Complex & Intelligent Systems (2024) 10:577–594
https://doi.org/10.1007/s40747-023-01170-9

ORIGINAL ARTICLE

Cross-modal knowledge guided model for abstractive summarization

Hong Wang[1] · Jin Liu[1] · Mingyang Duan[1] · Peizhu Gong[1] · Zhongdai Wu[2,3] · Junxiang Wang[2] · Bing Han[3]

Received: 6 September 2022 / Accepted: 29 June 2023 / Published online: 27 July 2023

Abstract

Abstractive summarization (AS) aims to generate more flexible and informative descriptions than extractive summarization. Nevertheless, it often distorts or fabricates facts in the original article. To address this problem, some existing approaches attempt to evaluate or verify factual consistency, or design models to reduce factual errors. However, most of the efforts either have limited effects or result in lower rouge scores while reducing factual errors. In other words, it is challenging to promote factual consistency while maintaining the informativeness of generated summaries. Inspired by the knowledge graph embedding technique, in this paper, we propose a novel cross-modal knowledge guided model (CKGM) for AS, which embeds a multimodal knowledge graph (MKG) combining image entity-relationship information and textual factual information (FI) into BERT to accomplish cross-modal information interaction and knowledge expansion. The pre-training method obtains rich contextual semantic information, while the knowledge graph supplements the textual information. In addition, an entity memory embedding algorithm is further proposed to improve information fusion efficiency and model training speed. We elaborately conducted ablation experiments and evaluated our model on the Visual Genome, FewRel, MSCOCO, and CNN/DailyMail datasets. Experimental results demonstrate that our model can significantly improve the FI consistency and informativeness of generated summaries.

Keywords تلخيص النص . نموذج لغوي مدرب مسبقًا . دمج المعلومات متعددة الوسائط . رسم المعرفة البياني . الشبكة العصبية العميقة

Introduction

With the further development of various deep learning technologies, how to efficiently extract the primary information from the original massive data in the 5G era, where text resources are growing exponentially and communication technologies are developing by leaps and bounds [1–3], has become a pressing problem. Automatic summarization is a task to make a concise and fluent summary of the source text, which has two main types of generation technologies: extractive and abstractive. The former directly copies some words from the original text [4], while the latter can flexibly generate new words and phrases that are not found in the original one [5].

Deep neural network-based sequence-to-sequence (Seq2Seq) methods are gaining tractions in various applications due to their feature representing capacity [6, 7]. Most of the existing studies on abstractive summarization (AS) also use Seq2Seq architecture [8] and have achieved promising results. However, along with the increasing number of AS generation tasks, researchers found that nearly 30% of the AS distort or fabricate factual information (FI) in the articles,

✉ Jin Liu
jinliu@shmtu.edu.cn

Hong Wang
202030310109@stu.shmtu.edu.cn

Mingyang Duan
duanmyer@163.com

Peizhu Gong
gongpeizhu012@163.com

Zhongdai Wu
wu.zhongdai@coscoshipping.com

Junxiang Wang
wang.junxiang@coscoshipping.com

Bing Han
han.bing@coscoshipping.com

1 College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

2 COSCO Shipping Technology Co., Ltd, Shanghai 200127, China

3 Shanghai Ship and Shipping Research Institute, Shanghai 200135, China

making even the generated summaries unusable. As shown in the example in Fig. 1, the FI refers to the basic facts described or implied in the article, similar to the theorems in mathematics, which cannot be denied and tampered with. Otherwise, there will be obvious ambiguities, i.e., factual errors.

In fact, some researchers have taken it into account in their work and have pushed the frontiers of ensuring consistency of FI in AS. Cao et al. [9] leveraged open information extraction and dependency parse techniques to extract factual descriptions and then guide AS. Falke et al. [10] proposed to rearrange candidate summary sentences through textual entailment models. Kryscinski et al. [11] proposed a weakly supervised approach to verify the factual consistency of generated summaries. Wang et al. [12] designed a framework for automatically detecting factual errors in supervised generated texts and use this it to develop QAGS, a metric used to measure factual consistency in AS. These models focus on proposing metrics to evaluate factual consistency, but they do not explore how to reduce factual errors. Later, some work looked at model designs to further address this problem in AS. For example, Zhang et al. [13] proposed to optimize the summary model with factual correctness reward by reinforcement learning. Dong et al. [14] used the knowledge learned from the question–answer model to correct the system-generated summaries through the span selection strategy. Zhu et al. [15] proposed a fact-aware summary model FASUM and a fact-corrector model FC, which frames the correction process as a seq2seq problem to make the corrected summary more realistically consistent with the article. These models are indeed effective in improving the factual consistency of AS, but often at the cost of greatly reducing the rouge scores of the generated summaries.

On the other hand, in scenarios such as news, where text and images often appear in parallel, but most of the current AS models are based on a single text modality, ignoring the interaction and supplementation of multimodal knowledge. Knowledge graph (KG) is a network structured knowledge base containing rich semantic information and knowledge; however, general KG is often prone to problems, such as insufficient entity and relationship information, and researchers often use techniques, such as knowledge graph embedding and knowledge graph complementation to solve it. These two types of approaches enhance the informativeness and optimize the structure of the KG by reasoning new information on the basis of existing information and constructing entity and relationship nodes into a continuous vector space, respectively, so that various models and downstream tasks can better utilize the KG. In our opinion, the knowledge embedding approach using graph networks is an effective strategy to represent rich FI in different modalities.

In this paper, we design a novel cross-modal knowledge guided model (CKGM) that incorporates external KG in the upstream encoding part of the AS task to accomplish cross-modal information interaction and knowledge augmentation. Specifically, we first introduce a multi-granularity entity-relation representation graph to extract structured factual graph elements, followed using a multi-scale semantic and spatial feature extraction method to obtain a scene graph containing image entity information, then design a multimodal knowledge graph (MKG) incorporating the image entity relationship information and textual factual information, and finally embedded it into BERT as external knowledge during the training of the model. The pre-training approach obtains rich contextual semantic information, while the knowledge graph complements the lack of knowledge of textual information. In addition, an entity memory embedding algorithm is further proposed to improve information fusion efficiency and model training speed.

Our contribution can be summarized as follows:

(1) Inspired by the knowledge graph embedding strategy, a novel pre-trained language model CKGM incorporating multimodal knowledge is constructed for knowledge supplementation and factual information enhancement in AS task.
(2) An entity memory embedding algorithm is designed to improve information fusion efficiency and model training speed, which achieves good performance.
(3) We analyze the contribution of each component of model and elaborately conduct extensive experiments on multiple datasets. Experimental results show that our model can significantly improve the FI consistency and informativeness of generated summaries.

The remainder of this article is structured as follows. In the section "Related works", related work is presented. In the section "Method", we introduce the architecture and technical details of the proposed model. In the section "Experiment", the experiment process is introduced, and the experimental results are analyzed. Finally, some conclusions are drawn in the section "Conclusion".

Related works

Abstractive summarization

Sequence-to-sequence (Seq2Seq) [8] is a widely used model that has greatly facilitated the development of AS [16]. In 2015, Rush et al. [17] first proposed an attention-based mechanism and a neural network language model (NNLM) for AS, which was the first time to apply the Seq2Seq model structure to the AS. Later work improved the above structure with better encoders, such as LSTM and GRU, to further capture long-range dependencies [18]. After that, to alleviate problems, such as unregistered words, generation of duplicates,

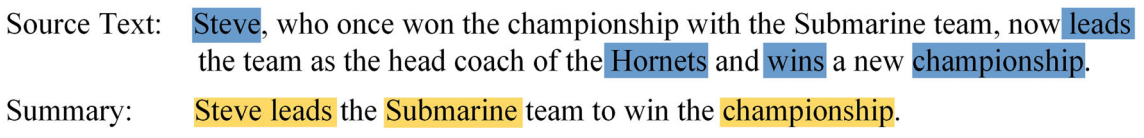

Fig. 1 Factual errors in the summarization

etc., which are prone to occur when using only Seq2Seq to generate summaries, See et al. [19] proposed adding Copy and Coverage mechanisms to Seq2Seq and Paulus et al. [20] further incorporated a novel intra-attention and reinforcement learning. Then, Li et al. [21] proposed sentence-level attention and sentence-level Coverage mechanism.

More recently, researchers have found that nearly 30% of the abstractive summaries contain factual errors. In fact, some studies have attempted to address this problem. On the one hand, some models have focused on proposing metrics to assess factual consistency [9–12], but they have not further explored how to reduce factual errors. Subsequently, some work has examined model design to further address this issue in AS. Zhang et al. [13] proposed an approach to optimize the summary model with factual correctness reward by reinforcement learning. Dong et al. [14] used the knowledge learned from the question–answer model to correct the system-generated summaries through the span selection strategy. Zhu et al. [15] proposes a fact-aware summary model FASUM and a fact-corrector model FC, which frames the correction process as a seq2seq problem, making the revised summary more truly consistent with the article. These above models are indeed effective in improving the factual consistency of AS, but often at the cost of greatly reducing the carmine score of the generated summaries. In this paper, we introduce a multi-granularity entity-relation representation graph to extract the structured fact tuples and select the most important FI by defining a scoring function. Moreover, we believe that cross-modal knowledge graph embedding based on factual consistency is an effective strategy that can not only represent rich FI in different modes, but also ensure the information richness contained in the generated summary through cross-modal information interaction and supplementation.

Multimodal knowledge fusion

Cross-modal fusion aims to fuse the information contained in different modalities to realize knowledge interaction and complementary. Therefore, it can solve the problem of insufficient and inaccurate summaries generated by relying only on a single text modality. In recent years, several multimodal fusion techniques, including weighted summation, direct splicing, graph attention mechanisms, and bilinear pooling, have been used to guide various generation tasks [22–24]. In this paper, we do not use the above methods, but through the construction of KG to fusion the features in the image and text, to achieve multimodal knowledge fusion. KG is a semantic network knowledge base, which usually represents knowledge in terms of triples and is stored as directed graphs. In recent years, two techniques, knowledge graph embedding [25] and knowledge graph supplementation, are commonly used to enhance the information content of the knowledge graph and optimize the structure of the knowledge graph by inferring and supplementing new information on top of the existing information and by constructing entity nodes and relationship nodes as a continuous vector space, respectively, so that various models and downstream tasks can make better and fuller use of the KG. Inspired by knowledge graph embedding, this paper extracts multimodal entities and relations (ERs) by fusing text and image information, and embed the constructed MKG into the pre-trained language model BERT to perform the AS task. In addition, an entity memory embedding algorithm is proposed in our model to improve the information fusion efficiency and model training speed.

