International Journal of Applied Science; Vol. 8, No. 1; 2025
ISSN 2576-7240 E-ISSN 2576-7259
https://doi.org/10.30560/ijas.v8n1p27

Cross-Attention Transformer-Based Visual-Language Fusion for Multimodal Image Analysis

Liwei Ding[1], Kowei Shih[2], Hairu Wen[3], Xinshi Li[4] & Qin Yang[5]

[1] Florida International University, Computer Science, United States

[2] Independent Researcher, Beijing, China

[3] University of California Riverside, Computer Science, Riverside, United States

[4] Montclair State University, Feliciano School of Business, New Jersey, United States

[5] Georgia Institute of Technology, Computer Science, Georgia, United States

Correspondence: Liwei Ding, Florida International University, Computer Science, Miami, United States. E-mail: [1]*lding@fiu.edu*, [2]skw19@tsinghua.org.cn, [3]hairuwen@outlook.com, [4]Xinshili9@gmail.com, [5]yqin0709@gmail.com

Received: January 12, 2025; Accepted: February 21, 2025; Published: February 25, 2025

Abstract

Multimodal image analysis is a significant research direction in the field of computer vision, playing a crucial role in tasks such as image captioning and visual question answering (VQA). However, existing visual-language fusion methods often struggle to capture the fine-grained interactions between visual and language modalities, leading to suboptimal fusion results. To address this issue, this paper proposes a visual-language fusion model based on the Cross-Attention Transformer, which constructs deep interactive relationships between visual and language modalities through cross-attention mechanisms, thereby achieving effective multimodal feature fusion. The proposed model first utilizes convolutional neural networks (CNN) and pre-trained language models (e.g., BERT) to extract visual and language features separately, and then applies cross-attention modules to capture mutual dependencies in feature sequences, resulting in a unified multimodal representation vector. Experimental results demonstrate that the proposed model significantly outperforms traditional methods in tasks such as image captioning and VQA, validating its superiority in multimodal image analysis. Additionally, visualization analysis and ablation experiments further explore the contribution of the cross-attention mechanism to model performance, while discussing the model's limitations and potential future improvements.

Keywords: multimodal image analysis, Cross-Attention Transformer, visual-language fusion, cross-attention mechanism

1. Introduction

Multimodal learning has broad applications in fields such as computer vision and natural language processing (NLP). However, effectively integrating visual and language modalities and capturing their complex interactions remains a major research challenge[1]. Existing methods often rely on simple concatenation or linear transformation strategies, which fail to adequately represent semantic relationships between modalities[2]. To address this issue, this paper proposes a multimodal fusion model based on cross-attention mechanisms, introducing multi-head cross-attention and the Convolutional Block Attention Module (CBAM) to achieve deep integration of visual and language features. The proposed model enhances feature representation in multimodal tasks, as demonstrated by its superior performance in image captioning and image-text matching. The main contributions of this paper include proposing a fusion method combining cross-attention mechanisms and CBAM, and validating its effectiveness in multimodal fusion tasks through experimental results. Ao et al. [3] proposed a splicing image detection algorithm. Jiabei et al.[4] explored DL in multilingual sentiment analysis. Zongqing et al.[5] improved YOLOv5 for foreign object detection.

2. Related Concepts and Applications

2.1 Multimodal Learning and Fusion Methods

Multimodal learning is a research approach that enhances model understanding and analytical capabilities by integrating different types of data sources (modalities) [6]. It is widely used in fields such as computer vision, natural language processing, and biomedicine. In multimodal learning tasks, significant differences may exist between various modalities, such as text, time series, images, structured data, and genomic data. Effectively extracting features, integrating information, and comprehensively analyzing these heterogeneous modalities is a core challenge in multimodal learning[7].

