Extractive Summarization of Documents with Images Based on Multi-Modal RNN

Jingqiang Chen, Hai Zhuge*

Nanjing University of Posts and Telecommunications, Nanjing, China
Guangzhou University, China
KLIIP, *ICT*, *University of Chinese Academy of Sciences, Chinese Academy of Sciences, China*
Aston University, UK

*Correspondence author: haizhuge@gmail.com

ABSTRACT

Rapid growth of multi-modal documents containing images on the Internet expresses strong demand on multi-modal summarization. The challenge is to create a computing method that can uniformly process text and image. Deep learning provides basic models for meeting this challenge. This paper treats extractive multi-modal summarization as a classification problem and proposes a sentence-image classification method based on the multi-modal *RNN* model. Our method encodes words and sentences with the hierarchical *RNN* models and encodes the ordered image set with the *CNN* model and the *RNN* model, and then calculates the selection probability of sentences and the sentence-image alignment probability through a logistic classifier taking text coverage, text redundancy, image set coverage, and image set redundancy as features. Two methods are proposed to compute the image set redundancy feature by combining the important scores of sentences and the hidden sentence-image alignment. Experiments on the extended *DailyMail* corpora constructed by collecting images and captions from the Web show that our method outperforms 11 baseline text summarization methods and that adopting the two image-related features in the classification method can improve text summarization. Our method is able to mine the hidden sentence-image alignments and to create informative well-aligned multi-modal summaries.

Keywords: Summarization; extractive summarization; multi-modal summarization; *RNN*; document summarization

1. INTRODUCTION

With the rapid growth of multi-modal documents with text and images such as news, blogs on the Internet, demands on multi-modal summarization increase rapidly. Most previous studies on text summarization focus on texts [1] [2] [3]. Image summarization is a research direction for creating an image summary to represent a collection of images [4] [5]. Multi-modal summarization is different from pure text summarization and image summarization in that multi-modal summarization creates multi-modal summaries from multi-modal documents.

In summarization application, incorporating image into summary can bridge incoherent sentences as an image can convey rich information (an image is worth thousands of words [6]). So, both text information and image information should be taken into account in realizing multi-modal summarization. Figure 1 - 2 are examples of multi-modal summaries of news.

Finnish military drop DEPTH CHARGES on 'foreign submarine' months after Swedish Navy accuse Russia of sending similar vessel into its waters

The Finnish military says it has dropped depth charges onto a suspected submarine in the sea outside Helsinki after twice detecting the presence of a foreign object. The navy said it noticed an underwater target yesterday and again this morning and fired some warning charges the size of grenades.

Finland, which shares an 833-mile border with Russia, has been increasingly worried about its powerful neighbour after a year of Russian air force sorties and military border exercises.

A Finnish coastguard ship tracks the underwater object - believed to have been a Russian submarine - in the waters near the capital Helsinki

Border patrol boats first identified an underwater target yesterday, before detecting it again early this morning and dropping 'warning' depth charges

It comes just months after Sweden suspected Russia of sending a vessel into waters close to the capital Stockholm.

In what was Sweden's biggest mobilisation since the Cold War, its navy hunted unsuccessfully for a week for what they believed to be a foreign submarine after several observations were made.

It comes just months after Sweden suspected Russia of sending a vessel into waters close to the capital Stockholm. In what was Sweden's biggest mobilisation since the Cold War, its navy hunted unsuccessfully for a week for what they believed to be a foreign submarine after several observations were made. Swedish officials never blamed any country, though most defense analysts said Russia was a likely culprit. Today Finland defence minister Carl Haglund did not say whether Russia was involved but told local media that the target could have been a submarine, and that it has likely left the area, adding that Finland has rarely used such warning charges.

Finland defence minister Carl Haglund did not say whether Russia was involved but told local media that it was extremely rare for the military to use such warning charges. Pictured is a Finnish navy boat

The Finland incident comes just months after Sweden's armed forces hunted unsuccessfully for what they believed to have been a foreign submarine close to Stockholm

He said:'We strongly suspect that there has been underwater activity that does not belong there. 'Of course it is always serious if our territorial waters have been violated,' he told Finnish news agency STT. Moscow retorted immediately, saying moves by Finland and Sweden towards closer ties with NATO were of 'special concern'. In a statement, the Finland Ministry of Defence said its surveillance system first alerted its navy to a 'possible underwater target' within territorial waters about midday yesterday. A second detection was then made early this morning after the navy began searching for the object, and underwater depth charges were fired at 3am. Commodore Olavi Jantunen told Helsingin Sanomat newspaper: 'The bombs are not intended to damage the target, the purpose is to let the target know that it has been noticed.'

Reports of a submarine spotted off Stockholm last year led to Sweden's biggest mobilization since the Cold War.

Regional tensions were reflected earlier in April after an unprecedented hawkish joint statement by Nordic countries - Sweden, Norway, Finland, Denmark and Iceland - that directly cited the Russian 'challenge' as grounds to increase defense cooperation.

Figure 1. An example of multi-modal news taken from *DailyMail*.

The charges released were warning shots, about the size of hand grenades.
Experts believe the object is likely to have been a Russian submarine.

The Finnish military has dropped depth charges on a possible submarine.
Its navy twice detected a foreign object within Helsinki territorial waters.

Figure 2. The multi-modal summary for the news in Figure 1.

Figure 1 is an example of news with images taken from the Daily Mail. The news has 22 sentences (about 548 words) and 4 images each of which has an accompanying caption. Figure 2 is the manually created multi-modal summary of the news. The summary has 4 sentences (46 words) and 2 images, all of which are generated or extracted from the original news. Each summary sentence is aligned with the most relevant image.

To create such a multi-modal summary from the document with images, the following three problems should be considered and solved:

1) How to generate or extract text from the original documents? Can image information improve text summarization?
2) How to score images in the original documents and select the images to create the image summary?
3) How to align the sentences and the images in the created summary?

The created multi-modal summaries can be abstractive or extractive, which corresponds to two text-image summarization methods for solving the above three problems:

1) The abstractive text-image summarization can generate text using the words that often do not appear in the original documents. The previous work treats the abstractive text-image summarization problem as a text generation problem and proposes a generation method based on the multi-modal attentional encoder-decoder model [7]. The method first encodes text and images, and then adopts the hierarchical decoder to generate abstractive text summaries by attending original sentences, images, and captions in the decoding steps.

2) The extractive text-image summarization creates summaries by extracting sentences and images from the original multi-modal document. This paper focuses on the extractive text-image summarization problem, which is treated as a sentence-image classification problem. The proposed classification method is based on the multi-modal *RNN* model. This paper extends our previous work [8] by adding captions to representing images, adding the image redundancy as the features, adding detailed evaluations of text summarization, adding evaluations of image summarization, and adding quantitative studies.

