16
Decoder 2 layers	6	Decoder 2 heads	16
Encoder dimension	768	Encoder feedforward dimension	3,072
Decoder 1 dimension	768	Decoder 1 feedforward dimension	3,072
Decoder 2 dimension	768	Decoder 2 feedforward dimension	3,072

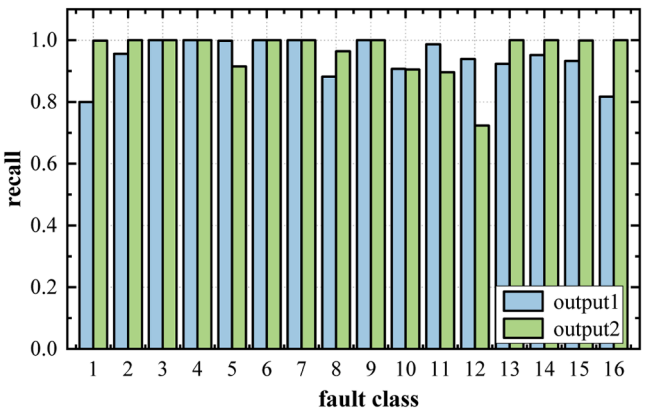


Fig. 6. Recalls of the two FDD outputs.

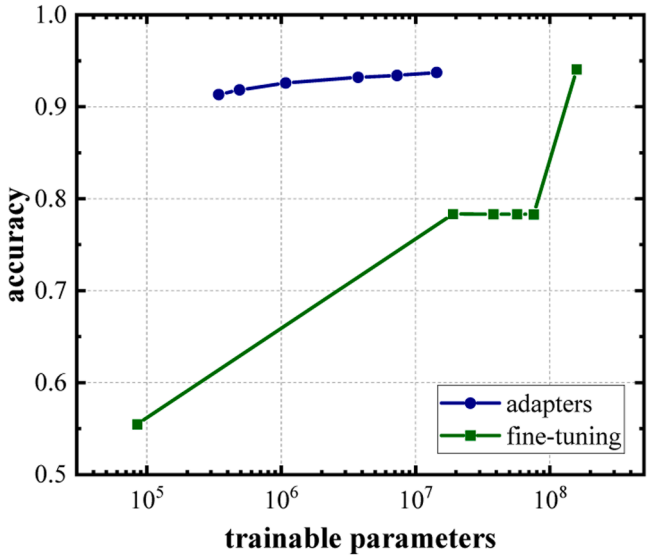


Fig. 7. Overall accuracy for scenario 1 with different trainable parameters.

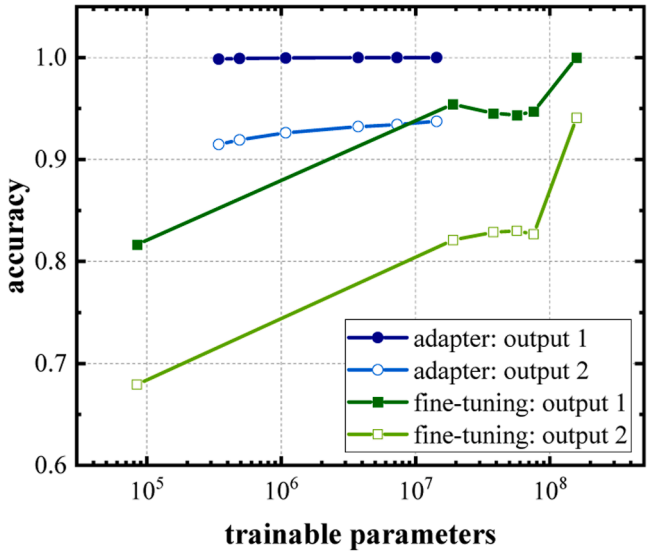


Fig. 8. Accuracies for scenario 1 with different trainable parameters.

not that easy to be performed. Specific layers to be fine-tuned may be decided according to the task, or experimental results. In contrast, for adapter-based method, simply increasing the adapter size proves effective in achieving better results, highlighting its ease of use.

