doi:10.3772/j.issn.1006-6748.2025.01.008

MKGViLT: visual-and-language transformer based on medical knowledge graph embedding①

CUI Wencheng(崔文成), SHI Wentao②, SHAO Hong
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, P.R. China)

Abstract

Medical visual question answering (MedVQA) aims to enhance diagnostic confidence and deepen patients' understanding of their health conditions. While the Transformer architecture is widely used in multimodal fields, its application in MedVQA requires further enhancement. A critical limitation of contemporary MedVQA systems lies in the inability to integrate lifelong knowledge with specific patient data to generate human-like responses. Existing Transformer-based MedVQA models require enhancing their capabitities for interpreting answers through the applications of medical image knowledge. The introduction of the medical knowledge graph visual language transformer (MKGViLT), designed for joint medical knowledge graphs (KGs), addresses this challenge. MKGViLT incorporates an enhanced Transformer structure to effectively extract features and combine modalities for MedVQA tasks. The MKGViLT model delivers answers based on richer background knowledge, thereby enhancing performance. The efficacy of MKGViLT is evaluated using the SLAKE and P-VQA datasets. Experimental results show that MKGViLT surpasses the most advanced methods on the SLAKE dataset.

Key words: knowledge graph(KG), medical vision question answer (MedVQA), vision-and-language transformer

0 Introduction

Medical visual question answering (MedVQA) integrates medical imaging and natural language processing(NLP) techniques to analyze medical images and provide accurate responses to natural language queries. This endeavor seeks to enhance the quality of healthcare services for both physicians and patients.

In modern medicine, it is necessary to assess the condition of the human body accurately; physicians often utilize various non-invasive medical sensors, including computed tomography and magnetic resonance imaging (MRI), to acquire essential body data. Advances in medical imaging technology have significantly contributed to the progress of the medical field[1-2]. Analyzing medical images is a crucial skill for medical assistive systems, and MedVQA systems[3] have been developed for this purpose. Existing MedVQA methods usually consist of four parts: extraction of visual features from queried medical images, extracting text features from queried medical problems, combining the previously proposed two modal features, and subsequently predicting the answer. Most MedVQA solutions leverage advanced deep learning techniques, commonly relying on methodologies like recursive neural networks[4-6] for text embedding and feature extraction, and convolutional neural networks (CNN) for visual feature extraction. The bidirectional encoder representations from Transformer (BERT)[7] has emerged as a preeminent model for representing textual information, enjoying widespread usage. BERT employs a bidirectional attention mechanism and a large-scale unsupervised corpus to generate contextual representations for each word in a phrase, ensuring relevance. The remarkable success of the Transformer in NLP has sparked significant interest in the field of computer vision.

Consequently, recent research in computer vision has generated heightened interest in the Transformer architecture and its components. Vision transformer (ViT)[8] is the first vision architecture to treat image chunks as words and encode them using Transformers. When trained on large datasets, ViT achieved impressive results in image recognition, as reported in Ref. [9].

① Supported by the National Natural Science Foundation of China (No. 62001313), the Liaoning Professional Talent Protect (No. XLYC2203046) and the Shenyang Municipal Medical Engineering Cross Research Foundation of China (No. 22-321-32-09).

② To whom correspondence should be addressed. E-mail:1737638110@qq.com.

Received on Mar. 15, 2024

Nevertheless, access to adequate training data can be challenging in the medical domain, especially when dealing with rare cases and limited sample sizes. This limitation can result in performance degradation when ViT is applied to MedVQA tasks. The underlying reason is that, despite its power, the Transformer model has a limited inductive bias that can facilitate overfitting. To normalize model capacity and enhance scalability, subsequent studies have explored customized sparse Transformer models tailored for visual tasks like local attention[10-12]. By reintroducing hierarchical architecture, these approaches aim to compensate for non-locality loss. The Transformer architecture can capture global features comparable to convolutional neural networks.

As Wang et al.[13] suggested, humans can utilize their past experiences when responding to questions related to visual content. To address this, the Facted-VQA benchmark was introduced to present a series of more complex questions, ensuring that the desired answers necessitate external knowledge beyond image-derived textual descriptions. Recently, knowledge graphs (KGs) have demonstrated their effectiveness in information retrieval and found broad applications across diverse domains, including search engines[14], as well as natural language comprehension[15], among others. KG representation learning aims to project entities and relations into a continuous low-dimensional vector space, which can implicitly facilitate the computation of inference between entities and is a knowledge-intensive task[16-18]. Like Facted-VQA, the MedVQA, SLAKE[19] and P-VQA[20] datasets have presented new research challenges and opportunities. These datasets emphasize the dependence of answers on external knowledge. By providing a knowledge graph, these datasets facilitate the extraction of answers by representing complex relationships between entities in a multi-relational manner.

To achieve an ideal MedVQA system, it is necessary to emulate the expertise of a professional doctor who can seamlessly integrate the knowledge accumulated throughout their career to address image-based queries. Nevertheless, in the proposed MedVQA task, the central challenge lies in retrieving the most pertinent and accurate entities from the knowledge graph.

The primary contributions of this work are as follows.

(1) This paper presents the medical knowledge graph visual language transformer (MKGViLT), a visual language Transformer specially tailored for federating medical knowledge graphs. MKGViLT seamlessly integrates visual entities from images, textual entities from questions, and external knowledge about named entities and their relationships within a medical image. This integration fosters a deeper understanding of the visual content, offering a unified representation of modal resources.

(2) The medical knowledge graph modules (MKGM) model has been engineered for knowledge retrieval, utilizing a Transformer architecture that presents a unique approach compared with traditional models. In contrast to conventional methods, the MKGM model demonstrates reduced vulnerability to significant fluctuations during training. The Transformer facilitates model convergence and alleviates oscillations during the training process.

(3) Extensive experiments are conducted on the benchmark datasets SLAKE[19] and P-VQA[20], and experimental results show that the model outperformed state-of-the-art methods on the SLAKE dataset.

1 Related work

1.1 Medical vision question answer

Conventional MedVQA, building upon the advancements in VQA technology for natural images, is gradually emerging in the medical domain. Since the inaugural VQA-Med task was introduced in the 2018 ImageCLEF 2018 competition[1], VQA has found application in the medical domain. Most MedVQA methods[21-22] are straightforward adaptations of cutting-edge generalized VQA models to the medical domain. These methods and their outcomes are summarized in the 2018 ImageCLEF-Med Challenge report. Attentional mechanisms such as multimodal compact bilinear pooling(MCB)[23], stacked attention networks (SAN)[24] and bilinear attention networks (BAN)[25] have been utilized to address the problem of learning joint representations between medical visual and textual information. Furthermore, Refs. [26,27] employed transfer learning techniques to extract medical image features. Recently, methods that directly tackle various aspects of MedVQA have been introduced, including methods for diagnosing model behavior, specialized framework design, and techniques for generating models to handle anomalies.