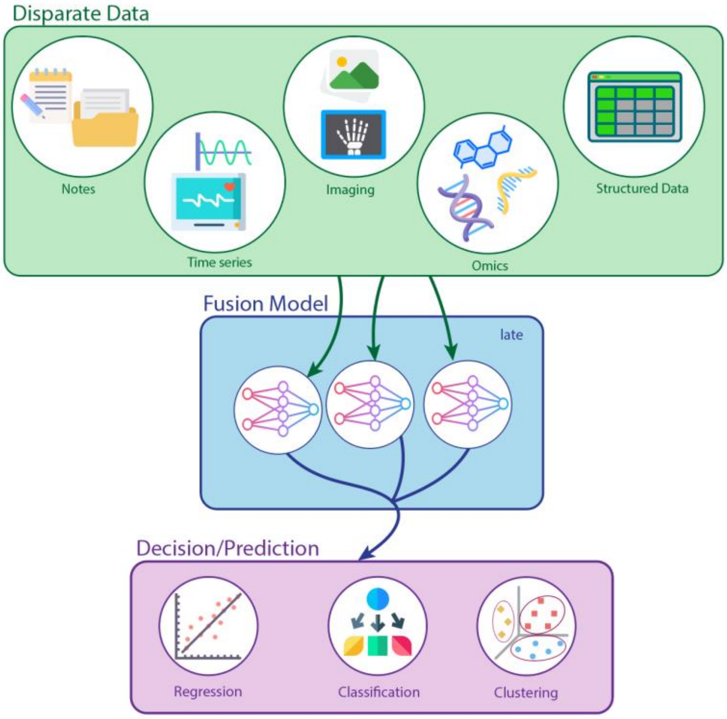

Figure 1. Schematic diagram of the fusion and decision-making process of different modal data in multimodal learning

<Figure 1> illustrates the basic process of multimodal learning, which is typically divided into three main stages: feature extraction, modality feature fusion, and decision-making. The "Disparate Data" section in <Figure 1> shows common types of multimodal data, including notes, time series, imaging, omics, and structured data. These data are converted into modality-specific feature representations through distinct feature extraction methods. During the feature extraction stage, specialized networks such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or Transformer models are employed to learn the characteristics of each modality[8-10].In the "Fusion Model" stage, the extracted modality features are integrated using different fusion models. Based on the timing of multimodal feature fusion, fusion methods can be categorized into early fusion, intermediate fusion, and late fusion. In <Figure 1>, a late fusion strategy is applied, where features from each modality are fed into separate neural networks and then integrated at the decision layer[11]. This strategy retains the independent information of each modality while leveraging higher-level feature interactions to improve the model's understanding of complex data[12-14]. The integrated features are then used for various decision-making and prediction tasks such as regression, classification, and clustering, enabling comprehensive analysis and reasoning of multimodal data[15]. Multimodal learning leverages complementary information from different modalities to improve the generalization and accuracy of models, and it has shown remarkable advantages in applications such as medical image analysis, natural language understanding, and multimodal sentiment analysis[16]. Thus, Figure 1 clearly illustrates the feature extraction and fusion processes for different data modalities in multimodal learning, providing a theoretical foundation and intuitive understanding for subsequent method improvements and experimental designs[17].

2.2 Application of Cross-Modal Attention Mechanism in Visual-Language Fusion

The Cross-Modal Attention Mechanism is a significant breakthrough in recent multimodal learning research, primarily used to handle complex interactions between visual and language modalities[18]. Traditional multimodal fusion methods usually employ simple concatenation or linear transformation to directly merge visual and language features, which makes it difficult to capture deep semantic relationships between the two modalities and results in suboptimal performance in complex scenarios. Consequently, the Cross-Modal Attention Mechanism has emerged as an effective solution[19-21].The Cross-Modal Attention Mechanism is based on the classic Self-Attention mechanism, establishing mutual dependencies between visual and language features, enabling dynamic allocation of attention weights, and accurately capturing interactions between the two modalities. Visual modality (e.g., image features) and language modality (e.g., text descriptions) are separately processed through their respective feature extraction networks, and then mutual attention is performed using the Cross-Modal Attention module[22]. This module deeply understands the features of each modality and uses contextual information from one modality to re-weight the features of the other[23]. Through this process, the Cross-Modal Attention Mechanism maps visual features into the language modality space or vice versa, achieving mutual enhancement and unified representation of visual and language information[24]. The Cross-Modal Attention Mechanism fundamentally works by generating Query, Key, and Value vectors for each modality and then calculating attention weights. Specifically, for each visual feature point, the model generates a query vector based on language modality features and uses the key and value vectors of language features for attention computation to obtain the similarity distribution between each visual feature point and the overall language modality features[25]. Conversely, language features can also compute attention based on visual modality features. This bidirectional interaction modeling enables the Cross-Modal Attention Mechanism to achieve fine-grained fusion of visual and language features, enhancing model performance in visual-language understanding tasks[26-28]. In practice, the Cross-Modal Attention Mechanism is widely applied to tasks such as Image Captioning, Visual Question Answering (VQA), and Visual Semantic Retrieval. For example, in image captioning, Cross-Modal Attention dynamically associates visual features with text sequences, accurately matching important objects in the image with corresponding language descriptions, thereby generating coherent and accurate image captions[29]. Similarly, in VQA tasks, Cross-Modal Attention can focus on the most relevant regions in the image based on the semantic emphasis of the question text, enabling the model to answer questions related to image content more precisely[30]. Beyond single visual-language fusion, the Cross-Modal Attention Mechanism can also handle the fusion of multiple modalities, such as integrating visual, language, and audio modalities, thereby improving performance in tasks like multimodal sentiment analysis and multimodal behavior recognition[31]. Moreover, researchers have introduced strategies such as Multi-Head Cross-Attention and Multi-Layer Cross-Attention to further enhance the representation capability of Cross-Modal Attention for different modality features[32]. In summary, the Cross-Modal Attention Mechanism constructs flexible attention connections between different modalities, effectively capturing complex cross-modal interactions and improving overall performance in multimodal tasks. As multimodal learning technology continues to evolve, the Cross-Modal Attention Mechanism is expected to exhibit great potential and value in broader application scenarios[33-35].