4.4. Transfer learning FDD in simulation-reality scenario

Similarly, the modified transformer and the transfer learning strategies are tested in the second scenario. The simulated AHU dataset is the source domain and the experimental dataset is the target domain. Direct fault diagnosis of the pre-trained model without transfer learning is tested at first. The overall accuracy is 0.19%, with the first accuracy of 22.39% and the second accuracy of 0.86%, much lower than the first cross-system scenario. This indicates that the two domains have bigger difference compared with the first cross-system scenario. Two transfer learning strategies, fine-tuning and adapter-based method, are tested. The adapter size ranges from 4, 12, 48, 96, 192, 384 to 768. In fine-tuning, the tuned layers will be added starting from the top layers. The outcomes of the testing are shown in Figs. 9 and 10.

The results after transfer learning remain lower than the first scenario, consistent with the outcomes of direct testing. This may be due to the difference between source and target domain and the available data amount. The figures illustrate an overall accuracy improvement for adapter-based method, increasing from 0.7 to approximately 0.95. This growth is much larger than in the first scenario. However, compared to the first scenario, it cannot achieve good results with a limited number of trainable parameters. This could be the results of the limited total available data amount in this dataset. It takes about more than 10% of the total trainable parameters of full fine-tuning to achieve almost the same effect as full fine-tuning. Nevertheless, the performance of adapter-based method remains significantly better than fine-tuning when employing a limited number of trainable parameters. As for the separate accuracies for the first output and the second output, similar phenomenon occurred. More layers in decoder would not help, or even slightly harm the transfer learning process. In this scenario, the accuracies of two outputs for the fine-tuning are similar to each other, with minimal difference. This is slightly different from the previous results.

4.5. Impact of available data amounts

It should be noted that the target domain in the first scenario has more data than the target domain in the second scenario. The target domain in the first scenario has 6,614,529 pieces of data, while the target domain in the second scenario has just 25,879 pieces of data. Thus, the influence of the data size in target domain for two transfer learning strategies is investigated. Data in the first scenario is used and the available data in target domain is limited manually according to the experimental settings. For adapter-based method, the adapter size is 48 and the trainable parameters are 3,736,342. For fine-tuning, the output layer and the first top layer in two decoders are tuned and the trainable parameters are 18,988,054. The test results are shown in Fig. 11. Overall, the performance of two strategies gets better when the available data increase. As shown in the figure, the performance of both methods will suddenly drop when the data amount is too small. However, the adapter methods are more robust to the changes of data amount. When the data amount is not too small, its changing magnitude is relatively smaller compared to the fine-tuning method.

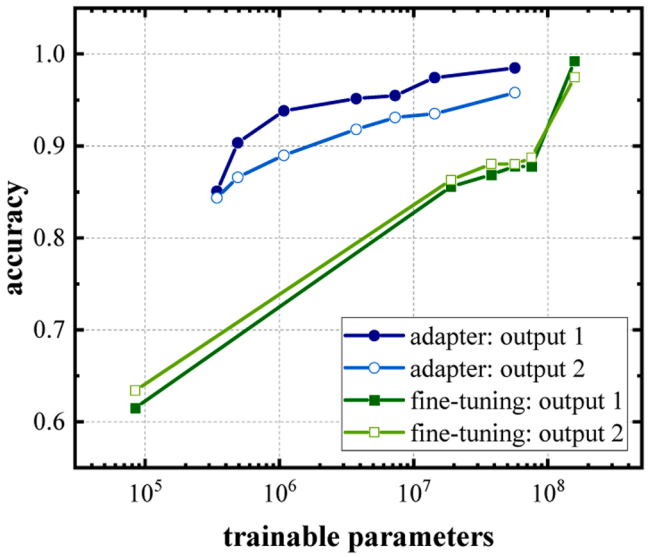


Fig. 10. Accuracies for scenario 2 with different trainable parameters.