Our classification method processes the multi-modal document twice. During the first time of processing, our method encodes the input text and the images within text. A bi-directional hierarchical *RNN* is adopted to encode the text. *VGGNet* is adopted to extract vector representation of the images [9], and then a bi-directional *RNN* model is used to encode the ordered image set. During the second time of processing, our method visits sentences one by one and adopts a logistic classifier to compute the summary probability of each sentence and each image, and the alignment probability of the sentences and the images. Four features including the text coverage, the text redundancy, the image coverage, and the image redundancy are computed and used in the classifier.

Image information plays an important role in our classification method. Three approaches are proposed to represent image information: vector representations of images, vector representations of captions, and the concatenation of vector representations of images and captions.

For experiments, the *DailyMail* dataset is extended by collecting images and captions from the Web through parsing the *HTML*-formatted news documents. Our model is trained with using the manually created text summary as target summaries. The hidden sentence-image alignment relationships can be discovered automatically.

Our main contributions can be summarized as follows:

1) We propose a multi-modal *RNN*-based extractive text-image summarization method by treating the problem as a classification problem, and taking the text coverage, the text redundancy, the image coverage, and the image redundancy as classification features to calculate the summary probability and the alignment probability of sentences and images.

2) We propose three approaches to represent image information in the classification method, i.e., represent image information with images, captions, or both. Experiments show using vector representation of images in our method performs the best.

3) We propose two approaches to calculating the image redundancy feature in the classification method by combining the summary probability of sentences and the alignment probability of sentences with images. Experiments show that considering image redundancy as a feature of our classification method outperforms the methods without considering the feature.

Experiments show that our model outperforms 11 existing state-of-the-art text summarization methods, including the classification-based extractive text summarization method *SummaRuNNer* [10], which does not consider image information, the abstractive text-image summarization methods based on the multi-modal hierarchical attentional Encoder-Decoder model [7], the graph-based attentional abstractive text summarization method [3], and others state-of-the-art text summarization methods, which indicates that considering image information can improve the classification method for text summarization. Experiments also show that our model can create good sentence-image alignment and good image summaries.

2. RELATED WORK

Since this work is on summarization of documents with text and images, related work concerns text summarization, image captioning, and multi-modal summarization.

Text summarization can be classified into extractive summarization and abstractive summarization. Recent work focuses on neural-based summarization. A neural abstractive sentence summarization method was proposed to summarize a long sentence and generate a shorter sentence as the resultant summary [11]. The work was based on the neural language model [12] [13] [14] and the attentional Encoder-Decoder model proposed for machine translation [15] [16].

A neural document summarization method was proposed to extract sentences and words based on the Encoder-Decoder model [1]. The work used the DailyMail/CNN corpora that have 300 thousands of news with manual text summaries as training and test data [17]. The sentences were encoded with the *CNN* model and the documents were encoded with the *RNN* model, sentences were extracted to create extractive summaries, and words were extracted to create abstractive summaries. A graph-based attentional hierarchical Encoder-Decoder model was proposed for neural abstractive document summarization [3]. The hierarchical Encoder-Decoder was first proposed in [18] to encode and decode document hierarchically conserving the structure of documents. The graph-based attentional mechanism computed the attention scores of the original sentences with the PageRank algorithm [19] in the decoding process. The method outperformed the most neural summarization models in the DailyMail/CNN corpora. Other abstractive models include hierarchical attentional models [20], and distracting attentional models [2].

The latest neural extractive summarization model is *SummaRuNNer* proposed in [10]. *SummaRuNNer* visited each sentence twice. At the first visit, it encoded the sentences and the documents with the bi-directional *RNN*. At the second visit, it computed the summary probability of each sentence based on the logistic classifier using coverage, redundancy, and sentence position as features. Our method is an extension of *SummaRuNNer* for multi-modal summarization.

Image captioning generates captions for an image, which is highly related to our task. Vector representation is first extracted from images using the *CNN* models such as *AlexNet* [21], *GoogleNet* [22], *VGGNet* [23], and *ResNet* [24]. *RNN* models can be used for image captioning [25] [26]. And the attention mechanism can improve image captioning. The image was split into multiple parts which are aligned with words when decoding [27] [28] [29]. An advanced method was to recognize objects from the image and encode the objects with the *RNN* model to represent the image [30]. The attention mechanism used in image caption can also be applied in our task. Our task treats the images in the multi-modal document as an ordered image set, and encodes the image set with the *RNN* model.

Multi-modal summarization is an important research branch of summarization. Traditional multi-modal summarization input multi-modal documents and output text summaries or multi-modal summaries such as [31] [32]. It can also input text documents and output multi-modal summaries, such as [33] [34]. An approach to summarizing texts with images based on citation is proposed in [35]. Image information was used to score and extract sentences in [36]. In the work, images was first retrieved from

Yahoo ! for the sentences in the document, and the *CNN* model was used to extract features from the images. And the matrix factorization algorithm was applied to score sentences. An abstractive summarization method for documents with images based on the multi-modal attentional Encoder-Decoder model was proposed in [7]. Three multi-modal attention mechanisms in the decoding stage were proposed in the work, i.e. caption attention, image attention, and image-caption attention.

3. OUR MODEL

Our model treats extractive multi-modal summarization as a classification problem which computes the probability of candidate sentences and images in the original multi-modal documents, and then selects the sentences and images with the highest probability as the multi-modal summary. Since the multi-modal summary covers both text information and image information of the original multi-modal document, our model uses text coverage, text redundancy, image set coverage, and image set redundancy as the features for classification. The text-related features are calculated by the *RNN* model, and the image-related features are calculated by the *CNN* model and the *RNN* model.