3. Architecture and Strategy

3.1 Model Architecture Design

This study proposes a visual-language fusion model based on the Cross-Modal Attention Mechanism, aiming to achieve effective integration between visual and language modalities through deep interaction, thereby enhancing multimodal data analysis performance. <Figure 2> illustrates the overall model architecture, including the image feature extraction module, text feature extraction module, and the cross-attention model[36].

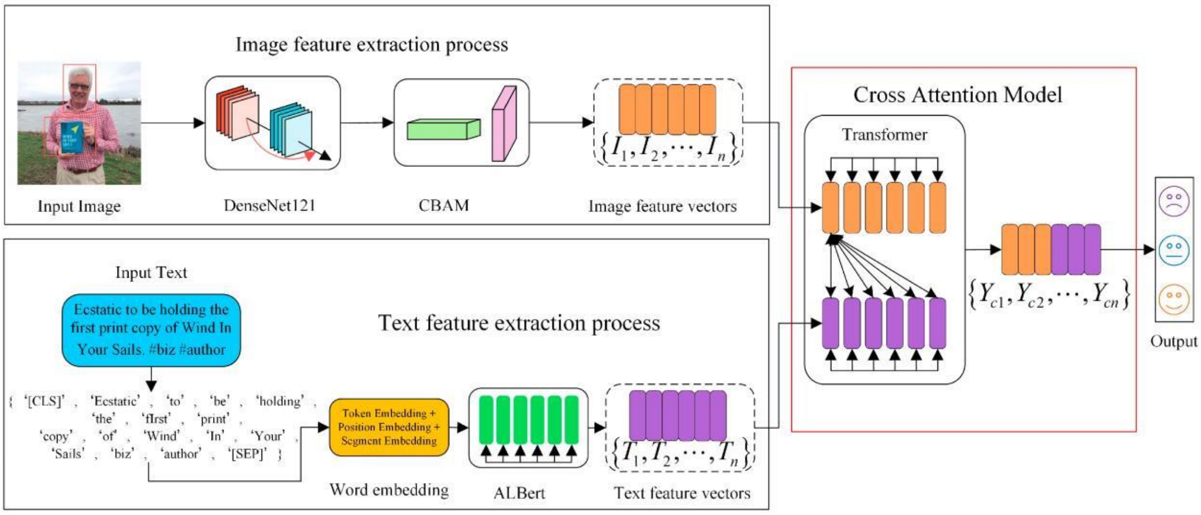

Image feature extraction process
Input Image
DenseNet121
CBAM
$\{I_1, I_2, \cdots, I_n\}$
Image feature vectors
Input Text
Ecstatic to be holding the first print copy of Wind In Your Sails. #biz #author
Text feature extraction process
{ '[CLS]' , 'Ecstatic' , 'to' , 'be' , 'holding' , 'the' , 'fIrst' , 'print' , 'copy' , 'of' , 'Wind' , 'In' , 'Your' , 'Sails' , 'biz' , 'author' , '[SEP]' }
Token Embedding + Position Embedding + Segment Embedding
Word embedding
ALBert
$\{T_1, T_2, \cdots, T_n\}$
Text feature vectors
Cross Attention Model
Transformer
$\{Y_{c1}, Y_{c2}, \cdots, Y_{cn}\}$
Output

