Extractive Text-Image Summarization with Relation-Enhanced Graph Attention Network

Feng Xie
Nanjing University of Posts and Telecommunications
JingQiang Chen (cjq@njupt.edu.cn)
Nanjing University of Posts and Telecommunications
Kejia Chen
Nanjing University of Posts and Telecommunications

Research Article

Keywords: Summarization, Extractive Summarization, Multi-modal Summarization, Graph Neural Networks

Posted Date: August 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1894502/v1

Extractive Text-Image Summarization with Relation-Enhanced Graph Attention Network*

Feng Xie[1†], Jingqiang Chen[1*†] and Kejia Chen[1]

[1*]Nanjing University of Posts and Telecommunications, NanJing, 210049, JiangSu, China.

*Corresponding author(s). E-mail(s): cjq@njupt.edu.cn;
Contributing authors: thexfonline@gmail.com;
chenkj@njupt.edu.cn;
†These authors contributed equally to this work.

Abstract

Multi-modal summarization with multi-modal output (MSMO) aims to generate multi-modal summaries for a multi-modal document to improve readability of summaries by making use of information of different modalities. Most existing Seq2Seq-based MSMO models cannot well capture multi-modal relations which are significant for generating high-quality multi-modal summaries. To address this issue, this paper proposes a relation-enhanced graph attention network for extractive text-image summarization (ReGAT-Summ) to capture inter-modal and intra-modal relations in the multi-modal document. Firstly, a multi-modal graph is constructed from the document. Then, node representations are calculated by proposed graph neural network. Finally, a sentence-image selector is trained to select salient sentences and images, which are further aligned by training. To our knowledge, we are the first to explore the graph-based model for MSMO. Experiments on two news datasets E-DailyMail and NYTime800k demonstrate that ReGAT-Summ achieves the state-of-the-art performance in terms of automatic metrics and human evaluations.

Keywords: Summarization, Extractive Summarization, Multi-modal Summarization, Graph Neural Networks

*This research was sponsored by the National Natural Science Foundation of China (No.61806101).

1 Introduction

Multi-modal summarization with multi-modal output (MSMO) can use data of different modalities to create more readable multi-modal summaries, which is different from the traditional text summarization that only handles plain text and outputs pure text summaries. Recently, with the development of deep learning in multi-modal tasks and the explosive growth of multi-media data, MSMO has attracted more and more researchers' attention. Most existing MSMO model [1, 2] are based on the advanced Seq2Seq models which were originally designed for machine translations [3]. These models summarize news documents with unaligned images to create *extractive* or *abstractive* summaries with aligned images and sentences.

However, traditional Seq2Seq-based MSMO methods cannot well capture long-distance multi-modal relations such as sentence-image relations, word-image relations, and word-sentence relations. These relations widely exist in multi-modal documents, and making use of these relations are significant for generating high-quality text-image summaries. Take the news in Fig.1 as an example. There are cross-modal semantic relations around the theme of "*violent video games*", which are marked by different colors. The cross-modal information can be incorporated into single-modal information as a supplement.

Intuitively, graphs can be used to model long-distance multi-modal relations for MSMO due to their ability to model relations between objects. As shown in the left part of Fig.1, the phrase "*Grand Theft Auto*" and "*Call Of Duty*" are instances of "*violent games*". The first and second images are semantically related to "*Grand Theft Auto*" and "*Call Of Duty*" respectively. The first sentence and last sentence of the document semantically match with the first image and its caption. These relation between words, sentences and images are important for summarization but have not been well utilized in previous work. And in the right part of Fig.1, green, blue and orange boxes represent sentence, word and image nodes respectively. **S1** consists of the word "*violent*" and "*game*" while **Img1** contains the word "*violent*", "*game*" and "*theft*" since the caption of **Img1** consists of these words. As a relay node, the relation of image-image, sentence-sentence, and sentence-image can be built through the common word nodes. For example, sentence **Img1** and **Img2** share the same word "*violent*" and "*game*", which connects them across sentence.

Currently, graph-based models are mainly used in pure text summarization and achieve considerable performance, such as the early works of TextRank [4] and LexRank [5], and the recent summarization models based on Graph Neural Network (GNN). For the MSMO task, relations among different modalities are more complicated than the cross-sentence relations in pure text summarization but are not well exploited yet. This paper proposes a graph-based extractive text-image summarization model. Firstly, an unified multi-modal graph is constructed and initialized, which contains three types of nodes, i.e. sentence nodes, word nodes and image nodes, and two types of relational edges, i.e. word-sentence edges and word-image edges. Secondly,