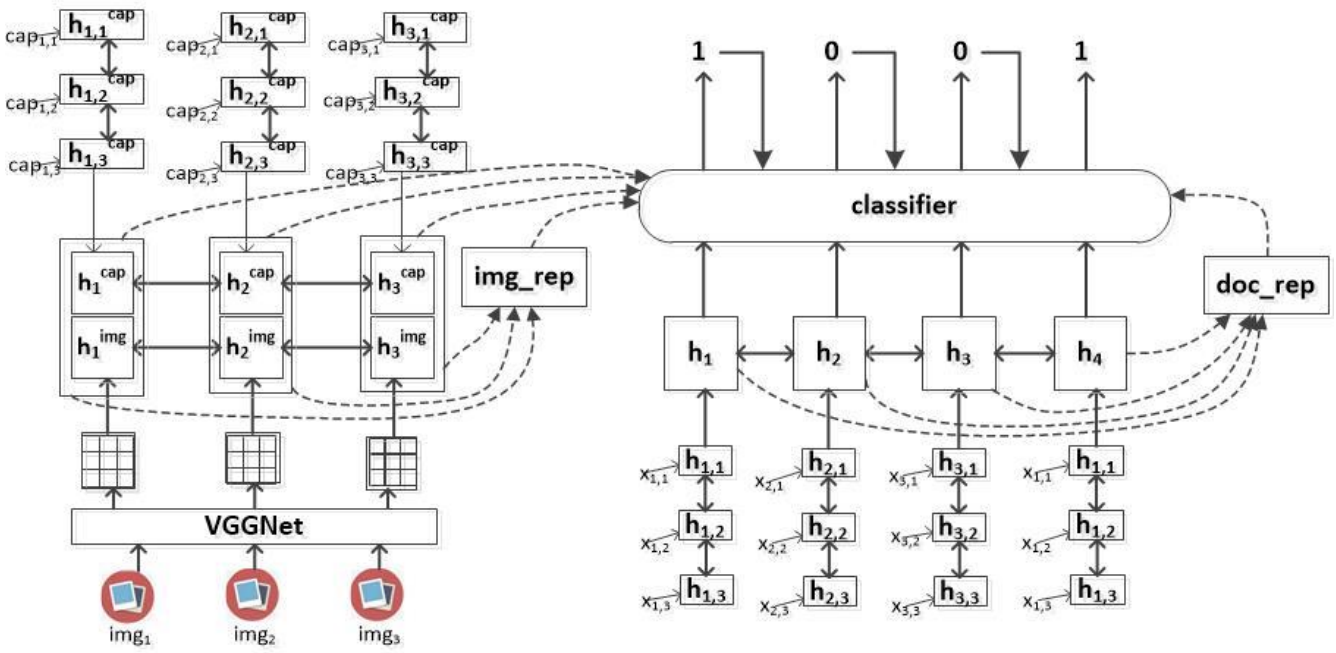

Figure 3. The framework of our neural multi-modal extractive summarization model

Figure 3 shows the architecture of our summarization model. The model consists of four parts: a document encoder for encoding sentences and the text into vector representations, an image set encoder

for encoding the image set into vector representations, a caption set encoder for encoding the caption set into vector representations, a logistic classifier for computing the summary probability of sentences and the alignment probability of sentences with images.

The input of our model is a multi-modal document $MD=\{D, PicSet\}$, where D is the text document and *PicSet* is the picture set.

In our model, the Gated Recurrent Unit (*GRU*) [37] is used as the *RNN* unit because *GRU* is effective and efficient.

3.1 Document Encoding

The text document is denoted as $D=\{s_1, s_2, \ldots, s_{|D|}\}$. s_i is the i^{th} sentence, and $s_i=\{x_{i,1}, x_{i,2}, \ldots, x_{i,|si|}\}$. $x_{i,j}$ is the word embedding of the j^{th} word of the i^{th} sentence.

A two-level bi-directional *RNN* is adopted to encode the text document D. The first level is the word-level encoder, a bi-directional *RNN* which encodes the i^{th} sentence using $\overrightarrow{h}_{i,j} = \overrightarrow{GRU}^{s}(\overrightarrow{h}_{i,j-1}, x_{i,j})$ computing from $x_{i,1}$ to $x_{i,|si|}$, and using $\overleftarrow{h}_{i,j} = \overleftarrow{GRU}^{s}(\overleftarrow{h}_{i,j+1}, x_{i,j+1})$ computing from $x_{i,|si|}$ to $x_{i,1}$. The vector representation of the sentence s_i is the concatenation of $\overrightarrow{h}_{i,1}$ and $\overleftarrow{h}_{i,|si|}$ computed by $s\_rep_i = [\overrightarrow{h}_{i,1}, \overleftarrow{h}_{i,|si|}]$, which is used as the input of the sentence-level encoder. The second level is the sentence-level bi-directional *RNN*-based encoder which encodes the document using $\overrightarrow{h}_i = \overrightarrow{GRU}^{d}(\overrightarrow{h}_{i-1}, s\_rep_i)$ computing from s_1 to $s_{|D|}$, and $\overleftarrow{h}_i = \overleftarrow{GRU}^{d}(\overleftarrow{h}_{i+1}, s\_rep_{i+1})$ computing from $s_{|D|}$ down to s_1. The document representation *d_rep* is modeled as a non-linear transformation of the average pooling of the concatenated hidden states of the sentence-level *RNN* using equation (1) and (2).

$$d\_rep = \tanh(W_d \frac{1}{|D|} \sum_{i=1}^{|D|} h_i + b) \quad (1)$$

$$h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i] \quad (2)$$

where $|D|$ is the number of sentences in the text document D, and h_i is the concatenated bi-directional hidden states.

3.2 Image Set and Caption Set Encoding

The ordered picture set is denoted as $PicSet$=$\{\{img_1, cap_1\}, \{img_2, cap_2\}, \ldots, \{img_{|PicSet|}, cap_{|PicSet|}\}\}$. Each picture in $PicSet$ consists of an image img_i and a caption cap_i, and is ordered by the occurring order in the original multi-modal document. The image occurring orders makes sense because images are often put near the most related sentences, and the sentences have strict order in the document. The vector representation of the image img_i is extracted with the 19-layer *VGGNet* [23]. For simplicity, img_i is used to denote the vector representation of the i^{th} image in the following. The caption is denoted as cap_i=$\{cap_{i,1}, cap_{i,2}, \ldots, cap_{i,|capi|}\}$ where $cap_{i,j}$ is the word embedding of the j^{th} word of the caption cap_i.

The images, the captions, or both can be used to represent the image set, accordingly there are three methods to encode and generate the vector representation of the ordered image set:

1) The first method uses images to represent the image set and adopt the bi-directional *RNN* model as the encoder which takes the *CNN* features of the images as input. The equations $\overrightarrow{h}_i^{img} = \overrightarrow{GRU}^{img}(\overrightarrow{h}_{i-1}^{img}, img_i)$ and $\overleftarrow{h}_i^{img} = \overleftarrow{GRU}^{img}(\overleftarrow{h}_{i+1}^{img}, img_i)$ compute the bi-directional hidden states when encoding the i^{th} image. As with document representation, the representation of the image set *is_rep* is modeled as a non-linear transformation of the average pooling of the concatenated bidirectional hidden states using equation (3) and (4).

$$is\_rep = \tanh(W_{is}^{img} \frac{1}{|PicSet|} \sum_{i=1}^{|PicSet|} h_i^{img} + b) \quad (3)$$

$$h_i^{img} = [\overrightarrow{h}_i^{img}, \overleftarrow{h}_i^{img}] \quad (4)$$

