Journal of Intelligent Information Systems (2024) 62:191–208
https://doi.org/10.1007/s10844-023-00812-1

RESEARCH

Self-supervised opinion summarization with multi-modal knowledge graph

Lingyun Jin[1] · Jingqiang Chen[1]

Received: 24 May 2023 / Revised: 21 August 2023 / Accepted: 21 August 2023 /
Published online: 1 September 2023

Abstract

Multi-modal opinion summarization aims at automatically generating summaries of products or businesses from multi-modal reviews containing text, image and table to present clear references for other customers. To create faithful summaries, multi-modal structural knowledge should be well utilized, which is neglected by most existing work on multi-modal opinion summarization. Thus, we propose an opinion summarization framework based on multi-modal knowledge graphs (MKGOpinSum) to utilize structural knowledge in multi-modal data for opinion summarization. To construct a multi-modal knowledge graph, we first build a textual knowledge graph from review text and then enrich it by linking detected image objects to its corresponding entities. Our method obtains each modality representation from their own encoders, and generates the summary from the text decoder. To address the issue of heterogeneity of multi-modal data, we adopt a multi-modal training pipeline. In the pipeline we first pretrain text encoder and decoder with only text modality data. Then we respectively pretrain table and MKG modality by taking text decoder as a pivot. Finally, we train the entire encoder-decoder architecture and fuse representations of all modalities to generate the summary text. Experiments on Amazon and Yelp dataset show the framework has satisfactory performances when compared to ten baselines.

Keywords Opinion summarization · Multi-modal knowledge graph · Multi-modal data · Training pipeline

Lingyun Jin and Jingqiang Chen contributed equally to this work.

✉ Jingqiang Chen
cjq@njupt.edu.cn

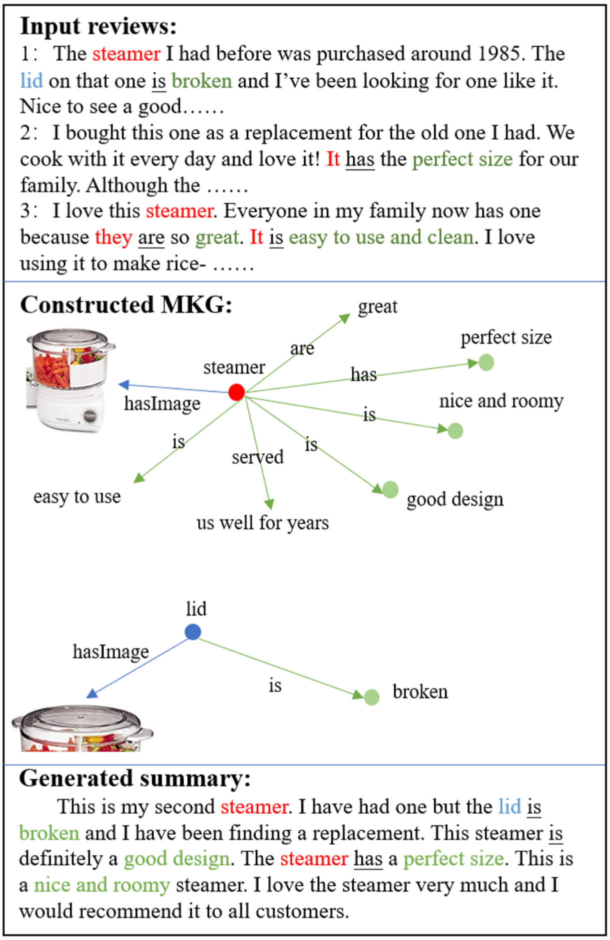
Lingyun Jin
jinlingyun19@gmail.com

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210049, Jiangsu, China

1 Introduction

Recently E-commerce industry such as online shopping is in full swing and has the tendency to replace offline stores. Being unable to touch the real products, other customers' comments become a significant reference to reach the purchase decision. However, it is quite difficult for one to quickly obtain reliable and useful information from numerous reviews. Thus, opinion summarization aiming at automatically generating summaries from several reviews arouses wide attention (Hu and Liu, 2004; Medhat et al., 2014).

Most reviews are multi-modal reviews with text, image, table, and e.t.c. Existing work makes use of multi-modal information to generate summary reviews. However, these methods utilize multi-modal data but ignore structural knowledge in reviews. Figure 1 shows an example of how multi-modal structural knowledge helps summary generation. With structural knowledge, meaningful characteristics that could offer intuitive information to customers are extracted, the target of reviews "steamer" is extracted and used to resolve ambiguity as well. The words colored red in Fig. 1 all point to the same product, while methods only



Input reviews:

1: The steamer I had before was purchased around 1985. The lid on that one is broken and I've been looking for one like it. Nice to see a good……

2: I bought this one as a replacement for the old one I had. We cook with it every day and love it! It has the perfect size for our family. Although the ……

3: I love this steamer. Everyone in my family now has one because they are so great. It is easy to use and clean. I love using it to make rice- ……

Constructed MKG:

Generated summary:

This is my second steamer. I have had one but the lid is broken and I have been finding a replacement. This steamer is definitely a good design. The steamer has a perfect size. This is a nice and roomy steamer. I love the steamer very much and I would recommend it to all customers.

Fig. 1 Sample of constructed multi-modal knowledge graph

with reviews may misunderstand these pronouns, which can cause a negative effect on the generated summary.

To make full use of multi-modal structural knowledge for opinion summarization, we propose to build and use multi-modal knowledge graph (MKG) to improve opinion summarization. To build MKG, we first construct text knowledge graphs, and then detect objects from pictures that are in the image set. After calculating the similarities between entities in textual knowledge graphs and detected objects, we connect those related pairs to form the final MKG. For instance, Fig. 1 shows the contribution of MKG. Apart from the benefits of knowledge graphs we mentioned before, with the help of detected objects in MKG the generated summary accurately catches the reason of replacement "lid is broken", which is missed by other multi-modal opinion summarization methods.