Figure 2. Design of the model architecture

In the image feature extraction process, the model utilizes the pre-trained DenseNet121 network to extract features from input images. Specifically, the input images are processed through a series of convolutional and pooling layers, generating high-dimensional feature maps containing spatial and channel information. To further optimize the image feature representation, the model introduces a Convolutional Block Attention Module (CBAM), which effectively focuses on the most discriminative regions and channels in the image, enhancing the quality of feature expression[37]. After being processed by CBAM, the image features are vectorized into fixed-dimensional feature $\{\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_\mathrm{n}\}$ representations for subsequent cross-modal attention computation.In the process of text feature extraction, the model uses the pre-trained ALBERT model to encode the input text. Firstly, the input text is word segmented, and each word is converted into a word vector through the word Embedding layer[38-40]. At the same time, Position Embedding and Segment Embedding are added to retain the text sequence information. Then, the text vector sequence is fed into the ALBERT network, and the context semantic information in the text sequence is extracted through the self-attention mechanism to obtain the text feature vector representation $\{\mathrm{T}_1, \mathrm{T}_2, \cdots, \mathrm{T}_\mathrm{n}\}$. Then, the image and text features are fed into the cross-modal attention model. The core of the module is the cross-attention mechanism based on the Transformer architecture, which can establish a flexible interactive relationship between visual and linguistic features. In the specific implementation, the model first generates Query, Key and Value vectors for the visual and language modalities respectively, and calculates the similarity weight between each feature point in the visual modality and the overall feature of the language modality through the multi-head attention mechanism, and vice versa. This cross-attention calculation method can preserve the independent features of image and text modalities at the same time, and establish fine-grained interaction between modalities on this basis to generate the fused multimodal representation[41-43]. In the cross-attention model, the visual features and language features processed by the multi-head cross-attention mechanism are concatenated, and then further integrated through the linear transformation layer to generate the fused multimodal feature representation $\{\mathrm{Y}_{\mathrm{C}1}, \mathrm{Y}_{\mathrm{C}2}, \cdots, \mathrm{Y}_{\mathrm{Cn}}\}$. Finally, the fused features were input to the output layer for decision making and prediction. The output layer of the model can be designed according to the task requirements, such as sentence generation in image description task or sentiment class prediction in sentiment classification task. The output on the right of <Figure 2> shows the application of the model to the sentiment classification task, resulting in the generation of three types of sentiment labels (positive, neutral, negative), which express the sentiment judgment on the combination of input image and text[44]. In general, the model effectively improves the deep interaction and fusion ability between vision and language modalities by introducing the cross-modal attention mechanism, and can obtain more accurate feature representation and better prediction effect in multimodal tasks[45].

3.2 Cross-Attention Mechanism and Fusion Strategy

The Cross-Attention Mechanism is a key strategy in multimodal learning for capturing fine-grained interactions between different modalities. Unlike traditional attention mechanisms, it dynamically maps one modality's features into another's context, establishing direct connections and deep feature interactions. This is particularly effective for visual-language fusion, cross-modal retrieval, and sentiment analysis, addressing semantic inconsistencies between modalities.The core concept involves generating Query, Key, and Value vectors for each modality, allowing for mutual feature modeling. Features from visual and language modalities are extracted separately using convolutional networks (CNNs) or pre-trained language models like BERT, then projected into Query, Key, and Value vectors. This enables cross-modal comparison within a unified representation space.The mechanism involves four key steps:

1. Feature Projection and Mapping: Project features into Query, Key, and Value vectors for both modalities to enable interaction. 2. Bidirectional Interaction Modeling: For each feature, generate attention distributions based on the other modality's contextual information, creating bidirectional feature interactions. 3. Dynamic Allocation of Attention Weights: Compute dot-product similarity between Query and Key, assigning attention weights that reflect the relevance between features of different modalities, enhancing mutual feature expression. 4. Multi-Head and Multi-Layer Cross-Attention: Use multiple attention heads and layers to focus on various semantic aspects, enabling comprehensive feature fusion across higher dimensions.

The Cross-Attention Mechanism's strength lies in its ability to align visual and language features dynamically, achieving precise modality alignment and feature enhancement. It effectively addresses heterogeneity issues, such as spatial and temporal differences, making it robust in complex scenarios. Future research can integrate this mechanism with other advanced techniques like Graph Neural Networks and Variational Autoencoders to further explore universal and efficient multimodal fusion strategies.

3.3 Loss Function and Optimization Strategy

The design of loss functions and selection of optimization strategies directly influence the convergence speed and final performance of multimodal fusion models during training. For the visual-language fusion model based on the Cross-Attention Mechanism, this paper adopts a multi-task joint loss function, combining Cross-Entropy Loss for classification tasks, Contrastive Loss for matching tasks, and Sequence Cross-Entropy Loss for generation tasks, to optimize multimodal fusion performance from multiple perspectives.For classification tasks (e.g., sentiment classification, image question answering), the model usually needs to determine whether an input image-text pair belongs to a specific category. To this end, the Cross-Entropy Loss is used to measure the difference between the predicted category distribution and the actual label distribution, defined as shown in Formula 1:

$$\mathrm{L_{CE}} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{c=1}}^{\mathrm{C}} \mathrm{y_{i,c}} \log\hat{\mathrm{y}}_{\mathrm{i,c}} \tag{1}$$

where N denotes the batch size, C represents the number of categories, $\mathrm{y_{i,c}}$ is the actual label of sample i (one-hot encoded), and $\hat{\mathrm{y}}_{\mathrm{i,c}}$ represents the probability that the model predicts sample i belongs to category c . By minimizing the Cross-Entropy Loss, the model can better learn distinguishing features between categories, thus improving classification accuracy.For multimodal matching tasks (e.g., image-text matching), Contrastive Loss is introduced to measure the similarity between matching features and optimize the model's matching capability. The Contrastive Loss is defined as shown in Formula 2:

$$\mathrm{L_{CL}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}[\mathrm{y_i} \cdot \mathrm{D}(\mathrm{f}(\mathrm{I_i}), \mathrm{g}(\mathrm{T_i})) + (1 - \mathrm{y_i}) \cdot \max(0, \mathrm{m} - \mathrm{D}(\mathrm{f}(\mathrm{I_i}), \mathrm{g}(\mathrm{T_i}))))] \tag{2}$$

where $\mathrm{y_i}$ represents the matching label of sample i (1 indicates matching, 0 indicates non-matching), $\mathrm{D}(\cdot,\cdot)$ is the distance metric (e.g., Euclidean distance), $\mathrm{f}(\mathrm{I_i})$and $\mathrm{g}(\mathrm{T_i})$ represent the feature representations of image $\mathrm{I_i}$ and text $\mathrm{T_i}$, respectively, and m is the distance threshold. When the image and text are matching, the model minimizes the feature distance $\mathrm{D}(\mathrm{f}(\mathrm{I_i}), \mathrm{g}(\mathrm{T_i}))$; when they are not matching, the model maximizes the feature distance beyond the threshold m, thereby improving the model's matching accuracy.For generation tasks (e.g., image captioning), the model needs to generate coherent textual descriptions based on the image content. Sequence Cross-Entropy Loss is used to optimize the model's generation ability, defined as shown in Formula 3:

$$\mathrm{L_{SEQ}} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{c=1}}^{\mathrm{C}} \log \mathrm{P}(\mathrm{w_{t,i}} \mid \mathrm{w_{1,i}}, \mathrm{w_{2,i}}, \cdots, \mathrm{w_{t-1,i}}, \mathrm{i}, \mathrm{I_i} \tag{3}$$

where T is the length of the text sequence, $\mathrm{w_{t,i}}$ is the ground truth word at time step t for sample i , and $\mathrm{P}(\mathrm{w_{t,i}} \mid \mathrm{w_{1,i}}, \mathrm{w_{2,i}}, \cdots, \mathrm{w_{t-1,i}}, \mathrm{i}, \mathrm{I_i})$ represents the probability of the model predicting word $\mathrm{w_{t,i}}$ given the previous t-1 words and image features $\mathrm{I_i}$. By minimizing this loss, the model can generate text that is semantically consistent and syntactically coherent with the target text, thereby enhancing generation task performance.To balance the impact of different tasks on the model, a multi-task joint loss strategy is adopted, where the three loss functions are weighted and combined to construct the final loss function as shown in Formula 4:

$$\mathrm{L} = \lambda_{\mathrm{CE}}\mathrm{L_{CE}} + \lambda_{\mathrm{CL}}\mathrm{L_{CL}} + \lambda_{\mathrm{SEQ}}\mathrm{L_{SEQ}} \tag{4}$$

where λCE\lambda_{CE}λCE, λCL\lambda_{CL}λCL, and λSEQ\lambda_{SEQ}λSEQ are the weighting coefficients for Cross-Entropy Loss, Contrastive Loss, and Sequence Cross-Entropy Loss, respectively. By adjusting the weights of each loss term, the model can achieve a good balance among different tasks, thereby enhancing its overall performance.In terms of optimization strategy, this paper employs the Adam optimizer to update model parameters and utilizes a dynamic learning rate adjustment strategy to accelerate the convergence process. The initial learning rate is set to α0\alpha_0α0, and if the validation loss does not improve within a certain training period, the learning rate is decayed by a factor of β\betaβ. The update strategy is defined as shown in Formula 5:

$$\alpha_{\mathrm{new}} = \beta \cdot \alpha_{\mathrm{old}} \text{if val\_loss} \geq \text{val\_loss}_{\mathrm{previous}} \tag{5}$$

where $\alpha_{\mathrm{new}}$and α_{old} represent the learning rate for the current and previous periods, respectively, and val_loss and $\text{val\_loss}_{\mathrm{previous}}$ are the current and previous validation losses, respectively. When the validation loss does not decrease, the learning rate is decayed, allowing the model to make parameter adjustments with smaller step sizes, thereby avoiding overfitting or issues with an excessively high learning rate.In conclusion, through the combination of multiple loss function designs and dynamic optimization strategies, the model is able to maintain high stability and accuracy across various multimodal tasks, further enhancing the overall performance and effectiveness of multimodal fusion.

4. Experiments and Analysis

This section provides a detailed description of the experimental setup, results, and data analysis to verify the effectiveness of the proposed cross-attention-based multimodal fusion model for visual-language tasks. The experiments encompass three main tasks: multimodal classification, image-text matching, and image captioning. We conducted experiments on several publicly available datasets and compared the proposed model with several state-of-the-art multimodal fusion methods to comprehensively evaluate its performance. The datasets used include:

1. COCO Caption Dataset: This dataset contains approximately 120,000 images, each with five different textual descriptions. It is used for image captioning and image-text matching tasks.