2) The second method uses captions to represent the image set. Each caption is treated as a sentence and the caption set is treated as a document, and then the document encoder is applied to the caption document. The representation of the image set is computed using equation (5) and (6), where $\overrightarrow{h}_i^{cap}$ and $\overleftarrow{h}_i^{cap}$ are the bi-directional sentence-level hidden states.

$$is\_rep = \tanh(W_{is}^{cap} \frac{1}{|PicSet|} \sum_{i=1}^{|PicSet|} h_i^{cap} + b) \quad (5)$$

$$h_i^{cap} = [\overrightarrow{h}_i^{cap}, \overleftarrow{h}_i^{cap}] \quad (6)$$

3) The third method uses both images and captions to represent the image set. Images are encoded using the first method and captions are encoded using the second method

independently, and then the respective hidden states are concatenated, and the representation is computed using equation (7) and (8), where h_i^{IC} is the hidden state of this method.

$$is\_rep = \tanh(W_{is}^{IC} \frac{1}{|PicSet|} \sum_{i=1}^{|PicSet|} h_i^{IC} + b) \quad (7)$$

$$h_i^{IC} = [h_i^{img}, h_i^{cap}] \quad (8)$$

3.3 Classifier

For classification, each sentence is visited the second time with a logistic classifier to calculate the summary probability, taking the text coverage, the text redundancy, the image set coverage, and the image set redundancy as features. Equation (9) is the classifier.

$$\begin{aligned} P(y_i = 1 | h_i, d\_rep_i, is\_rep_i) = \sigma(& \\ & W_{con} h_i && \#content \\ + \; & h_i^T W_{t\_cov} d\_rep_i && \#text\_coverage \\ - \; & h_i^T W_{t\_red} Sum\_T_i && \#text\_redundancy \\ + \; & h_i^T W_{is\_cov} is\_rep_i && \#image\_coverage \\ - \; & h_i^T W_{is\_red} Sum\_IS_i && \#image\_redundancy \\ + \; & W_{pos} Pos_i && \#sentence\_position \\ + \; & b \quad) & & \end{aligned} \quad (9)$$

Equation (9) calculates the summary probability of the i^{th} sentence conditioned at h_i, d_rep_i, and is_rep_i. y_i=1 means the sentence is extracted as a summary sentence. σ is the sigmoid function. $W_{con} h_i$ represents the information of the i^{th} sentence. $h_i W_{t\_cov} d\_rep_i$ represents the information coverage of the sentence with respect the text document. $h_i W_{t\_red} Sum\_T_i$ represents the information coverage of the sentence with respect to the current text summary when visiting the i^{th} sentence. Sum_T_i is the representation of the text summary computed by non-linear transformation of the summation of the weighted representation of previous sentences using equation (10). $W_{pos} Pos_i$ represents the salience of the sentence position in the document.

$$Sum\_T_i = \tanh(\sum_{j=1}^{i-1} h_j P(y_j = 1 | h_j, d\_rep_j, is\_rep_j)) \quad (10)$$

In equation (9), $h_i^T W_{is\_cov} is\_rep_i$ represents the information coverage of the sentence with respect to the image set. $h_i^T W_{is\_red} Sum\_IS_i$ represents the information redundancy with respect to the image summary, where Sum_IS_i is the current image summary when visiting the i^{th} sentence. The image set can be represented by images, captions, or both, as introduced in the previous subsection.

The idea of the attention mechanism of the Encoder-Decoder model is borrowed to compute the representation of the image summary when visiting the i^{th} sentence. So the following two methods are proposed to compute the image summary Sum_IS_i and the image redundancy, named by ***IR*1** and ***IR*2**:

1) ***IR*1:** This method weights the attention by the summary probability of the previously visited sentences, and then calculates the image summary based on the attentions. Equation (11), equation (12) and equation (13) are the equations for *IR*1. Equation (11) calculates $att(h_i, h_k^{is})$ which is the attention of h_i to h_k^{is}. Equation (12) calculates the normalized attention $\alpha^{IR1}(h_i, h_k^{is})$ when visiting the i^{th} sentence by summing up the attention of each previously visited sentence to h_k^{is} weighted by the summary probability. Equation (13) calculates the image set representation $Sum\_IS^{IR}(h_i)$ when visiting the i^{th} sentence by summing up h_k^{is} weighted by the attentions.

$$att(h_i, h_k^{is}) = h_j W^{att} h_k^{is} \quad (11)$$

$$\alpha^{IR1}(h_i, h_k^{is}) = \frac{\sum_{j=1}^{i-1} \exp(att(h_j, h_k^{is})) P(y_j = 1 | h_j, d\_rep_j, is\_rep_j)}{\sum_{k=1}^{|PicSet|} \sum_{j=1}^{i-1} \exp(att(h_j, h_k^{is})) P(y_j = 1 | h_j, d\_rep_j, is\_rep_j)} \quad (12)$$

$$Sum\_IS^{IR}(h_i) = \sum_{k=1}^{|PicSet|} h_k^{is} \alpha^{IR1}(h_i, h_k^{is}) \quad (13)$$

2) ***IR*2:** This method calculates the image context for each visited sentence, weights the image context by the summary probability of the corresponding sentence, and sums up the weighted context to get the image summary. Equation (14), equation (15) and equation (16) are the equations for *IR*2. Equation (14) calculates $\alpha(h_i, h_k^{is})$ which is the normalized attention of h_i to h_k^{is}. Equation (15) calculates image_$context(h_i)$, which is the image context for the i^{th} sentence. Equation (16) calculates the image set representation $Sum\_IS^{IR2}(h_i)$ when visiting the i^{th} sentence by summing up image_$context(h_j)$ of each previously visited sentence j weighted by the summary probability.

$$\alpha(h_i, h_k^{is}) = \frac{\exp(att(h_i, h_k^{is}))}{\sum_{k=1}^{|PicSet|} \exp(att(h_i, h_k^{is}))} \quad (14)$$

$$image\_context(h_i) = \sum_{k=1}^{|PicSet|} \alpha(h_i, h_k^{is}) h_k^{is} \quad (15)$$

$$Sum\_IS^{IR2}(h_i) = \sum_{j=1}^{i-1} P(y_j = 1 | h_j, d\_rep_j, is\_rep_j) image\_context(h_j) \quad (16)$$

The two image redundancy representing methods will be discussed in the following experiments.

3.4 Training

Since there are no ground-truth target multi-modal summaries, the extractive text summaries are used to train our model. The hidden sentence-image alignment can be mined in training and inferring.