For there can hardly find an off-the-peg summary for reviews, recent studies prefer self-supervised way (Bražinskas et al., 2020; Amplayo and Lapata, 2020; Elsahar et al., 2021) that they select a review text from the entire review set as the pseudo summary. We propose our opinion summarization framework called MKGOpinSum by following these self-supervised studies. Our framework is an encoder-decoder architecture that each modality has its own encoder to obtain representations, and generates summaries through the text decoder. To address the issue of heterogeneity of multi-modal data, we adopt and modify a multi-modal training pipeline based on (Im et al., 2021). This pipeline takes text modality as a pivot, so that we pretrain text encoder and decoder with the entire review set first. Then we respectively pretrain table encoder and MKG encoder. To fully utilize the structural knowledge in table data, we create a graph structure to encode table modality data. In order to make better use of constructed MKG, we modify graph attention network (GAT) (Velickovic et al., 2017) by taking relation information into consideration to acquire MKG embeddings. During the training process of table modality and MKG modality, we froze the text decoder pretrained before to train their encoders' ability of obtaining homogenous representation with text encoder. Finally, we combine the multi-modal information obtained from each modality by training the entire model. Following are the main contributions of our work:

- Our work is the first to apply MKG to the task of opinion summarization.
- We propose the summarization framework MKGOpinSum that first builds MKG from reviews and contains a training pipeline to issue the problem of heterogeneity. We propose to modify graph attention network to make use of our constructed MKG as well.
- Experiments show that our model outperforms baselines such as Self&Control and MultimodalSum on Yelp and Amazon datasets according to ROUGE score and BERT score, which proves the effectiveness of our model.

2 Related work

2.1 Multi-modal knowledge graph

Knowledge graph can be seen as a semantic network used to describe entities and concepts in the real world and the relationships between them. As its extension, Multi-modal Knowledge Graph(MKG) is able to fully integrate and utilize data from multi-modal sources such as texts, images and videos. Recently, several studies have applied MKG to their own task and achieved remarkable results. For instance, (Pezeshkpour et al., 2018) uses different neural encoders to learn embeddings of entities and multi-modal data in the knowledge graph, and then employs them to the knowledge base completion task. (Wilcke et al., 2020) learns knowledge from both the structure of graphs and the possible divers set of multi-modal node features. (Chen et al., 2020, 2022) integrate the embeddings of relational, visual and attribute modality as a total embedding and leverage it to align entities between different MKGs. (Li et al., 2023) feeds the constructed attritube-consistent KG into graph neural network for relation and entity representations. (Zhao et al., 2022) modifys Graph Convolutional Network (GCN) to leverage cross-modal relation information for multi-modal NER task. Besides, there are also lots of studies putting MKG into real world applications. (Sun et al., 2020) enhances recommender systems by utilizing a multi-modal graph attention technique over MKG. (Sacenti et al., 2022) leverages Graph Summarization to recommend movies with help of knowledge graph. (Ma et al., 2022) proposes a special multi-modal event knowledge graph that bridges and complements different modalities of knowledge for better understanding. So far, there hardly exists studies that focus on applying MKG to opinion summarization.

2.2 Summarization

2.2.1 Multi-modal summarization

Different from text summarization, multi-modal summarization has the ability of extracting information from different modalities and generating more reliable summaries. (Li et al., 2018) creates single-modal summary by taking other modalities data as additional input. (Xiao et al., 2023) takes the contribution of images into consideration by a contribution network, which helps the generation of summary. (Liang et al., 2023) generates more accurate summaries by capturing the summary-oriented visual features. Apart from the output of single-modal, there are also studies that provide multi-modal outputs (Zhu et al., 2018; Chen and Zhuge, 2018; Zhu et al., 2020; Zhang et al., 2022; Xie et al., 2022). In this paper, we focus on generating single-modal summaries with the help of text modality data, table modality data and the constructed MKG.

2.2.2 Opinion summarization

Existing studies on opinion summarization can be divided into extractive and abstractive methods. The extractive methods aim to select a subset of useful sentences from the input review to get a concise summary. (Ku et al., 2006); (Paul et al., 2010); (Angelidis and Lapata, 2018) collect cluster opinions about the same side and then choose the text that represents each cluster. (Basu Roy Chowdhury et al., 2022) leverages dictionary learning to obtain semantic information from the review and learns the potential representation of each sentence to identify representative opinions. (Amplayo et al., 2021) conducts a synthetic training dataset and leads summary generation towards the specified aspects. The abstractive methods generate summaries based on the reviews, and the generated words do not have to exist in the original reviews. (Chu and Liu, 2019) generates summaries from the information aggregated from the encoder-decoder architecture. (Bražinskas et al., 2020) randomly selects one review as the pseudo summary while others as source reviews. (Zhang et al., 2023) builds a heterogeneous graph consisting of reviews and opinion clusters as nodes, and design an attention mechanism to select aspects that are most likely to appear in summary.

The most related work for our study is (Im et al., 2021). They use different encoders to respectively obtain representation of each modality, and a text decoder to generate summaries. They propose a training pipeline to address the heterogeneity problem of multi-modal data. (Im et al., 2021) shows good performance for using multi-modal data. However, for image modality they aggregate all image representations to generate summary which neglects structural knowledge in images and review text. For table modality, their study simply concentrates representations of field name and field value, ignoring structural knowledge between fields and source text. Our study is different from (Im et al., 2021). We model text and image modality data by constructing MKG from text and image data to make use of multi-modal structural knowledge. We use the graph structure to model table modality data, providing more structured information. Hence, we are able to deeply mine the knowledge among text, image and table, creating more detailed and related summaries.

3 Our method

The goal of our method is to generate the summary with review texts, tables and images as input. Figure 2 shows our opinion summarization framework based on multi-modal knowledge graphs called MKGOpinSum. Firstly, MKG is constructed from input review texts and images. Secondly, the multi-step training pipeline is proposed to generate review summaries by separately pretraining text, table and MKG modality, and then we train the entire framework for summary generation. Details are given in following subsections.