In recent years, the Transformer architecture has become pivotal in the multimodal domain due to its superior capability in processing sequential data. In MedVQA, the Transformer is primarily divided into two main models: (1) the single-stream model, which utilizes self-attention to establish connections within and between various modalities, exemplified by Refs. [28,29];

(2) the dual-stream model, which emphasizes the establishment of connections between modalities through techniques such as cross-attention and co-attention. Relevant work also includes those conducted by Ref. [28]. However, these models need to address knowledge-based capabilities. Before the advent of single-channel Transformer models, the dual-flow model was the primary model employed. Nevertheless, the dual-stream model's excessive parameter count led to overfitting on most medical datasets, resulting in suboptimal resource utilization and limited benefits. Consequently, the single-stream model has gradually gained broader application.

1.2 Knowledge graph embedding methods

Based on spatial distance methods, the transE approach, introduced by Bordes et al. [30], depicts entities and relationships within a d-dimensional vector space $h, t, r \in R^d$. It ensures the embedding follows the translation principle rule: $h + r \approx t$. To address the limitation of transE in representing both entities and relations within a single space, Wang et al. [31] introduced individual hyperplanes for each relation. Each entity can be mapped to these hyperplanes, enabling a more comprehensive representation of the roles associated with each relation. This approach is instrumental in handling scenarios where there are multiple relationships between a pair of entities. Lin et al. [32] proposed transR to separate the space of entities and relationships by projecting the entity ($h, t \in R^k$) into the space of relationships $r \in R^d$ through the projection matrix $\boldsymbol{m}_r \in R^{k \times d}$. Although transR is based on linear transformation, it still faces limitations in effectively representing nonlinear or more complex relationships.

Based on semantic matching methods, Bordes et al. [33] conducted a semantic-level integration by aligning two sets of entity-relation pairs: (h, r) and (r, t). The scoring function used by Bordes et al. [33] consists of two distinct matching components: the linear and bilinear blocks. To effectively capture intricate interactions in relational data and promote efficient computation, Nickel et al. [34] employed a mechanism known as embedded cyclic correlation. The mechanism of embedded cyclic correlation can be interpreted as a compressed tensor product that enables the learning of combined representations for entities and relations. Zhou et al. [35] proposed a novel rotational combination mechanism for considering additional path information and symmetric relations. However, approaches based on translational distance and semantic matching typically have limited capacity to learn comprehensive feature and interaction information due to their simplified architectures, thereby limiting further enhancements in model performance.

2 Methods

The problem is defined as follows: given a medical image I, a related question Q, and a knowledge graph G containing head entities, relations, and tail entities, the objective of the MKGViLT task is to predict the answer Y. This prediction can be formally represented using mathematical notation, as shown in Eq. (1).

$$Y = \operatorname{argmax} F(AQ, I, G, \theta), A \in \pounds \tag{1}$$

where, $\pounds$ represents the set of potential answers; A denotes one of the answers; F and θ signify the MKGViLT framework and its parameters, respectively. Within the MKGViLT model, the Transformer encoder is employed to gather and merge information from the question, image, and knowledge graph.

2.1 Medical vision-and-language transformer for joint knowledge graph

Architecture overview: MKGViLT consists of five main components as shown in Fig. 1. (1) Image feature extraction using CNN-based ResNet-101 [36] to capture low-level features. (2) Problem feature extraction using a BERT-based model to capture contextual information. (3) Feeding of the problem into the MKGM model to capture the problem-related entity information and the relationship information. (4) A Transformer encoder is used to fuse the extracted image (visual) and problem (linguistic) features with graph information (textual) features and to model the high-level global features. (5) The encoded features are averaged, sampled, and regularized by the Transformer encoder and fed into the dense layer for final prediction.

Pre-trained models such as VGG-19 [37], DenseNet [38], and ResNet-101 are employed to extract image characteristics. The dimensions of image I are adjusted to match those of ResNet-101 (224, 224, 3). Since ResNet-101 is not utilized for classification, only two fully connected layers and the final average pooling layer output were retained as image features. The image I is fed into the ResNet101 network to get the medical image features as shown in Eq. 2.

$$\boldsymbol{V}_I = \text{ResNet101}(I) \tag{2}$$

The image feature matrix $\boldsymbol{V}_I$ is inputted into another kernel, a 3×3 two-dimensional(2D) convolutional layer, and passed to a dense layer for channel reduction. Reshaping and flattening are performed to maximize information retention and output an image feature matrix $\boldsymbol{V}_I'$. This design aligns the first dimension of the image

feature matrix with the first dimension of the question feature matrix while maintaining as much information as possible.

$$V_I^{'} = \text{Convolution2D}(V_I) \tag{3}$$

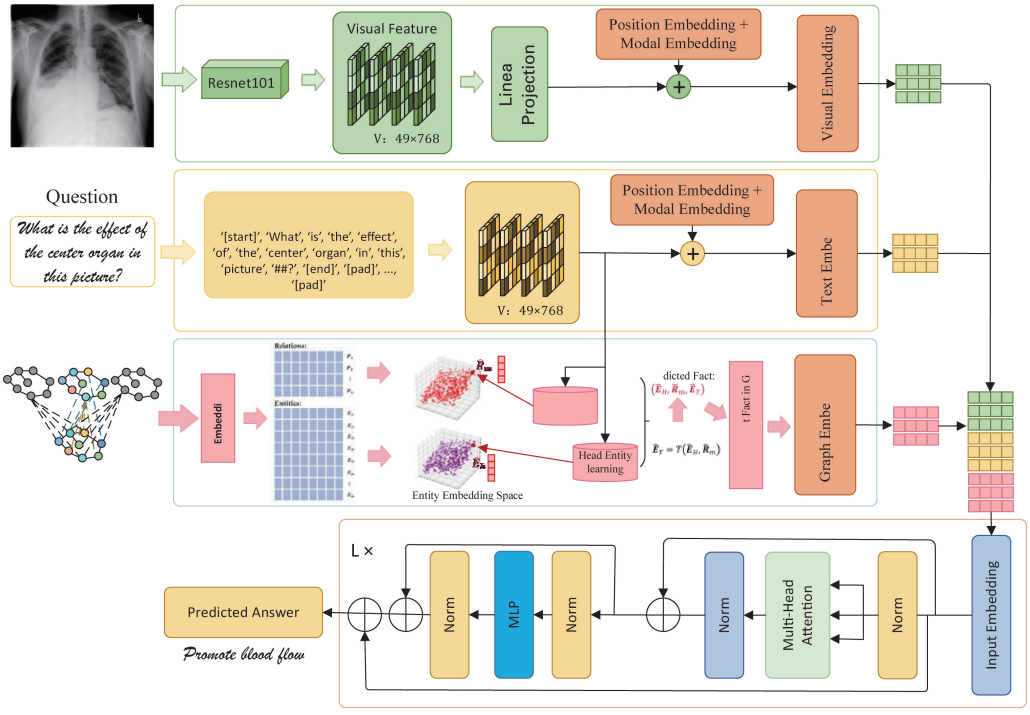

Fig. 1 Overview of the proposed MKGViLT model

All medical problems are lowercased to avoid having two identical words with different cases in the response. The transformer model requires that the input vector be one-dimensional (1D). Consequently, each input problem $q \in \delta^{S\times|V|}$ is pre-trained using BERT. This pre-training ensures a richer, contextually relevant representation of the word vectors. In this study, S represents the length of the problem sequence, and V denotes the dimensionality of the BERT word vector. Given that the Transformer cannot capture positional information and involves two modalities, positional embeddings $Q_{\text{pos}} \in \delta^{S\times D}$ and modal-type embeddings $Q_{\text{type}} \in \delta^{D}$ are introduced. The position embedding matrix and the corresponding modal type embedding vectors form the problem embedding $Q' \in \delta^{S\times D}$, with D indicating the number of dimensions required for input to the Transformer encoder:

$$Q' = \text{BERT}(q) + Q_{\text{pos}} + Q_{\text{type}} \tag{4}$$

When the MKGM module cannot directly process the textual information, the entity information E or relationship information R is transferred to the corresponding embedding spaces. The closest relevant information points are then identified using 1-nearest neighbor aggregation. This information is then represented using BERT embedding with a dimension size of $\delta^{B\times D}$, the parameter B denotes the amount of graph information.

Since there are two different modal features, an additional modal X_{class} needs to be added for differentiation, the question embedding Q', X_{class}, the image embedding $V_I^{'}$, and the knowledge graph information G are then concatenated into a sequence Z_0.

$$Z_0 = [Q^{'}; X_{\text{class}}; V_I^{'}; G] \tag{5}$$

The Transformer encoder is introduced next to acquire the ultimate coded sequence z_t. This procedure requires input to the Transformer encoder, and as shown in Eq. 6.

$$z_t = \text{Transformer Encoder}(z_0) \tag{6}$$

L deep Transformer layers iteratively update the context vector z_0 to achieve z_t. This Transformer encoder differs from the standard Transformer, as presented in subection 3.3. Finally, to predict the answer z_t, it is aggregated and averaged to derive the context sequence.

$$\bar{z}_t = \frac{1}{S+1+N}\sum_{i=0}^{S+N} z_t \tag{7}$$

Given that batch normalization (BN)[39] accelerates model convergence, it has been implemented in the output layer. $\bar{z}_t$ is fed into MLP and BN, resulting in the final predicted answer y. Lastly, cross-entropy is utilized as the classification loss:

$$y = \mathrm{MLP}(\mathrm{BN}(\mathrm{MLP}(\bar{z}_t))) \tag{8}$$

2.2 Medical knowledge graph modules

In MedVQA, answers are concise, and each tail entity can respond when the appropriate head entity and relationship are correctly identified. An MKGM model that leverages the GRU method is introduced to accurately predict the head and relation entities.

The knowledge graph embedding module relies on the embedding representation of all relations and entities, denoted as R and E, respectively. Existing KG embedding algorithms are employed to acquire entity and relation embeddings. Low-dimensional vectors represent each relation/entity in the KG, preserving its original structure and relations. BERT is utilized to generate the embedding representation (e_h, r_l, e_t) for every fact (h, l, t) in G. Subsequently, a function $\varphi(\cdot)$ is established to evaluate the relation of the fact (h, l, t) within the embedding space, $e_t \approx f(e_h, r_l)$. TransR[31] defines it as $e_t \boldsymbol{M}_l \approx e_h \boldsymbol{M}_l + p_l$, where $\boldsymbol{M}_l$ is the transformation matrix of relation l. For all facts in G, the embedding algorithm minimizes the overall distance between e_t and $f(e_h, r_l)$. One common approach involves establishing marginal ranking criteria by utilizing positive and negative samples, which encompass facts absent in G and composite facts, and subsequently training them, as depicted in Fig. 2, the goal of MKGM is not to infer head entities and relations directly, but to jointly recover the head entity, relation and tail entity representations of the problem in the knowledge graph embedding space. If the answer cannot be obtained directly by MKGM, the existing information is used to find the closest graph to it for encoding and then join it with visual embedding and text embedding. The space where the learned relational representation of $R_i (i = 1, 2, \ldots, m)$ is located is defined as the relational embedding space, and the space for $E_i (i = 1, 2, \ldots, N)$ is termed the entity embedding space.

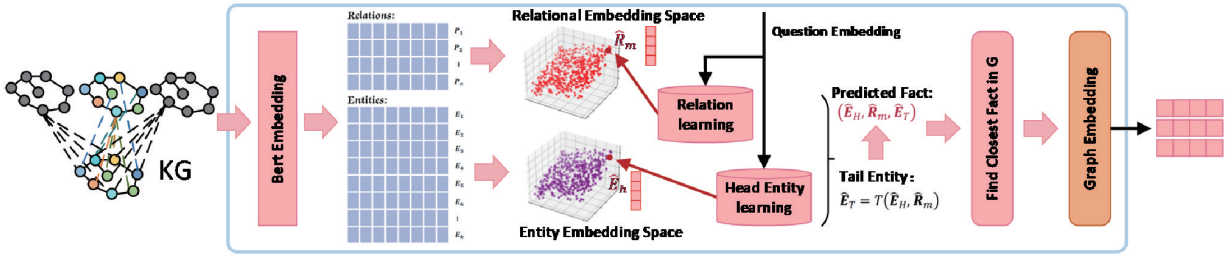

Fig. 2 Overview of the MKGM architecture

The goal is to identify a point in the relational embedding space, designated as $\hat{\boldsymbol{R}}_m$, that represents the relationships of the given question, and a corresponding point in the entity embedding space, labeled as $\hat{\boldsymbol{E}}_h$, representing its head entity. When a question can be answered by G, its relational vector representation must occupy a position within the relational embedding space. If G cannot answer a question, its entity vectors or relational vectors must exist within the entity embedding space or the relational embedding space. Consequently, the objective is to create a model that inputs a question and outputs either a vector $\hat{\boldsymbol{R}}_m$ with relationally embedded representations or entity-embedded representations $\hat{\boldsymbol{E}}_h$, closely aligned with the question.