4.6. Ablation experiments

Ablation experiments are conducted to evaluate the importance of adapters at different locations. Both scenarios are tested. The adapter size used in the first scenario is 48, and 96 for the second scenario. There are 6 layers in encoder and 6 layers in each decoder, forming 12

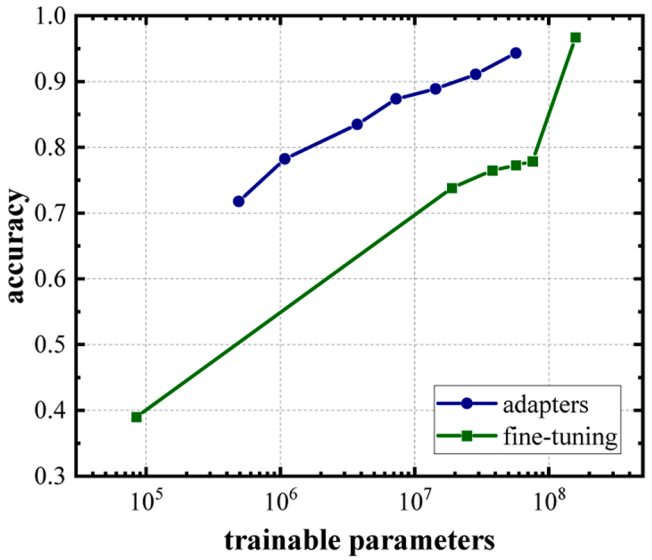


Fig. 9. Overall accuracy for scenario 2 with different trainable parameters.

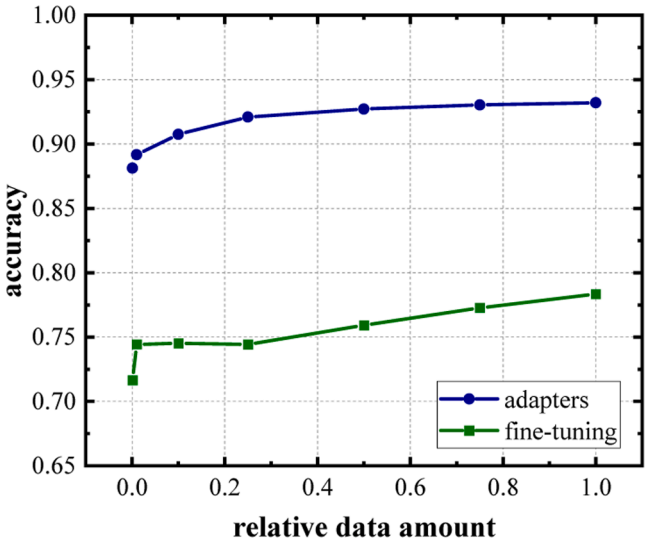


Fig. 11. Test accuracy with different data amount in target domain.

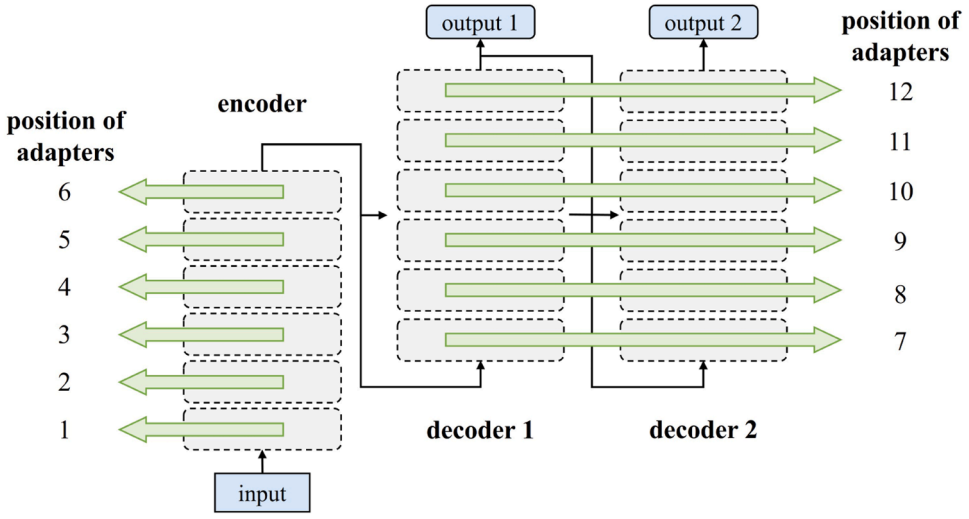


Fig. 12. Position of ablated adapters.

positions in total, as shown in Fig. 12. In each time, adapters in the same encoder layer or two layers in the same position in two decoders are removed. There are two adapters in an encoder layer, and three adapters in a decoder layer. After the adapters removed, the model will not be re-trained and will be tested directly.