3.1 Multi-modal knowledge graph construction

Common knowledge graphs are usually in the form of barely text. MKG add information of additional modalities such as image, which enhances representation of images and text. We denote a MKG as $G = \{E, C\}$, where E represents the entity set and C represents the relation set. The entity set E contains both entities from review text and detected object from image. To construct a MKG, we first build a textual knowledge graph and then enrich it by

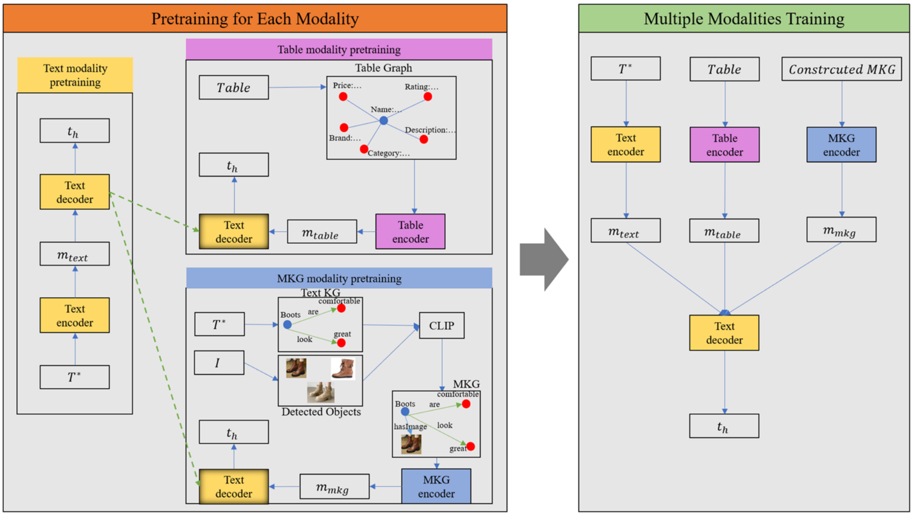


Fig. 2 The structure of our proposed method. The whole framework can be divided into two parts: The first part is pretraining for text, table and MKG modality. We firstly pretrain text modality and use the pretrained text decoder to train table and MKG decoder; The second part is multiple modalities training, we further train the whole framework for generation of summaries

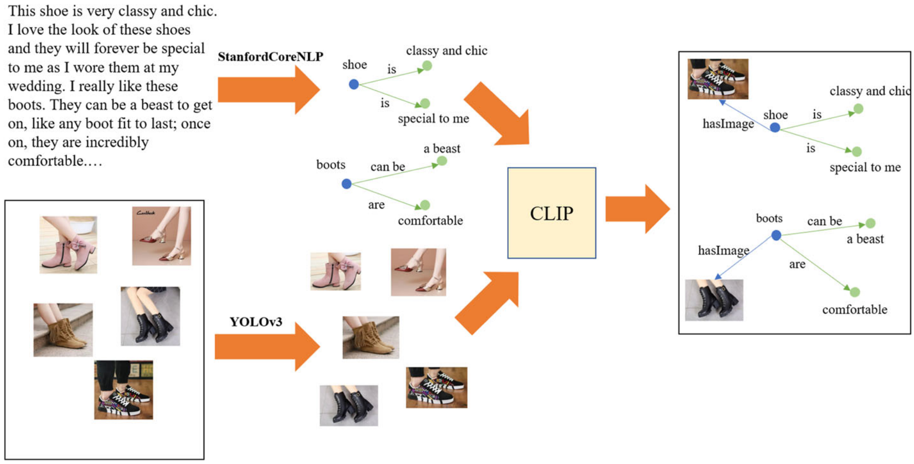

This shoe is very classy and chic. I love the look of these shoes and they will forever be special to me as I wore them at my wedding. I really like these boots. They can be a beast to get on, like any boot fit to last; once on, they are incredibly comfortable.…
StanfordCoreNLP
shoe
is
classy and chic
is
special to me
boots
can be
a beast
are
comfortable
YOLOv3
CLIP
hasImage
shoe
is
classy and chic
is
special to me
boots
can be
a beast
hasImage
are
comfortable

Fig. 3 The construction process of MKG

linking detected objects to their corresponding text entities. Figure 3 shows an example of how we construct our MKG.

To build the text knowledge graph $G^t = \{E^t, C^t\}$ where E^t represents the entities in the review text and C^t represents the relations, we integrate several reviews of one product or business as a document, and then utilize the Stanford CoreNLP toolkit (Manning et al., 2014) along with spaCy (Honnibal et al., 2020) for Named Entity Recognition (NER), entity extraction and coreference resolution. The output is a set of triples in the form of $<$ head, relation, tail $>$. For each triple, we take its head and tail as nodes and connect them with a directed edge whose attribute is the relation.

To enrich the knowledge graph, we apply YOLOv3 (Redmon and Farhadi, 2018) to detect objects from the image set. The output is a set of detect objects called E^i. The similarity between each entity in E^t and E^i is calculated through CLIP (Radford et al., 2021) model. CLIP is a pre-training neural network model for matching images and texts, which can be said to be a classic work in the field of multi-modal research in recent years. For the entity-object pairs whose similarities are higher than 0.75, we create a directed edge from text entity to the detected object with the attribute ‘hasImage’.

3.2 Training pipeline

Pretraining has been widely used in the field of machine learning because of its ability of making model less burdensome for specific tasks. Thus, inspired by (Im et al., 2021), we employ a training pipeline which contains four steps. The first is text modality pretraining that we pretrain the text encoder and decoder with only text modality data. Secondly, we pretrain the table encoder with the pretrained text decoder as a pivot. The third step is similar to the second step that we pretrain the MKG encoder. Finally, we train the whole framework by using all modality data.

3.2.1 Text modality pretraining

Given a dataset, we take $R = \{r_1, r_2, r_3, ..., r_N\}$ to present the review set related to one product or business. For each review r_i, it is made up of the review text t_i and the review rating g_i. Here the review rating means the general grade of evaluation. Since there is no ground-of-the-truth summaries for supervising, we use a self-supervised way for training by choosing a review from the review set as the hypothetic summary. The review is denoted as t_k which is the k-th review in the review set. We pretrain our text encoder and decoder based on BART (Lewis et al., 2020) with a denoising autoencoder for our model. BART is a Transformer-based model (Vaswani et al., 2017) with contextual information and autoregressive features, and shows great performance in the area of generation task and text comprehension task. Our text encoder and decoder work as follows:

$$m_{text} = BART_{encoder}(T^*),$$

$$t_k = BART_{decoder}(m_{text}),$$

where $m_{text} \in R^{(N-1)*l_T*d_T}$ is the text representation acquired from the text encoder, l_T means the number of tokens a review text contains, and $T^* = \{t_1, t_2, ..., t_{k-1}, t_{k+1}, ..., t_N\}$ represents the review text set.