Method

The architecture of CKGM is shown in Fig. 2. We first construct a MKG fusing image entity-relationship information and textual FI, specifically, (a) precoding the input textual information using the benchmark BERT, (b) we simultaneously extract multi-scale features from images and generating entity and entity-relationship graphs, and then use graph convolutional networks and graph attention mechanisms to infer and generate entity precoding, (c) we also propose an entity memory embedding algorithm to further improve the information fusion efficiency and model training speed, and (d) further, the text precoding is fused with the image entity and entity-relationship precoding. Finally, we complete the model training around multiple subtasks based on BERT. The abbreviations in the section "Method" are described in Table 1. Next, we will introduce the details of each part in the following subsections.

Textual FI representation

Multi-granularity entity-relation representation

In the coding layer, we use the benchmark BERT to encode the basic semantics and feature representation of the text.

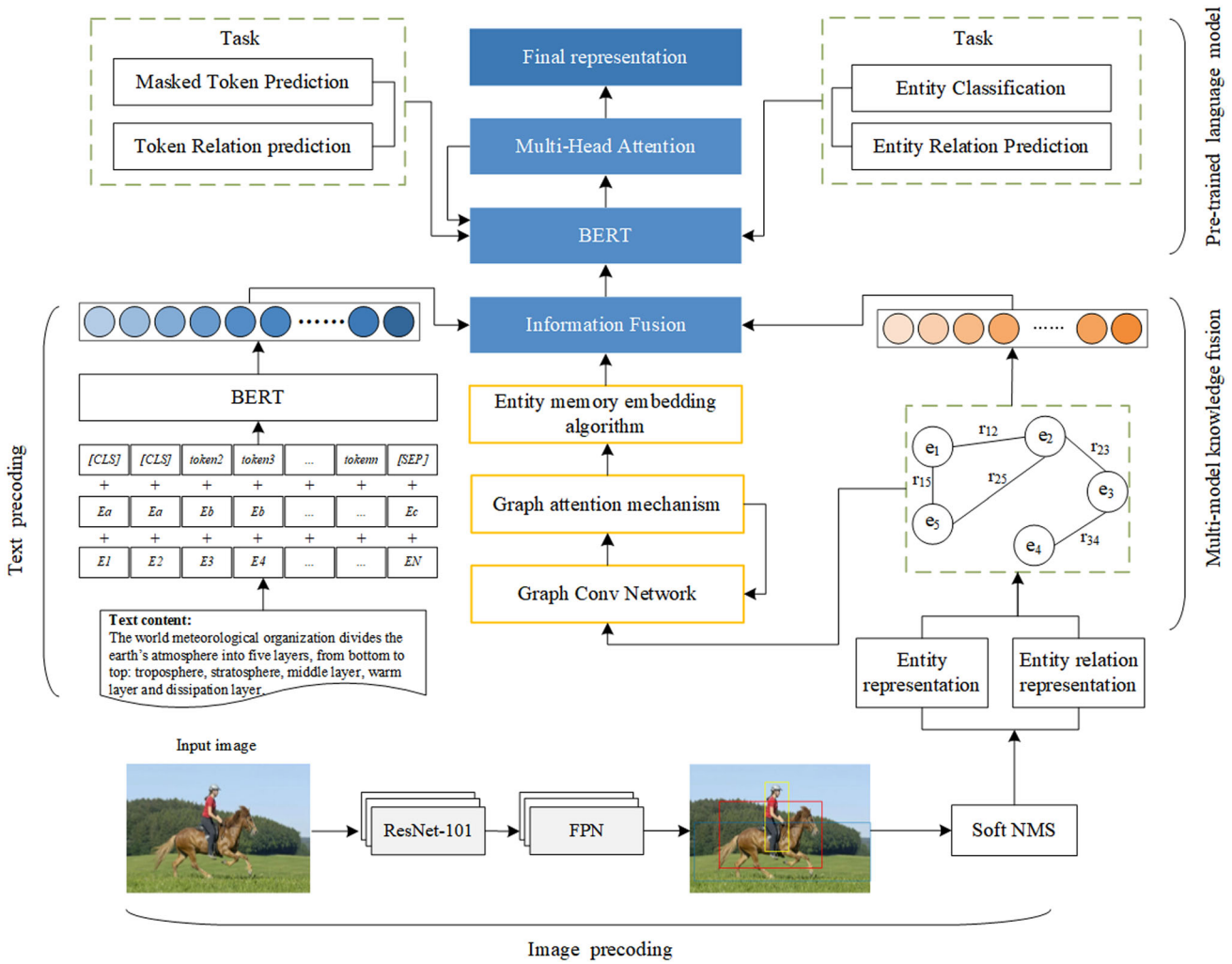

Fig. 2 CKGM architecture

Table 1 Description of abbreviations in Method

Abbreviation	Description
C_i	Common Eno
B_i	Basic Eno
S_i	Sentence Eno
$l_{B;C}$	The edge between B_i and C_i
$l_{B;B}$	The edge between two B_i
$l_{C;S}$	The edge between C_i and S_i
$l_{S;S}$	The edge between two S_i
FI	Factual information
Eno	Entity node
ER	Entity relation
ERs	Entities and relations
ERG	Entity-relation graph
KG	Knowledge graph
MKG	Multimodal knowledge graph

As shown in Fig. 3, since the overall idea of this part is to uncover relevant factual information in the text at different granularities, we embed chapter-level, sentence-level, and word-level sequences into the encoding layer as the input of feature embedding, which can be formalized as follows:

$$Chap\_Level = \{[CLS], \mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n, [SEP]\} \quad (1)$$

$$Sent\_Level = \{[CLS], \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n, [SEP]\} \quad (2)$$

$$Word\_Level = \{[CLS], \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n, [SEP]\} \quad (3)$$

$$\mathcal{L} = \{Chap\_Level, Sent\_Level, Word\_Level\}, \quad (4)$$

where $[CLS]$ and $[SEP]$ denote the start and end markers of the token sequence, respectively, $\mathcal{C}_n$ denotes the word encoding of chapter-level sequences, $\mathcal{S}_n$ denotes the word encoding

of sentence-level sequences, $\mathcal{W}_n$ denotes the word-level encoding, n denotes the sequence length of each sequence, and the input $\mathcal{L}$ of the final BERT model is the stitching of sequences at different levels. Then, through BERT coding, the hidden states of words in different levels h^N_{Cj}, h^N_{Sj}, h^N_{Wj} $\in 1 \times \mathcal{R}^{hw}$ can be obtained, where j represents the number of word positions in the current sequence of levels. We pool the output encoding vectors of the hidden layers, and then concatenate the pooled hidden states of different levels.

We further construct a graph structure to represent ERs built upon the above text encoding. As shown in Fig. 3, first, we use different types of nodes and edges to represent different ERs in a global feature dimension, and then fuse hierarchical labels for the ERs in the graph to further enhance their representation. Specifically, three entity node (Eno) types are defined: common Eno, basic Eno, and Sentence Eno. Among them, common Eno is defined as C_i $i = 1, 2, .., n$, which is directly extracted from the encoded semantic vector and is the most obvious entity in the original text. Basic Eno is defined as B_i $i = 1, 2, .., n$, which can be derived as a node of common Enos, calculated by averaging the vectors of all common Enos related to it. Sentence Eno, as the name implies, denotes the node representing a sentence or a paragraph, defined as S_i $i = 1, 2, .., n$. In addition, four different types of edges are defined to represent the relationship between entities: $l_{B;C}$, $l_{B;B}$,$l_{C;S}$, $l_{S;S}$. Among them, $l_{B;C}$ connects a common entity and a basic entity that can be derived from this common entity; $l_{B;B}$ connects two basic entities that appear in the same sentence; $l_{C;S}$ connects the common entity and the sentence it belongs to; $l_{S;S}$ connects all sentences into paragraphs or chapter.

Based on the above constructed graph structure, the lth layer Graph Convolutional Network (GCN) [26] is used for convolution and feature learning, and the result of (l+1)-th layer is formalized as follows:

$$n_i^{l+1} = \varphi\left(\sum_{y\in\mathcal{Y}}\sum_{j\in\mathcal{N}_i^y}\frac{1}{|\mathcal{N}_i^y|}\mathcal{M}_y^l n_j^l + \mathcal{M}_0^l n_i^l\right), \tag{5}$$

where φ represents the activation function, $\mathcal{N}_i^y$ represents the set of adjacent nodes connected with n_i^l, and $\mathcal{Y}$ represents the set of relationships of the edges contained in adjacent nodes. $\mathcal{M}$ is the weight matrix between different layers with dimension $\mathbb{R}^{d_n\times d_n}$.

After the above iteration, the final representation obtained is the global representation of each common Eno and entity relation (ER), converted to higher level features. Next, we use the graph attention network (GAT) to update the edges of the relationship between entities, and gradually learn the FI contained in the graph. Then, we use the attention mechanism to make the updating of the edges between entities quickly converge, and finally obtain an entity and entity-relationship graph containing FI. Specifically, taking all Enos in the graph as input to GAT, based on the set of adjacent entities $\mathcal{N}_i^y$ mentioned above, we can learn the hidden value of each Eno through attention mechanism.

FI extraction and importance evaluation

Through the work above, we have obtained the initial representation $\mathcal{H}$, the updated representation $\mathcal{H}_i$, and the global representation of the entity $\mathcal{N}_i$, as shown in Fig. 4. The multi-head attention weights are calculated as follows:

$$\begin{aligned} &MultiHead\left(\mathcal{H},\ \mathcal{H}_i,\ \mathcal{N}_i\right)\\ &\quad = Concat\left(head_1,\ head_2,\ head_3\right)\mathcal{W} \end{aligned} \tag{6}$$

$$head_i = softmax\left(\frac{\mathcal{N}_i\mathcal{W}_i^{\mathcal{N}_i}\cdot\mathcal{H}\mathcal{W}_i^{\mathcal{H}}}{\sqrt{d_{\mathcal{H}_i}}}\right)\mathcal{H}_i\mathcal{W}_i^{\mathcal{H}_i}, \tag{7}$$

where $\mathcal{W}_i^{\mathcal{H}}$, $\mathcal{W}_i^{\mathcal{H}_i}$, $\mathcal{W}_i^{\mathcal{N}_i}$ are weight matrices, then the multi-head attention weights of a pair of entities connected by an edge are calculated, and the entity pair or the adjacent entity set with higher weight values has higher relevance.