The ground truth needed is the sentence-level binary labels. However, most summarization corpora only provide abstractive human-written summaries. The extractive summaries are generated from the abstractive summaries using a greedy approach because it is expansive to select a global optimal subset of sentences [10]. The algorithm works as follows: at each time, select the sentence that can maximize the *Rouge* score of the summary using the abstractive summary as the reference until there are no such sentences [38].

Based on the labels, the negative log-likelihood of the labels is use as the objective function:

$$\begin{aligned} l(W, b) = -\sum_{d=1}^{N_{DS}} \sum_{j=1}^{N_d} & y^d{}_j \log P(y^d{}_j = 1 | h^d{}_j, d\_rep^d{}_j, is\_rep^d{}_j) \\ + \quad & (1 - y^d{}_j) \log(1 - P(y^d{}_j = 1 | h^d{}_j, d\_rep^d{}_j, is\_rep^d{}_j)) \end{aligned} \quad (17)$$

where W, b is the parameters, N_{DS} is the number of document in the training set, x is representation of the document, and y is the vector of labels. At test time, the model emits the summary probability of the sentences sequentially.

3.5 Image Selection and Alignment

The sentence-image alignment relationships can be calculated through the training of the attention mechanism in the classifier by equation (14) for both *IR*1 and *IR*2 image redundancy representation methods. In equation (14), $\alpha(h_i, h_k^{is})$ is the normalized attention of h_i to h_k^{is}.

The images are scored based on the summary probability of the sentences and the attention of sentences to images using equation (18).

$$score(img_k) = \sum_{i=1}^{|D|} \alpha(h_i, h_k^{is}) P(y_i = 1 | h_i, d\_rep_i, is\_rep_i) \quad (18)$$

The top-scored images are selected as the image summary *Sum_IS*. The image *k* is chosen for each sentence *i* in the text summary using the equation $k = \arg\max_{k \in Sum\_IS} \alpha(h_i, h_k^{is})$.

4. EXPERIMENTS

4.1 Corpora

Since there are no large-scale existing multi-modal documents corpora for training, our corpora are created by extending the *DailyMail* corpora through collecting images and captions. The *DailyMail*/*CNN* corpora are originally constructed by Hermann et al. [17] for question answer, and are re-purposed for document summarization by Cheng and Lapata [1]. Only the *DailyMail* corpora are used and extended because the news in the *DailyMail* corpora have more images and it is easier to collect images for the DailyMail Corpors as the images of the *CNN* corpora are not available. The HTML-formatted documents provided in the original *DailyMail* corpora are parsed, and the captions and the image *URL* links are extracted through which the images are downloaded. The corpora are called as the *E-DailyMail* corpora. We find that the text documents provided in the original *DailyMail* corpora have already contained captions, so we don't change the original text documents in *E-DailyMail*. As with *DailyMail*, *E-DailyMail* is partitioned into 90% training, 5% dev, and 5% test datasets. The statistics is shown in Table 1. Since there are more than 210K multi-modal documents in the corpora, the scale of the corpora is sufficient for training and testing.

Table 1. The statistics of the E-DailyMail Corpora. D.L and S.L indicate the average number of sentences in the document and summary. I.N indicates the average number of images in the story. Sent.L and Cap.L indicates the average length of sentences and captions respectively.

Train	Dev	Test	D.L.	S.L.	I.N	Sent.L	Cap.L
196557	12147	10396	26.0	3.84	5.42	26.86	24.75

4.2 Implementation and Settings

The texts of the *E-DailyMail* corpora are preprocessed by tokenizing the text and replacing the digits with the <NUM> token. Entity labels are not used and all texts are treated as ordinary texts. The 150k most frequent words in the corpora are kept as the vocabulary. Other words are replaced with the *OOV* token.

Our model is implemented with Tensorflow in Python. For the *RNN* cell, one layer of *GRU* is used. The dimension of hidden state of the bi-directional *RNN* encoder is 200. The dimension of the word embedding vector is 100. The word embeddings are initialized with Google's word2vec tools [11] in the whole text of *DailyMail*/*CNN* corpora. The 4096-dimension full-connected layer of 19-layer *VGGNet* pre-trained on ImageNet is extracted as the vector representation of the image resized to 224 by 224. Adam is adopted as the optimizer, and the parameters of Adam are set to those provided by Kingma and Ba [39]. Gradient clipping is employed to regularize our model and an early stopping criterion based on validation cost. The batch size is set to 5 multi-modal documents. The training set is shuffled to train our models. It takes about one day for training 170k ~ 180k steps depending on the models on a *GTX*-1080 *TI GPU* card. The early stopping criterion is met at about 300k steps.

At test time, the top sentences sorted by the predicted summary probability are picked until the summary length limit is reached when limited-length *Rouge* is used for evaluation.

4.3 Evaluations of Document Summarization

4.3.1 Comparison with existing methods

In the following, three versions of our model are compared and discussed to show which version is most suitable for text summarization, and then the models are compared with 11 existing state-of-the-art text summarization methods to show the strength of our models in text summarization. *ROUGE* is adopted to evaluate the result [38].

There are three versions of our model, named by *MRNN-I*, *MRNN-C*, *MRNN-IC*, depending on the representation methods of the image set. *MRNN-I* is the multi-modal *RNN* based summarization model using images to represent the image set, *MRNN-C* is the one using captions to represent the image set, and *MRNN-IC* is the one using both images and captions to represent the image set. Two image

redundancy methods *IR*1 and *IR*2 are also compared for *MRNN-I*, *MRNN-C*, and *MRNN-IC*. For example, *MRNN-I* with *IR*1 is the summarization method *MRNN-I* with the image redundancy representing method *IR*1. The first 6 lines in Table 2 are the results of the three models with two image redundancy representing methods at the summary length of 75 bytes, and the first 6 lines in Table 3 are the results at the length of 275 bytes. The results show that the model *MRNN-I* with *IR*1 performs the best. An interesting observation is that *MRNN-C* and *MRNN-IC* are not better than *MRNN-I* though *MRNN-IC* has more features than *MRNN-I* does because the text document also contains the captions and thus images are new features with respect to the text document.

Table 2. Comparison results on the E-DailyMail corpora using rouge recall at the summary length of 75 bytes.

Method	*Rouge*-1	*Rouge*-2	*Rouge*-L
MRNN-I with *IR*1	**27.6**	12.6	**17.1**
MRNN-C with *IR*1	27.0	12.3	15.4
MRNN-IC with *IR*1	27.2	12.4	15.6
MRNN-I with *IR*2	27.4	**12.7**	15.7
MRNN-C with *IR*2	26.6	12.1	15.2
MRNN-IC with *IR*2	27.2	12.0	15.4
SummaRuNNer	26.2	11.1	14.5
HNNattTI	24.84	8.7	16.99
HNNattTC	18.61	6.7	13.44
HNNTattTIC	21.17	8.1	15.24
Lead-3	21.9	7.2	11.6
TextRank	24.7	9.5	12.2
NN-SE	22.7	8.5	12.5
Tan et al.'17	27.4	11.3	15.1
LREG(500)	18.5	6.9	10.2
NN-ABS(500)	7.8	1.7	7.1
NN-WE(500)	15.7	6.4	9.8

Table 3. Comparison results on the E-DailyMail corpora using rouge recall at the summary length of 275 bytes.