For a review text set with N items, the loss function is adopted as follows: $L = \sum_{k=1}^{N} \log p(t_k|R^*)$ where $R^* = \{r_1, r_2, ..., r_{k-1}, r_{k+1}, ..., r_N\}$. The text decoder integrates the representation of $N - 1$ review texts to generate the summary. The integration process is carried out in the multi-head self-attention layer of the text decoder. In order not to left information unconcerned, we take the mean of all $N - 1$ single-head attention results for each encoded representation at each head (Elsahar et al., 2021).

Since we use the hypothetic summary, we utilize review ratings in review set as the extra features. We apply rating deviation (Im et al., 2021) to reduce the disparities between the training task and the generation task. Here is the definition of the rating deviation: $gt_k = \sum_{i \neq k}^{N} \frac{g_i}{N-1} - g_k$. During training, rating deviation operates normally while we let it be 0 when generating summaries in testing. Figure 4 shows how rating deviation is applied. We make some changes to how Transformer gets the input embeddings. Similar to positional embeddings, we add gt_k*deviation embeddings which has the same dimension with token embeddings to the token embeddings together with positional embeddings.

3.2.2 Table modality pretraining

The table in reviews consists of several pairs of field names and field values that are used to provide detailed information of the product or business. Previous work (Im et al., 2021) uses table data by simply encoding field names and values with BART and concentrating the representations as the output of table encoder, ignoring the structural knowledge in table. Different from previous work , we propose to construct graphs for tables. We convert the table to the graph structure by taking the field of product name as the center node of the graph, and connecting other fields to the center node with undirected edges, as is shown in Fig. 2. To obtain representations of the constructed graphs, we adopt graph attention network(GAT) (Velickovic et al., 2017) to encode the nodes. Finally, we get the overall table representation m_{table} by stacking every node representation into U. To fit the dimensionality and feature distribution with text modality representation acquired from text decoder, we add an additional linear layer:

$$u_i = BART_{encoder}(n_i; v_i),$$

$$m^i_{table} = GAT(u_i),$$

$$m_{table} = U \cdot W_{table},$$

where n and v are d_{Table}-dimensional representations that respectively represents the field name and the field value, u_i is d_{Table}-dimensional representation that represents the i-th node, l_{Table} is the number of fields that one product has, $U \in R^{l_{Table}*d_{Table}}$ denotes the stacked representations of all nodes, W_{Table} is the weight matrix of the additional linear layer, $m_{table} \in R^{l_{Table}*d_T}$

For using text modality as the main modality, we thus pretrain the GAT-based table encoder by taking the pretrained text decoder as a pivot to obtain the table representation m_{table}, and then use the text decoder to generate a summary from it. We simply create N single reference pairs from each product or business and shuffle pairs to construct the training dataset (Zheng et al., 2018).

3.2.3 MKG modality pretraining

As is well-known, GAT (Velickovic et al., 2017) solves the problems existing in GCN and achieves the state-of-art performance in many tasks such as node classification and link prediction. However, common GAT leaves the node modality aside, causing loss of significant information. To fully utilize cross-modal information, we propose to modify GAT to encode constructed MKGs for pretraining. For preparation, we use the pretrained BART text encoder to encode the text modality data in MKGs. For image modality data, we apply ResNet101 (He et al., 2016) pretrained in ImageNet and add an additional linear layer to obtain image representations. Following are how we extract information from the edges which represent relations in MKGs:

$$p_{ij} = LeakyReLU(W_2(W_1 c_{ij} + b_1) + b_2),$$

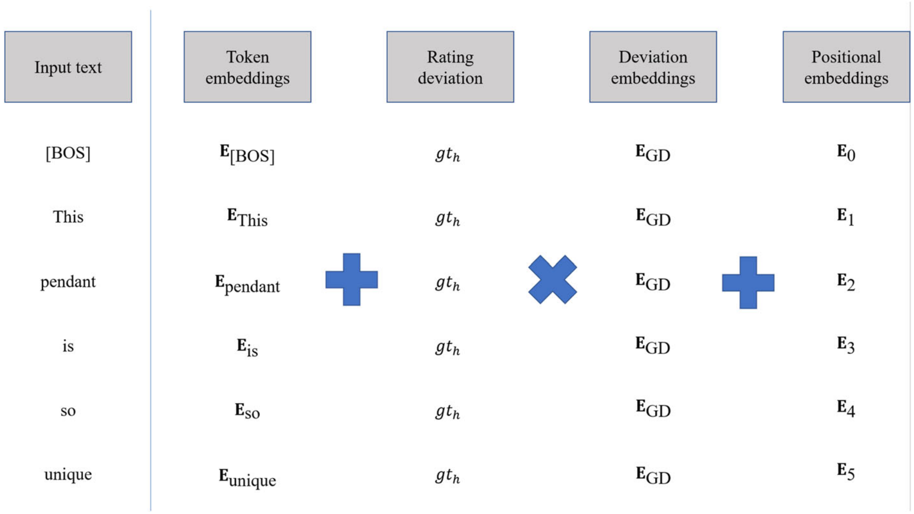


Fig. 4 Illustration of how rating deviation works

$$a_{ij} = \frac{exp(p_{ij})}{\sum_{j \in N_i} exp(p_{ij})},$$

$$h^l_{c_i} = \sigma(\sum_{j \in N_i} a_{ij} W_c h^{l-1}_j),$$

where c_{ij} denotes the d_T-dimensional relation embedding between node i and node j, N_i is the set of adjacent nodes of node i, $h^l_{c_i}$ represents the relational-attentional head in l-th layer,and W_1, W_2, W_c, b_1 and b_2 are trainable parameters.

The node-attentional head $h^l_{node_i}$ and the final representation of MKG is computed as follows:

$$e_{ij} = LeakyReLU(W_3 h_i || W_4 h_j),$$

$$z_{ij} = \frac{exp(e_{ij})}{\sum_{j \in N_i} exp(e_{ij})},$$

$$h^l_{node_i} = \sigma(\sum_{j \in N_i} z_{ij} W_{node} h^{l-1}_j),$$

$$h^l_i = ||_{i \in E^n} \sigma(W(h^l_{node_i} || h^l_{c_i}) + b),$$

$$H^l_i = FFN(h^l_i),$$

$$m_{mkg} = (||_{i \in E^n} H^l_i) W_{mkg},$$

where h_i and h_j are representations of node i and node j, $||$ is the concatenation operation; E^n is the entity set of MKG, W_3, W_4, W_{node}, W and b are learnable parameters, FFN denotes a Feed Forward Network layer, H^l_i is the node embedding, m_{mkg} is the aggregated representations of MKG, W_{mkg} is the weight matrix of the additional linear layer.