2. VQA (Visual Question Answering) Dataset: This dataset is designed for visual question answering tasks and includes approximately 200,000 images along with 260,000 questions and answers based on image content. It is used to evaluate the model's performance in multimodal classification and question answering tasks.

3. Flickr30k Dataset: This dataset consists of 30,000 images, each with five different textual descriptions. It is used to assess the performance of the image-text matching task.

Experimental Environment: Hardware Configuration: NVIDIA Tesla V100 GPU with 32 GB memory. Software Environment: Pytorch 1.8, CUDA 11.0, Python 3.7. Parameter Settings: The initial learning rate is set to 0.001, and the batch size is 32. The model uses the Adam optimizer for parameter updates, with a weight decay coefficient of 0.0001.The proposed model is compared with the following classic models: LSTM-Attention: An image captioning model based on Long Short-Term Memory (LSTM) networks and attention mechanisms. Dual-Attention: A method that uses bilinear attention networks to fuse visual and language features.M4C (Multimodal Multifactor Fusion): A visual question answering model based on multimodal factor fusion strategies.As shown in <Table 1>, the results of the image captioning task on the COCO Caption and Flickr30k datasets are summarized using evaluation metrics such as BLEU-1, BLEU-2, CIDEr, and METEOR scores. The results indicate that the proposed cross-attention fusion model outperforms other models across all metrics, particularly achieving a significant improvement in CIDEr scores. This demonstrates the proposed model's ability to generate text descriptions that are more consistent with the target descriptions.

Table 1. Performance comparison of various models on the COCO Caption and Flickr30k datasets for the image captioning task

Model	Dataset	BLEU-1	BLEU-2	CIDEr	METEOR
LSTM-Attention	COCO Caption	73.2	55.4	112.8	27.1
Dual-Attention	COCO Caption	76.5	58.9	123.4	28.3
M4C	COCO Caption	79.1	61.2	134.7	30.2
Proposed Model	COCO Caption	82.4	64.7	145.3	32.5
LSTM-Attention	Flickr30k	68.9	52.1	97.6	26.2
Dual-Attention	Flickr30k	72.8	55.4	108.3	27.4
M4C	Flickr30k	74.5	57.1	115.7	29.1
Proposed Model	Flickr30k	77.8	60.2	126.5	30.8

In the image-text matching task, <Table 2> presents the Top-1 and Top-5 matching accuracies on the COCO Caption and Flickr30k datasets. The experimental results show that the proposed model achieves a Top-1 matching accuracy of 64.3%, which is a significant improvement compared to M4C's 59.2%. This indicates that the introduction of the cross-attention mechanism enables the model to better capture fine-grained interactions between image and text features.

Table 2. Top-1 and Top-5 matching accuracy comparison of various models on the COCO Caption and Flickr30k datasets for the image-text matching task

Model	Dataset	Top-1 Accuracy	Top-5 Accuracy
LSTM-Attention	COCO Caption	53.1	78.4
Dual-Attention	COCO Caption	58.7	81.3
M4C	COCO Caption	59.2	83.5
Proposed Model	COCO Caption	64.3	87.2
LSTM-Attention	Flickr30k	48.7	73.9
Dual-Attention	Flickr30k	52.9	76.5
M4C	Flickr30k	55.1	79.8
Proposed Model	Flickr30k	60.4	83.1

To further evaluate the contribution of each module in the image captioning task, ablation experiments were conducted on the COCO Caption dataset. <Table 3> summarizes the ablation results using CIDEr scores as the evaluation metric. The results show that after removing the cross-attention mechanism, the CIDEr score drops by 8.5 points, while removing the CBAM module results in a 5.3-point decrease in CIDEr score. This indicates that both the cross-attention mechanism and CBAM module play critical roles in enhancing model performance.

Table 3. Ablation analysis of different model configurations on the CIDEr scores for the image captioning task

Model Configuration	CIDEr Score
Full Model	145.3
Remove Cross-Attention	136.8
Remove CBAM Module	140.0
Visual Features Only	121.4
Text Features Only	118.9

Based on the experimental results, the proposed cross-attention fusion model demonstrates superior performance compared to existing multimodal fusion methods in multimodal tasks. The specific analysis is as follows:

1. Image Captioning Task: The proposed model significantly outperforms other models in BLEU and CIDEr evaluation metrics, indicating that the cross-attention mechanism effectively captures fine-grained interactions between visual and language modalities, generating more coherent and precise text descriptions.

2. Image-Text Matching Task: The proposed model achieves higher performance in Top-1 and Top-5 matching accuracies compared to existing methods, validating the cross-attention mechanism's dynamic alignment capability between different modality features, thereby improving matching accuracy.