Next, we defined a quadruple $F_i = [\mathcal{H}_i, \mathcal{H}_j, l_{ij}, \mathcal{K}]$, where i represents the sequence number of FI, $\mathcal{H}_i$ and $\mathcal{H}_j$ represent the entity pairs with high correlation, and l_{ij} represents the relationship between entity pairs. $\mathcal{K}$ represents the entity with the highest weight in the adjacent entity set $\mathcal{N}_i^y$. We directly concatenate these four elements into an FI encoding, formalized as $e_i = [\mathcal{H}_i \oplus \mathcal{H}_j \oplus l_{ij} \oplus \mathcal{K}]$, where $\oplus$ represents the concatenation operation. Determining whether the FI is important can be defined as a binary problem

$$\mathcal{Y}_i = \varphi\left(We_i + b\right)\ \left(\mathcal{Y}_i \in \{0, 1\}\right), \tag{8}$$

where W represents the parameter that can be iteratively updated, b is a bias term, and φ represents activation function. For an FI e_i, a key value and a query value are set, respectively

$$k_i = W_k e_i \tag{9}$$

$$q_i = W_q e_i, \tag{10}$$

where W_k and W_q are parameter matrices. Then, it can be classified based on the attention weight between facts, and the equation is as follows:

$$\mathcal{Y}_i = \varphi\left(\sum_{j=1,j\neq i}^{n} q_i^T k_i\right); \tag{11}$$

the higher the score, the more important the FI is. In this way, we obtain the most important FI.

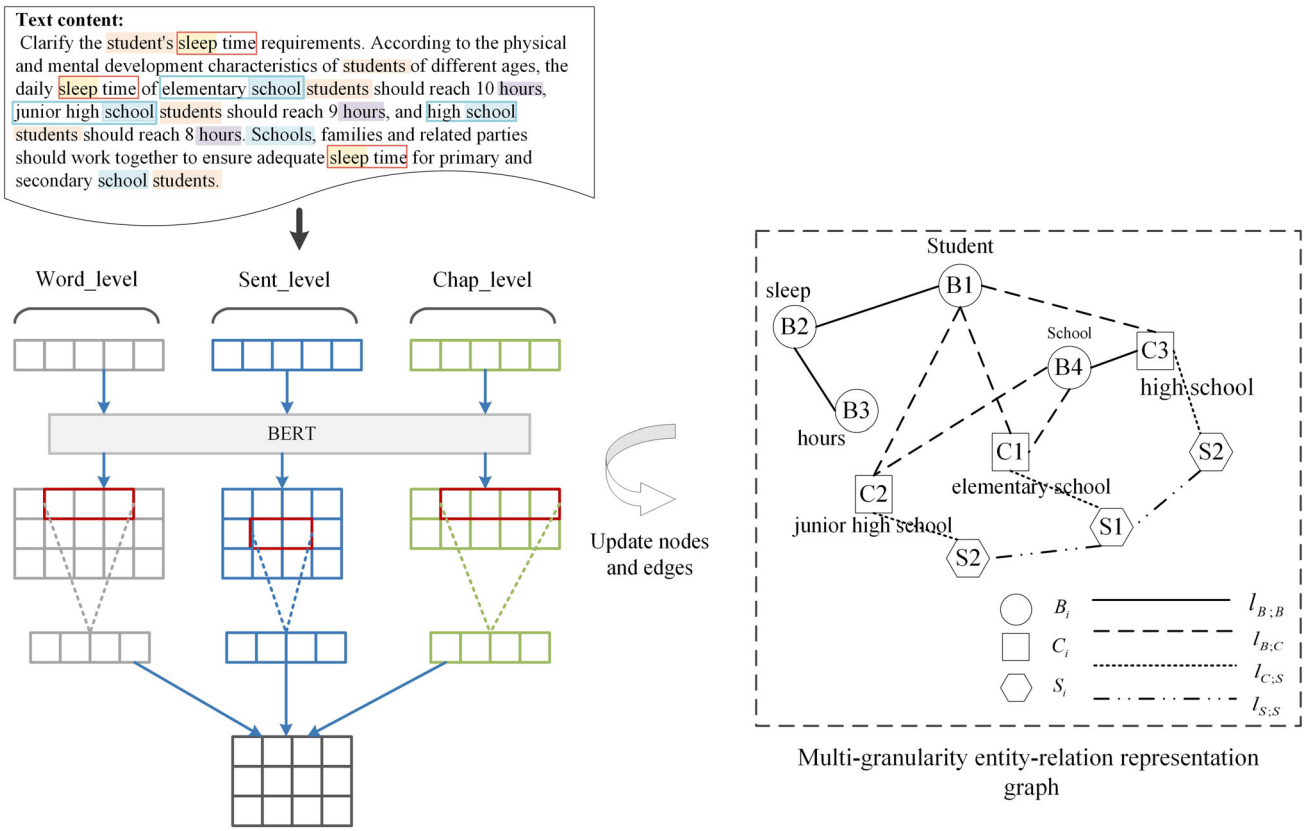

Fig. 3 Multi-granularity entity-relation representation graph

Image knowledge representation

Since the sizes of the images are different, the images need to be normalized first to avoid the possible loss of image information in the convolution operation. After pre-processing, multi-scale semantic and spatial feature are extracted, as shown in Fig. 5.

To fully extract the ERs of the image, it is necessary to recognize object category and object position simultaneously. First, on the basis of the ResNet-101 network [27], we added additional hidden layers to make the number of layers the same as the Feature Pyramid Network (FPN) [28]. In particular, the fifth and sixth hidden layers are composed of three identical convolution modules, each including two layers of 256 1×1 convolution kernels and one layer of 256 3×3 convolution kernels. The added 1×1 convolution kernels help to dynamically maintain the dimension of the weights to affirm the consistency of the features.

FPN is used to extract image features further. We set the n^{th} convolution module to C_n; each convolution module contains convolution, pooling, and activation operations. Taking m_n as the feature map generated at the n^{th} layer, the output set M_{global} can be calculated as follows:

$$M_{global} = \{m'_{n-k}, \ldots, m\} \tag{12}$$

$$m'_n = m_n \tag{13}$$

$$m'_{n-1} = m_n + m_{n-1}\ (n > k > 0) \tag{14}$$

$$m'_{n-k} = m_n + m_{n-1} + \cdots + m_{n-k}. \tag{15}$$

Next, we extracted semantic feature ERs based on the Soft NMS [29] model, and the obtained feature map is utilized to identify the candidate regions containing important entities in the image. Whether the entity is important or not can be regarded as a classification problem cls, and the circle of the candidate regions can be regarded as a coordinate regression problem reg, the loss function is as follows:

$$L\left(\{p_i\}, \{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p'_i\right) + \frac{1}{N_{reg}} \times \sum_i p'_i L_{reg}\left(t_i, t'_i\right), \tag{16}$$

where p_i represents the predicted probability of the object entity, t_i is a vector representing the coordinates of the predicted bounding box, and p'_i, t'_i are the ground-truth value of p_i, t_i, respectively. N_{cls} and N_{reg} are the normalization

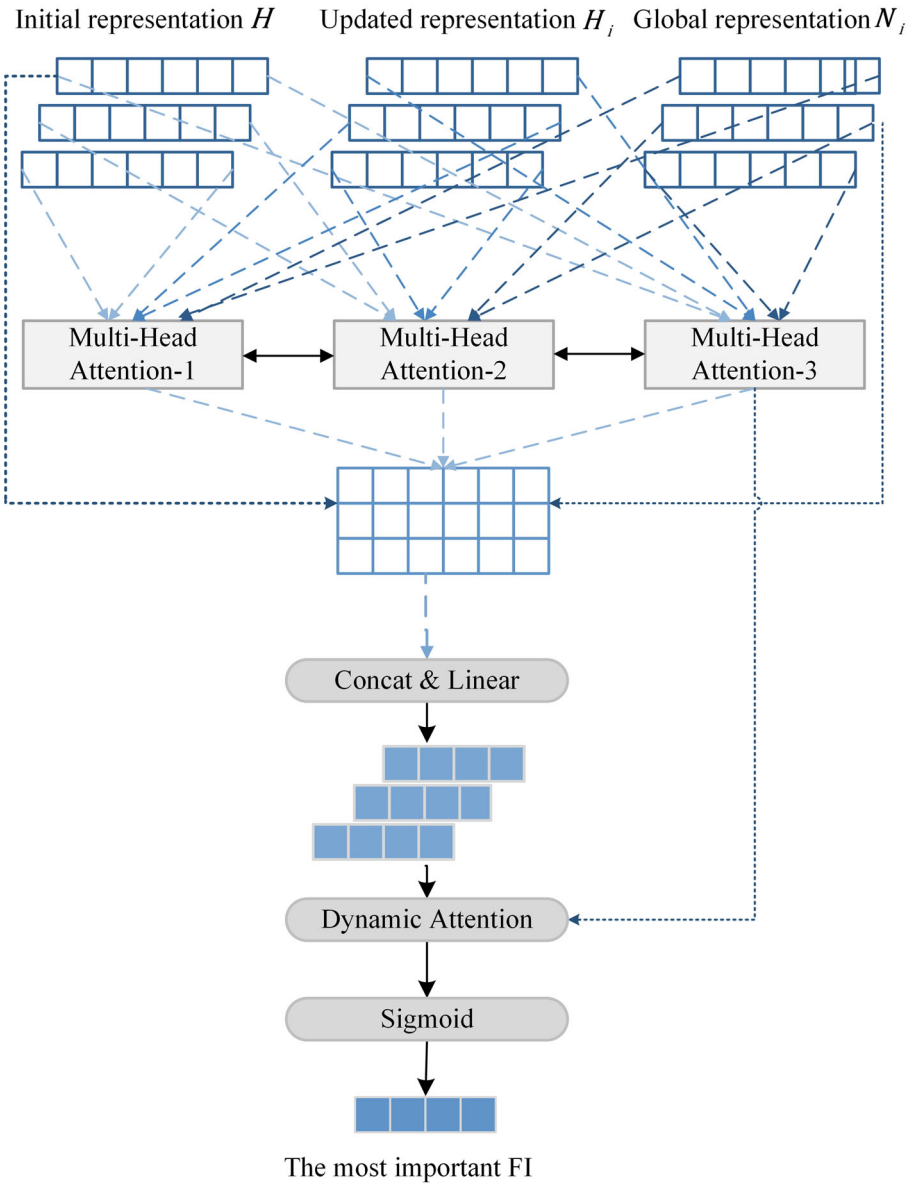

Fig. 4 Multi-granularity entity representation fusion based on multi-head attention