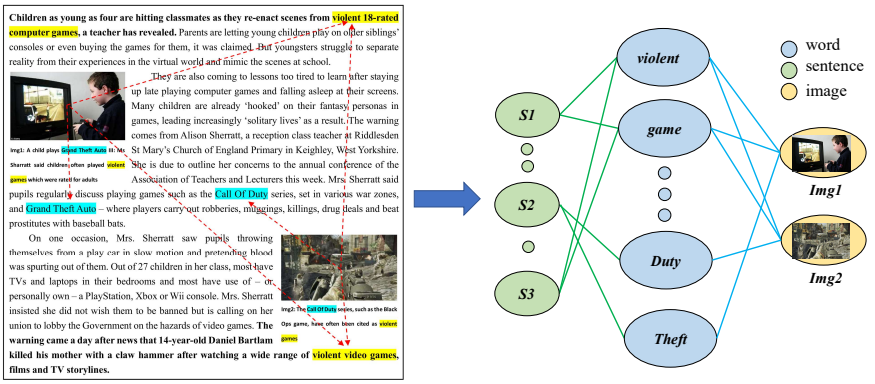

Fig. 1 Example of conversion process from a multi-modal news document to a multi-modal graph structure.

a relation-enhanced graph attention network (ReGAT) is proposed by introducing relation-attentional heads and node-attentional heads into GAT [6] to calculate node representations. Relation-attentional heads collect information from adjacent relational edges, and node-attentional heads collect informaiton from adjacent nodes. Thirdly, a multi-task selector is trained with node representations as input to select salient sentences and images, which are then aligned by training with a contrastive loss. The contributions of our work are summarized as follows:

- To our best knowledge, it is the first attempt to exploit graph-based models to capture various semantic relations between multi-modal semantic units for MSMO. And our proposed model is flexible and can be extended to other modalities (e.g. videos) for other multi-modal tasks.
- A relation-enhanced graph attention network is proposed for text-image summarization to better utilize multi-modal relations to fill semantic gaps between different modalities.
- Experimental results on two datesets E-DailyMail and NYTime800k show that our model not only outperforms both traditional text summarization baselines and MSMO baselines in terms of ROUGE scores, but also achieves impressive performance in image selection and image-sentence alignment.

2 Related Works

2.1 Extractive Text Summarization

In recent years, text summarization has achieved great progress with the development of neural networks. There are two types of text summarization: abstractive summarization and extractive summarization. The former concentrates on generating a summary word-by-word after encoding the entire document [7, 8], while the latter directly select salient sentences from original documents[9, 10].

Recently, various models for extractive summarization are developed. The reinforcement learning framework is introduced to optimize the evaluation metric with the rewards from policy gradient for text summarization [11]. The pre-trained language models are employed to improve text summarization due to their robust text representation ability [12] . The GNN-based summarization models [13] achieve competitive performance on benchmark datasets via building graphs consisting of different semantic units from documents. In this paper, we focus on extractive multi-modal summarization.

2.2 Multi-Modal Summarization

Different from pure text summarization, multi-modal summarization is a task to utilize information of different modalities to enhance the quality of summaries. According to whether the output summaries contain one or more modalities of input data, multi-modal summarization can be categorized into single-modal output [14] and multi-modal output[2]. The latter is more complicated and there are only limited studies. Chen et al. [1] and Zhu et al. [2] propose multi-modal encoders and a multi-modal attentional hierarchical decoder to capture cross-modal relations for jointly generating a textual summary and selecting the most relevant images from a collection of images in the input multi-modal document. [15] introduce a multi-modal objective function to effectively train their model by optimizing text summary generation and image selection. Following their work, Li et al. [16] propose the VMSMO model to select a frame as the video cover of news and meanwhile generate a textual summary of the article by multi-modal dual-interaction mechanism. Despite their success, how to better capture multi-modal relations remains an open problem. This paper constructs a multi-modal graph to address this issue.

2.3 Graph Neural Networks for NLP

Recently, GNN and its variants like gated graph neural network [17], graph convolutional network [18] and graph attention network [6] are effectively applied in many NLP tasks such as text generation [19], text representation [20] and text classification [21]. In the text summarization area, GNNs are also effectively used to summarize pure text documents [13]. since they can model various relations between sentences or words. For multi-modal documents, there are more complicated relations among different modalities, which

can also be modeled by GNNs. Hence, we extend the graph attention network (GAT) with relation-enhanced mechanism to fully exploit these relations for the MSMO task.

3 Problem Formulation

Let $\mathcal{D}$ denote the source document consisting of a sequence of sentences $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ and a collection of image-caption pairs $\mathcal{P} = \{(p_1, c_1), (p_2, c_2), ..., (p_m, c_m)\}$, where s_i is the i-th sentence of the input document and (p_j, c_j) is the j-th image-caption pair. Let $\mathcal{T}$ denote the ground-truth textual summary. Extractive MSMO is defined to predict two sequences of labels $\{y_1, y_2, ..., y_n\}$ and $\{z_1, z_2, ..., z_m\}$ $(y_i, z_j \in \{0, 1\})$ for sentences and images respectively, where $y_i = 1$ indicates the sentence s_j should be considered as a summary sentence, and $z_j = 1$ indicates that the image p_j should be considered as a summary image. Finally, each summary sentence is aligned with the most relevant summary image in the output summary. We employ ORACLE [7] to iteratively extract sentences as the ground-truth summary that obtains the highest ROUGE score calculated by $\mathcal{S}$ and $\mathcal{T}$. Similarly, we label images by calculating the ROUGE score between the corresponding captions and $\mathcal{T}$, and regard the original image-caption pairs in the document as the ground truth of multi-modal alignment.

