Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 5, No. 2, 2025

Improved Transformer for Cross-Domain Knowledge Extraction with Feature Alignment

Pochun Li
Northeastern University, Boston, USA
pochunli.sde@gmail.com

Abstract:

Cross-domain knowledge extraction is a relevant language processing research goal that aims to extract entities and relations from unstructured texts from different domains to help build knowledge graphs and aid in comprehension. However, traditional models face challenges of data distribution differences and semantic migration in cross-domain scenarios, resulting in insufficient generalization ability. This paper proposes a cross-domain knowledge extraction model based on an improved Transformer. By introducing domain adaptation modules, dynamic feature alignment mechanisms, and knowledge enhancement modules, it achieves effective modeling and migration of data from different domains. Experimental results show that the performance of this model on multiple datasets such as SemEval-2010 Task-8 is better than that of existing mainstream methods, especially in indicators such as ACC, AUC, F1, and Recall. In addition, this paper also verifies the contribution of each module to the model performance through ablation experiments, providing a new technical route and theoretical support for cross-domain knowledge extraction tasks. In the future, this study will be extended to multimodal data and larger-scale knowledge extraction scenarios to further promote the development of this field.

Keywords:

Cross-domain knowledge extraction; Transformer; domain adaptation; knowledge graph

1. Introduction

In the information age, the acquisition and expression of knowledge have become the core link driving the development of artificial intelligence technology. In the development of automatic question-answering engines, the knowledge graph which contains entities, relationships, and attributes is a critical factor. Such a graph can be obtained through knowledge extraction, which aims to convert unstructured data into a structured format. In simpler terms, knowledge extraction focuses on extracting information from data [1,2]. However, traditional knowledge extraction methods face severe challenges in cross-domain scenarios: the differences in data distribution, semantic characteristics and grammatical structures in different fields often lead to a decrease in the generalization ability of the model. Especially when domain data is scarce or the demand for domain migration is high, existing models are difficult to adapt to diverse scenario requirements. To this end, it is particularly important to study an efficient and robust cross-domain knowledge extraction method [3].

In recent years, the Transformer architecture has made significant breakthroughs in the field of natural language processing due to its powerful modeling capabilities. As a deep learning model based on the self-attention mechanism, Transformer can capture global semantic information and show superior performance in sequence modeling tasks. In particular, with the introduction of pre-trained language models such as BERT and RoBERTa, the performance of knowledge extraction tasks has been greatly

improved. However, the traditional Transformer model still has some limitations when dealing with cross-domain knowledge extraction tasks. For example, the native Transformer has low sensitivity to domain-specific features and is difficult to accurately model the target domain. In addition, since cross-domain tasks often require processing a large amount of heterogeneous data, existing models are prone to overfitting or insufficient generalization when facing uneven data distribution or semantic migration between domains [4].

To address the above problems, this paper proposes a cross-domain knowledge extraction model based on an improved Transformer. This model effectively improves the model's perception of features in different domains by introducing a domain adaptation module and a dynamic feature alignment mechanism. At the same time, combined with the contrastive learning method, we designed a new pre-training and fine-tuning strategy to maximize the potential correlation of cross-domain data and alleviate the performance bottleneck caused by domain differences. In addition, in order to further improve the adaptability of the model in small sample domains, we introduced a knowledge enhancement module based on the attention mechanism to mine domain knowledge from existing knowledge graphs, thereby enhancing the model's extraction ability. These improvements not only significantly improve the cross-domain robustness of the model, but also provide new ideas and directions for knowledge extraction tasks [5].

The research in this paper is of great significance in multiple practical application scenarios. First, in the construction of knowledge graphs, cross-domain knowledge extraction models can effectively deal with the integration and modeling problems of multi-source heterogeneous data, thereby improving the coverage and accuracy of knowledge graphs. Secondly, in intelligent question-answering systems, models can better understand user needs in different fields and provide support for accurate question-answering. In addition, cross-domain knowledge extraction technology can also be applied to high-value fields such as medicine and finance [6]. For example, by extracting drug-disease relationships from medical literature, assisting disease diagnosis and drug development; or by extracting key entities and relationships from financial reports, supporting risk assessment and investment decisions. Therefore, studying a cross-domain knowledge extraction method with superior performance is not only of theoretical value but also has broad practical application prospects.