As with the text encoder and the table encoder, we pretrain the GAT-based MKG by obtaining MKG representations with the encoder and then feeding the representations to the text decoder to generate the summary.

3.2.4 Multiple modalities training

After the pretraining for each modality we get three pretrained encoders for text, tables, and MKG respectively. Then we train the whole model in this step. We obtain the representation of each modality m_{text}, m_{table}, m_{mkg} from encoders, and generate the hypothetic summary t_k based on them from the text decoder. Like using text decoder as a pivot, we also take text modality as the main modality in our model, which means that our model should work without table and MKG modality. To achieve that goal, we modify the multi-head self-attention layer of BART model with a multi-modality fusion method. Each layer would provide us with the attention result of each modality, and we fuse these attention results as follows:

$$fa_{fuse} = fa_{text} + \alpha \odot fa_{table} + \beta \odot fa_{mkg},$$

where fa_{text}, fa_{table}, fa_{mkg} represents the attention result of each modality, $\odot$ is the mathematical operator that means element-wise multiplication; α and β are d_T-dimensional multimodal gates that is calculated by $\alpha = \phi([fa_{text}; fa_{table}]W_\alpha)$, $\beta = \phi([fa_{text}; fa_{mkg}]W_{beta})$ where ϕ is activation function. It is obvious that α and β should be zero vectors when table and MKG modalities do not exist. Thus, we use ReLU function as ϕ instead of common use of sigmoid function. The values of α and β are initialized at approximately 0.5.

4 Experimental setup

4.1 Datasets

Our experiments are mainly conducted on two review datasets: Yelp Dataset Challenge and Amazon product reviews (He and McAuley, 2016). The data statistics are shown in Table 1. The Yelp dataset focuses more on reviews that reflect personal preference of a specific product or business. Also, the Yelp dataset provides various images and meta data including several characteristics of businesses such as "good for kids" or "good for meal dessert". The Amazon dataset focuses more on providing objective reviews, and for one product it usually offers only one image. The Amazon dataset also provides the more limited metadata than the Yelp dataset.

4.2 Evaluation criteria

We adopt ROUGE-1, 2, L (Lin, 2004) and BERT-score (Zhang et al., 2019) to evaluate the quality of generated summaries. ROUGE is a common evaluation criterion in Machine Translation, automatic summarization and QA generation. ROUGE-1, 2, L actually splits the result generated by the model and the standard result by 1, 2, and L-gram to calculate the recall rate. BERT-score is an assessment criterion based on pretrained BERT context embedding which is for language generation task. BERT-score calculates the similarity of two sentences as the sum of the cosine similarity between their mark embeddings. In Machine Translation, BERT-score correlates more strongly on multiple common benchmarks with system-level and segmental levels of human judgement than existing criteria do.

4.3 Compared baselines

We compare our method with several baselines including extractive and abstractive methods. For extractive methods, Clustroid (Bražinskas et al., 2020) selects the review with the highest ROUGE score. Lead (Bražinskas et al., 2020) selects the leading sentences from review texts and concatenate them to construct a summary. Random (Bražinskas et al., 2020) randomly

Table 1 Data statistics of Yelp and Amazon dataset

Yelp	Train	Dev	Test
businesses	50113	100	100
reviews/business	8	8	8
summaries/business	1	1	1
max images	10	10	10
max fields	47	47	47
Amazon	Train	Dev	Test
products	60935	28	32
reviews/product	8	8	8
summaries/product	1	3	3
max images	1	1	1
max fields	5+128	5+128	5+128

selects a review text as the summary. LexRank (Erkan and Radev, 2004) constructs a similarity graph of sentences to select important sentences.

For abstractive methods, MeanSum (Chu and Liu, 2019) is an end-to-end unsupervised model that generates the summary with the mean of the representations of input reviews. DenoiseSum (Amplayo and Lapata, 2020) selects a review and creates the noisy version of it, then learns to denoise the text and generates the summary. Copycat (Bražinskas et al., 2020) is a self-supervised method that hierarchically extends the variational autoencoder model and its decoder has direct access to the text of input to include specifics in the summary. Self&Control (Elsahar et al., 2021) can be seen as an extension of Transformer architecture and makes use of some control codes to generate the more fluent and correlative summary. COOP (Iso et al., 2021) is a latent vector aggregation framework that considers convex combinations of the latent vectors of input reviews for summary generation. MultimodalSum (Im et al., 2021) takes the modality of image and table into consideration as well and proposes a multi-modal training pipeline to fully extract useful information.

4.4 Implementation

Our model was implemented in PyTorch (Paszke et al., 2019) and the Transformer comes from the Hugging Face (Wolf et al., 2020) library. We utilized BART-Large as our text encoder and decoder. We optimized our model by the Adam optimizer (Kingma and Ba, 2014) with a linear learning rate decay. The dimensions of d_T and d_{Table} were both set to 1024. The flattened image feature map size obtained from ResNet101 was set to 14*14. The size of FFN layer was set to 512. The hyperparameters of each modality pretraining are shown in Table 2. For Table pretraining, MKG pretraining and multimodal training, we set label smoothing as 0.1 and the maximum gradient norm as 1. The generated summaries were limited to at most 144 tokens and at least 56 tokens.

5 Experimental results

5.1 Comparison with state-of-the-art models

We compare MKGOpinSum with several state-of-the-art opinion summarization methods including extractive and abstractive methods on the Amazon and Yelp datasets to evaluate our model. Table 3 demonstrates the evaluation results of the methods. For Yelp dataset, it shows that MKGOpinSum performs the best in the token level and the segment level according to the scores of R-L, and F_{BERT}. We ascribe the higher R-2 score of Self&Control method in Yelp dataset to that it utilizing inferred control tokens which enrich reviews. For Amazon dataset, MKGOpinSum has the highest R-2, R-L and F_{BERT} scores. When compared to that of the

Table 2 The hyperparameters of training

Pipeline step	batch	epochs	warmup	lr
Text pretraining	16	5	0.5	5e-05
Table pretraining	32	20	1	1e-04
MKG pretraining	32	20	1	1e-04
Multimodal training	8	5	0.25	1e-05

Table 3 Experimental results of opinion summarization on Yelp and Amazon datasets