3. Ablation Study Analysis: The ablation study results confirm that the cross-attention mechanism and CBAM module are essential for improving model performance. Removing these modules results in a significant performance drop, further demonstrating their critical roles in the model.

In conclusion, through a series of quantitative and qualitative experiments, the proposed model achieves excellent performance in multimodal fusion tasks, validating the effectiveness of the cross-attention mechanism in enhancing feature fusion capability and overall performance. Future research can further explore the combination of other attention mechanisms and multimodal feature representation methods to advance multimodal learning.

5. Conclusion

This paper proposes a visual-language multimodal fusion model based on the cross-attention mechanism, and its superior performance is validated through various experiments on image captioning, image-text matching, and other tasks. Experimental results demonstrate that the proposed model effectively captures fine-grained interactions between visual and language modalities, significantly improving classification accuracy and the quality of generated text descriptions. Compared to traditional methods, the introduction of the cross-attention mechanism enhances semantic alignment between modalities. Additionally, the ablation study further validates the critical role of the cross-attention mechanism and the CBAM module in improving model performance. Future research can explore more efficient attention mechanisms and feature fusion strategies to further advance multimodal learning.

References

[1] Tan, C., Zhang, W., Qi, Z., et al. (2025). Generating multimodal images with GAN: Integrating text, image, and style. *arXiv preprint arXiv:2501.02167*.

[2] Zhang, J., Xiang, A., Cheng, Y., et al. (2024). Research on detection of floating objects in river and lake based on AI image recognition. *Journal of Artificial Intelligence Practice, 7*(2), 97-106.

[3] Xiang, A., Zhang, J., Yang, Q., et al. (2024). Research on splicing image detection algorithms based on natural image statistical characteristics. *arXiv preprint arXiv:2404.16296*.

[4] Liu, J., et al. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. *Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8*(2), 243-260.

[5] Qi, Z., Ma, D., Xu, J., et al. (2024). Improved YOLOv5 based on attention mechanism and FasterNet for foreign object detection on railway and airway tracks. *arXiv preprint arXiv:2403.08499*.

[6] Wang, T., Cai, X., & Xu, Q. (2024). Energy market price forecasting and financial technology risk management based on generative AI. *Applied and Computational Engineering, 100*, 29-34.

[7] Wu, X., Liu, X., & Yin, J. (2024). Multi-class classification of breast cancer gene expression using PCA and XGBoost. *Preprints*, 2024101775. https://doi.org/10.20944/preprints202410.1775.v3

[8] Min, L., et al. (2024). Financial prediction using DeepFM: Loan repayment with attention and hybrid loss. In *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*. IEEE.

[9] Wu, Z. (2024). An efficient recommendation model based on knowledge graph attention-assisted network (kgatax). *arXiv preprint arXiv:2409.15315*.

[10] Wang, H., Zhang, H., & Lin, Y. (2024). RPF-ELD: Regional prior fusion using early and late distillation for breast cancer recognition in ultrasound images. *Preprints*. https://doi.org/10.20944/preprints202411.1419.v1

[11] Qi, Z., Ding, L., Li, X., et al. (2024). Detecting and classifying defective products in images using YOLO. *arXiv preprint arXiv:2412.16935*.

[12] Yan, H., et al. (2024). Research on image generation optimization based deep learning. In *Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering*.

[13] Tang, X., et al. (2024). Research on heterogeneous computation resource allocation based on data-driven method. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*. IEEE.

[14] Wu, Z., Chen, J., Tan, L., et al. (2024). A lightweight GAN-based image fusion algorithm for visible and infrared images. In *2024 4th International Conference on Computer Science and Blockchain (CCSB)*. IEEE, 466-470.

[15] Mo, K., Chu, L., Zhang, X., et al. (2024). DRAL: Deep reinforcement adaptive learning for multi-UAV navigation in unknown indoor environment. *arXiv preprint arXiv:2409.03930*.

[16] Zhang, W., Huang, J., Wang, R., et al. (2024). Integration of Mamba and Transformer--MAT for long-short range time series forecasting with application to weather dynamics. *arXiv preprint arXiv:2409.08530*.

[17] Zhao, Y., Hu, B., & Wang, S. (2024). Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition. *arXiv preprint arXiv:2409.12376*.

[18] Zhao, Y., Hu, B., & Wang, S. (2024). Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition. *arXiv preprint arXiv:2409.12376*.

[19] Diao, S., et al. (2024). Ventilator pressure prediction using recurrent neural network. *arXiv preprint arXiv:2410.06552*.

[20] Gao, D., et al. (2023). Synaptic resistor circuits based on Al oxide and Ti silicide for concurrent learning and signal processing in artificial intelligence systems. *Advanced Materials, 35*(15), 2210484.