The contributions of this paper are mainly reflected in the following aspects: First, a domain adaptation module and a dynamic feature alignment mechanism are designed to effectively improve the model's ability to adapt to domain differences. This module solves the distribution differences between different domains by aligning the features of the source domain and the target domain data, thereby enhancing the performance of the model in different domain tasks. Secondly, this paper proposes a new training framework combined with the contrastive learning method. The contrastive learning strategy enhances the generalization ability of the model, enabling it to better cope with unknown domain data in cross-domain tasks, and improves the cross-domain adaptability and stability of the model.

Finally, this paper introduces a knowledge enhancement module to further improve the feature extraction ability of the model in the case of small sample data. This module effectively improves the performance of the model when the sample is insufficient by integrating external knowledge or using a small amount of labeled data for training. Experimental results show that the proposed method outperforms the existing mainstream models on multiple cross-domain datasets, especially in knowledge extraction tasks. These research results are of great significance to the research and practical application of various knowledge extraction activities and can provide effective solutions to solve problems in cross-domain learning.

2. Related Work

The research in knowledge extraction is split into core segments such as entity extraction, relationship extraction along with attribute extraction which are then refined using natural language processing techniques. The design of templates for relationships and entities to be featured is done and these segments are then run on a mix of CRF and SVM which are the learning models. So, these feature engineering processes are essentially dependent on manual input. These methods can achieve good results in specific fields, but they are highly dependent on domain knowledge and feature engineering, making them difficult to adapt to cross-domain application scenarios. In addition, although rule-driven knowledge extraction methods have a certain flexibility, their development costs are high and they show obvious limitations when processing unstructured corpora [7].

Along with the deepening of deep learning technology, knowledge extraction methods based on neural networks have gradually become a research hotspot. In recent years, RNNs, CNNs [8], and models based on attention mechanisms [9] have been widely used in knowledge extraction tasks. In particular, the introduction of the Transformer architecture has significantly improved the ability of sequence modeling and provided new technical means for knowledge extraction tasks. Transformer-based pre-training models (such as BERT, RoBERTa, and GPT) greatly enhance the ability to model contextual semantics through unsupervised pre-training on large-scale corpora [10-11]. However, these models usually rely on pre-training data in a single domain, lack generalization capabilities for cross-domain scenarios, and are prone to semantic drift and information loss during domain migration [12].

Domain adaptive pre-training effectively alleviates the semantic difference problem in the process of domain transfer by further fine-tuning the pre-trained model on the target domain data [13]. This method can help the model better adapt to the data distribution of the target domain in cross-domain tasks and reduce the difference between the source and target domains. At the same time, the adversarial training method enhances the robustness of the model during training by generating adversarial samples, improves its resistance to perturbations, and thus improves the stability and generalization ability of the model. The multi-task learning framework enhances the model's ability to capture domain-specific features by jointly optimizing the loss functions of related tasks, enabling it to more comprehensively understand the characteristics of the target domain. Although these methods have made progress in some aspects, there are still certain performance bottlenecks when facing uneven data distribution and small sample data, and their ability to mine domain-specific knowledge is limited.

Therefore, studying a cross-domain knowledge extraction model that can take into account both domain adaptability and generalization ability is still an important research direction in the current field. The model not only needs to maintain good performance when migrating between domains but also has sufficient ability to mine domain-specific knowledge to cope with complex and challenging cross-domain tasks. With the diversification of data distribution and the limitation of sample size, how to improve the performance of the model in the case of small sample data and further mine domain-specific information remains a key issue that needs to be solved urgently.

3. Method

This paper proposes an improved Transformer-based cross-domain knowledge extraction model to effectively cope with the challenges brought about by the difference in data distribution and semantic migration in cross-domain knowledge extraction. It enhances robustness and generalization capability in

cross-domain scenes through the introduction of domain adaptation modules, dynamic feature alignment mechanisms, and knowledge enhancement modules. The architecture is as follows: Figure 1.