Model	Yelp				Amazon			
	R-1	R-2	R-L	F_{BERT}	R-1	R-2	R-L	F_{BERT}
MeanSum (Chu and Liu, 2019)	28.86	3.66	15.91	86.5	29.20	4.70	18.15	-
Clustroid (Bražinskas et al., 2020)	26.28	3.48	15.36	85.8	29.27	4.41	17.78	86.4
Lead (Bražinskas et al., 2020)	26.34	3.72	13.86	85.1	30.32	5.85	15.96	85.8
Random (Bražinskas et al., 2020)	23.04	2.44	13.44	85.1	28.93	4.58	16.76	86.0
LexRank (Erkan and Radev, 2004)	34.90	2.76	14.28	85.4	29.46	5.53	17.74	86.4
DenoiseSum (Amplayo and Lapata, 2020)	30.14	4.99	17.65	85.9	-	-	-	-
Copycat (Bražinskas et al., 2020)	29.47	5.26	18.09	87.4	31.97	5.81	20.16	87.7
Self & Control (Elsahar et al., 2021)	32.76	**8.65**	18.82	86.8	-	-	-	-
COOP (Iso et al., 2021)	**35.37**	7.35	19.94	-	**36.57**	7.23	21.24	-
MultimodalSum (Im et al., 2021)	33.00	6.63	19.84	87.7	34.19	7.05	20.81	87.9
MKGOpinSum(Ours)	35.11	7.11	**20.26**	**89.1**	36.25	**7.37**	**21.64**	**89.5**

The first block shows the results of extractive methods while the second shows the results of abstractive methods. The third one shows the results of multi-modal methods. The best results are bolden

Yelp dataset, the results on Amazon dataset have the better performance because of Amazon dataset providing the more objective and targeted reviews and pictures. It's worth noting that as a single-modal method, COOP shows the excellent performance on R-1 score in both datasets for its exquisite processing of latent vector aggregation. Although the encoding strategy of text modality data seems relatively simple when comparing to COOP, MKGOpinSum still get a similiar R-1 score, highlighting the effectiveness of MKG modality. MKGOpinSum outperforms the multi-modal opinion summarization method MultimodalSum because of structural information obtained from the table encoder and the MKG encoder. Thus, the conclusion is that MKGOpinSum outperforms most kinds of opinion summarization methods, and has its specific advantages when compared with the well designed text opinion summarization methods.

5.2 Ablation study

To investigate the effect of each part in MKGOpinSum, we carry out ablation study on the Amazon dataset.

Firstly, we respectively remove table modality and MKG modality to evaluate their effects. The results are shown in the upper part of Table 4. It shows that both table and MKG modalities can improve the performance of our model, comparing with the model using only

Table 4 Ablation study on Amazon dataset

Models	R-L
MKGOpinSum	21.64
w/o table modality	21.32
w/o mkg modality	20.98
w/o table & mkg modality	20.48
w/o table modality pretraining	20.37
w/o mkg modality pretraining	20.34
w/o text modality pretraining	20.29
w/o all modalities training	20.24

text modality. Concretely, the model with MKG modality performs better than the model with table modality. This is because that MKGs fully utilize the inner relations in review texts and the relations between text and images. The image encoder we use also extract some important details like the appearance of the product that table data does not offer. However, table data also provide comprehensive information which may not be mentioned in the reviews and images such as their brands, categories and prices, etc. All these narrow the gap between MKG and table modalities, and proves the effectiveness of the table modality.

Secondly, for the training pipeline, we compared the performance by removing text pretraining part, table pretraining part, MKG pretraining part, or multiple modalities training part to investigate the effect of each part. The results in Table 4 show that there are large degrees of decline on performances when removing each pretraining part. In detail, the lack of table pretraining and MKG pretraining leads to different degrees of performance decline, which shows the interference of heterogeneity of multi modalities. MKG pretraining performs better, indicating that MKG has the potential of better performance with sufficient pretraining. Since we take the text modality as the major modality, removing text modality pretraining cause the maximal performance decline. In general, we can conclude that the training pipeline helps a lot with the summarization model.

5.3 Evaluation of MKG

To further evaluate the effectiveness of our MKG, we conduct several comparative experiments. The results are shown in Table 5. The upper part focuses on the use of knowledge graphs and the multi-modal information in them. It can be seen that text KG is able to truly help the summary generation task with its inner relational information, and with image data being added the newly formed MKG promotes the better performance. The lower part offers

Table 5 Evaluation results of MKG

Models	R-1	R-2	R-L
MKGOpinSum	36.25	7.37	21.64
MKGOpinSum(w/ text KG)	34.96	7.01	21.17
MKGOpinSum(w/o MKG)	33.95	6.93	20.98
MultimodalSum(w/o table modality)	33.84	6.91	20.53
MKGOpinSum(w/o table modality)	34.61	7.07	21.32

the specific comparison results between MultimodalSum and MKGOpinSum. For the reason that we also adjust table modality, we respectively remove the table modality of these two models to compare the performance of their image modality and our MKG modality. Result shows that taking context and structural information into consideration by MKG performs better than simply extracting features from images. Thus, we can conclude that the above studies reveal the significant effect of our MKG.

5.4 Human evalutaion

We utilize the Best-Worst Scaling (BWS) (Louviere et al., 2015) which is widely used in the field of opinion summarization to evaluate the generated summaries. We invite 5 participants to evaluate 50 examples where one example contains the gold summary randomly selected from the Amazon test set, the summary generated by MultimodalSum, and the summary generated by MKGOpinSum. Participants are asked to compare summaries according to three criteria, Gramma (the summary should be grammatical and readable), Coherence (the structure of summary should be well organized) and Overall (judging the summary by participants' own opinions). For each example, the participants make their judgement and give the best summary with 3 points, the worst with 1 point and the middle one with 2 points. By averaging the points of all examples from all participants we obtain the final points for each method. Results of human evaluations in Table 6 show that as multi-modal methods, MKGOpinSum is more favored by participants than MultimodalSum, which agrees with the automatic evaluation results.