Method	*Rouge*-1	*Rouge*-2	*Rouge*-L
MRNN-I with *IR*1	**43.5**	**18.2**	**35.4**
MRNN-C with *IR*1	42.5	17.2	34.8
MRNN-IC with *IR*1	42.9	17.9	35.0
MRNN-I with *IR*2	43.1	17.9	35.2
MRNN-C with *IR*2	42.8	16.9	34.7

MRNN-IC with *IR*2	43.0	17.3	34.9
SummaRuNNer	42.0	16.9	34.1
HNNattTI	33.40	12.65	26.75
HNNattTC	27.93	10.67	20.33
HNNattTIC	31.74	11.83	24.47
Lead-3	40.5	14.9	32.6
TextRank	42.0	17.0	34.2
NN-SE	42.4	17.3	34.8

The first competitive state-of-the-art summarization method is the state-of-the-art extractive text summarization method *SummaRuNNer*, which treats text summarization as a sentence classification problem and adopts a *RNN*-based classification model using text information coverage and text information redundancy as features to compute the important scores of the sentences. As an extractive multi-modal summarization method, our method treats multi-modal summarization as a sentence-image classification and alignment problem to create multi-modal summaries, and considers image information and caption information by adding image coverage and image redundancy as features. Results in Table 2 and Table 3 show that our models have considerable improvement over *SummaRuNNer* for text summarization. This shows that considering image information in text summarization by adding image-related features in the classification method can improve extractive text summarization.

The second competitive methods are three abstractive text-image summarization methods proposed in [7], which treats multi-modal summarization as a text generation problem, and adopts the multi-modal attentional *Encoder-Decoder* model to generate text summaries and align sentences with images. The three abstractive methods named *HNNattTI*, *HNNattTC* and *HNNattTIC* attend images, captions, or both in the decoding stage respectively. Our models are also extractive text-image summarization methods creating multi-modal summaries by extracting sentences and images. Table 2 and Table 3 show the results of comparison with the three abstractive text-image summarization models. Our extractive models outperform the three abstractive models. This is because the quality of the extracted sentences is usually higher than that of the sentences generated by the abstractive methods.

The third competitive method is the state-of-the-art neural abstractive model based on the graph-based attentional hierarchical *Encoder-Decoder* model [3], which uses a graph-based attention mechanism to compute the attention scores in the decoding steps. It first decodes sentences and then decodes words. Our methods are extractive multi-modal summarization methods. Table 2 shows our

method *MRNN-I* outperforms the model proposed in [3]. The cause is that our models consider image information in text summarization.

The fourth state-of-the-art text summarization method is *TextRank* [40], which builds a graph with sentences within text as nodes and applies the *PageRank* algorithm to rank the sentences [41]. An extension of *TextRank* in [42] uses the *Semantic Link Network* to build the graph by splitting the document into paragraphs and considering the relationships between paragraphs in the ranking [43]. Table 2 and Table 3 show the *Rouge* scores of *TextRank* on the *DailyMail* corpora. According to Table 2 and Table 3, our models outperform *TextRank*.

Other existing text summarization methods include the extractive models *Lead*-3 and *NN-SE* [1], and the abstractive methods *LREG, NN-ABS* [12] and *NN-WE* [1]. *Lead*-3 is a strong baseline using the leading 3 sentences as summary. *NN-SE* is the neural extractive summarization method based on the *Encoder-Decoder* model. *LREG* is a feature-based method using linear regression. *NN-ABS* is a simple hierarchical extension of the sentence summarization model proposed in [12]. *NN-WE* is an abstractive model restricting the generation of words from the original document. Table 2 and Table 3 show the *Rouge* scores of these methods on the *DailyMail* corpora. *LREG*, *NN-ABS* and *NN-WE* are only tested on 500 samples of the test set. Table 2 and Table 3 show that our models outperform *Lead*-3, *NN-SE, LREG, NN-ABS* and *NN-WE* for the summary length 75 bytes and 275 bytes. The cause is that our methods consider both text information and image information while those other text summarization methods only consider text information.

In short, our method *MRNN-I* with *IR*1 performs the best. *MRNN-I* performs better than *MRNN-C* and *MRNN-IC*, implying that using images are better than using captions or using both to representing the image set. Our models with *IR*1 perform is better than our model with *IR*2, implying that the image redundancy method *IR*1 is more suitable for text summarization than *IR*2. Our methods outperform *SummaRuNNer* because our models consider the features of image information coverage and image information redundancy in the classification method. Our methods outperform the abstractive text-image summarization methods *HNNattTI, HNNattTC* and *HNNattTIC* because the quality of the extracted sentences is usually higher than that of the text generated by the abstractive methods. Our methods also outperform the state-of-the-art neural abstractive text summarization methods, *TextRank, Lead*-3, *NN-SE, LREG, NN-ABS* and *NN-WE*, showing that considering image information can improve text summarization.

4.3.2 Effect of using image redundancy feature

To check the effect of using the feature of the image set redundancy in equation (9), we remove this feature from the models. The models without this feature are named *MRNN-I* w/o *IR*, *MRNN-C* w/o *IR*, and *MRNN-IC* w/o *IR*, *which* are then compared with the models with the image set redundancy. Table 4 and Table 5 show that our models without considering the image set redundancy perform is worse than the ones considering the feature. That is, considering the feature of image set redundancy in the classification method can improve our models in text summarization.

Table 4. The Results of our models without the factor of the image set redundancy at the length of 75 bytes.

Method	*Rouge*-1	*Rouge*-2	*Rouge*-L
MRNN-I w/o *IR*	27.1	12.5	15.6
MRNN-C w/o *IR*	26.8	12.1	15.2
MRNN-IC w/o *IR*	26.9	12.2	15.3

Table 5. The Results of our models without the factor of the image set redundancy at the length of 275 bytes.

Method	*Rouge*-1	*Rouge*-2	*Rouge*-L
MRNN-I w/o *IR*	42.9	17.8	35.0
MRNN-C w/o *IR*	42.6	16.9	34.5
MRNN-IC w/o *IR*	42.8	17.4	34.2

4.4 Evaluations of Image Summarization

To evaluate the image summarization, the ground truth is created using a greedy algorithm on the captions as follows: at each time i, choose img_k to maximize $Rouge(\{cap_1,\ldots cap_{i-1},cap_k\}, Abs\_Sum)$ - $Rouge(\{cap_1,\ldots cap_{i-1}\}, Abs\_Sum)$) where *Abs_Sum* is the gold text summary and cap_k is the caption of img_k. The average number of images in summaries is 2.15.