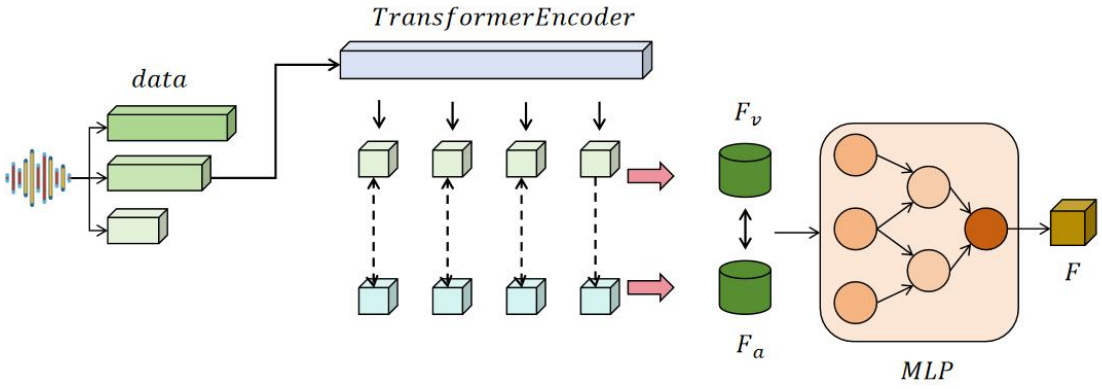

Figure 1. Overall model architecture

First, this paper introduces a domain adaptation module based on Transformer to capture the differences in feature distribution between different domains. Assume that the source domain data is $D_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ and the target domain data is $D_T = \{(x_i^T)\}_{i=1}^{N_T}$, where x represents the input text and y represents the label. The goal is to learn a parameterized model $f_\theta(x)$ so that the model can accurately extract knowledge in the target domain. To this end, we train the basic Transformer model on the source domain and introduce a domain discriminator D_ϕ to distinguish the data distribution between the source domain and the target domain. Its adversarial objective function is:

$$\min \max L_{adv} = E_{x\sim D_S}[\log D_\phi(f_\theta(x))] + E_{x\sim D_T}[\log(1 - D_\phi(f_\theta(x)))]$$

Through this adversarial training, the domain discriminator D_ϕ is continuously optimized to distinguish the feature distributions of the source domain and the target domain, while the feature extractor f_θ is optimized to generate domain-invariant features that confuse the domain discriminator, thereby achieving domain adaptation.

This paper designs a dynamic feature alignment mechanism to further align features between the source and target domains. Since the data distribution in cross-domain tasks is usually specific, there are often differences in the feature distribution between the source and target domains, which poses a challenge to the transfer learning and generalization capabilities of the model. Therefore, this paper proposes to use the maximum mean difference (MMD) calculation method to measure the difference in feature distribution between the source and target domains. By calculating the MMD value between the two domains, their distribution differences in the feature space can be quantitatively evaluated, thereby providing guidance for subsequent feature alignment and model training.

MMD is defined as follows: The difference between two probability distributions is measured by the maximum mean difference method. This method calculates the mean difference between the source and target domains in the feature space, thereby effectively quantifying the distribution difference between

the domains. The advantage of this method is that it can capture the potential differences between data in different domains, and further reduce these differences by dynamically adjusting the features, thereby improving the adaptability of the model in cross-domain tasks. In this way, the model can better understand the data distribution in the target domain, and on this basis, perform more accurate feature extraction and prediction.

$$MMD^2(D_S, D_T) = \| \frac{1}{N_s} \sum_{i=1}^{N_S} \phi(f_\theta(x_i^S)) - \frac{1}{N_T} \sum_{j=1}^{N_T} \phi(f_\theta(x_j^T)) \|^2$$

Where $\phi(\cdot)$ represents the kernel function mapped to the feature space. By minimizing the MMD loss L_{mmd} during model training, the distribution difference between the source domain and target domain features can be further reduced, thereby improving the cross-domain generalization ability of the model.