4 The Proposed Model

This section introduces the proposed relation-enhanced graph attention network for text-image summarization (ReGAT-Summ) consisting of three modules (Figure 2).

- *Graph construction and Initialization.* It builds a multi-modal graph and initializes node representations with a word encoder, a sentence encoder and an image encoder.
- *Relation-Enhanced Graph Attention Layer.* It updates node representations by iteratively aggregating information from adjacent nodes through different types of relational edges, with relation-attentional heads and node-attentional heads to control multi-modal information flow.
- *Multi-Modal Selection and Alignment.* It uses fused representations of sentence nodes and image nodes in the joint embedding space as features to train a multi-modal selector, which can select salient sentences and images to form the output summary. And, each selected sentence is aligned to its most relevant image.

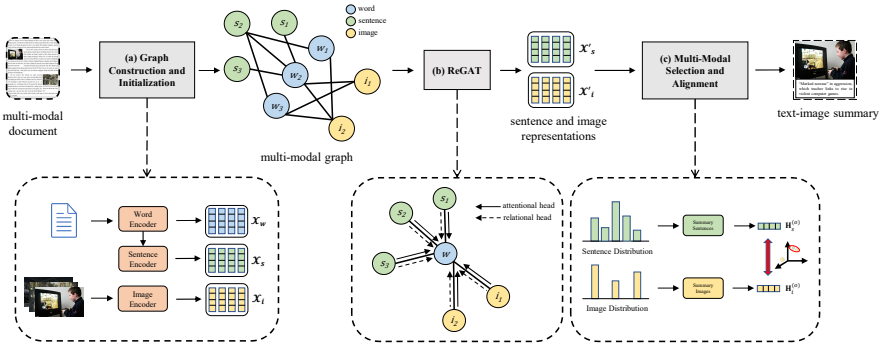

Fig. 2 Overview of the ReGAT-Summ model. It can be divided into three modules: (a) *Graph Construction and Initialization*, where a multi-modal graph is constructed and initialized through encoding nodes; (b) *Relation-Enhanced Graph Attention Layer*, which iteratively aggregates information from different modalities to learn fused representations; (c) *Multi-Modal Selection and Alignment*, which selects salient sentences and images and then aligns them to form a text-image summary output.

4.1 Graph construction and Initialization

4.1.1 Graph Construction

This multi-modal graph contain three types of nodes i.e. image nodes, sentence nodes and word nodes, and two types of edges i.e. sentence-word edges and image-word edges. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected multi-modal graph, where $\mathcal{V}$ represents a node set and $\mathcal{E}$ stands for edges between nodes. $\mathcal{V}$ and $\mathcal{E}$ are defined as follows:

- $\mathcal{V} = \mathcal{V}^w \cup \mathcal{V}^s \cup \mathcal{V}^p$, where $\mathcal{V}^w = \{w_1, ..., w_n\}$ denotes n unique words in the whole document, $\mathcal{V}^s = \{s_1, ..., s_m\}$ represents the m sentences in the article, and $\mathcal{V}^p = \{p_1, ..., p_t\}$ corresponds to the t images (pictures) in the document.
- $\mathcal{E} = \mathcal{E}^{wp} \cup \mathcal{E}^{ws}$, where $\mathcal{E}^{wp} \in \mathbb{R}^{n \times t}$ is a bi-value matrix of the word-image subgraph and $\mathcal{E}^{ws} \in \mathbb{R}^{n \times m}$ is a TF-IDF valued matrix of the word-sentence subgraph, where $e^{wp}_{ij} = 1$ indicates that the caption of the j-th image contains the i-th word, and $e^{ws}_{qt} \neq 0$ represents that the t-th sentence of the article contains the q-th word.

4.1.2 Node Embedding Initialization

In order to encode the words, we use GloVe [22] to obtain the word embedding matrix for the news texts including captions. Then we follow the method of [13] to encode sentences by using Bi-LSTM and CNN. Due to the limited computational resource, we do not use pre-trained contextualized encoders (i.e. BERT [23]), and we regard it as our future work. As for image nodes, we apply ResNet-152 [24] to extract 2048-dimensional global feature vectors for all image nodes. Formally, let $\mathcal{X}^w \in \mathbb{R}^{n \times d_w}$, $\mathcal{X}^s \in \mathbb{R}^{m \times d_s}$ and $\mathcal{X}^p \in \mathbb{R}^{t \times d_p}$ represent embedding matrices of word nodes, sentence nodes and image nodes respectively.

4.1.3 Edge Embedding Initialization

In order to exploit relational information between different semantic units, we map the two types of edges into two multi-dimensional embedding spaces. For word-sentence edges, we use the method of [13], to map each corresponding TF-IDF value into the relation embedding space to get $\mathbf{r}_{ij}^{ws}$, which represents the relation embedding between the word node i and the sentence node j. For word-image edges, since they are built from image captions contain corresponding words, we directly use the caption embeddings as the edge embeddings. The captions are encoded using the sentence encoder mentioned above to get vector representation $\mathbf{r}_{qt}^{wp}$, which denotes the embedding of the relational edge between the word node q and the image node t.