To predict the relational terms within a problem, conventional methods typically depend on semantic parsing techniques paired with manually created thesauri to forge the requisite relationships[39]. Alternatively, each class of relation term could be classified into a labeled category, effectively turning it into a classification task[40]. However, these methods falter when dealing with the medical domain due to its unbounded nature, meaning relations within a new problem could potentially differ from all existing ties within the training data Q. Additionally, it has been identified that the global relational data stored in entities R and E is not only accessible but also potentially beneficial to the overall accuracy of MedVQA. In response, a neural network-based model for relationship learning has been proposed. The aim is to detect the representation of a problem within the knowledge graph (KG) embedding space. Therefore, it doesn't matter whether it's a head entity or a relational entity in question. The entity learning model aims to ascertain a vector $\hat{\boldsymbol{E}}_h$ that correlates as closely as possible with the embedded representation of the head entity related to that problem. The neural network architecture is used to predict both the head entity representation $\hat{\boldsymbol{E}}_h$ and relational entity representation $\hat{\boldsymbol{R}}_m$. The proposed solution, using gated recurrent unit (GRU)[41] as an example of a recurrent neural network, is illustrated in Fig. 3. Given a problem of length L, a pre-trained BERT model is used to

map an L-token problem into a sequence of word embedding vectors $\{\boldsymbol{x}_j\}$, $j = 1, \ldots, L$. Then GRU is utilized to learn the forward hidden state sequence $(\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_L)$ and the backward hidden state sequence $(\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_L)$. The backward direction, for example, $\{\overleftarrow{h}_j\}$ is computed by the following equations, where $\boldsymbol{W}$ and b represent the weight matrix and bias term.

$$\boldsymbol{f}_g = \sigma(W_{xf}\boldsymbol{x}_j + W_{hf}\overleftarrow{h}_{j+1} + b_f) \tag{9}$$

$$\boldsymbol{i}_g = \sigma(W_{xi}\boldsymbol{x}_j + W_{hi}\overleftarrow{h}_{j+1} + b_i) \tag{10}$$

$$\boldsymbol{o}_g = \sigma(W_{xo}\boldsymbol{x}_j + W_{ho}\overleftarrow{h}_{j+1} + b_o) \tag{11}$$

$$\boldsymbol{c}_g = \boldsymbol{f}_g \odot \boldsymbol{c}_{j+1} + i_g \tanh(W_{xc}\boldsymbol{x}_j + W_{hc}\overleftarrow{h}_{j+1} + b_c) \tag{12}$$

$$\overleftarrow{h}_j = \boldsymbol{o}_g \odot \tanh(\boldsymbol{c}_g) \tag{13}$$

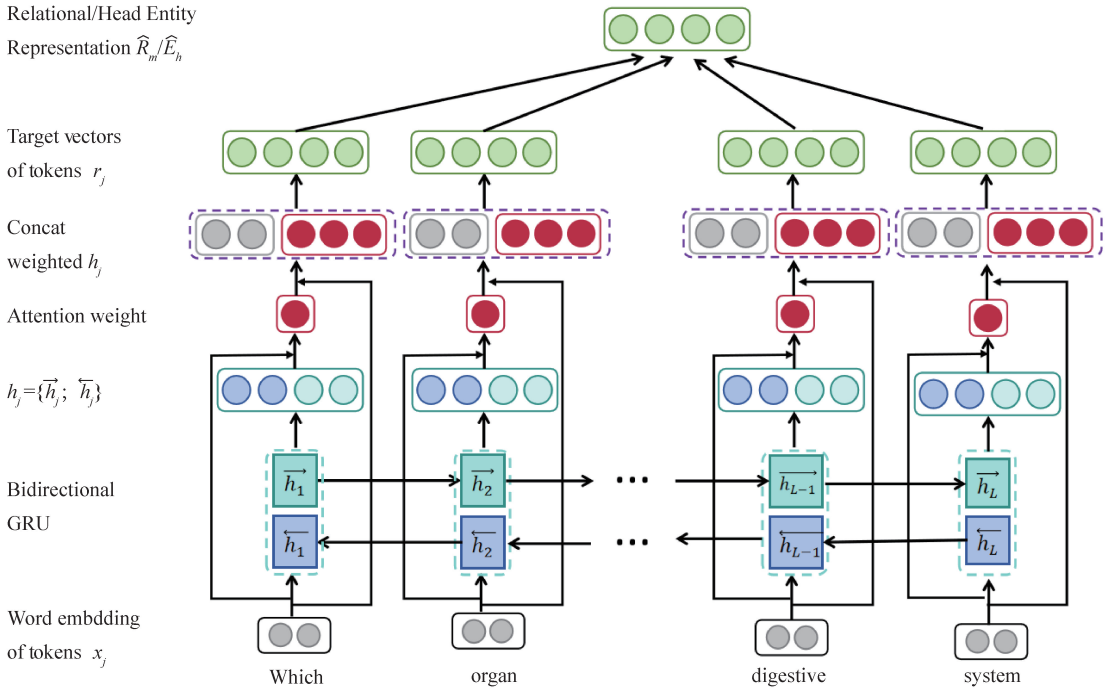

Fig. 3 Proposed learning model for entity embedding space and relationship embedding space

where $\boldsymbol{f}_g$, $\boldsymbol{i}_g$ and $\boldsymbol{o}_g$ are the activation vectors of the oblivion, input and output gates respectively; $\boldsymbol{c}_g$ is the cell state vector; σ and tan$h$ are the Sigmoid and Hyperbolic tangent functions respectively; [$\odot$] is the Hadamard operation. Splice the forward and backward hidden state vectors to obtain $h_j = \{\overrightarrow{h}_j, \overleftarrow{h}_j\}$. The attentional weight of the jth token, i. e., a_j is calculated based on the following equations.

$$a_j = \frac{\exp(q_j)}{\sum_{i=1}^{L}\exp(q_i)} \tag{14}$$

$$q_j = \tanh(\boldsymbol{w}^{\mathrm{T}}\{\boldsymbol{x}_j; h_j\} + b_q) \tag{15}$$

The attention weight a_j is applied to h_j and concatenated with the embedding vectors $\boldsymbol{x}_j$ to obtain hidden state $s_j = \{x_j; a_j; h_j\}$. A fully-connected layer is then applied to s_j, with the result $\boldsymbol{r}_j \in \delta^{d\times 1}$ denoted as the target vector of the j-th token. The predicted relation denotes $\widehat{R}_m$ is computed as the average of the target vectors of all tokens, i. e:

$$\widehat{R}_m = \frac{1}{L}\sum_{j=1}^{L}\boldsymbol{r}_j^{\mathrm{T}} \tag{16}$$

The weight matrices, weight vectors $\boldsymbol{w}$, and bias terms are calculated from the training data, which is the problem and the relational embedding representation of the problem in Q.

2.3 Transformer encoder

Develop a modified Transformer encoder that differs from the standard version to extract visual, textual, and fusion features using a unified approach. As depicted in Fig. 4, the modified Transformer incorporates increased LayerNorm (LN)[42] and residual connections between layers. Like the standard Transformer, the modified encoder accepts a 1D token embedding sequence as input, with the context vector $\boldsymbol{z}_0$ fed into the layer Transformer encoder. After several itera-

tions of updates, z_L is obtained. This iterative process can be formally represented as

$$z''_l = \mathrm{LN}(\mathrm{LN}(\mathrm{MSA}(z_{l-1}))) + z_{l-1} \quad l = 1,2,\cdots,L \quad (17)$$

$$z'_l = \mathrm{LN}(\mathrm{MLP}(\mathrm{LN}(z''_l))) + z''_l \quad l = 1,2,\cdots,L \quad (18)$$

$$z_L = z'_l + z_{l-1} \quad l = 1,2,\cdots,L \quad (19)$$

The enhanced Transformer structure improves model stability and convergence speed by incorporating additional LayerNorms and introducing residual connections between layers. The inclusion of LayerNorm strengthens the data distribution, while integrating residual layers effectively addresses network degradation and gradient vanishing issues.