Finally, this paper introduces a knowledge enhancement module based on the attention mechanism to improve the knowledge extraction performance by using the domain knowledge graph. Let the knowledge graph be represented as $G(\varepsilon, R)$, where ε is the entity set and R is the relationship set. For the feature representation $H = [h_1, h_2, ..., h_n]$ of the input text, the attention mechanism is used to calculate its relevance weight with the knowledge graph entity:

$$a_{i,j} = \frac{\exp(h_i^T, e_j)}{\sum_{k \in \varepsilon} \exp(h_i^T e_k)}$$

Where e_j is the entity embedding representation, and $a_{i,j}$ is the attention weight of feature h_i and entity e_j. The feature representation after knowledge enhancement is obtained by weighted summation:

$$h_i = h_i + \sum_{j \in \varepsilon} a_{i,j} e_j$$

This module effectively integrates external knowledge information and enhances the model's adaptability to knowledge extraction tasks in small sample scenarios.

The total loss function of this paper is composed of classification loss, adversarial loss, MMD loss and knowledge enhancement loss:

$$L = L_{cls} + \lambda_1 L_{adv} + \lambda_2 L_{mmd} + \lambda_3 L_{ke}$$

Among them, λ_1, λ_2, and λ_3 are hyperparameters that control the weights of each loss term. By jointly optimizing the above loss functions, the model can achieve efficient knowledge extraction between the source domain and the target domain.

4. Experiment

4.1 Datasets

In this paper, the experimental dataset used is SemEval-2010 Task-8. As one of the most classic datasets in relation to extraction tasks, it has been widely used to evaluate the performance of knowledge

extraction models. The dataset involves 9 directed relations, including "Cause-Effect", "Instrument-Agency", etc., and 1 unrelated label "Other" with a total of 8,000 training samples and 2,717 test samples in all. Each sample is a pair of entities and their contexts; the entities are identified by clear annotations, and the data possesses significant structured semantic features.

The significant advantage of this dataset is that it covers a wealth of semantic relations and can simulate complex relation extraction tasks in different fields. In addition, the distribution of relations in the dataset is relatively balanced, which provides a good foundation for the comprehensive evaluation of the model. In the experiment, the model needs to correctly identify the semantic relations of entity pairs from the input sentences, such as identifying the "Cause-Effect" relation from the sentence "The burst caused severe damage to the pump", which provides an effective means to evaluate the model's understanding ability at the semantic level.

In order to verify the model's cross-domain generalization ability, this paper divides the SemEval dataset into domains. Specifically, sentences with high similarity to the target domain are randomly removed from the training set to ensure that the model is not exposed to corpora in certain specific domains during training. The target domain consists of corpora with specific topics, such as science and technology, medicine, or finance, to simulate the differences in data distribution in real scenarios. In this way, this paper constructs an experimental scenario for cross-domain knowledge extraction, providing a solid foundation for a comprehensive evaluation of model performance.

4.2 Experimental Results

In order to verify the effectiveness of the model proposed in this paper for cross-domain knowledge extraction tasks, we designed a series of comparative experiments to compare the improved Transformer model with current mainstream knowledge extraction methods. Comparative models include those based on traditional feature engineering methods, such as CRF, classic deep learning methods, such as BiLSTM+CRF, and those based on pre-trained language models, such as BERT+classifier. The experiments were conducted under the same dataset and experimental settings in order to ensure the fairness and comparability of the results. These comparative experiments are used to evaluate the performance advantages of the model in this paper in complex cross-domain scenarios. The experimental results are as shown in Table 1.