4.2 Relation-Enhanced Graph Attention Layer

The self-attention mechanism in GAT [6] computes the attention coefficient for each node, which allows every node to attend on its neighborhood with different attention weights. However, this aggregation fails to take the node modality into consideration, thus may lose important cross-modal relational information. In the multi-modal graph, there are two modalities of adjacent nodes (image nodes and sentence nodes) and two types of relational edges for each intermediate word node.

To make use of the above information, we propose ReGAT by introducing the relation-attentional head to collect information from adjacent edges, and the node-attentional head to collect information from adjacent nodes.

4.2.1 Relation-Attentional Head

Eq. (1) to (5) compute the relation-attentional head $\mathbf{h}_{rel_i}^{(l)}$ in l^{th} layer for the node i:

$$u_{ij} = \text{LeakyReLU}(\mathbf{W}_2(\mathbf{W}_1\mathbf{r}_{ij} + \mathbf{b}_1) + \mathbf{b}_2) \tag{1}$$

$$\alpha_{ij} = \frac{\exp(u_{ij})}{\sum_{j\in\mathcal{N}_i^s}\exp(u_{ij}) + \sum_{k\in\mathcal{N}_i^p}\exp(u_{ik})} \tag{2}$$

$$\mathbf{h}_{rel_i}^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}_j^s}\alpha_{ij}\mathbf{W}_{rel}^{s}\mathbf{h}_j^{(l-1)} + \sum_{k\in\mathcal{N}_j^p}\alpha_{ik}\mathbf{W}_{rel}^{p}\mathbf{h}_k^{(l-1)}\right) \tag{3}$$

In Eq. (1), $\mathbf{r}_{ij} \in \mathbb{R}^d$ is training parameters, which represents the relation-specific embedding between the node i and the sentence node j, and d is the embedding size. In Eq. (2), $\mathcal{N}_i^s$ and $\mathcal{N}_i^p$ are adjacent sentence nodes and image nodes of the word node i respectively. $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_{rel}^s, \mathbf{W}_{rel}^p$ and $\mathbf{b}_1, \mathbf{b}_2$ are trainable parameters.

4.2.2 Node-Attentional Head

Eq. (4) and (5) compute the node-attentional head $\mathbf{h}_{nod_i}^{(l)}$ for the node i. In Eq. (5), $\mathbf{h}_j$ and $\mathbf{h}_k$ are the representations of the adjacent nodes j and k of

the node i, and β_{ij} is computed using e_{ij} as with α_{ij} in Eq. (2).

$$e_{ij} = \text{LeakyReLU}(\mathbf{a}^T \cdot [\mathbf{W}_w \mathbf{h}_i \parallel \mathbf{W}_s \mathbf{h}_j]) \tag{4}$$

$$\mathbf{h}_{nod_i}^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}_i^s} \beta_{ij} \mathbf{W}_{nod}^s \mathbf{h}_j^{(l-1)} + \sum_{k \in \mathcal{N}_i^p} \beta_{ik} \mathbf{W}_{nod}^p \mathbf{h}_k^{(l-1)}\right) \tag{5}$$

Then the multi-head concatenation is used for the combination of the two heads, denoted as:

$$\mathbf{h}_i^{(l)} = \Big\|_{m=1}^{M} \sigma\Big(\mathbf{W}(\mathbf{h}_{nod_i}^{(l)} \parallel \mathbf{h}_{rel_i}^{(l)}) + \mathbf{b}\Big) \tag{6}$$

where $\parallel$ represents the concatenation operation.

Finally, a layer of Feed Forward Network (FFN) is used to obtain the embedding of the node i:

$$\mathbf{H}_i^{(l)} = \text{FFN}(\mathbf{h}_i^{(l)}) \tag{7}$$

4.3 Multi-Modal Selection and Alignment

4.3.1 Multi-Task Sentence-Image Selector

In order to jointly select salient sentences and images to form the multi-modal summary, a multi-task sentence-image selector is trained using node embeddings computed by ReGAT as input. The binary cross-entropy objective function is defined as follows:

$$\mathcal{L}_{sent} = \sum_{i=1}^{n} (y_i \log(P_{sent_i}) + (1 - y_i) \log(1 - P_{sent_i})) \tag{8}$$

$$\mathcal{L}_{img} = \sum_{j=1}^{m} (z_j \log(P_{img_j}) + (1 - z_j) \log(1 - P_{img_j})) \tag{9}$$

$$P_{sent_i}, P_{img_j} \sim Softmax(FC(\mathbf{H}^{(L)})) \tag{10}$$

where P_{sent} and P_{img} are extractive probabilities of sentence and image respectively calculated by Eq. (13), where FC is the full connection operation and L is the last layer of ReGAT. We carry out binary classification on all sentence nodes and all image nodes, and obtain $\mathcal{S}^s = \{s_1, s_2, ..., s_N\}$ and $\mathcal{S}^p = \{p_1, p_2, ..., p_M\}$ as outputs.