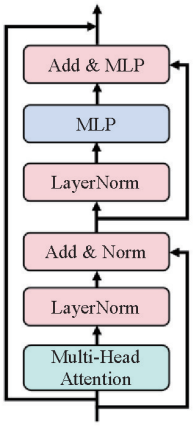

Fig. 4 The proposed Transformer encoder

3 Experiments

3.1 Datasets

SLAKE[19] dataset includes questions which are classified by response type: 'Closed' questions with fixed options and 'Open' questions allowing free-text responses. The English version, SLAKE-EN, is used, containing two radiology images and 7 033 question-answer pairs. SLAKE categorizes questions into two types: knowledge graph-based questions (abbreviated as 'KG') and vision-based questions. The SLAKE knowledge graph encompasses two primary attributes: disease-relationship and organ-relationship.

P-VQA[20] dataset has been developed to establish a VQA system tailored to patients, encompassing the entire treatment process, which includes medical consultations and diagnostic imaging. It comprises 20 prevalent diseases, 2 169 medical images, 24 800 question-answer pairs, and a medical knowledge graph containing 419 entities. It is essential to highlight that 34% of the image-question pairs within the P-VQA dataset feature two or more answers, indicating a multiple-choice format.

3.2 Experimental setup

3.2.1 Dataset splitting

SLAKE-EN[19] dataset consists of 4 919 question-answer pairs and 550 medical images. Validation and testing of the dataset utilized 92 medical images and 1 061 question-answer pairs. For the P-VQA[20] dataset, 1 518 medical images and 13 360 question-answer (Q&A) pairs were used for training, and 3 720 Q&A pairs and 325 medical images were used for testing.

3.2.2 MKGM

To evaluate the MKGM method's performance, the traditional setup[43] is followed, using the same training, validation, and testing splits as originally provided in SimpleQuestions[44]. 2010 i2b2/VA[45] is used as G in KG. The KG embedding algorithm, TransR[31], is applied to G to learn R and E. The MKGM method is used to predict the head entities and relations for each question in the test split.

3.2.3 MKGViLT implementation details

The experiment is conducted in an Ubuntu environment using an NVIDIA RTX 3090 24 GB GPU to train various models, including CNN variant networks, MKGM, and Transformer. The hyper-parameter settings for the MKGViLT model trained on different datasets are shown in Table 1. Accuracy serves as the evaluation metric. Additionally, the mixup data enhancement method[46] is employed to enhance the model's generalization performance on images. All models are trained on SLAKE-EN[19] and P-VQA[20] to obtain the final results.

Table 1 The model hyper-parameters settings

Hyper-parameters	Batch size	Weight decay	Iteration	Epoch	Lr
SLAKE-EN[19]	128	0.001	1 060	1 600	0.000 1
P-VQA[20]	64	0.002	702	1 600	0.000 1

3.3 Result

To address the problem, the pre-trained BERT model is used to obtain word embeddings, and the text embedding parameters Q_{pos} and Q_{type} are trained from scratch. Employ ResNet-101[36] to process the images to extract fundamental feature maps. Similar to the previous study, V_{pos} and V_{type} are trained without any pre-

training. The additional modal distinction X_{class} is set to either 0 or 1. The extracted image and text features, along with the distinguished representations, are fed into the Transformer for further feature extraction and fusion. Finally, the output of the Transformer is averaged and fed into the fully connected layer for making predictions. The results of SLAKE-EN [19] are summarized in Table 2, where ± represents an interval value derived from the three initialization methods: random initialization, Xavier initialization, and He initialization. As shown in Table 2, the MKGViLT model surpasses the recently published baseline CPRD + BAN + CR[47], which incorporates external image pretraining and achieves state-of-the-art (SOTA) performance. The MKGViLT model employs ResNet-101 [36] to extract local features and uses a custom Transformer for global feature extraction, leading to excellent results. The MKGViLT model performs better in addressing 'closed' problems and KG accuracy, achieving improvements of 2.3% and 2.7%, respectively, compared with the state-of-the-art baseline MKBN_MGE[20] on the SLAKE-EN datasets.

Table 2 Accuracy of the baseline and the proposed MKGViLT tested on SLAKE-EN[19] (%)

Methods	Overall	Open	Closed	KG
MFB[48]	73.30	72.20	75.00	/
SAN[19]	75.40	72.20	79.80	70.30
BAN[47]	76.30	74.60	79.10	/
MEVF + BAN + CR[47]	80.00	78.80	82.00	/
CPRD + BAN[47]	81.10	79.50	83.40	/
CPRD + BAN + CR[47]	82.10	81.20	83.40	/
MKBN_MGE[20]	80.60	77.70	85.10	75.70
MKGViLT	83.60 ± 0.34 ↑	79.90 ± 0.27 ↑	87.40 ± 0.18 ↑	78.40 ± 0.46 ↑

The results of P-VQA are shown in Table 3. The results indicate that the MKGViLT model significantly outperforms traditional models in terms of performance, and shows comparable performance compared to methods that use external data for pre-training.

Table 3 Performance of the baseline and the proposed MKGViLT tested on P-VQA[20] (%)

Methods	*Accuracy*	*Recall*	*Precision*	*F*1
MKBN_TransE[20]	94.29 ± 0.71	96.70 ± 0.17	96.44 ± 0.63	96.46 ± 0.35
MKBN_TransH[20]	94.30 ± 0.90	96.58 ± 0.37	96.39 ± 0.44	96.38 ± 0.42
MKBN_ConvKB[20]	94.97 ± 0.34	97.36 ± 0.40	97.13 ± 0.34	97.18 ± 0.15
MKBN_KBGAT[20]	94.69 ± 0.67	97.01 ± 0.44	96.80 ± 0.44	96.81 ± 0.46
MKBN_MGE[20]	95.45 ± 0.13	97.49 ± 0.26	97.43 ± 0.24	97.40 ± 0.23
MKGViLT	96.68 ± 0.42	97.94 ± 0.63	96.48 ± 0.84	97.55 ± 0.48

3.4 Comparison of different Transformer architectures

This section compares the performance of various Transformer architectures, focusing on model stability and convergence efficiency. Fig. 5 illustrates the experimental outcomes, highlighting the enhanced stability and accelerated convergence of the Transformer structure in test set evaluations. The comparative analysis encompasses the original Transformer architecture[49], the vision transformer (ViT)[8], and the proposed structure, ensuring uniform configuration across all experiments. The findings reveal that the model has superior accuracy compared to the other two architectures. Compared with the original Transformer architecture, the proposed model demonstrates superior performance in terms of both convergence speed and accuracy. This improvement is attributed to the incorporation of LayerNorm and Layer Residual connections, which enhance model stability during training and contribute to superior results.