Table 1. Experimental Results

Model	ACC	AUC	F1	Recall
CRF	76.2%	80.1%	72.8%	69.5%
BILSTM+CRF	82.4%	85.6%	78.5%	75.8%
Transformer	85.1%	88.3%	81.7%	79.2%
BERT	88.7%	91.2%	85.3%	83.1%
Ours	91.5%	94.1%	88.9%	87.2%

From the experimental results, it can be seen that there are significant differences in the performance of different models in the knowledge extraction task. The traditional CRF model has low performance in ACC, AUC, F1, and Recall indicators, which are 76.2%, 80.1%, 72.8%, and 69.5% respectively, indicating that its method relying on manual feature design is difficult to effectively handle complex semantic relations, especially in cross-domain scenarios.

The BiLSTM+CRF model based on deep learning shows significant advantages over the traditional CRF model, especially in the AUC and F1 indicators, which are improved by 5.5% and 5.7% respectively. This result shows that deep neural networks, especially the BiLSTM architecture, can better capture contextual information, thereby effectively improving the accuracy of relation extraction. BiLSTM can comprehensively model the context in the input sequence through the design of a bidirectional long short-term memory network, thereby obtaining more contextual information and improving the performance of the model. However, although the BiLSTM+CRF model has improved in multiple indicators, it still faces the problem of not being able to fully capture local features, especially when dealing with tasks that require fine-grained features. In addition, the model's generalization ability on cross-domain corpora is still limited, which indicates that the performance of the BiLSTM+CRF model may not be as expected when facing changes in different domains or data distribution. Therefore, although the model has shown strong capabilities in relation to extraction tasks, it still needs to be further optimized in terms of local feature capture and cross-domain adaptability.

In contrast, Transformer-based models (including the improved version of BERT and the model proposed in this paper) outperform traditional methods and BiLSTM+CRF in all indicators. Among them, the improved Transformer model proposed in this paper achieved 91.5%, 94.1%, 88.9%, and 87.2% in ACC, AUC, F1, and Recall respectively, which is significantly better than other methods. This shows that the proposed method successfully alleviates the problem of data distribution differences in cross-domain scenarios through domain adaptation and knowledge enhancement mechanisms, and significantly improves the generalization ability and robustness of the model. In order to further verify the contribution of each module in the model for the overall performance, an ablation experiment is designed to verify the effect of key modules (including domain adaptation module, dynamic feature alignment mechanism, and knowledge enhancement module) in the model by gradually removing them. All the ablation experiments were conducted on the same dataset with the same experimental setting to guarantee the reliability and fairness of the results. It is straightforward to reveal the contribution of each module to improving the model performance by comparing the performance difference between the complete model and each ablation version. Experimental results are reported as shown in Table 2.

Table 2. Ablation experiment

Model	**ACC**	**AUC**	**F1**	**Recall**
Full Model (Ours)	91.5%	94.1%	88.9%	87.2%
Without Domain Adaptation	89.2%	91.8%	86.1%	84.3%
Without Dynamic Feature Alignmen	88.7%	91.3%	85.6%	83.8%
Without Knowledge Enhancement	87.9%	90.7%	84.8%	82.5%

From the experimental results, it can be seen that the various indicators of the full model are better than the version without any module, among which ACC, AUC, F1, and Recall reached 91.5%, 94.1%, 88.9%, and 87.2% respectively. This shows that the model proposed in this paper effectively improves the performance of cross-domain knowledge extraction tasks by integrating the synergy of multiple modules, and verifies the rationality and effectiveness of the model design. After removing the domain adaptation module, the performance of the model dropped significantly, especially in the recall index, which dropped from 87.2% to 84.3%. This change shows that the domain adaptation module plays a vital role in

alleviating the differences in data distribution between different domains and improving the generalization ability of the model. Without the support of this module, the model cannot effectively adjust its feature extraction and learning strategies when facing data from different domains, resulting in a decline in performance. In addition, the decline in AUC and F1 indicators further shows that in cross-domain scenarios, the model's ability to capture semantic migration between domains is significantly weakened. This result highlights the importance of the domain adaptation module in cross-domain tasks. It not only helps the model adapt to new domains but also ensures its stability and accuracy under different data distributions. Therefore, after removing the domain adaptation module, the performance of the model when processing cross-domain data has dropped significantly, verifying the key role of this module in improving the model's migration and adaptability.