parameters, and $L_{cls} = -p_i' log(p_i)$ is the cross-entropy loss. After completing the extraction of semantic features, we also extract the spatial feature between image entities, and the spatial location of two entities can be formalized as follows:

$$D_m = \begin{cases} 1, & Floor\left(\frac{\theta_{ij}}{45^\circ}\right) + 1 = m \\ 0, & Floor\left(\frac{\theta_{ij}}{45^\circ}\right) + 1 \neq m, \end{cases} \tag{17}$$

where D consists of a set of 8 directions in space, D_m indicates whether entity j is in the m^{th} direction region of reference entity i, $Floor()$ is the downward integral function, and θ_{ij} represents the angle between the central line of the two entities and the horizontal line.

Based on the above semantic and spatial feature information, we further utilize GRU [30] to construct the ERG of the image. In ERG, nodes represent entities and edges represent relations between entities. To generate weights using the relations between entity-relation pairs, the entities are clustered, so that related entities are in adjacent positions and a relation group, as shown in the following formulas:

$$N_{cut} = \sum_{i=1}^{k} \frac{\text{cut}\,(X_i, X_{i+1})}{\text{ass}\,(X_i, T)} \tag{18}$$

$$\text{cut}\,(X_i, X_{i+1}) = \sum_{p \in X_i, q \in X_{i+1}} w(p, q) \tag{19}$$

$$\text{ass}\,(X_i, T) = \sum_{p \in X_i, t \in T} w(p, t), \tag{20}$$

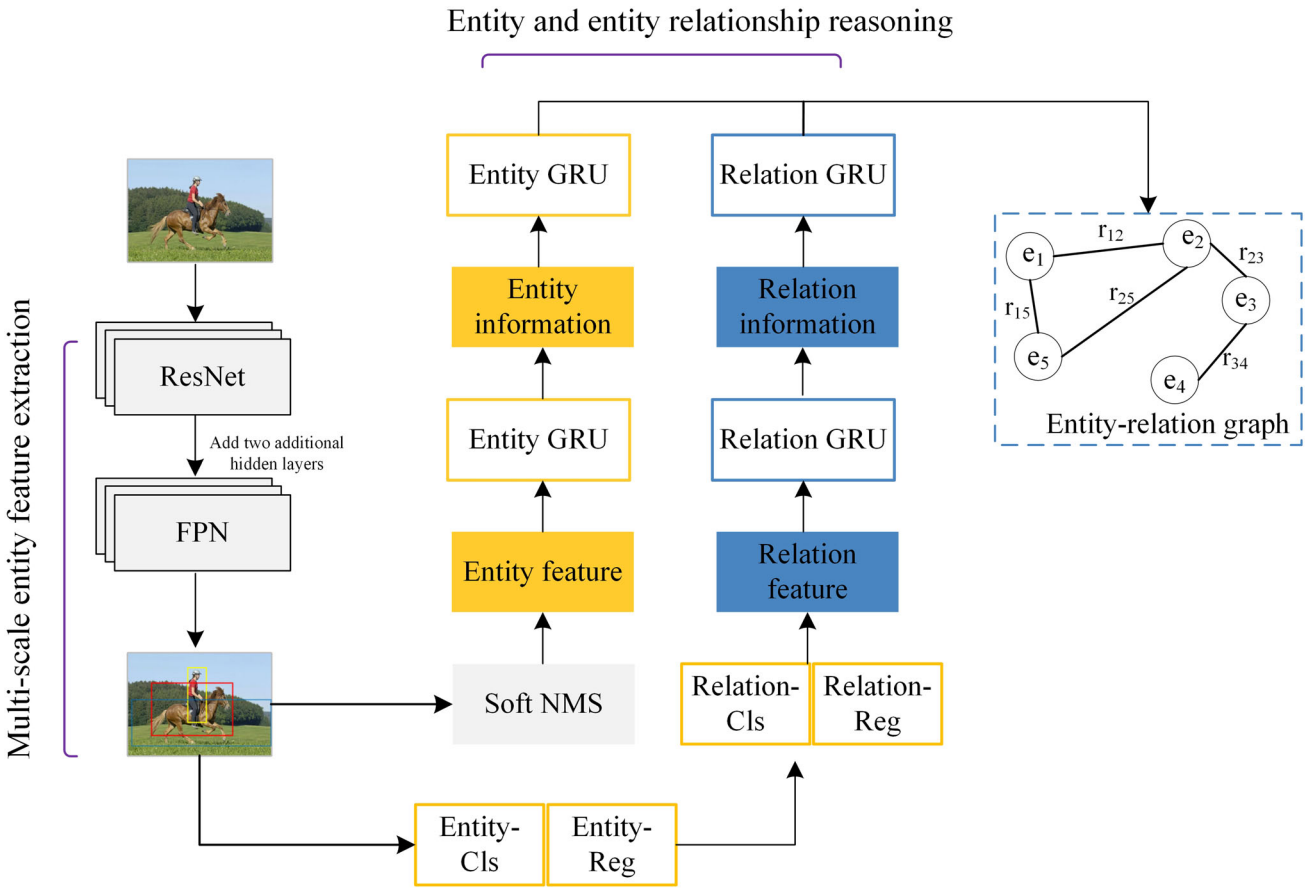

Fig. 5 Extraction and reasoning of entity relation of the image

where X_i represents the set of all entities at the ith layer. T denotes the set of all entities, $cut(X_i, X_{i+1})$ is the sum of the weights of the entities between X_i and X_{i+1}, and $ass(X_i, T)$ is the sum of the weights of the entity in X_i to all the entities connected to it. p, q, t represent the entities in the set. The function $w()$ calculates the weight between entities.

Knowledge fusion module

We have completed the extraction of ERs in the image and obtained FI in the text. To achieve cross-modal knowledge fusion, we need to fuse image features with text features to construct an MKG that will further guide the training of the pre-trained model and the task of AS. Specifically, we fuse the obtained entity-relation graph with the multi-granularity entity-relation representation graph containing textual FI, and finally, we get the knowledge graph that integrates the image entity-relation information and textual fact information.

A knowledge representation of the KG is denoted as a triplet $K_i = (E_i, E_j, R_{ij})$, where E_i and E_j represent entities. R_{ij} denotes the relationship between the two entities. In this paper, we hope to re-determine ERs through a method to realize the construction of MKG. To this end, we use a neural network to learn the representation of ERs, with the first layer projecting a pair of input entities into a low-dimensional vector and the second layer combining the two vectors into a scalar that is compared by a scoring function with specific relation parameters to achieve the final knowledge graph construction. In particular, first, we construct a high-dimensional feature sequence of entities from text and images by One-Hot encoding, and learn the entity representation by a linear function, as shown in the following formula:

$$y_{E_i} = f\left(\mathcal{W} x_{E_i}\right), \tag{21}$$

where x_{E_i} represents the entities after One-Hot encoding, which represents each entity as the average of its word vectors. $\mathcal{W}$ is the weight matrix of the neural network layer. $\mathcal{Y}_{E_i}$ represents the entity representation learned by the network layer.

Next, we define $\mathcal{Y}_{E_i} \in \mathbb{R}^n$ as the entity vector, and $\mathcal{W}_{E_i} \in \mathbb{R}^n$ as the parameter matrix, and then, a scoring function of the ER can be defined as follows:

$$G = \mathcal{W}_{E_i} - \mathcal{W}_{E_j} = -\left(2g_r^{E_i}\left(y_{E_i}, y_{E_j}\right) - 2g_r^{E_j}\left(y_{E_i}, y_{E_j}\right) + \mathcal{T}_{E_i}\right) \quad (22)$$

$$g_r^{E_i}\left(y_{E_i}, y_{E_j}\right) = y_{E_i}^T \mathcal{M}_r y_{E_j}, \quad (23)$$

where $g_r^{E_i}$ represents a bilinear function and the corresponding tensor parameter matrix is $\mathcal{T}_{E_i} \in \mathbb{R}^n$. $\mathcal{M}_r$ represents the parameter matrix.

Then, we can update the knowledge by getting higher scores with stronger relation, and the loss function is defined as follows:

$$Loss = \sum_{(E_i, E_j, R_{ij}) \in \{T \bigcup T'\}} ln(exp(-\vartheta \cdot g(E_i, E_j, R_{ij}) + 1), \quad (24)$$

where T and T' represent positive strong relation and negative weak relation, respectively. ϑ is a parameter that takes the value of 1 when the triplet belongs to the positive strong relation and -1 otherwise.

Multimodal knowledge graph embedding

We designed a novel pre-trained language model CKGM based on the above MKG embedding and benchmark BERT, which contains two submodules: the knowledge embedding module and language module. The knowledge embedding module adopts GAT to realize structured information perception and entity embedding. The language module generates text semantic encoding based on contextual information.

As mentioned in the section "Knowledge fusion module", we define the knowledge as a triple $K_i = (E_i, E_j, R_{ij})$, and then, the argument $\mathcal{K}$ in the textual FI quadruple $F_i = (\mathcal{H}_i, \mathcal{H}_j, l_{ij}, \mathcal{K})$ and the adjacent entities in the ERG are jointly defined as a new entity set $\mathcal{N}_v$. We define $\mathcal{V} = \{[MASK], [CLS], [EOS], w_1, \ldots, w_v\}$ as a vocabulary to mark contextual information, where $[MASK]$ represents the special mark that masks the token and $[CLS]$, $[EOS]$ represents the start and end marks of the token sequence, respectively. In addition, we define a token sequence $\mathcal{X} = [x_1, x_2, \ldots, x_v]$, and a reference entity set of $\mathcal{X}$ is $\mathcal{M} = [m_1, m_2, \ldots, m_m]$, where for each $m_i = (e_{mi}, s_{mi}, o_{mi})$ of the sets, e_{mi} represents the corresponding entity, while s_{mi}, o_{mi} represent the start and end marks, respectively. To better integrate external knowledge with text, we also extended the entity description $x^{e_{mi}}$ to the entity e_{mi} in the MKG.