[21] Shi, X., Tao, Y., & Lin, S. C. (2024). Deep neural network-based prediction of B-cell epitopes for SARS-CoV and SARS-CoV-2: Enhancing vaccine design through machine learning. *arXiv preprint arXiv:2412.00109*.

[22] Wang, B., Chen, Y., & Li, Z. (2024). A novel Bayesian Pay-As-You-Drive insurance model with risk prediction and causal mapping. *Decision Analytics Journal, 13*, 100522.

[23] Li, Z., Wang, B., & Chen, Y. (2024). Incorporating economic indicators and market sentiment effect into US Treasury bond yield prediction with machine learning. *Journal of Infrastructure, Policy and Development,*

8(9), 7671.

[24] Zhao, R., Hao, Y., & Li, X. (2024). Business analysis: User attitude evaluation and prediction based on hotel user reviews and text mining. *arXiv preprint arXiv:2412.16744*.

[25] Guo, H., Zhang, Y., Chen, L., et al. (2024). Research on vehicle detection based on improved YOLOv8 network. *arXiv preprint arXiv:2501.00300*.

[26] Xu, Q., Wang, S., & Tao, Y. (2025). Enhancing anti-money laundering detection with self-attention graph neural networks. *Preprints*. https://doi.org/10.20944/preprints202501.0587.v1

[27] Ziang, H., Zhang, J., & Li, L. (2025). Framework for lung CT image segmentation based on UNet++. *arXiv preprint arXiv:2501.02428*.

[28] Weng, Y., & Wu, J. (2024). Fortifying the global data fortress: A multidimensional examination of cyber security indexes and data protection measures across 193 nations. *International Journal of Frontiers in Engineering Technology, 6*(2), 13-28.

[29] Wu, Z. (2024). Large language model-based semantic parsing for intelligent database query engine. *Journal of Computer and Communications, 12*(10), 1-13.

[30] Wang, Z., et al. (2024). Improved Unet model for brain tumor image segmentation based on ASPP-coordinate attention mechanism. In *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*. IEEE.

[31] Liu, D. (2024). Mt2st: Adaptive multi-task to single-task learning. *arXiv preprint arXiv:2406.18038*.

[32] Luo, D. (2024). Enhancing smart grid efficiency through multi-agent systems: A machine learning approach for optimal decision making. *Preprints preprints, 202411*, v1.

[33] Luo, D. (2024). Quantitative risk measurement in power system risk management methods and applications. *Preprints*. https://doi.org/10.20944/preprints202411.1636.v1

[34] Luo, D. (2024). Decentralized energy markets: Designing incentive mechanisms for small-scale renewable energy producers. *Preprints*. https://doi.org/10.20944/preprints202411.0696.v1

[35] Li, Z., Wang, B., & Chen, Y. (2024). Knowledge graph embedding and few-shot relational learning methods for digital assets in USA. *Journal of Industrial Engineering and Applied Science, 2*(5), 10-18.

[36] Weng, Y., & Wu, J. (2024). Leveraging artificial intelligence to enhance data security and combat cyber attacks. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 5*(1), 392-399.

[37] Liu, Dong, et al. (2024). Graphsnapshot: Graph machine learning acceleration with fast storage and retrieval. *arXiv preprint arXiv:2406.17918*.

[38] Tan, C., Li, X., Wang, X., et al. (2024). Real-time video target tracking algorithm utilizing convolutional neural networks (CNN). In *2024 4th International Conference on Electronic Information Engineering and Computer (EIECT)*. IEEE, 847-851.

[39] Liu, D. (2024). Contemporary model compression on large language models inference. *arXiv preprint arXiv:2409.01990*.

[40] Li, Z., Wang, B., & Chen, Y. (2024). A contrastive deep learning approach to cryptocurrency portfolio with US treasuries. *Journal of Computer Technology and Applied Mathematics, 1*(3), 1-10.

[41] Weng, Y., Wu, J., Kelly, T., et al. (2024). Comprehensive overview of artificial intelligence applications in modern industries. *arXiv preprint arXiv:2409.13059*.

[42] Huang, B., Lu, Q., Huang, S., et al. (2024). Multi-modal clothing recommendation model based on large model and VAE enhancement. *arXiv preprint arXiv:2410.02219*.

[43] Li, Z., Wang, B., & Chen, Y. (2024). Knowledge graph embedding and few-shot relational

learning methods for digital assets in USA. *Journal of Industrial Engineering and Applied Science, 2*(5), 10-18.

[44] Zhao, P., & Lai, L. (2024). Minimax optimal q learning with nearest neighbors. *IEEE Transactions on Information Theory*.

[45] Feng, J., Wu, Y., Sun, H., Zhang, S., & Liu, D. (2025). Panther: Practical secure 2-party neural network inference. *IEEE Transactions on Information Forensics and Security*.

Copyrights