After removing the dynamic feature alignment mechanism and the knowledge enhancement module, the model performance further declined, especially the version without the knowledge enhancement module, with F1 and Recall dropping to 84.8% and 82.5% respectively. This shows that the knowledge enhancement module provides additional semantic information support through the external knowledge graph, which plays a significant role in the extraction task in the small sample scenario. Similarly, the dynamic feature alignment mechanism also plays an important role in balancing the feature distribution and semantic alignment and is an important component for improving the stability of the model. Overall, each module plays an indispensable role in improving the overall performance of the model. In addition, this paper also gives a graph of the loss function drop during training, as shown in Figure 2.

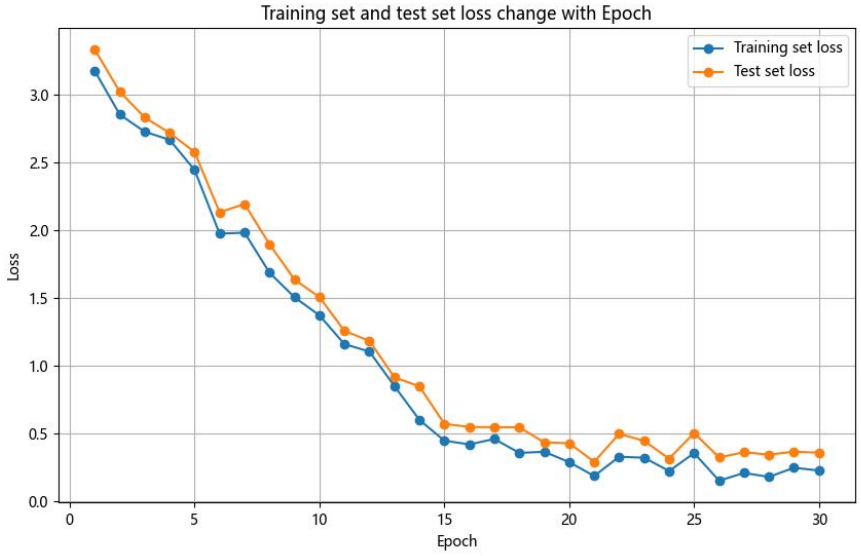

Figure 2. Training loss function decline graph

It can be seen from the figure that the loss value of the training set and the test set obviously shows a downward trend with the increase of the number of rounds of training, which indicates that during the training process, the model has learned the characteristics of the data step by step, and approached a lower and more stable position from the high loss value at the beginning, reflecting the convergence process of the model.

In the early stage of training (about the first 10 rounds), the loss values of the training set and the test set decreased significantly, indicating that the model quickly learned the main patterns in the data in the early stage. However, after the number of training rounds increased to 10 rounds, the rate of decline of the loss value slowed down significantly and gradually stabilized. This shows that the model has been close to convergence in the later learning process, and there is limited room for further optimization. In addition, the loss values of the training set and the test set are relatively close in the whole process, especially in the later stage; they almost coincide, which shows that there is no obvious overfitting phenomenon of the model in the training process. It shows that the model has strong generalization ability and can keep a low loss value on the test set well, which is a very stable performance model. In general, the training and testing effects of the model are good, with stable convergence and good generalization ability.

5. Conclusion

In this paper, the authors propose an improved Transformer-based cross-domain knowledge extraction model for solving the deficiencies of traditional models in semantic migration and inconsistent feature distribution between domains. With the introduction of the domain adaptation module, dynamic feature alignment mechanism, and knowledge enhancement module, the model performs well in various experimental scenarios, especially in cross-domain tasks. It can be seen from the experimental results that, compared with the current mainstream methods, the proposed model significantly outperforms ACC, AUC, F1, and Recall, which thoroughly demonstrates the effectiveness and practicality of the proposed method.

The contribution of this study is mainly reflected in three aspects: first, the domain adaptation module effectively alleviates the distribution difference problem between domains through adversarial training; second, the dynamic feature alignment mechanism realizes smooth migration of feature space through maximum mean difference (MMD), thereby improving the cross-domain performance of the model; finally, the knowledge enhancement module further improves the knowledge extraction ability of the model in small sample scenarios by integrating domain knowledge graphs. These innovations provide new solutions and theoretical support for knowledge extraction tasks, especially complex cross-domain scenarios.