The models are compared at the image summary length of 1-image or 2-images. The top 1-2 ranked images are selected as the image summaries. Results in Table 6 show that our models *MRNN-I*, *MRNN-C*, *MRNN-IC* outperform the random baseline by the Recall metric, and that our models outperform the three abstractive text-image summarization models at the image summary length of 1-image. In particular, the model *MRNN-I* with *IR*1 performs the best at the summary length of 1-image. Our models can generate roughly acceptable image summaries for the text summaries to align with.

Table 6. Image summarization using the recall for the 1-image or 2-images summary.

Method	1-image	2-images
MRNN-I with *IR*1	**0.5031**	0.4696
MRNN-C with *IR*1	0.5029	0.4693
MRNN-IC with *IR*1	0.5017	0.4677
MRNN-I with *IR*2	0.5027	0.4662
MRNN-C with *IR*2	0.4363	0.4172
MRNN-IC with *IR*2	0.4853	0.4562
HNNattTI	0.4978	**0.4783**
HNNattTC	0. 4137	0. 3998
HNNattTIC	0. 4362	0. 4230
Random	0. 4721	0. 4417

4.5 A Case Study

A case study carried out on the model *MRNN-I* with *IR*1 for the multi-modal news shown in Figure 1.

For the sake of explanation, the original values of the six abstract features (i.e., the content, the text coverage, the text redundancy, the image set coverage, the image set redundancy, the position respectively) in equation (9) are normalized and shown in Table 7. Especially, the feature of text redundancy and the feature of image redundancy are transformed into the text novelty feature and the image novelty feature respectively. Table 7 also shows the summary probability for the Top-5 ranked sentences. The normalized values of text novelty and image novelty are calculated with the equation $1-exp(v)/(exp(max\_v)-exp(min\_v))$. The normalized values of other four features are calculated with the equation $exp(v)/(exp(max\_v)-exp(min\_v))$. Herein v, max_v, min_v are the original value, the maximum original value, and the minimum original value of an abstract feature.

The five sentences in Table 7 are identified as *S*1, *S*2, *S*3, *S*4 and *S*5 respectively. According to the values in Table 7, *S*1 has the highest value of text novelty which means that *S*1 is a surprising sentence and contains the most novel text information. *S*3 has the highest value of image novelty which means that *S*3 is also a surprising sentence and contains the most novel image information. *S*5 has the lowest value of content and text coverage which means that the content of *S*5 is not salient and it does not cover much text information of the document. A surprising observation is that *S*1 gets the highest summary probability and it is also the first sentence in the original document but the value of its position feature is not the highest. The first sentence of a document is not always the most important sentence.

Table 7. The normalized values of each feature in equation (9) for five sentences of highest summary probability for the example news in Figure 1.

Sentences	Content	Text-Covera	Text-Novel	Image-Coverage	Image-Novelty	Position	Probability

		ge	ty				
S1: the finnish military says it has dropped depth charges onto a suspected submarine in the sea outside helsinki after twice detecting the presence of a foreign object.	0.2256	0.1665	1.0	0.3225	0.2355	0.5019	0.6183
S2: the navy said it noticed an underwater target yesterday and again this morning and fired some warning charges the size of grenades.	0.1827	0.1480	0.9220	0.9030	0.3314	0.1028	0.3980
S3: finland , which shares an 833-mile border with russia, has been increasingly worried about its powerful neighbour after a year of russian air force sorties and military border exercises.	0.2140	0.2016	0.7493	1.0	0.1407	0.4010	0.2937
S4: the finland incident comes just months after sweden 's armed forces hunted unsuccessfully for what they believed to have been a foreign submarine close to stockholm.	0.1968	0.2111	0.7107	0.3517	0.8074	0.6646	0.2622
S5: it comes just months after sweden suspected russia of sending a vessel into waters close to the capital stockholm.	0.0	0.0	0.8038	0.5274	0.7722	0.6089	0.2439

Table 8. The sentence-image alignment scores and the total scores of images.

Sentence No.	IMG1	IMG2	IMG3	IMG4
S1	0.2254	0.3142	0.1532	0.3072
S2	0.2709	0.2784	0.1875	0.2633
S3	0.2630	0.3249	0.2588	0.1533
S4	0.3528	0.1434	0.1321	0.3717
S5	0.4367	0.1082	0.1043	0.3509
total scores	1.2796	1.1316	0.8261	1.3771

Table 8 shows the alignment probability of the five sentences with each image and the total scores of each image. The images in the example news in Figure 1 are numbered from left to right and from top to bottom as IMG1, IMG2, IMG3 and IMG4 respectively. The total score of each image is calculated by summing up the weighted alignment scores of the image and all the sentences in the original document. The alignment scores are weighted by the probability of the sentences selected in summary.

Figure 4 shows the created multi-modal summary. The summary has three sentences and two images, and each sentence is aligned with an image. The Top-3 scored sentences *S*1, *S*2, *S*3 in Table 7 are extracted as the text summary. The top-2 scored images *IMG*4 and *IMG*1 in Table 8 are selected as the image summary. *S*1 is aligned with *IMG*4, and S1 and S3 are aligned with *IMG*1 according to the alignment scores in Table 8. The generated multi-modal summary is highly overlapped with the manual summary shown in Figure 2. The F-measure scores for Rouge-1, Rouge-2, Rouge-L of the text summary with respect to the manually generated text summary is 42.91, 14.63, 32.95 respectively.

S3: finland, which shares an 833-mile border with russia, has been increasingly worried about its powerful neighbour after a year of russian air force sorties and military border exercises.
S2: the navy said it noticed an underwater target yesterday and again this morning and fired some warning charges the size of grenades.

S1: the finnish military says it has dropped depth charges onto a suspected submarine in the sea outside helsinki after twice detecting the presence of a foreign object.

Figure 4. The extractive summary with images created by *MRNN-I* with *IR*1 for the example news in Figure 1.

5. DISCUSSION AND FUTURE WORK

To discover the sentence-image alignment relationships in the source document and in the multi-modal summary is important for multi-modal summarization. High-quality sentence-image alignment can improve text scoring and image scoring and also improve the generated multi-modal summaries. There has been much research on word-word alignment for machine translation [15] [16] and word-pixel alignment for image captioning [28] [29]. Sentence-image alignment is also useful for information retrieval and has been studied in information retrieval.