4.3.2 Contrastive Sentence-Image Alignment

The selected images should semantically matche the selected sentences in the multi-modal summary. To guarantee that similar sentences and images are

close in the embedding space, a triplet contrastive loss function, which is commonly used to measure the sentence-image relevance, formulated as:

$$\mathcal{L}_c = \sum_{\hat{p}} \max(0, \delta - \mathrm{c}(p, s) + \mathrm{c}(\hat{p}, s)) \tag{11}$$

In Eq. (11), δ represents the margin, s and p denote the positive sentence-image pair, and $\hat{s}$ and $\hat{p}$ correspond to the negative pair. Denote $\mathbf{H}^{(o)}$ as the node embedding in the output layer. The similarity measure is defined as $c(p, s) = \cos\langle \mathbf{H}_p^{(o)}, \mathbf{H}_s^{(o)} \rangle$. Faghri et al. [25] discovered that using the hardest negative in a mini-batch during training rather than all negatives samples can boost performance. Therefore, we follow that in this study and define the loss function as:

$$\mathcal{L}_c^+ = \max(0, \delta - \mathrm{c}(p, s) + \mathrm{c}(p', s)) \tag{12}$$

where $p' = \arg\max_{j \neq p} \mathrm{c}(j, s)$ is the hardest negatives in the mini-batch.

We create a positive image-sentence pair by selecting the summary sentence with the highest ROUGE score referring to the caption of the image. Negative pairs are created by randomly selecting a sentence for a image. The sentence-image alignment task can be seen as an image retrieval task, which consider sentences in the $\mathcal{S}^s$ as queries and rank the images set $\mathcal{S}^p$ with respect to each query according to the scoring function. For $s_i \in \mathcal{S}^s$, we align it with p^* denoted as:

$$p^* = \underset{p_j \in \mathcal{S}^p}{\arg\max} \quad \cos\langle \mathbf{H}_{s_i}^{(o)}, \mathbf{H}_{p_j}^{(o)} \rangle \tag{13}$$

4.3.3 Final Loss

The final loss of our model is the linear combination of these three parts:

$$\mathcal{L} = \mathcal{L}_{sent} + \mathcal{L}_{img} + \lambda \mathcal{L}_c^+ \tag{14}$$

where λ is the hyperparameter.

5 Experiments

5.1 Datasets

We employ two datasets E-DailyMail [1] and NYTimes800k [26] both of which contain news articles and images, and each image is paired with a caption. The statistics of these two datasets is shown in Table 1.

- **E-DailyMail** is an extended version of the standard DailyMail dataset for single-document summarization, which is constructed by collecting images from the DailyMail website for each document in original DailyMail corpora. The dataset is split into 187,921/11,410/9,821 for training, validation, and

testing. Each sample contains a piece of news article, at least one image-caption pair and a multi-sentence summary.

- **NYTimes800k** is a long document dataset initially constructed for the image captioning task, which contains articles and images with captions from The New York Times spanning 14 years. In order to adapt this dataset to the MSMO task, we select the samples containing a news article, at least one image-caption pair and a summary. Following Tran et al. [26], we split the dataset into 156,988/3,052/8,495 for training, validation and testing.

Table 1 Statistics of the two datasets.

	E-DailyMail	**NYTimes800k**
NumDocs	209,152	168,535
AvgDocsLen	26.4	46.1
AvgSumLen	3.8	1.8
AvgImgCaps	5.4	3.1
AvgSentTokens	25.2	20.9
AvgCapTokens	24.7	18.3

NumDocs denotes the number of documents. AvgDocsLen and AvgSumLen denote the average number of sentences in a article and in a summary respectively. AvgImgCaps denotes the number of image-caption pairs. AvgSentTokens and AvgCapTokens denote the average number of tokens in a sentence and in a caption respectively.

5.2 Models for Comparison

We compare ReGAT-Summ with 10 text summarization baselines and 3 multimodal summarization baselines. And we add all image captions to the dataset for training and testing:

- **LEAD** selects the first several sentences of article as the text summary [10].
- **ORACLE** achieves the approximate maximum ROUGE scores with human reference summary, using the extractive summary which results from greedily selection [12].
- **ABS** is a classic abstractive summarizaion method besed on the encoder-decoder architecture with an attention mechanism [27].
- **PGC** is a Seq2Seq attentional model for abstractive summarization with the pointer network and a coverage mechanism [8].
- **SummaRuNNer** is an extractive summarization model by defining a sentence classfication model taking as features the content salience, the sentence novelty, and the position of each sentence to select salient sentences. [10].
- **NeuSum** integrates the selection strategy into the scoring model and jointly learning to score and select sentences for extractive summarization [28].