Our model first reconstructs the MKG to generate knowledge-based entity representations by optimizing the problem that the GAT can only exploit ERs on a single edge. Specifically, by taking the co-embedding of ERs, we embed the entity e_i into the l-th hidden layer of the model and encode it as $Emb_{e_i}^l$

$$Emb_{e_i}^l = LN\left(\sum_{k=1}^{K} \delta\left(\sum_{(r,j)\in\mathcal{N}_v} \xi_{v,r,j}^k W^k f(Emb_{e_j}^{l-1}, R_r)\right)\right) + Emb_{e_i}^{l-1} \quad (25)$$

$$\xi_{v,r,j}^k = \frac{exp(LeakyReLU(a^T[W^k Emb_{e_i}^{l-1} \bigoplus W^k Emb_{e_j}^{l-1}, R_r]))}{\sum_{(r',j')\in\mathcal{N}_v} exp(LeakyReLU(a^T[W^k Emb_{e_j}^{l-1} \bigoplus W^k Emb_{e_{j'}}^{l-1}, R_{j'}]))}, \quad (26)$$

where K denotes the number of multi-head attentions. W^k is the parameter between each layer of the model, R_r represents the embedding vector of the relation r between entities, and the function $f()$ can embed entities and relationships into a representation; by repeating the above operations, the final output of entity embedding based on GAT is $Emb_{e_i}^{\text{final}}$.

Algorithm 1 Entity memory embedding algorithm

Start:
1: The entity e_i is embedded and encoded as $Emb_{e_i}^l$:
2: Transform entity e_i into the corresponding entity description x_i^e
3: The language module calculates the context embedding $Emb_j^t = \frac{\mathcal{B}_{s_{mi}}^{e_i} + \mathcal{B}_{o_{mi}}^{e_i}}{2}$ of all entity descriptions x_i^e
4: The knowledge embedding module selects the suitable Emb_{e_i} from Emb_{all}
5: Update the representation of the Emb_{e_i}:
6: Update the entity description embedding based on the current language module before each time step:
7: Calculate the time steps $\mathcal{J}_{\text{initial}}$ before the first update of the embedded representation
8: Set the increase rate of the update interval $\lambda^{[i/r]}$
9: $\mathcal{T}(i) = \min\left(\mathcal{J}_{\text{initial}} * \lambda^{[i/r]}, \mathcal{J}_{\text{now}}\right)$
10: Dynamic update entity embedding $Emb_{new} = \xi Emb_{e_i} + (1-\xi) Emb_{\mathcal{T}(i)}$
Update completed

The above embedding of ERs is complementary to the initial semantic representation generated by the BERT model. However, such a combination tends to have the problem of cycle dependence that affects the model's convergence, so we further optimize the BERT-based pre-training language model. To this end, we divide the BERT hiding layers into two relatively small encoding models with equal numbers. The former completes the basic encoding of the input sequence and the initial encoding for the output $Emb_{e_i}^{\text{final}}$ of the above knowledge embedding, while the latter fuses the basic contextual encoding with the initial encoding of the knowledge embedding. Specifically, the former initializes the sequence $\mathcal{X} = [x_1, x_2, \ldots, x_v]$ as $\mathcal{B}^{emb}$ and sends it to the next, which also precedes the entity description x^{e_i} as $\mathcal{B}^{e_i}$, and integrating the mean value of encoding between s_{mi} and o_{mi} in the entity set $m_i = (e_{mi}, s_{mi}, o_{mi})$ into the entity embedding encoding

process. The operation of the current time step is as follows:

$$\mathcal{B}^{emb} = \mathcal{G}\left(x_v\right) \tag{27}$$

$$\mathcal{B}^{e_i} = \mathcal{G}\left(x^{e_i}\right) \tag{28}$$

$$Emb_j^t = \frac{\mathcal{B}_{smi}^{e_i} + \mathcal{B}_{omi}^{e_i}}{2}. \tag{29}$$

Through n time steps of iteration, we obtain the final embedded representation $Emb_{e_i}^{\text{final}}$. We use a series of subtasks such as entity classification and relation classification to complete the pre-training, and take random sampling in each training s step to further enrich the semantic representation capability of the language model. In addition, due to the high computational power and training cost when MKG is embedded in BERT, we design an entity memory embedding algorithm to speed up the model training process. Algorithm 1 outlines the proposed approach's process.

Experiment

In the experiment, we adopt ablation strategy to gradually carry out experiments for different modules of our model to verify the effectiveness of each module, including image feature extraction and entity-relation graph generation, multimodal knowledge graph embedding, and finally to validate the performance of the proposed model on the text summarization task.

Dataset and evaluation metrics

Dataset

We conducted experiments on the visual Genome dataset [31] for validation of image feature extraction and ERG generation, which contains images, as well as annotations, attributes, and relationships between entities in the images.

For the experiment of MKG embedding and pre-training modeling, we used the lightweight dataset FewRel [32] and MSCOCO dataset [33]. The lightweight dataset FewRel [32] includes 70K samples on 100 relations. The MSCOCO dataset [33] is an open-source dataset built by Microsoft for tasks such as detection, segmentation, including over 120K images and 5 descriptive texts for each image.

We used the CNN/DailyMail dataset [34] for text summarization, which is one of the most comprehensive datasets for this task, containing millions of news articles and human-edited summaries covering different topics and styles. In this paper, we further preprocess this dataset. In detail, we anonymized all the documents in the dataset, segmented the words using the Stanford-CoreNLP parser, and then divided them into a training set and a test set after converting all words to lowercase letters.

Evaluation metrics

We evaluated our model with three evaluation metrics. The first is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [35], which is one of the international common text summary evaluation standards and provides a method to quickly evaluate a model's ability to produce summaries closer to those written by humans. In this paper, we use the standard ROUGE-1, ROUGE-2 and ROUGE-L metrics to measure summary quality. These three metrics evaluate the accuracy on unigrams, bigrams, and the longest common subsequence.

In addition, in the three downstream subtasks of the validation module for image feature extraction and ERG generation, and the two image processing subtasks of the knowledge graph embedding module, we choose to use the metric Recall@N, which is a commonly used metric that measures the relationship score between the real entity relationship triad and the predicted top N triads.

In the knowledge graph embedding module, we evaluate the classification task using the commonly used precision, recall, and F1 values, whose evaluation metrics are calculated as shown below, where TP represents the number of samples correctly predicted in a classification, FP represents the number of samples incorrectly predicted as that classification in other classifications, and TN is the number of samples incorrectly predicted as other classifications in that classification

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{30}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}} \tag{31}$$

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{32}$$

Experimental results and analysis

Validation of image feature extraction and ERG generation

Since the semantic feature of the image in our model depends on the ERG, we conduct experiments to verify the effectiveness of image feature extraction and ERG generation in this section. The scale and parameter settings are shown in Table 2.

We use three mainstream image processing tasks on the visual Genome dataset [31] to verify the performance of the model, namely the Entity-Pre task, which predicts the

Table 2 Parameter settings of image feature extraction and ERG generation

Feature extraction		ERG generation	
Parameter	Value	Parameter	Value
Learning rate	0.004	Learning rate	0.004
Decay rate	0.4	Decay rate	0.4
Convolution step length	1000	Random gradient descent impulse	0.9
Anchor frame size	[64, 128, 256]	Random gradient descent optimizer	SGD
Random gradient descent impulse	0.8	Loss function	Cross entropy loss
Random gradient descent optimizer	SGD		
Loss function	Cross entropy loss		

Table 3 There seems to be a problem with the layout of table 3. Recall@10 and Recall@30 are evaluation indicators and should not be included in the "Tasks" column. Please adjust the format to match the original layout.

Task	VRD	IMP	CKGM (Img)
		Recall@10	
Entity-Pre	29.3	46.5	64.1
Relation-Clas	12.6	26.4	37.8
Relation-Pre	0.5	5.1	21.7
		Recall@30	
Entity-Pre	38.2	55.7	71.3
Relation-Clas	15.5	26.9	39.4
Relation-Pre	0.7	7.2	25.3

relationship between entities, the Relation-Clas task, which predicts the relationship between entity categories and entities, and the Relation-Pre task, which predicts the relationship between entity locations, entity categories and entities. Our baseline models are IMP [36] and VRD [37], we used the Recall@N evaluation metric to measure the relationship scores between the actual ER triples and the predicted first N triples. As can be seen from Table 3, whether it is recall-10 or recall-30 evaluation index, the performance of our proposed model is better than the other two models. In detail, compared to the VRD model, our model scores increased by 34.8, 25.2, and 21.2 for the three tasks on Recall-10, and by 33.1, 23.9, and 24.6 for the three tasks on Recall-30, respectively. Compared to the IMP model, our model scores increased by 17.6, 11.4, and 16.6 for the three tasks on Recall-10, and 15.6, 12.5, and 18.1 for the three tasks on Recall-30, respectively.

Multi-model knowledge graph embedded in pre-trained language model

In this section, we further implemented the MKG embedding pre-training language model experiment through a series of downstream tasks, and the process and results of the experiment are shown below.