The introduction of residual connections between network layers can impact data processing stability. To address this issue, LayerNorm is introduced to maintain consistent data distribution. LayerNorm stabilizes the network's data distribution, enhancing the model's generalization capabilities. Applying LayerNorm after multi-head attention ensures the stability of the Transformer structure and makes the model more reliable in processing data. Table 4 lists the results of the Transformer structure with two other structures. It is evident that the model's accuracy has improved by 2 to 3 percentage points, indicating that the Transformer structure is better suited for processing small-scale datasets. Compared with other models, this model exhibits better performance and generalization ability when dealing

with small-scale data. This improvement can be attributed to the Transformer structure's enhanced ability to capture data features more effectively and possess a more robust generalization capability, enabling more effective handling of small-scale data.

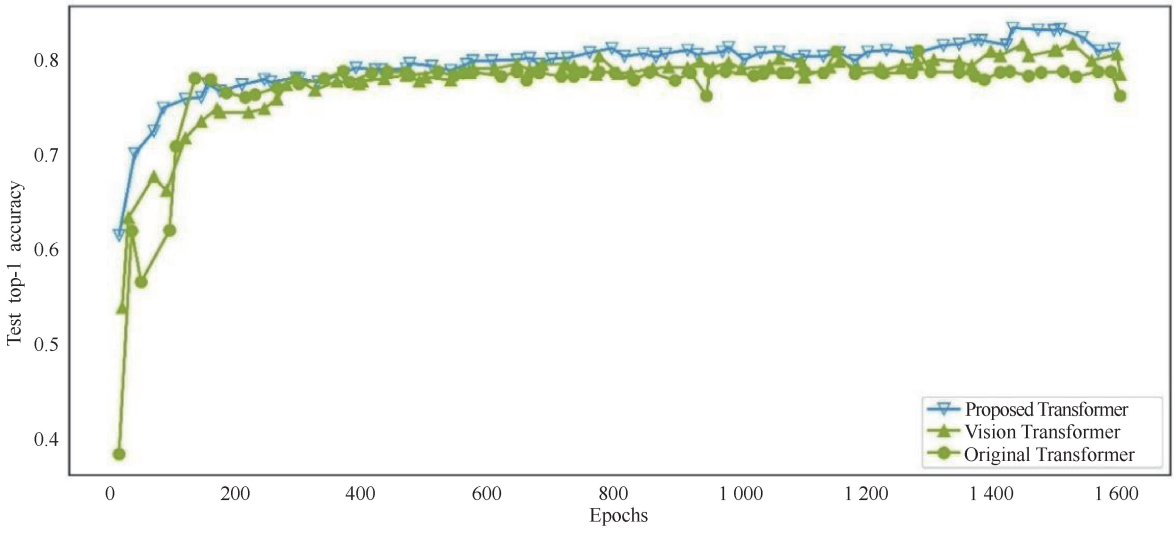

Fig. 5 Comparison of the proposed Transformer encoder with other encoders

Table 4 Testing the impact of Transformers with different structures on the SLAKE-EN[19] (%)

Methods	Overall	Open	Closed
Original Transformer	81.80	77.80	87.20
ViT Transformer	81.50	77.60	86.30
Proposed Transformer	83.60 ± 0.34	79.90 ± 0.27	87.40 ± 0.18

3.5 Ablation

This section assesses the effectiveness by investigating the influence of each submodule on MKGViLT. To enhance ablation comparison, the experiments focus narrowly on SLAKE-EN. The same configuration as previously is used for SLAKE-EN, encompassing 4 919 Q&A pairs, 550 medical images for training, and 92 medical images with 1 061 Q&A pairs for validation and testing.

3.5.1 Impact of convolutional module use on MKGViLT models

This study assesses the impact of the convolution module on MKGViLT by comparing scenarios with and without its use. Experimental results are detailed in Table 5. 'None' indicates that images are segmented directly into small blocks and then mapped to the required Transformer dimensions via fully connected layers, similar to operations in ViT. Integrating established feature extraction networks like VGG-16[37] or DenseNet[38] into MKGViLT significantly enhances model performance. Mainly, there is a substantial accuracy increase of 16% when utilizing ResNet-101 compared with ViT without employing the convolution framework, as highlighted in Fig. 6. The experimental outcomes suggest that the model with the integrated CNN module not only demonstrates greatly improved performance, addressing the deficiency in training resources needed for the Transformer, but also exhibits faster convergence. ViT's method of slicing and scaling images needs to be more complex and crude, disrupting the original image's structure and overall content and hindering the model's ability to learn more comprehensive visual features. Incorporating ResNet-101 into MKGViLT effectively extracts local image features and preserves image relationships when focusing directly.

Table 5 Testing the impact of ResNet-101 on MKGViLT performance on the SLAKE-EN[19] (%)

CNN-model	Overall	Open	Closed
None	67.6	66.4	72.3
DenseNet[38]	78.8	77.3	83.2
VGG-16[37]	81.1	79.4	86.7
ResNet-101[36]	83.6	79.9	87.4

3.5.2 Impact of using the MKGM module on MKGViLT

The efficacy of MKGM is assessed by conducting a comparative analysis of MKGViLT's performance with and without MKGM integration. After accurately predicting the entities and relations in the question, the relevant information of the 1-nearest neighbor in the constructed knowledge graph is located. This information was integrated with previous visual and textual data

before being input into the network for fusion-based answer prediction. Experimental results, as shown in Table 6, indicate that incorporating the MKGM module improved accuracy on the SLAKE-EN dataset. Fig. 7 further illustrates that the overall stability and accuracy of MKGViLT have been enhanced through the integration of MKGM.

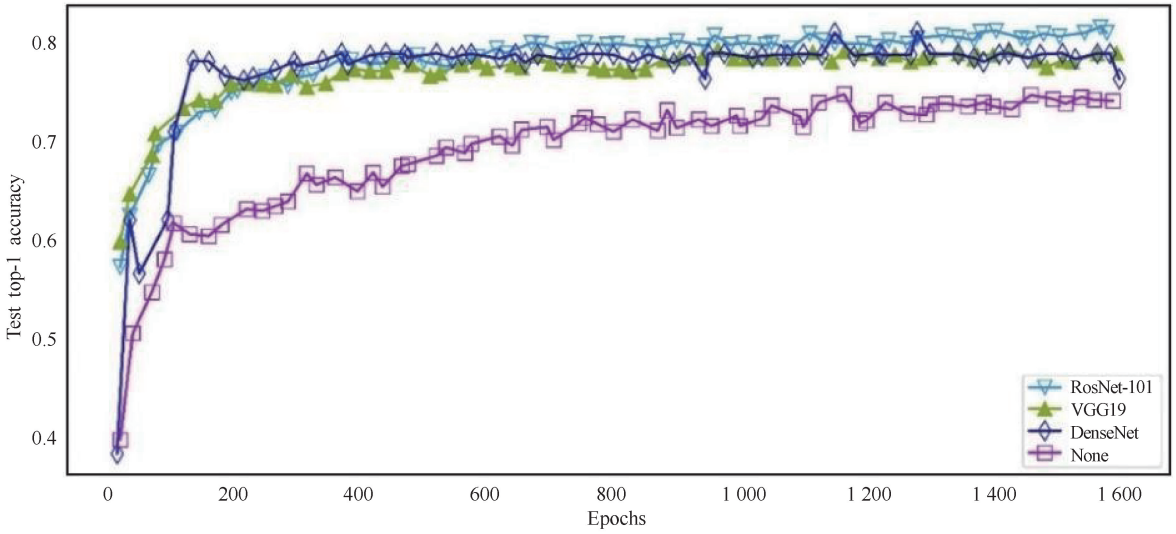

Fig. 6 Evaluating the performance of MKGViLT with and without ResNet-101

Table 6 Testing the impact of using MKGM structures on the SLAKE-EN[19] (%)