5.5 Case study

Table 7 shows the summary generated by MultimodalSum, the summary generated by MKGOpinSum, and the gold summary for us to compare the two multi-modal opinion summarization methods more intuitively. All the summaries are generated from the source review of steamer on the Amazon dataset with production id B00006IUVM. MultimodalSum misunderstoods the information related to time and generates the wrong expression "it is still going". Also, it generates the characteristic "color" that does not exist in the source review. Thanks to the constructed MKG shown in Fig. 1, MKGOpinSum is able to provide a more accurate summary. Contrary to the mistake of MultimodalSum, MKGOpinSum catches the reason of replacing the old steamer "broke", and it further recognizes the lid from the picture, thus generates the more detailed expression "but the lid is broken". What's more, it can be seen from Fig. 1 and the summary that MKGOpinSum makes full use of the information extracted by MKG. On the whole, the summary generated by MKGOpinSum is closer to the source review and has fewer mistakes, which highlights the effectiveness of our model, especially the MKG.

Table 6 Human evaluation results on Amazon dataset

Models	Gramma	Coherence	Overall
MutimodalSum	1.87	1.72	1.38
MKGOpinSum	1.96	2.03	2.05
Gold	2.25	2.26	2.57

Table 7 Summaries generated by different models

Gold	This is our second B/D steamer. The first served us well for years before the plexiglass bowl cracked and finally broke. It was impossible to find a replacement in any major chain stores. Lots of rice steamers but no small in size vegetable steamers. Thanks Jeff!
MultimodalSum	I love this steamer. I have had one for years and it is still going. I love the color and the size of the steamer. It’s a little smaller than I expected, but I still love it so much. It’s a replacement of my old one.
MKGOpinSum	This is my second steamer. I have had one but the lid is broken and I have been finding a replacement. This steamer is definitely a good design. The steamer has a perfect size. This is a nice and roomy steamer. I love the steamer very much and I would recommend it to all customers.

6 Conclusion

In this paper, we propose the multi-modal opinion summarization method MKGOpinSum by making use of multi-modal structural knowledge in the self-supervised way. There are two types of structural knowledge we used. The one is multi-modal knowledge graphs constructed from the input texts and images. The other is the graphs constructed from tables. We further propose a training pipeline that pretrains text, table and multi-modal knowledge graph to deal with the interference caused by heterogeneity of multi-modal data. Extensive experiments on Yelp Dataset Challenge and Amazon product reviews show that our method performs better than most baselines such as Self&Control and MultimodalSum on ROUGE score and BERT score. Ablation study and further researches also show effectiveness of each part of our model, especially the MKG modality. In the future, we expect to link MKG to external data such as Wikipedia to make use of knowledge beyond reviews, and to apply our method to other NLP tasks such as image captioning and text-image summarization.

Acknowledgements This research was sponsored by the National Natural Science Foundation of China (No.61806101), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 21KJB520017). Jingqiang Chen is the corresponding author

Author Contributions Lingyun Jin and Jingqiang Chen contributed equally to this work.

Funding This research was sponsored by the National Natural Science Foundation of China (No.61806101), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 21KJB520017).

Availability of data and material The authors declare that the all supporting data are available.

Declarations

Ethical Approval and Consent to participate Not Applicable.

Consent for publication The authors declare that they consent for publication.

Human and Animal Ethics Not Applicable.

Competing interests The authors declare that they have no confict of interest.

References

Amplayo, R.K., & Lapata, M. (2020). Unsupervised opinion summarization with noising and denoising. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. *Association for Computational Linguistics*, Online, pp 1934–1945. https://doi.org/10.18653/v1/2020.acl-main.175

Amplayo, R.K., Angelidis, S., & Lapata, M. (2021). Aspect-controllable opinion summarization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics, Online and Punta Cana, Dominican Republic*, pp 6578–6593. https://doi.org/10.18653/v1/2021.emnlp-main.528

Angelidis, S., & Lapata, M. (2018). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics, Brussels, Belgium*, pp 3675–3686. https://doi.org/10.18653/v1/D18-1403

Basu Roy Chowdhury, S., Zhao, C., & Chaturvedi, S. (2022). Unsupervised extractive opinion summarization using sparse coding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). *Association for Computational Linguistics, Dublin, Ireland*, pp 1209–1225. https://doi.org/10.18653/v1/2022.acl-long.86

Bražinskas, A., Lapata, M., & Titov, I. (2020). Unsupervised opinion summarization as copycat-review generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. *Association for Computational Linguistics*, Online, pp 5151–5169. https://doi.org/10.18653/v1/2020.acl-main.461

Chen, J., & Zhuge, H. (2018). Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics, Brussels, Belgium*, pp 4046–4056. https://doi.org/10.18653/v1/D18-1438

Chen, L., Li, Z., & Wang, Y., et al. (2020). Mmea: Entity alignment for multi-modal knowledge graph. In: Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, pp 134–147. https://doi.org/10.1007/978-3-030-55130-8_12

Chen, L., Li, Z., & Xu, T., et al. (2022). Multi-modal siamese network for entity alignment. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. *Association for Computing Machinery*, New York, NY, USA, KDD '22, pp 118–126. https://doi.org/10.1145/3534678.3539244

Chu, E., & Liu, P. (2019). Meansum: a neural model for unsupervised multi-document abstractive summarization. In: Chaudhuri K, Salakhutdinov R (Eds.) *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol 97. PMLR, Long Beach, California, USA, pp 1223–1232. https://doi.org/10.48550/arXiv.1810.05739

Elsahar, H., Coavoux, M., & Rozen, J., et al. (2021). Self-supervised and controlled multi-document opinion summarization. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. *Association for Computational Linguistics*, Online, pp 1646–1662. https://doi.org/10.18653/v1/2021.eacl-main.141

Erkan, G., & Radev, D.R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. textitJ Artif Int Res 22(1), 457–479. https://doi.org/10.48550/arXiv.1109.2128

He, K., Zhang, X., & Ren, S., et al. (2016). Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 770–778. https://doi.org/10.1109/CVPR.2016.90

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th International Conference on World Wide Web. *International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '16,* pp 507–517. https://doi.org/10.1145/2872427.2883037

Honnibal, M., Montani, I., & Landeghem, S.V., et al. (2020). spacy: Industrial-strength natural language processing in python. 1. https://doi.org/10.5281/zenodo.1212303

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *Association for Computing Machinery*, New York, NY, USA, KDD '04, pp 168–177. https://doi.org/10.1145/1014052.1014073

Im, J., Kim, M., & Lee, H., et al. (2021). Self-supervised multimodal opinion summarization. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint

Conference on Natural Language Processing (Volume 1: Long Papers).*Association for Computational Linguistics*, Online, pp 388–403. https://doi.org/10.18653/v1/2021.acl-long.33

Iso, H., Wang, X., & Suhara, Y., et al. (2021). Convex Aggregation for Opinion Summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2021. *Association for Computational Linguistics, Punta Cana, Dominican Republic*, pp 3885–3903. https://doi.org/10.18653/v1/2021.findings-emnlp.328

Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv e-prints, arXiv:1412.6980, https://doi.org/10.48550/arXiv.1412.6980

Ku, L.W., Liang, Y.T., & Chen, H.H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI, pp 100–107. https://cdn.aaai.org/Symposia/Spring/2006/SS-06-03/SS06-03-020.pdf

Lewis, M., Liu, Y., & Goyal, N., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. *Association for Computational Linguistics*, Online, pp 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, H., Zhu, J., & Liu, T., et al. (2018). Multi-modal sentence summarization with modality attention and image filtering. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. *AAAI Press, Louisiana, USA, IJCAI'18*, pp 4152–4158. https://dl.acm.org/doi/abs/10.5555/3304222.3304347

Li, Q., Guo, S., & Luo, Y., et al. (2023). Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In: Proceedings of the ACM Web Conference 2023. *Association for Computing Machinery, New York, NY, USA, WWW '23*, pp 2499–2508. https://doi.org/10.1145/3543507.3583328

Liang, Y., Meng, F., & Xu, J., et al. (2023). Summary-oriented vision modeling for multimodal abstractive summarization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). *Association for Computational Linguistics, Toronto, Canada* pp 2934–2951. https://doi.org/10.48550/arXiv:2212.07672

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. *Association for Computational Linguistics*, Barcelona, Spain, pp 74–81. https://aclanthology.org/W04-1013

Louviere, J., Flynn, T., & Marley, A. A. J. (2015). *Best-Worst Scaling: Theory*. Methods and Applications: Cambridge University Press. https://doi.org/10.1017/CBO9781107337855

Ma, Y., Wang, Z., Li, & M., et al. (2022). MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. *Association for Computational Linguistics*, Dublin, Ireland, pp 231–239. https://doi.org/10.18653/v1/2022.acl-demo.23

Manning, C., Surdeanu, M., & Bauer, J., et al. (2014). The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. *Association for Computational Linguistics*, Baltimore, Maryland, pp 55–60. https://doi.org/10.3115/v1/P14-5010

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Paszke, A., Gross, S., & Massa, F., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library, Curran Associates Inc., Red Hook, NY, USA, chap 1, pp 8026–8037. https://doi.org/10.48550/arXiv.1912.01703

Paul, M., Zhai, C., & Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics,* Cambridge, MA, pp 66–76. https://aclanthology.org/D10-1007

Pezeshkpour, P., Chen, L., Singh S (2018) Embedding multimodal relational data for knowledge base completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics*, Brussels, Belgium, pp 3208–3218. https://doi.org/10.18653/v1/D18-1359

Radford, A., Kim, J.W., & Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In: Meila, M., & Zhang, T. (Eds.) *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol 139. PMLR, Online, pp 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv e-prints https://doi.org/10.48550/arXiv.1804.02767,

Sacenti, J. A. P., Fileto, R., & Willrich, R. (2022). Knowledge graph summarization impacts on movie recommendations. *J Intell Inf Syst, 58*(1), 43–66. https://doi.org/10.1007/s10844-021-00650-z

Sun, R., Cao, X., & Zhao, Y., et al. (2020). Multi-modal knowledge graphs for recommender systems. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management.

Association for Computing Machinery, New York, NY, USA, CIKM '20, pp 1405–1414. https://doi.org/10.1145/3340531.3411947

Vaswani, A., Shazeer, N., & Parmar, N., et al. (2017). Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, pp 6000–6010. https://doi.org/10.48550/arXiv.1706.03762

Velickovic, P., Cucurull, G., & Casanova, A., et al. (2017). Graph Attention Networks. arXiv e-prints https://doi.org/10.48550/arXiv.1710.10903

Wilcke, W.X., Bloem, P., & de Boer, V., et al. (2020). End-to-End Entity Classification on Multimodal Knowledge Graphs. arXiv e-prints arXiv:2003.12383, https://doi.org/10.48550/arXiv.2003.12383

Wolf, T., Debut, L., & Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. *Association for Computational Linguistics*, Online, pp 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xiao, M., Zhu, J., & Lin, H., et al. (2023). CFSum coarse-to-fine contribution network for multimodal summarization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). *Association for Computational Linguistics,* Toronto, Canada, pp 8538–8553. https://doi.org/10.18653/v1/2023.acl-long.476

Xie, F., Chen, J., & Chen, K. (2022). Extractive text-image summarization with relation-enhanced graph attention network. *Journal of Intelligent Information Systems* pp 1–17. https://doi.org/10.21203/rs.3.rs-1894502/v1

Zhang, L., Zhang, X., & Pan, J. (2022). Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 11676–11684. https://doi.org/10.1609/aaai.v36i10.21422

Zhang, M., Zhou, G., Huang, N., et al. (2023). Asu-osum: Aspect-augmented unsupervised opinion summarization. *Information Processing and Management, 60*(1), 103–138. https://doi.org/10.1016/j.ipm.2022.103138

Zhang, T., Kishore, V., & Wu, F., et al. (2019). Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1904.09675

Zhao, F., Li, C., & Wu, Z., et al. (2022). Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In: Proceedings of the 30th ACM International Conference on Multimedia. *Association for Computing Machinery*, New York, NY, USA, MM '22, pp 3983–3992. https://doi.org/10.1145/3503161.3548228

Zheng, R., Ma, M., & Huang, L. (2018). Multi-reference training with pseudo-references for neural translation and text generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics*, Brussels, Belgium, pp 3188–3197. https://doi.org/10.18653/v1/D18-1357

Zhu, J., Li, H., Liu, T., et al. (2018). MSMO: Multimodal summarization with multimodal output. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. *Association for Computational Linguistics,* Brussels, Belgium, pp 4154–4164. https://doi.org/10.18653/v1/D18-1448

Zhu, J., Zhou, Y., Zhang, J., et al. (2020). Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 9749–9756. https://doi.org/10.1609/aaai.v34i05.6525

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.