- **GPG** is proposed by [29] to generate a text summary by "editing" pointed tokens instead of hard copying.
- **JECS** is an extractive summarization method that selects sentences and compresses them by pruning a dependency tree to reduce redundancy [30].
- **BERTSUM** inserts multiple segmentation tokens into documents to represent each sentence. It is the first BERT-based extractive summarization model [12].
- **HETERSUMGRAPH** is an extractive model proposed by [13] to model relations between sentences based on their common words, which select salient sentences to form an extractive summary through node classification.
- **HAMS** is an abstractive text-image summarization model using the attentional hierarchical Seq2Seq framework to summarize a textual summary and its accompanying images [1].
- **MSMO** is a multi-modal attention model to jointly generate text and select the most relevant image by multi-modal coverage mechanisms [2].
- **MOF** extends MSMO by introducing a multi-modal objective function to incorporate the multi-modal reference, which adds image accuracy as another loss [15].

5.3 Evaluation Metrics

Since our model outputs multi-modal summaries containing sentences and images, it needs to be evaluated from three aspects, i.e. selected sentences, selected images and sentence-image alignments. The quality of selected sentences is evaluated by ROUGE, which calculates the overlap lexical units of extracted sentences and the ground truth. We report the ROUGE-1, ROUGE-2, and ROUGE-L for all models. The quality of selected images is evaluated by precision, recall, and F1-score. The quality of sentence-image alignments is also evaluated by the ROUGE score calculated between the caption and the aligned sentence.

5.4 Implementation Details

We implement our model in Pytorch, and run on an NVIDIA RTX 2080Ti GPU for 10 epochs. We set the vocabulary to 50k the dimension of word embeddings to 300-dimensional in GloVe. The dimension of the hidden state of the BiLSTM is 128, and the number of layers is 2. The input images have been cropped and resized to 224×224 before encoding. The dimension of edge embedding $\mathbf{r}^{ws}$ and $\mathbf{r}^{wi}$ is all set to 128. The number of ReGAT layers is set to 2, and each GAT layer has 8 heads, The hidden size $d_h = 128$, and the size of FFN is 512. For training, we use the batch size of 16 and employ the Adam optimizer with a learning rate of 0.001. We also use gradient clipping with a range of $[-1, 1]$ and added a dropout of 0.1. Finally, we select top-3 sentences and top-2 images for E-DailyMail and top-2 sentences and images for NYTime800k according to the average length of their ground truth summaries and the average number of images in the document. The hyperparameter λ is set to 0.5.

Table 2 Evaluations of text summaries.

Models	E-DailyMail			NYTimes800k		
	R-1	R-2	R-L	R-1	R-2	R-L
LEAD	40.52	14.9	32.60	20.16	7.31	18.56
ORACLE	54.83	31.67	50.20	40.22	15.76	35.19
ABS	34.46	13.30	31.65	20.77	6.80	18.04
PGC	38.53	16.48	35.38	21.40	6.95	18.20
GPG	39.02	15.34	35.79	22.05	6.88	18.96
SummaRuNNer	42.05	16.96	34.15	22.05	6.98	18.31
NeuSUM	42.59	18.95	37.28	22.31	7.15	18.20
JECS	42.85	18.30	37.60	22.45	7.68	18.57
BERTSUM	**43.15**	19.23	39.60	25.94	8.94	19.89
HETERSUMGRAPH	42.65	19.07	39.22	25.07	8.78	19.33
HAMS	41.91	17.84	36.40	23.20	6.84	17.55
MSMO	40.76	18.13	37.41	22.92	6.70	18.85
MOF	41.02	18.35	38.70	23.15	7.04	19.20
ReGAT-Summ	43.09	**19.85**	**40.96**	**25.31**	**9.02**	**20.54**

5.5 Results and Analysis

5.5.1 Evaluations of Text Summaries

The experiment results in Table 2 shows the performance of different models on two multi-modal news datasets and examine effectiveness of our proposed ReGAT-Summ in terms of ROUGE. The first two lines are the Lead baseline and the ORACLE upper bound, the following eight lines are traditional text summarization baselines including extractive and abstractive, and the last four lines are multi-modal summarization methods. In addition to automatic evaluation, model performance was also evaluated by human judgments in Table 5. The results of our model are highlighted in boldface. From the results, we make the following observations:

- Our model almost outperforms all pure text summarization baselines, including HETERSUMGRAPH. The differences between our model and HETERSUMGRAPH are that our model considers image information and adds relation-attentional heads in GAT, which can improve text summarization as indicated by the results.
- Compared with three abstractive MSMO approaches including HAMS, our model also achieve considerable improvements. One reason for this is that ReGAT-Summ is an extractive approach which usually perform better than abstractive counterparts. The other reason is that the three baselines are all Seq2Seq-based models, and our model is a ReGAT-based model which can better make use of long-distance relations.
- The improvements of performance on E-DailyMail are lager than NYTime800K, because the number of image-caption pairs in a document on E-DailyMail is larger than that of NYTime800K as shown in Table 1.

Table 3 Evaluations of image summaries.

Models	E-DailyMail			NYTimes800k		
	P	**R**	**F1**	**P**	**R**	**F1**
Random	0.34	0.37	0.35	0.41	0.48	0.44
ReGAT-Summ	**0.58**	**0.79**	**0.68**	**0.65**	**0.74**	**0.69**

Table 4 Evaluations of sentence-image alignments.