Methods	Overall	Open	Closed
MKGViLT without MKGM	81.2	77.3	87.2
MKGViLT	83.6	79.9	87.4

3.5.3 Effect of using different normalization methods on MKGViLT at the output layer

This study evaluates the influence of normalization techniques at the output layer on the multimodal visual question-answering system's overall performance. In models without normalization, the output consists of two fully connected layers. Conversely, models that incorporate normalization apply LayerNorm[42] and BatchNorm[39] following the initial fully connected layer. Rigorous preprocessing of images for MKGViLT involves segmenting them into smaller units and subsequently mapping them onto the Transformer via fully connected layers. Results delineated in Table 7 demonstrate a marked improvement in system performance with the implementation of BatchNorm at the output layer. Specifically, implementing BatchNorm results

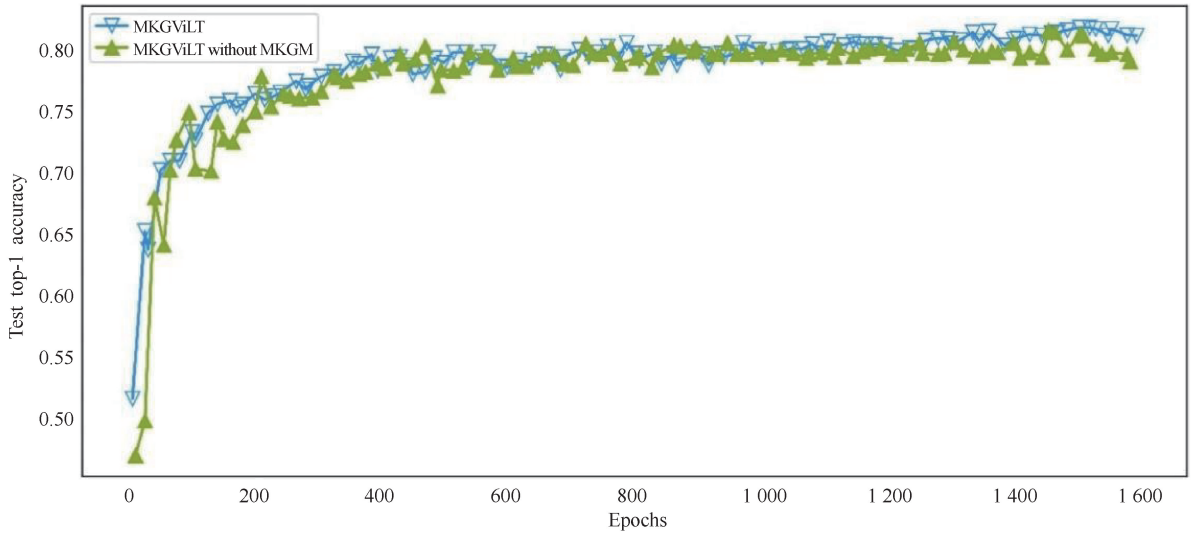

Fig. 7 Evaluating the performance of MKGViLT with and without MKGM

in an 18.5% increase in accuracy compared with LayerNorm, and a 25.8% increase compared with models without normalization. Fig. 8 visually represents these enhancements, highlighting the efficacy of BatchNorm in optimizing model performance.

Table 7 Evaluating the impact of various normalization methods on output layer performance in MKGViLT in the SLAKE-EN[19] (%)

Methods	BN	LN	SLAKE-EN		
			Overall	Open	Closed
MKGViLT			49.6	47.3	51.4
		√	56.9	56.4	57.6
	√		75.4	73.9	77.6

4 Conclusion

This study investigates the application of Transformer in MedVQA and introduces the MKGViLT model. The MKGViLT model incorporates external knowledge through the MKGM module and is specifically designed for MedVQA tasks. Compared with previous methods, this model employs an enhanced Transformer architecture for feature extraction and modal fusion. Integrating knowledge graphs in the MKGM model enhances both the convergence speed and performance stability for MedVQA tasks. Besides its applicability to the current task, the proposed MKGViLT architecture shows promising transferability to additional tasks. According to experimental evaluations, this model delivers top-tier performance that surpasses leading benchmarks without the need for external images.

However, several challenges remain. First, although the MKGViLT model speeds up convergence, achieving optimal performance still requires significant training time and resources. Second, the limited availability of pre-training data in the medical domain challenges MKGM models, which require ample data for pre-training. Lastly, the substantial parameter size of the MKGViLT model suggests that exploring methods to optimize its structure is a viable future research direction.

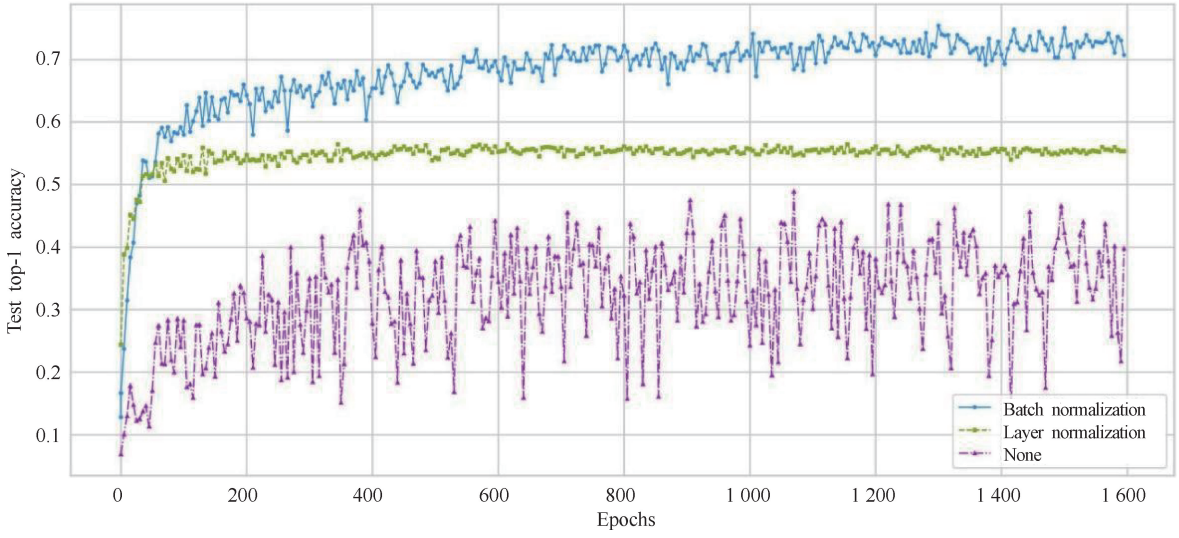

Fig. 8 Evaluating the effectiveness of different normalization techniques on the output layer performance in MKGViLT

References

[1] AMBATI R, DUDYALA C R. A sequence-to-sequence model approach for ImageCLEF 2018 medical domain visual question answering[C]//2018 15th IEEE India Council International Conference. Coimbatore, India: IEEE, 2018: 1-6.