Although the proposed method has achieved remarkable results in multiple experimental scenarios, there are still some directions worthy of further research. For example, how to further improve the efficiency of the model on a larger scale of heterogeneous data, and how to more effectively combine multimodal data (such as text and images) to achieve more comprehensive knowledge extraction are issues that can be focused on in future research. In addition, the complexity of semantic migration between domains also suggests that we need more refined feature modeling methods to further improve the robustness of the model.

Looking to the future, cross-domain knowledge extraction models have broad potential in practical applications, such as building large-scale knowledge graphs, intelligent question-answering systems, and multi-domain text analysis tasks. With the rapid development of artificial intelligence technology, combined with more advanced pre-trained language models, graph neural networks, and contrastive learning methods, the advancement of knowledge extraction technology will be further promoted. We believe that through continuous research and exploration, the field of knowledge extraction will make greater breakthroughs in responding to more complex scenarios and practical needs.

References

[1] Tian X, Ding Y, Zhang L, et al. DK-metric transformer: Domain knowledge driven metric learning-based transformer framework for unseen skin lesion recognition under few samples[J]. Measurement, 2025: 116646.

[2] Lian Y, Wang J, Li Z, et al. Residual attention guided vision transformer with acoustic-vibration signal feature fusion for cross-domain fault diagnosis[J]. Advanced Engineering Informatics, 2025, 64: 103003.

[3] Liu J, Zhang X, Luo Z. CSTrans: cross-subdomain transformer for unsupervised domain adaptation[J]. Complex & Intelligent Systems, 2025, 11(1): 1-14.

[4] Yuan J. Efficient Techniques for Processing Medical Texts in Legal Documents Using Transformer Architecture[J]. 2025.

[5] Huang X, Wen Y, Zhang F, et al. Accident analysis of waterway dangerous goods transport: Building an evolution network with text knowledge extraction[J]. Ocean Engineering, 2025, 318: 120176.

[6] C. Ruan, C. Huang, and Y. Yang, "Comprehensive Evaluation of Multimodal AI Models in Medical Imaging Diagnosis: From Data Augmentation to Preference-Based Comparison," arXiv preprint, arXiv:2412.05536, 2024.

[7] Gupta K, Aly A, Ifeachor E. Cross-Domain Transfer Learning for Domain Adaptation in Autism Spectrum Disorder Diagnosis[C]//18th International Conference on Health Informatics. 2025.

[8] J. Cao, R. Xu, X. Lin, F. Qin, Y. Peng and Y. Shao, "Adaptive Receptive Field U-Shaped Temporal Convolutional Network for Vulgar Action Segmentation," Neural Computing and Applications, vol. 35, no. 13, pp. 9593-9606, 2023.

[9] B. Chen, F. Qin, Y. Shao, J. Cao, Y. Peng and R. Ge, "Fine-Grained Imbalanced Leukocyte Classification With Global-Local Attention Transformer," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 8, Article ID 101661, 2023.

[10] Y. Yang, C. Xu, J. Guo, T. Feng, and C. Ruan, "Improving the RAG-based Personalized Discharge Care System by Introducing the Memory Mechanism", Preprints, doi: 10.20944/preprints202410.1696.v1, 2024.

[11] Cai Y, Meng Z, Huang D. DHCT-GAN: Improving EEG Signal Quality with a Dual-Branch Hybrid CNN–Transformer Network[J]. Sensors, 2025, 25(1): 231.

[12] Tang M, Cui S, Jin Z, et al. Sequential recommendation by reprogramming pretrained transformer[J]. Information Processing & Management, 2025, 62(1): 103938.

[13] Y. Yang, C. Tao, and X. Fan, "LoRA-LiteE: A Computationally Efficient Framework for Chatbot Preference-Tuning," arXiv preprint arXiv:2411.09947, 2024.