The first is the relation classification task, which mainly allows the model to predict the relationship between two entities. Given M relationships of N entities in each prediction, there will be a total of N × M ER pairs, and the model needs to complete classification of these ER pairs. The lightweight dataset FewRel [32] is applied to this task, we directly use the pre-trained language model and set a sequence classification layer outside the output layer of the model, by sending the connected sequence to the sequence classification layer, the scores of two samples representing the same relationship can be obtained. We have conducted comparative experiments with the benchmark Transformer [38], BERT [39], RoBERTa [40], and the above models with GNN added. As can be seen from Table 4, the precision scores of our model outperformed all other models when the size of the number of ER of the input model was set to 5 × 5, 10 × 5 and 10 × 10, respectively. Specifically, our model scores improve by 7.7%, 8.4%, and 8.6%, respectively, compared to the baseline Transformer, and further improve by 0.84%, 0.25%, and 0.38%, respectively, compared to the RoBERTa model with GNN added, which validates the effectiveness of our proposed method. The experimental results are intuitively shown in Fig. 6, where the input scales are 5 × 5 for the left image, 10 × 5 for the middle image, and 10 × 10 for the right image. It can be observed that our proposed model can effectively classify the relationships between entities regardless of the scale of the number of ER of the input model.

In addition, to validate the inference ability of our proposed model to ERs in the MKG, we also designed the entity classification task to predict the category labels corresponding to new entities in the test set. We conducted experiments using datasets of different sizes to verify the robustness of CKGM, namely, full dataset, random half dataset, and random quarter dataset. The experimental results are shown in Table 5. It can be seen that with the reduction of the size of the dataset, the score of the evaluation index decreases accordingly. However, the results of our model at the quarter dataset size are still close to the scores of Transformer on the full dataset. Specifically, on the full dataset, the precision,

Table 4 Experimental results of relation classification task

Model	5 × 5			10 × 5			10 × 10		
	P	R	F1	P	R	F1	P	R	F1
Transformer	77.6	76.7	75.9	75.4	74.8	74.2	73.6	72.8	71.7
BERT	79.2	78.5	77.8	77.9	77.2	76.4	77.1	76.1	75
Transformer+GNN	78.5	77.9	77.1	77.4	76.9	76	76.8	75.8	75.1
BERT+GNN	80.5	79.2	78.7	79.1	78.3	77.7	78.2	77.2	76.8
RoBERTa	81.8	80.3	79.9	80	79.2	78.8	79.1	78	77.2
RoBERTa+GNN	82.9	82	80.9	81.5	80.2	79.1	79.6	78.7	77.6
CKGM	83.6	83.2	82.4	81.7	81.1	80.7	79.9	79.2	78.1

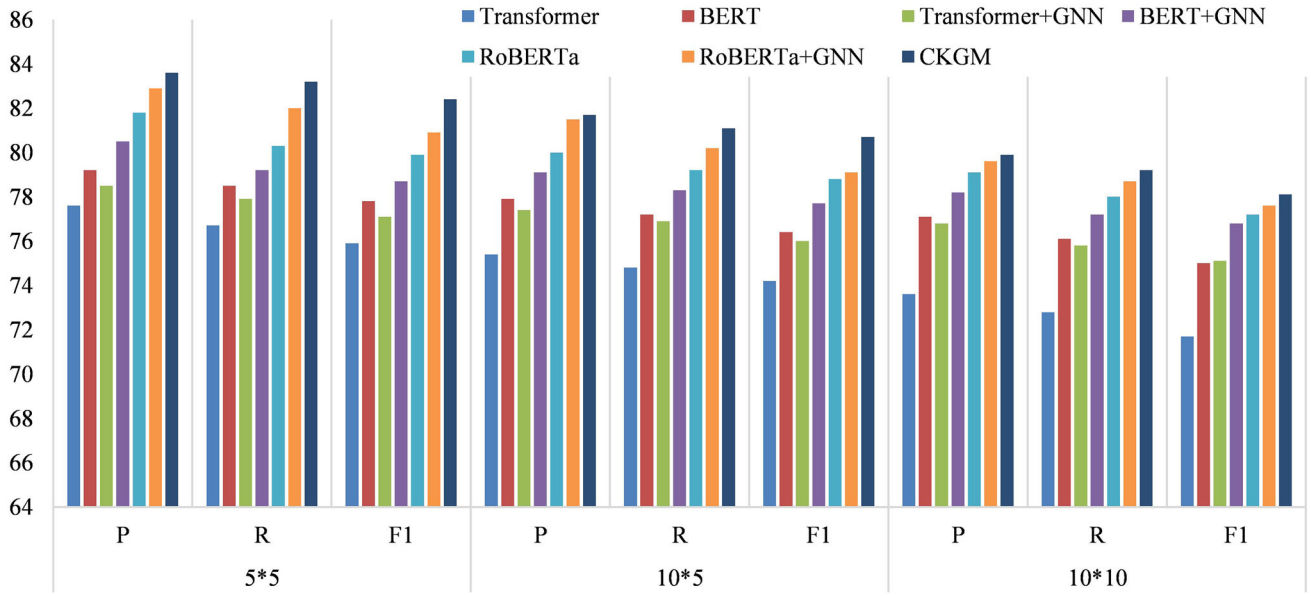

Fig. 6 Comparison results of different models on entity-relation classification subtask

recall, and F1 values of our models improve by about 30.1%, 30.0%, and 30.2%, respectively, compared to Transformer, and by about 4.0%, 3.9%, and 3.9%, respectively, compared to RoBERTa+GNN. Figure 7 shows the experimental results for each model on the full data set (left), half data set (middle), and quarter data set (right).

Then, we further proposed a downstream task of image-text retrieval task for experiments on the MSCOCO dataset [33], including two subtasks: image-to-text retrieval and text-to-image retrieval. For each text-image pairs, we need to retrieve the related image based on the text and retrieve the candidate text based on the image. We conducted comparative experiments on two types of subtasks with other related work, such as DVSA [41], m-CNN [42], DSPE [43], VSE++ [44], SCAN [45], SCG [46], PFAN [47], etc. The evaluation indexes include Recall@1, Recall@3, Recall@5, and Recall@10, and the experimental results are shown in Table 6. It can be seen that both CKGM and other models have better performance on text-based related image retrieval task than image-based related text retrieval task. Meanwhile, our model performs better than other models on different evaluation indicators of the two types of subtasks, and its scores are both above 70 points. In detail, compared with the [47], our model improved by 3.9%, 5.6%, 5.6%, and 6.3% in image-based related text retrieval task, and by 4.4%, 3.9%, 6.3%, and 7.3% in text-based related image retrieval task, respectively. Obviously, our model can match the most relevant series of image and text pairs in in such tasks, which reflects the effectiveness of our proposed model.

Text summary generation based on CKGM model

We compare our method with other state-of-the-art models on the CNN/Daily Mail dataset, including BanditSum [48], NeuSum [49], JECS [50], PG+SA [51], REFRESH [52], LATENT [53], BERTSum [54], PNBert [55], HiBert [56], ExtraPhrase [57], DRAS [58] etc. As can be seen from Table 7, our proposed CKGM performs more effectively than other models. Its scores on R-1, R-2, and R-L improved by 19.9%, 45.6%, and 16.8%, respectively, compared with the worst

Table 5 Experimental results of entity classification task

Model	Full			Half			Quarter		
	P	R	F1	P	R	F1	P	R	F1
Transformer	62.5	61.7	61	49.2	48.1	47.6	33.2	32.6	31.4
BERT	70.2	69.4	68.7	56.1	55.1	54.4	39.7	38.5	38
Transformer+GNN	65.1	64.2	63.7	51	50.1	48.9	34.1	33.4	31.8
BERT+GNN	73	72.6	72.1	60.4	59.7	58.9	42.6	41.4	40.7
RoBERTa	75.9	74.3	73.1	63.9	62.3	61.7	44.7	43.6	43.1
RoBERTa+GNN	78.2	77.2	76.4	67.2	66	65.3	47.2	46.4	45.6
CKGM	81.3	80.2	79.4	69.5	69.1	68.3	52.1	51.4	50.8

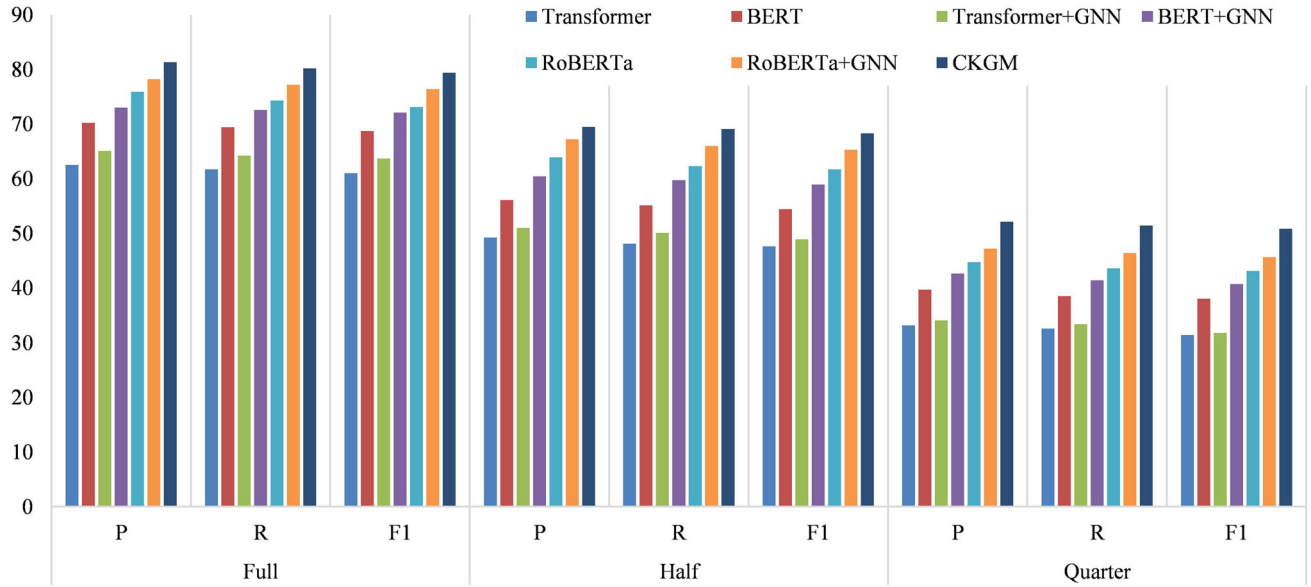

Fig. 7 Comparison results of different models on the entity label classification subtask

Table 6 Comparative experimental results of the two types of subtasks

Model	Image-based related text retrieval				Text-based related image retrieval			
	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10
DVSA [41]	41.4	42.5	42.9	43.3	42.5	43.6	43.8	44.3
m-CNN [42]	44.1	44.7	45.2	47.1	45.1	45.8	46.5	47.2
DSPE [43]	49.7	50.2	50.8	52	50.5	51.3	52.7	53.5
VSE++ [44]	52.4	53.2	53.9	55.1	54.3	54.9	55.6	56.2
SCAN [45]	57.9	58.6	59.3	60.1	59.4	60.8	61.3	62.1
SCG [46]	65.6	66.3	67.5	68.3	65.8	66.3	67.4	68.6
PFAN [47]	69.6	70.2	71.5	73	70.3	71.1	71.7	72.9
CKGM	72.3	74.1	75.5	77.6	73.4	73.9	76.2	78.2

model [48]. And our model improves 5.5% and 0.8% on R-2 and R-L, respectively, compared to the DRAS [58] with the highest R-1 score. In this paper, our model fuses multi-modal features, so the generated text encoding contains richer semantic information. In addition to the text, it also contains rich context representation and entity-relation information of the image, so as to improve the effect of text summary generation. To further visualize the experimental results, as shown in Fig. 8, it can be seen that our proposed model works well on all three metrics.