[2] KOVALEVA O, SHIVADE C, KASHYAP S, et al. Towards visual dialog for radiology[C]// Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. Online: ACL, 2020: 60-69.

[3] STANISLAW A, AISHWARYA A, JIASEN L, et al. VQA: visual question answering[C]//2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 2425-2433.

[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. Computer Science, Linguistics, 2018, 2018: 1-12.

[6] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014: 1724-1734.

[7] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-train-

ing of deep bidirectional transformers for language understanding[J]. North American Chapter of the Association for Computational Linguistics, 2019, 19: 4171-4186.
[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]// International Conference on Learning Representations. Vienna, Austria: OpenReview, 2021: 1-21.
[9] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]// Proceedings of the 38th International Conference on Machine Learning. Online: PMLR, 2021: 10347-10357.
[10] YANG J, LI C, ZHANG P, et al. Focal self-attention for local-global interactions in vision transformers[C]//Advances in Neural Information Processing Systems. Sydney, Australia: Curran Associates, 2021: 1-15.
[11] LI Y, ZHANG K, CAO J, et al. LocalViT: bringing locality to vision transformers[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. Detroit, USA: IEEE, 2023: 9598-9605.
[12] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 10012-10022.
[13] WANG P, WU Q, SHEN C, et al. FVQA: fact-based visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(10): 2413-2427.
[14] XIONG C Y, RUSSELL P, JAMIE C. Explicit semantic ranking for academic search via knowledge graph embedding[C] //Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: ACM, 2017: 1271-1279.
[15] SCHNEIDER P, SCHOPF T, VLADIKA J, et al. A decade of knowledge graphs in natural language processing: a survey[C]//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Online: ACL, 2022: 601-614.
[16] GUO Q, ZHUANG F, QIN C, et al. A survey on knowledge graph-based recommender systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(8): 3549-3568.
[17] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.
[18] XIE X, LI Z, WANG X, et al. PromptKG: a prompt learning framework for knowledge graph representation learning and application [EB/OL]. (2023-09-14) [2024-03-15]. https://arxiv.org/pdf/2210.00305.
[19] LIU B, ZHAN L M, XU L, et al. SLAKE: a semantically-labeled knowledge-enhanced dataset for medical visual question answering[C]// 2021 IEEE 18th International Symposium on Biomedical Imaging. Nice, France: Springer, 2021: 1650-1654.
[20] HUANG J, CHEN Y, LI Y, et al. Medical knowledge-based network for patient-oriented visual question answering[J]. Information Processing and Management, 2023, 60(2): 1-17.
[21] PENG, Y, LIU, F, ROSEN M P, et al. UMass at ImageCLEF medical visual question answering (MedVQA) 2018 task [EB/OL]. [2024-03-15]. https://ceur-ws.org/Vol-2125/paper_163.pdf.
[22] LAU J J, GAYEN S, BEN A, et al. A dataset of clinically generated visual questions and answers about radiology images[J]. Scientific Data, 2018, 5: 1-10.
[23] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA: ACL, 2016: 457-468.
[24] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 21-29.
[25] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2018: 1571-1581.
[26] GUPTA D, SUMAN S, EKBAL A. Hierarchical deep multi-modal network for medical visual question answering [J]. Expert Systems with Applications, 2021, 164: 1-18.
[27] VU M H, LOFSTEDT T, NYHOLM T, et al. A question-centric model for visual question answering in medical imaging[J]. IEEE Transactions on Medical Imaging, 2020, 39(9): 2856-2868.
[28] WANG J, HUANG S, DU H, et al. MHKD-MVQA: multimodal hierarchical knowledge distillation for medical visual question answering[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine. Las Vegas, USA: IEEE, 2022: 567-574.
[29] HUANG X, GONG H. A dual-attention learning network with word and sentence embedding for medical visual question answering[J]. IEEE Transactions on Medical Imaging, 2023, 43(2): 832-845.
[30] BORDES A, USUNIER N, ALBERTO G D, et al. Translating embeddings for modeling multi-relational data [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Taihao Lake, USA: MIT Press, 2013: 2787-2795.
[31] WANG Z, ZHANG J W, FENG J L, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec, Canada: AAAI Press, 2014: 1112-1119.
[32] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA: AAAI Press, 2015: 2181-2187.
[33] BORDES A, GLOROT X, WESTON J, et al. A semantic

matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(1): 233-259.
[34] NICKEL M, ROSASCO L, POGGIO T. Holographic embeddings of knowledge graphs[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix ,USA: AAAI Press, 2016: 1955-1961.
[35] ZHOU X, YI Y, JIA G. Path-RotatE: knowledge graph embedding by relational rotation of path in complex space [C]//2021 IEEE/CIC International Conference on Communications in China. Xiamen, China: IEEE, 2021: 905-910.
[36] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.
[37] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//The 3rd International Conference on Learning Representations. San Diego, USA: OpenReview, 2015: 1-14.
[38] HUANG G, LIU Z, MAATEN V D, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 4700-4708.
[39] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR, 2015: 448-456.
[40] TURE F, JOJIC O. No need to pay attention: simple recurrent neural networks work! [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: ACL, 2017: 2866-2872.
[41] DEY R, SALEM F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th International Midwest Symposium on Circuits and Systems. Boston, USA: IEEE, 2017: 1597-1600.
[42] BA J L, KIROS J R, HINTON G E. Layer normalization [EB/OL]. (2016-07-21) [2024-03-15]. https://arxiv.org/pdf/1607.06450.
[43] YIN W P, YU M, XIANG B, et al. Simple question answering by attentive convolutional neural network[C]// The 26th International Conference on Computational Linguistics. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 1746-1756.
[44] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks. [EB/OL]. (2015-06-05) [2024-03-15]. https://arxiv.org/pdf/1506.02075.
[45] UZUNER Ö, SOUTH B R, SHEN S Y, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.
[46] ZHANG H Y, CISSE M, DAUPHIN Y N. mixup: beyond empirical risk minimization[C]//The 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview, 2017: 1-13.
[47] LIU B, ZHAN L M, WU X M. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images[C]// Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference. Strasbourg, France: Springer, 2021: 210-220.
[48] YU Z, YU J, FAN J. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1839-1848.
[49] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2017: 6000-6010.

CUI Wencheng, born in 1973. He has been engaged in teaching and research work at Shenyang University of Technology since July 1997 and is a member of the China Computer Society. His current research interests include intelligent information processing, medical artificial intelligence, etc.