Models	E-DailyMail			NYTimes800k		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Random	35.98	13.01	35.25	24.21	5.05	12.38
ReGAT-Summ	**39.85**	**18.73**	**36.40**	**28.40**	**6.68**	**15.35**

This is another proof of the influence of visual information for multi-modal summarization.

5.5.2 Evaluations of Image Summaries

As mentioned, we employ three metrics: precision, recall, and f1-score to measure image summaries comparing with the ground-truth image labels. Results in Table 3 show that our model significantly outperforms the RANDOM baseline which randomly select images. This indicates ReGAT-Summ is able to select salient images, at least better than random selection.

5.5.3 Evaluations of Sentence-Image Alignments

To evaluate similarity of each sentence-image pair in the output summaries, we regard ROUGE scores between the sentence in a sentence-image pair and the caption corresponding to the image as alignment scores. Table 4 shows the scores of our model and the RANDOM baseline which randomly aligns sentences and images in the output summaries. Our model significantly outperform the RANDOM baseline for sentence-image alignment, implying our model can achieve acceptable text-image alignment in the output summaries.

5.5.4 Human Evaluation

It is not enough only relying on the ROUGE evaluation for a summarization system, although the ROUGE correlates well with human judgments. In order to verify how robust summarization models are to hallucinations and evaluate the performance of ReGAT more accurately, we design an experiment based on ranking method. Following Cheng and Lapata [9], we randomly select 50 samples from E-DailyMail test set. Each sample is annotated by three different participants separately.

Table 5 Human evaluation on E-DailyMail

Models	1st	2nd	3rd	4th	5th	Avg
SummaRuNNer	0.12	0.27	0.25	0.23	0.13	2.97
BERTSUM	0.25	0.28	0.30	0.12	0.05	2.78
MOF	0.34	0.27	0.18	0.11	0.10	2.65
ReGAT-Summ	0.45	0.34	0.15	0.06	0.00	2.32
Ground-Truth	0.72	0.19	0.04	0.05	0.00	1.42

This evaluation estimated the overall quality of the textual summaries by asking participants to rank these summaries according to their informativeness (can the summary capture the important information from the document) and fluency (is the summary fluent and grammatical). The human participants are presented with a original document and a list of corresponding summaries produced by different models. Participants were presented with the ground truth summaries and the summaries generated from four baseline models (SummaRuNNer, BERTSUM, MOF, ReGAT-Summ). From the results shown in Table 5, we can see that participants overwhelmingly prefer our model.

5.5.5 Ablation Study

In order to investigate the effectiveness of different components, including *relation-attentional head* (Rel), *node-attentional head* (Nod) and *contrastive loss* (CL), and the importance of using images (Img), we conduct ablation study using on E-DailyMail dataset. According to the results in Table 6, each module is necessary and combining them can help our model achieve the best performance:

- w/o Rel: In this variant, the relation-attentional head is removed from our model. Apparently, the performance degradation reported in line 1 demonstrates that ReGAT can well capture relational information between different semantic nodes in the message propagation process, which is essential for MSMO.
- w/o Nod: In this variant, we remove the node-attentional head from the model. The result in line 2 also shows an insignificant performance drop comparing to line 1. It indicates that relation-attentional head is more important than node-attentional head because there is abundant relational information in multi-modal document, which build a bridge between different semantic units.
- w/o CL: It is the variant removing the contrastive loss. The results in line 3 show that the performance improvement caused by CL is considerably significant. The underlying reason is that CL constrains the similarity score of the matched image-text pairs larger than the similarity score of the unmatched ones by a margin.
- w/o Img: We replace image features with corresponding caption features in our model and conduct the experiments in this variant. The results in line 4

verified that, compared to plain text summarization, usage of multi-modal information can improve summarization.

Table 6 Ablation study on E-DailyMail

Models	R-1	R-2	R-L
ReGAT-Summ	**43.09**	**19.85**	**40.96**
w/o Rel	42.76	19.27	40.33
w/o Nod	42.82	19.75	40.80
w/o CL	42.64	19.23	40.22
w/o Img	42.71	19.24	40.15

5.5.6 Case Study

We show a case study in Table 7, which includes the input source article, the ORACLE summary and the text-image summary created by our model. The summaries created by our model have three sentences S1, S2, S3 and two images Img1 and Img2. S1 and S3 are aligned with Img1, and S2 is aligned with Img2 according to the alignment scores in the Table 8, which are calculated by cosine similarity between the embeddings of sentence and image. It is obvious that our model select salient sentences and salient images from the source multi-modal document, and the sentences are aligned with relevant images. And compared to HAMS, the text-image pairs aligned by our model have higher relevance, which implies that our model can contribute to inter-modality retrieval. This case study also reveals that our model is able to generate more accurate and readable multi-modal summaries.

Table 7 Case study on an example taken from the E-DailyMail test set.