The drop amplitude of the overall accuracy after ablating adapters is shown in Fig. 13. Two most important locations are the bottom (near input) and the top (near output). And the first 4 layers (position 7, 8, 9 and 10) in decoders are the least important. This is not consistent with the general understanding for fine-tuning. While it is generally observed that during fine-tuning, the top layers of a pre-trained model tend to be more important, the specific behavior of adapters in different positions can vary. The importance of adapters in different positions within the modified transformer model can be influenced by various factors. Also, the removal of adapters in the first six positions (in encoders) actually eliminates two adapters, whereas removal in the last six positions removes six adapters. However, the drop amplitude observed when ablating adapters in encoders is still greater than when ablating adapters in decoders. This implies that the adapters in the encoders (those near the input) are the most important.

Here are some possible explanations. The bottom adapters, being close to the input, may be responsible for capturing the most essential information from the input data. By ablating these adapters, the model's ability to encode important features can be limited, resulting in a significant drop in performance. These adapters function like the feature extraction or transformation. Difference between features in different domains or feature absence can be addressed by these adapters. The top adapters near the output of the model are responsible for transforming the representations learned by the underlying encoder layers into task-specific outputs. These adapters have a direct impact on the final predictions and are critical for aligning the model's internal representations with the specific task at hand. Ablating these adapters might result in the model being unable to produce accurate or meaningful predictions for the task.

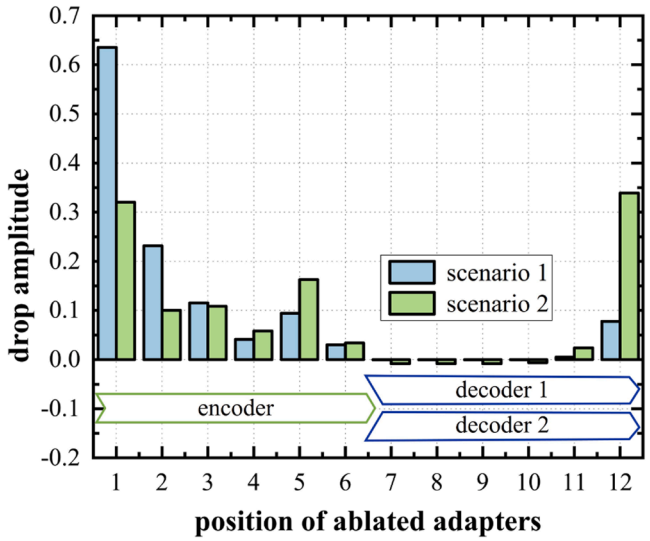


Fig. 13. Accuracy drop amplitude after ablating adapters.

In addition, one thing to be noticed is that the important positions of the two scenarios are slightly different. In scenario 1, which is a transfer learning between similar systems, bottom adapters are more important while the top adapters are less important. However, in scenario 2 which involves a transfer from simulation to real-world experiments, the top adapters are more important than the bottom adapters. This is consistent with the possible explanation mentioned above. In transfer learning scenarios involving similar systems, the emphasis is likely to be on feature transformation, as these systems may share identical, similar, or absent features that require alignment. Conversely, when transferring from simulation to real-world experiments on the same system, the significance of input features diminishes, highlighting the increased importance of the top adapters.

However, in practical applications, it is recommended to use all adapters. The significance of adapter locations can be affected by a lot of factors. Incorporating adapters throughout the entire structure ensures an ample level of flexibility for effective transfer learning. And the trainable parameters of adapters can be adjusted by changing the adapter size, allowing for further optimization and customization.