Table 7 Index scores of each model on CNN/Daily Mail dataset

Model	CNN/Daily Mail						
	R-1	R-2	R-L		R-1	R-2	R-L
BanditSum [48]	34.31	12.85	32.74	BERTSum [54]	38.45	16.85	37.17
NeuSum [49]	34.56	13.62	33.35	PNBert [55]	38.87	17.01	37.29
JECS [50]	36.82	14.75	34.03	HiBert [56]	39.91	17.31	37.64
PG+SA [51]	38.15	16.98	33.20	ExtraPhrase [57]	40.57	18.22	37.51
REFRESH [52]	38.27	16.51	35.85	DRAS [58]	41.35	17.73	37.91
LATENT [53]	38.64	16.46	36.14	CKGM	41.13	18.71	38.24

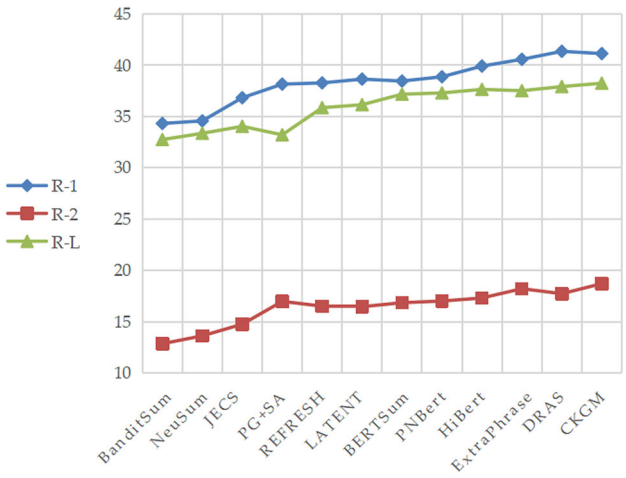

Fig. 8 Comparison results for text summaries on the CNN/Daily Mail dataset

Ablation studies

Evaluation of unimodal text summary

To verify the effectiveness of unimodal (text-only) summary generation, we set up ablation experiments on the multi-level encoding module, the multi-granularity entity representation module, and the multi-head attention of our proposed model. Specifically, as shown in Fig. 9, the multi-level encoding module contains three strategies: word-level, sentence-level, and chapter-level encoding. And the FI extraction module also contains three strategies: initial, updated, and global entity representation. Based on the above strategies, we set up six groups of experiments (including full model as a control group) on CNN/Daily Mail dataset and adopted ROUGE-1, ROUGE-2, and ROUGE-L as evaluation indexes. The experimental results are shown in Tables 8 and 9.

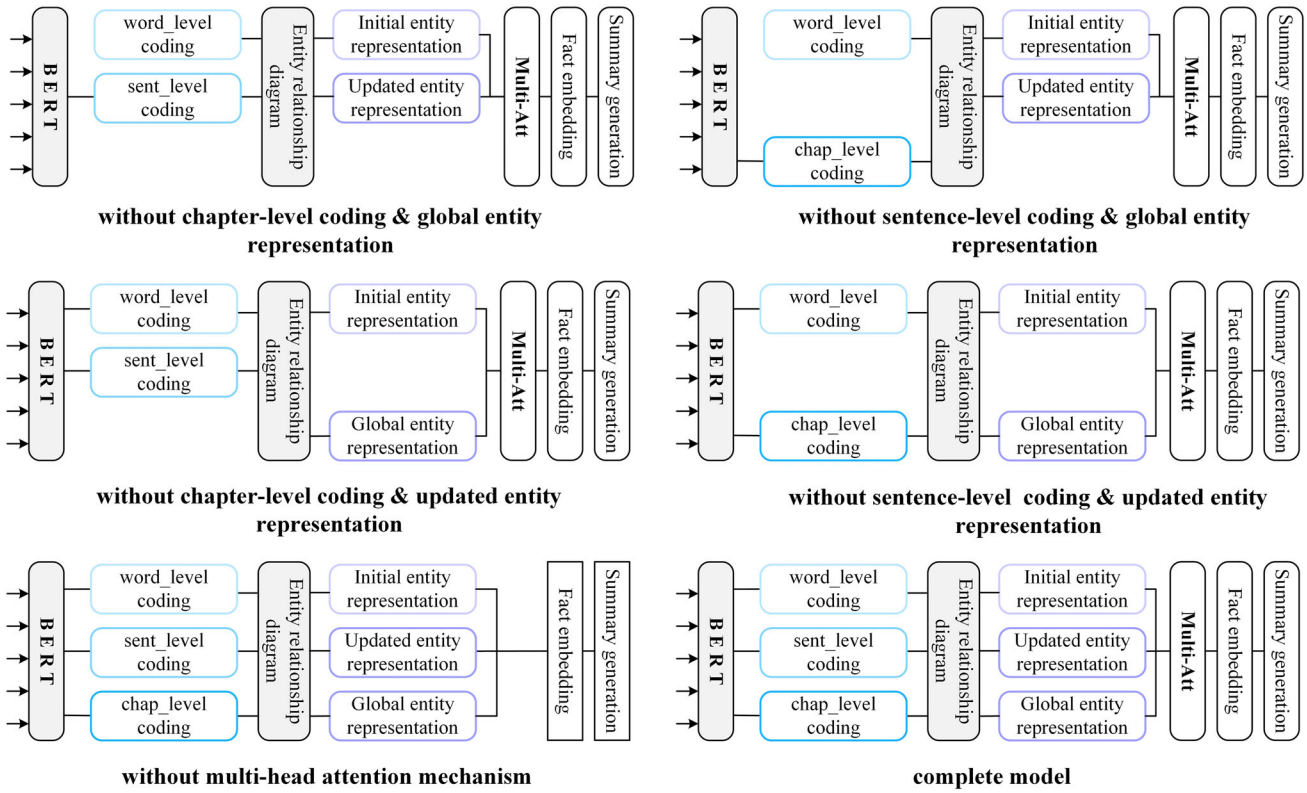

Fig. 9 Different strategies for ablation experiments

Table 8 The ablation results on the CNN dataset

Ablation strategy (without)									
word_ level coding	para_ level coding	text_ level coding	Initial entity representation	Updated entity representation	Global entity representation	Multi-head attention	R-1	R-2	R-L
✓	✓	–	✓	✓	–	✓	33.48	12.73	32.83
✓	–	✓	✓	✓	–	✓	34.15	13.13	33.91
✓	✓	–	✓	–	✓	✓	32.13	12.01	30.44
✓	–	✓	✓	–	✓	✓	33.52	12.46	31.05
✓	✓	✓	✓	✓	✓	–	38.24	16.11	36.87
✓	✓	✓	✓	✓	✓	✓	39.85	16.59	36.91

Table 9 The ablation results on the Daily Mail dataset

Ablation strategy (without)									
word_ level coding	para_ level coding	text_ level coding	Initial entity representation	Updated entity representation	Global entity representation	Multi-head attention	R-1	R-2	R-L
✓	✓	–	✓	✓	–	✓	34.37	13.49	33.72
✓	–	✓	✓	✓	–	✓	35.01	13.61	34.11
✓	✓	–	✓	–	✓	✓	32.89	12.5	31.81
✓	–	✓	✓	–	✓	✓	33.71	13.11	33.37
✓	✓	✓	✓	✓	✓	–	39.28	17.35	37.07
✓	✓	✓	✓	✓	✓	✓	40.29	17.51	37.57

We observed that the complete model achieved the best scores on all three indexes. When the model lacks the global entity representation, its scores of R-1, R-2, and R-L indexes on the CNN dataset decreased by 6.37, 3.86, and 4.08, respectively, compared with the complete model, which is due to the fact that the global entity representation integrates rich context information and contains FI that plays a key role in text summarization. When the model lacks the multi-head attention mechanism, it performs much better than without other modules, but it is still does not perform as well as the complete model. its scores of R-1, R-2, and RL indexes on the Daily Mail dataset decreased by 2.5%, 0.9%, and 1.3%, respectively, compared with the complete model, this is because although the multi-level coding and multi-granularity entity representation can integrate contextual features and relationships, the model cannot incorporate and embed FI well without multi-head attention. The experimental result proves the effectiveness of each module in the model, whether multi-level encoding module, multi-granularity entity representation, or multi-head attention, they all play an important role in the performance of the text summary generation.

Evaluation of CKGM for AS

To demonstrate and analyze the effect of our proposed model CKGM on AS task, we further set up ablation strategies and conducted experiments on it. Specifically, in addition to the multi-level encoding module and the multi-granularity entity representation module mentioned in "Evaluation of unimodal text summary", we further added the multi-scale image feature extraction module and the MKG embedding module to verify the effectiveness of our model.

As can be seen from Table 10, by comparing the experimental results of (A) and (B), the model using only multi-level coding and multi-granularity entity representation has increased by 0.78, 1.09, and 0.37, respectively, compared with only using multi-scale image feature and MKG embedding, and its better performance on the R-2 index confirms that the CKGM can extract the most relevant entity-relation information well. By comparing the experimental results of (C), (D), (E), (F), and (G), we observe that the models (E) and (F) obtained better scores than models (C) and (D), which indicates that multi-scale image feature and MKG embedding have more important gains for the model. Without multi-scale image feature extraction, the entity-relation information contained in the image cannot be fully identified and extracted. In the end, the model with all modules achieved the most outstanding experimental performance. The scores on the three indexes are 41.97, 19.36, and 39.65, respectively, which are 4.01%, 9.87%, and 4.70% higher than the model (F), respectively, which validates the effectiveness and advancement of our proposed model.