Article(truncated): The North Sea may seem a surprising location to discover a woolly mammoth skeleton, but Dutch fossil hunters have hauled ancient bones from its depths. (...) Mr Broch said: "Most weeks we go to the fishing ports to meet the fishing vessels and buy the fossils they caught."
ORACLE summaries: The North Sea may seem a surprising location to discover a woolly mammoth skeleton, but Dutch fossil hunters have hauled ancient bones from its depths. During the Ice Age, when mammoth roamed the Earth, lots of water that now makes up seas and oceans, was locked up in glaciers and huge sheets of ice, so sea levels were lower than they are today. Mr.Broch said it is "extremely rare" to find mammoth skulls and large bones on the seabed.
ReGAT-Summ: **S1**: The North Sea may seem a surprising location to discover a woolly mammoth skeleton, but Dutch fossil hunters have hauled ancient bones from its depths. **S2**: During the Ice Age, when mammoth roamed the Earth, lots of water that now makes up seas and oceans, was locked up in glaciers and huge sheets of ice, so sea levels were lower than they are today. **S3**: Mr.Broch said it is "extremely rare" to find mammoth skulls and large bones on the seabed.   Img1: **S1** and **S3** Img2: **S2**
HAMS: (1): The skeleton is composed of mammoth bones found off the coast of Rotterdam. (2): There is a vast tundra on an ancient land called Doggerland between Britain and Europe. (3): It is extremely rare to find a complete mammoth skeleton on the seabed.   (1)(3) (2)

Table 8 The sentence–image alignment scores.

	S1	**S2**	**S3**
Img1	0.39	0.14	0.55
Img2	0.24	0.43	0.42

6 Conclusion

In this paper, we focus on improving multi-modal summarization with multi-modal output by proposing the relation- enhanced GAT to leverage multi-modal semantic units and relations in multi-modal documents. Relation-attentional heads and node-attentional heads are defined in ReGAT-Summ to make use of multi-modal information of relations and nodes. Node representations are calculated by aggregating information from adjacent relational edges using relation-attentional heads, and by aggreagating information from adjacent nodes using node-attentional heads. A multi-task text-image selector is trained to select salient sentences and images, and a sentence- image alignment model is trained with a contrastive loss. Experiments demonstrate that our model outperforms pure text summarization baselines and multi-modal summarization baselines, and also performs well on sentence-image alignment. The Ablation study also shows the effectiveness of each module. As an independent module, ReGAT is also expected to be applied in other NLP tasks such as text classification and text-image matching, and its effectiveness will be further explored.

Statements and 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

Availability of supporting data

The authors declare that the all supporting data are available..

Competing interests

The authors declare that they have no conflict of interest.

Funding

This research was sponsored by the National Natural Science Foundation of China (No.61806101).

Authors' contributions

Feng Xie and JingQing Chen contributed equally to this work.

Acknowledgments

The research was sponsored by the National Natural Science Foundation of China (No.61806101). We thank the anonymous reviewers for helpful comments. JingQiang Chen is the corresponding author.

References

[1] Chen, J., Zhuge, H.: Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In: Proc. of EMNLP (2018)

[2] Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., Zong, C.: Msmo: Multimodal summarization with multimodal output. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)

[3] Calixto, I., Liu, Q., Campbell, N.: Doubly-attentive decoder for multi-modal neural machine translation (2017)

[4] Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)

[5] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research (2004)

[6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks (2017)

[7] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al.: Abstractive text summarization using sequence-to-sequence rnns and beyond (2016)

[8] See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks (2017)

[9] Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words (2016)

[10] Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proc. of AAAI (2017)

[11] Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning (2018)

[12] Liu, Y., Lapata, M.: Text summarization with pretrained encoders (2019)

[13] Wang, D., Liu, P., Zheng, Y., Qiu, X., Huang, X.: Heterogeneous graph neural networks for extractive document summarization (2020)

[14] Li, H., Zhu, J., Liu, T., Zhang, J., Zong, C., *et al.*: Multi-modal sentence summarization with modality attention and image filtering. In: IJCAI (2018)

[15] Zhu, J., Zhou, Y., Zhang, J., Li, H., Zong, C., Li, C.: Multimodal summarization with guidance of multimodal reference. In: Proc. of AAAI (2020)

[16] Li, M., Chen, X., Gao, S., Chan, Z., Zhao, D., Yan, R.: VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles (2020)

[17] Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.: Gated graph sequence neural networks (2015)

[18] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016)

[19] Song, L., Zhang, Y., Wang, Z., Gildea, D.: A graph-to-sequence model for AMR-to-text generation (2018)

[20] Xue, M., Cai, W., Su, J., Song, L., Ge, Y., Liu, Y., Wang, B.: Neural collective entity linking based on recurrent random walk network learning (2019)

[21] Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proc. of AAAI (2019)

[22] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)

[23] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)

[24] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[25] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)

[26] Tran, A., Mathews, A., Xie, L.: Transform and tell: Entity-aware news image captioning. In: Proc. of CVPR (2020)

[27] Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization (2015)

[28] Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T.: Neural document summarization by jointly learning to score and select sentences (2018)

[29] Shen, X., Zhao, Y., Su, H., Klakow, D.: Improving latent alignment in text summarization by generalizing the pointer generator. In: Proc. of EMNLP (2019)

[30] Xu, J., Durrett, G.: Neural extractive text summarization with syntactic compression (2019)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.zip