5. Conclusion

In this study, a novel transfer learning approach for HVAC FDD is presented. A modified transformer model is developed as the basic model for transfer learning. It is designed with one encoder and two decoders connected, enabling simultaneous outputs of fault types and severity levels. Then, an adapter-based parameter-efficient transfer learning method is investigated for the transfer learning of the FDD model. Its performance is compared with the fine-tuning method in two designed scenarios. Finally, some influencing factors of the transfer learning are analyzed, including available data amount and important positions. The conclusions are as follows:

(1) The proposed modified transformer model exhibits a good performance for the complex HVAC FDD tasks. The designed structure allows this model to effectively classify the fault types and severity levels. While most transfer learning methods for FDD require a same feature list for two domains, which is not common in reality, this model allows absent features in target domain when doing transfer learning. It is a suitable large model in the context of HVAC FDD tasks and it serves well as a basic model for transfer learning.
(2) The adapter-based parameter-efficient method emerges as an efficient transfer learning approach for HVAC FDD problems. Two scenarios, a transfer between similar HVAC systems and a transfer from simulation to real-world deployment, have been designed to investigate the transfer learning process. Leveraging the proposed modified transformer, this adapter-based method achieves comparable results to full fine-tuning with trainable parameters of a smaller order of magnitude. It demonstrates exceptional performance with limited trainable parameters.
(3) The adapter-based method is more stable to the change of available data amount in the target domain, compared to fine-tuning. It works well when the data in target domain are not too small. In addition, the most important adapters for transfer learning locate near the bottom and top layers. Bottom layers play a crucial role in feature transformation or extraction across different domains. Top layers aid in aligning latent representations with task-specific outputs, facilitating effective transfer learning.

The modified transformer model and an adapter-based transfer learning method are investigated in this study. However, the test scenario is currently limited by the available data. It would be beneficial to conduct broader tests in various scenarios. It would also be helpful to investigate different configurations of adapters, including positions and structures. Furthermore, to enhance the effectiveness of transfer learning and the practical application of HVAC FDD, future work could focus on a data-driven and domain knowledge hybrid approach to measure domain similarities and differences. This will facilitate the seamless transfer of knowledge between different domains.

CRediT authorship contribution statement

Zi-Cheng Wang: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft. **Dong Li:** Formal analysis, Investigation, Writing – review & editing. **Zhan-Wei Cao:** Formal analysis. **Feng Gao:** Formal analysis. **Ming-Jia Li:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The study is supported by the National Natural Science Foundation of China (Grant Nos.: 52293413 and 52076161).