Table 10 Ablation experiments with different module combinations

Models	Multi-level encoding	Multi-granularity entity representation	Multi-scale image feature	MKG embedding	R-1	R-2	R-L
(A)	✓	✓	–	–	40.35	17.62	37.87
(B)	–	–	✓	✓	41.13	18.71	38.24
(C)	✓	✓	✓	–	40.79	17.8	37.91
(D)	✓	✓	–	✓	41.02	18.53	38.08
(E)	✓	–	✓	✓	41.26	18.89	38.87
(F)	–	✓	✓	✓	41.62	19.18	39.36
(G)	✓	✓	✓	✓	41.97	19.36	39.65

Conclusion

In this work, we propose a novel Cross-modal Knowledge Guided Model (CKGM) for AS. First, we introduce a multi-granularity entity-relation representation and extract the FI of the text, and define a scoring function to complete the importance assessment of the FI. Second, we construct the ERG of the image by extracting semantic and spatial feature information, and finally design an MKG and embed it into BERT as external knowledge. In addition, we also propose an entity memory embedding algorithm further to improve the information fusion efficiency and model training speed. Our model can effectively increase the informativeness of the summaries while improving the factual consistency of the generated summaries. We conducted a number of comparative experiments and evaluated our model on multiple datasets. Experimental results demonstrate that our proposed model can improve the performance of text summaries compared to previous works.

In addition, our proposed model (CKGM) in this paper embeds an MKG to implement the AS task, which can also be used as a baseline model for other tasks of natural language processing, such as question and answer systems, and sentiment recognition. Therefore, we will continue to work on improving the training efficiency of the model, optimizing the quality of summary, and considering applying our model to other natural language generation tasks in the future.

Funding This study was supported by National Key Research and Development Program of China (No. 2021YFC2801000), National Natural Science Foundation of China (No. 61872231), Major Research plan of the National Social Science Foundation of China (No. 20&ZD130).

Data Availability Statement: The four open-access datasets, CNN/ DailyMail, MSCOCO, Visual Genome, and FewRel are used in our study. Their links are as follows: CNN/ DailyMail: https://cs.nyu.edu/\protect\unhbox\voidb@x\penalty\@M\kcho/DMQA/; MSCOCO: http://mscoco.org; Visual Genome: https://visualgenome.org/api/v0/api_home.html; FewRel: https://thunlp.github.io/fewrel.html.

Declarations

Conflict of interest The authors declare that there is no conflict of interest in this paper.

References

1. Fang H, Zhu G, Stojanovic V, Nie R, He S, Luan X, Liu F (2022) Adaptive optimization algorithm for nonlinear Markov jump systems with partial unknown dynamics. Int J Robust Nonlinear Control 31:2126–2140
2. Chang S, Liu J (2020) Multi-lane capsule network for classifying images with complex background. IEEE Access 99:1–1
3. Song X, Sun P, Song S, Stojanovic V (2022) Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance. J Franklin Inst 9:359
4. Ren P, Chen Z, Ren Z, Wei F, Nie L, Ma J, Rijke MD (2018) Sentence relations for extractive summarization with deep neural networks. ACM Trans Inf Syst 36(4):39–13932
5. Deng Z, Ma F, Lan R, Huang W, Luo X (2020) A two-stage Chinese text summarization algorithm using keyword information and adversarial learning. Neurocomputing 425:117–126
6. Liu J, Yang Y, Lv S, Wang J, Chen H (2019) Attention-based BiGRU-CNN for Chinese question classification. J Ambient Intell Humaniz Comput 2:1–12
7. Shang S, Liu J, Yang Y (2020) Multi-layer transformer aggregation encoder for answer generation. IEEE Access 8:90410–90419
8. Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. Comput Sci
9. Cao Z, Wei F, Li W, Li S (2017) Faithful to the original: fact aware neural abstractive summarization
10. Falke T, Ribeiro L, Utama PA, Dagan I, Gurevych I (2019) Ranking generated summaries by correctness: an interesting but challenging application for natural language inference. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics
11. Kryscinski W, Mccann B, Xiong C, Socher R (2020) Evaluating the factual consistency of abstractive text summarization. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)
12. Wang A, Cho K, Lewis M (2020) Asking and answering questions to evaluate the factual consistency of summaries
13. Zhang Y, Merck D, Tsai EB, Manning CD, Langlotz CP (2019) Optimizing the factual correctness of a summary: a study of summarizing radiology reports
14. Dong Y, Wang S, Gan Z, Cheng Y, Liu J (2020) Multi-fact correction in abstractive text summarization
15. Zhu C, Hinthorn W, Xu R, Zeng Q, Zeng M, Huang X, Jiang M (2020) Boosting factual correctness of abstractive summarization
16. Jin LA, Yy A, Hh B (2020) Multi-level semantic representation enhancement network for relationship extraction. Neurocomputing 403:282–293
17. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. Comput Sci
18. Chopra S, Auli M, Rush AM (2016) Abstractive sentence summarization with attentive recurrent neural networks. In: Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies
19. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks
20. Paulus R, Xiong C, Socher R (2017) A deep reinforced model for abstractive summarization
21. Li W, Xiao X, Lyu Y, Wang Y (2018) Improving neural abstractive document summarization with structural regularization. In: Proceedings of the 2018 conference on empirical methods in natural language processing
22. Zhang C, Zhang Z, Li J, Liu Q, Zhu H (2021) Ctnr: compress-then-reconstruct approach for multimodal abstractive summarization.

In: International joint conference on neural networks. https://doi.org/10.1109/IJCNN52387.2021.9534082
23. Li H, Zhu J, Zhang J, He X, Zong C (2020) Multimodal sentence summarization via multimodal selective encoding. In: International conference on computational linguistics
24. Zhu J, Xiang L, Zhou Y, Zhang J, Zong C (2021) Graph-based multimodal ranking models for multimodal summarization. Transactions on Asian and low-resource language information processing
25. Gong P, Liu J, Yang Y, He H (2020) Towards knowledge enhanced language model for machine reading comprehension. IEEE Access 8:224837–224851
26. Schlichtkrull M, Kipf TN, Bloem P, Berg R, Titov I, Welling M (2017) Modeling relational data with graph convolutional networks
27. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. IEEE
28. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
29. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-nms—improving object detection with one line of code
30. Dey R, Salemt FM (2017) Gate-variants of gated recurrent unit (gru) neural networks. In: IEEE international Midwest symposium on circuits and systems, pp 1597– 1600
31. Krishna R, Zhu Y, Groth O, Johnson J, Li FF (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1)
32. Han X, Zhu H, Yu P, Wang Z, Yao Y, Liu Z, Sun M (2018) Fewrel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Proceedings of the 2018 conference on empirical methods in natural language processing
33. Lin TY, Maire M, Belongie S, Hays J, Zitnick CL (2014) Microsoft coco: common objects in context. Springer International Publishing, Cham
34. Hermann KM, Koisk T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. MIT Press, Cambridge
35. Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Proceedings of the workshop on text summarization branches out (WAS 2004)
36. Xu D, Zhu Y, Choy CB, Fei-Fei L (2017) Scene graph generation by iterative message passing. IEEE Computer Society
37. Leibe B, Matas J, Sebe N, Welling M (2016) Lecture notes in computer science. Computer vision—ECCV 2016, vol 9905. Visual relationship detection with language priors (Chapter 51), pp 852–869. https://doi.org/10.1007/978-3-319-46448-0
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need
39. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding
40. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach
41. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Computer vision and pattern recognition
42. Ma L, Lu Z, Shang L, Li H (2015) Multimodal convolutional neural networks for matching image and sentence. In: IEEE international conference on computer vision
43. Wang L, Yin L, Lazebnik S (2016) Learning deep structure-preserving image-text embeddings. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR)
44. Faghri F, Fleet DJ, Kiros JR, Fidler S (2017) Vse++: improved visual-semantic embeddings
45. Lee KH, Xi C, Gang H, Hu H, He X (2018) Stacked cross attention for image-text matching
46. Shi B, Ji L, Lu P, Niu Z, Duan N (2019) Knowledge aware semantic concept expansion for image-text matching. In: Twenty-eighth international joint conference on artificial intelligence IJCAI-19
47. Wang Y, Yang H, Qian X, Ma L, Fan X (2019) Position focused attention network for image-text matching
48. Yue D, Shen Y, Crawford E, Hoof HV, Cheung, J (2018) Banditsum: extractive summarization as a contextual bandit. In: Proceedings of the 2018 conference on empirical methods in natural language processing
49. Zhou Q, Yang N, Wei F, Huang S, Zhou M, Zhao T (2018) Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics, vol 1: Long Papers
50. Xu J, Durrett G (2019) Neural extractive text summarization with syntactic compression
51. Chowdhury T, Kumar S, Chakraborty T (2020) Neural abstractive summarization with structural attention
52. Narayan S, Cohen SB, Lapata M (2018) Ranking sentences for extractive summarization with reinforcement learning
53. Zhang X, Lapata M, Wei F, Ming Z (2018) Neural latent extractive document summarization. In: Proceedings of the 2018 conference on empirical methods in natural language processing
54. Liu Y, Lapata M (2019) Text summarization with pretrained encoders
55. Zhong M, Liu P, Wang D, Qiu X, Huang X (2019) Searching for effective neural extractive summarization: what works and what's next. arXiv e-prints, arXiv:1907.03491 [cs.CL]
56. Zhang X, Wei F, Zhou M (2019) Hibert: document level pre-training of hierarchical bidirectional transformers for document summarization. In: Proceedings of the 57th annual meeting of the association for computational linguistics
57. Loem M, Takase S, Kaneko M, Okazaki N (2022) Extraphrase: efficient data augmentation for abstractive summarization
58. Liao W, Ma Y, Yin Y, Ye G, Zuo D (2021) Improving abstractive summarization based on dynamic residual network with reinforce dependency. Neurocomputing 448:228–237