References

[1] International Energy Agency, Buildings – Analysis – IEA, International Energy Agency, 2023. Paris.
[2] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (2008) 394–398.
[3] S. Katipamula, M.R. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems—A review, part I, HVAC&R Res. 11 (2005) 3–25.
[4] D. Lee, C.-C. Cheng, Energy savings by energy management systems: a review, Renew. Sust. Energ. Rev. 56 (2016) 760–777.
[5] A. Costa, M.M. Keane, J.I. Torrens, et al., Building operation and energy performance: monitoring, analysis and optimisation toolkit, Appl. Energy 101 (2013) 310–316.
[6] X. Zhu, Z. Du, X. Jin, et al., Fault diagnosis based operation risk evaluation for air conditioning systems in data centers, Build. Environ. 163 (2019) 106319.
[7] A.P. Rogers, F. Guo, B.P. Rasmussen, A review of fault detection and diagnosis methods for residential air conditioning systems, Build. Environ. 161 (2019) 106236.
[8] Y. Zhao, T. Li, X. Zhang, et al., Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future, Renew. Sust. Energ. Rev. 109 (2019) 85–101.
[9] F. Zhang, N. Saeed, P. Sadeghian, Deep learning in fault detection and diagnosis of building HVAC systems: a systematic review with meta analysis, Energy AI 12 (2023) 100235.
[10] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: a comprehensive review, Energy Build. 229 (2020) 110492.
[11] Z. Wang, S. Wang, D. Li, et al., An intelligent fault detection and diagnosis model for refrigeration systems with a comprehensive feature selection method, Int. J. Refrig. 160 (2024) 28.
[12] Z. Chen, Z. O'Neill, J. Wen, et al., A review of data-driven fault detection and diagnostics for building HVAC systems, Appl. Energy 339 (2023) 121030.
[13] Y. Himeur, M. Elnour, F. Fadli, et al., Next-generation energy systems for sustainable smart cities: roles of transfer learning, Sustain. Cities Soc. 85 (2022) 104059.
[14] G. Pinto, Z. Wang, A. Roy, A. Capozzoli, et al., Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives, Adv. Appl. Energy 5 (2022) 100084.
[15] J. Liu, Q. Zhang, X. Li, et al., Transfer learning-based strategies for fault diagnosis in building energy systems, Energy Build. 250 (2021) 111256.
[16] C. Fan, W. He, Y. Liu, et al., A novel image-based transfer learning framework for cross-domain HVAC fault diagnosis: from multi-source data integration to knowledge sharing strategies, Energy Build. 262 (2022) 111995.
[17] Q. Zhang, Z. Tian, J. Niu, et al., A study on transfer learning in enhancing performance of building energy system fault diagnosis with extremely limited labeled data, Build. Environ. 225 (2022) 109641.
[18] V. Martinez-Viol, E.M. Urbano, J.E. Torres Rangel, et al., Semi-supervised transfer learning methodology for fault detection and diagnosis in air-handling units, Appl. Sci. 12 (2022) 8837.
[19] X. Zhu, K. Chen, B. Anduv, et al., Transfer learning based methodology for migration and application of fault detection and diagnosis between building chillers for improving energy efficiency, Build. Environ. 200 (2021) 107957.
[20] G. Li, L. Chen, J. Liu, et al., Comparative study on deep transfer learning strategies for cross-system and cross-operation-condition building energy systems fault diagnosis, Energy 263 (2023) 125943.
[21] Q. Zhang, Z. Tian, Y. Lu, et al., Experimental study on performance assessments of HVAC cross-domain fault diagnosis methods oriented to incomplete data problems, Build. Environ. 236 (2023) 110264.
[22] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing System, in: , 2017, pp. 6000–6010. Curran Associates Inc., Long beach.
[23] T. Lin, Y. Wang, X. Liu, et al., A survey of transformers, AI Open 3 (2022) 111–132.
[24] S. Khan, M. Naseer, M. Hayat, et al., Transformers in vision: a survey, ACM Comput. Surv. 54 (2022) 1–41.
[25] F. Shamshad, S. Khan, S.W. Zamir, et al., Transformers in medical imaging: a survey, Med. Image Anal. 88 (2023) 102802.
[26] Q. Wen, T. Zhou, C. Zhang, et al., Transformers in time series: a survey, at https://doi.org/10.48550/arXiv.2202.07125.
[27] C. Fan, Y. Lei, Y. Sun, et al., Novel transformer-based self-supervised learning methods for improved HVAC fault diagnosis performance with limited labeled data, Energy 278 (2023) 127972.
[28] B. Wu, W. Cai, F. Cheng, et al., Simultaneous-fault diagnosis considering time series with a deep learning transformer architecture for air handling units, Energy Build. 257 (2022) 111608.
[29] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.
[30] A. De Santana Correia, E.L. Colombini, Attention, please! A survey of neural attention models in deep learning, Artif. Intell. Rev. 55 (2022) 6037–6124.
[31] F. Zhuang, Z. Qi, K. Duan, et al., A comprehensive survey on transfer learning, Proc. IEEE 109 (2021) 43–76.
[32] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.
[33] N. Houlsby, A. Giurgiu, S. Jastrzebski, et al., Parameter-efficient transfer learning for NLP, in: Proceedings of the Machine Learning Research, 2019, pp. 2790–2799.
[34] J. Granderson, G. Lin, A. Harding, et al., Building fault detection data to aid diagnostic algorithm creation and performance testing, Sci. Data 7 (2020) 65.
[35] L.B.N. Laboratory, LBNL fault detection and diagnostics data sets, at. https://faultdetection.lbl.gov/data/.