Our models used an unsupervised way by treating sentence-image alignment as hidden variables, which are mined through training the attention mechanism in the classification method. Experiments show that the image summaries created by our models are only better than random image summaries. To supervise the learning of sentence-image alignment in our models is a feasible way to improve the image summaries. However, fully supervised learning of sentence-image alignment needs manually aligned sentences and images in ground truth multi-modal summaries, which are labor- and time-consuming to create. A more feasible way is to use the image summaries greedily created as described in Section 4.4 to supervise the average attention scores of the attentional mechanism.

Our method can be applied to summarizing a scientific article to create multi-modal abstract or the multi-modal slides based on our previous work [44] [45]. So far, there is little research on multi-modal scientific article summarization. The training corpora can be built by collecting papers from the open-access online digital libraries. The images and the main texts can be extracted from scientific articles as the source data, and the abstracts of scientific articles can be extracted as the ground truth text summaries. The method can also be applied to other document summarization application fields like law.

6. CONCLUSIONS

This paper proposes a neural extractive multi-modal summarization method that summarizes documents containing images. Our method treats the summarization problem as a sequential sentence and image classification problem, and the logistic classifier is applied to the sentences in sequence by taking the text coverage, the text redundancy, the image set coverage, and the image set redundancy as features encoded by the *RNN* model and the *CNN* model. The sentence-image alignment probability is figured out in training and inferring. Experiments show that our models outperform the state-of-the-art neural summarization method that does not consider image and caption information. This shows that considering images and captions can improve extractive text summarization. In particular, our method using images to represent the image set perform the best in the extended *DailyMail* corpora because the text documents in the corpora already contain captions. Our models outperform the state-of-the-art abstractive neural text-image summarization methods because the quality of the sentences extracted from the original documents is usually higher than that of the text generated by the abstractive methods. Our models also outperform other state-of-the-art extractive and abstractive summarization methods and the graph-based attentional neural-based abstractive method because our methods consider both text information and image information. Experiments also show that the model considering the feature of image redundancy outperforms the model without considering the feature and our model can generate informative image summaries.

ACKNOWLEDGMENTS

The research was sponsored by the National Natural Science Foundation of China (No.61806101, No.61876048, No. 61602256, No.61876091), and the Open Foundation of Key Laboratory of Intelligent

Information Processing, ICT, CAS (IIP2019-2). Professor Hai Zhuge is the corresponding author of this paper.

REFERENCES

[1] Cheng, J. and Lapata, M., 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

[2] Chen, Q., Zhu, X., Ling, Z., et al. Distraction-Based Neural Networks for Document Summarization. *IJCAI*, 2016.

[3] Tan, J., Wan, X. and Xiao, J., 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1171-1181).

[4] Yang, C., Shen, J., Peng, J. and Fan, J., 2013. Image collection summarization via dictionary learning for sparse representation. *Pattern Recognition*, *46*(3), pp.948-961.

[5] Samani, Z.R. and Moghaddam, M.E., 2017. A knowledge-based semantic approach for image collection summarization. *Multimedia Tools and Applications*, *76*(9), pp.11917-11939.

[6] Rossiter, M. J., Derwing, T. M., Jones, V. M. L. O. Is a Picture Worth a Thousand Words? *Tesol Quarterly*, 2012, 42(2):325-329.

[7] Chen, J. and 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*.

[8] Chen, J., Zhuge, H. Extractive Text-Image Summarization Using Multi-Modal RNN. *Semantics, Knowledge and Grids (SKG), 2018 14th International Conference on*. IEEE, 2018.

[9] Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science*, 2014.

[10] Nallapati, R., Zhai, F. and Zhou, B., 2017, February. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI* (pp. 3075-3081).

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *International Conference on Neural Information Processing Systems* (Vol.26, pp.3111-3119). Curran Associates Inc.

[12] Rush, A. M., Chopra, S., Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. *Computer Science*, 2015.

[13] L. Liu, Y-J Liu, and S. C. Tong, "Neural networks-based adaptive finite-time fault-tolerant control for a class of strict-feedback switched nonlinear systems," *IEEE Transactions on Cybernetics*, DOI:10.1109/TCYB.2018.2828308, 2018.

[14] T. Gao, Y. J. Liu, L. Liu and D. Li, "Adaptive neural network-based control for a class of nonlinear pure-feedback systems with time-varying full state constraints," in *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 5, pp. 923-933, September 2018.

[15] Bahdanau, D., Cho, K., Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science*, 2014.

[16] Luong, M. T., Pham, H., Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. *Computer Science*, 2015.

[17] Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M. and Blunsom, P., 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems* (pp. 1693-1701).

[18] Li J., Luong M. T., Jurafsky D. A Hierarchical Neural Autoencoder for Paragraphs and Documents. *Computer Science*, 2015.

[19] Page, L. (1998). The pagerank citation ranking : bringing order to the web. *Stanford Digital Libraries Working Paper*, *9*(1), 1-14.

[20] Nallapati, R., Zhou, B., dos Santos, C., glar Gulçehre, Ç. and Xiang, B., 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. *CoNLL 2016*, p.280..

[21] Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2013, 60(2):2012.

[22] Szegedy, C., Liu W., Jia, Y., et al. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[23] Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science*, 2014.

[24] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv*:1512.03385, 2015.

[25] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A., 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

[26] Wang, C., Yang, H., Bartz, C. and Meinel, C., 2016, October. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 988-997). ACM.

[27] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, *39*(4), pp.652-663.

[28] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057).

[29] Liu, C., Mao, J., Sha, F. and Yuille, A.L., 2017, February. Attention Correctness in Neural Image Captioning. In *AAAI*(pp. 4176-4182).

[30] Liu, C., Sun, F., Wang, C., Wang, F. and Yuille, A., 2017. MAT: A multimodal attentive translator for image captioning. *arXiv preprint arXiv:1702.05658*.

[31] Wu, P., & Carberry, S. (2011, May). Toward extractive summarization of multimodal documents. In *Proceedings of the Workshop on Text Summarization at the Canadian Conference on Artificial Intelligence* (pp. 53-61).

[32] Greenbacker, C. F. (2011, June). Towards a framework for abstractive summarization of multimodal documents. In *Proceedings of the ACL 2011 Student Session* (pp. 75-80). Association for Computational Linguistics.

[33] Hu, Y., & Wan, X. (2015). Ppsgen: learning-based presentation slides generation for academic papers. *IEEE transactions on knowledge and data engineering*, *27*(4), 1085.

[34] Qiang, Y., Fu, Y., Guo, Y., Zhou, Z. H., & Sigal, L. (2016). Learning to generate posters of scientific papers. *arXiv preprint arXiv:1604.01219*.

[35] Zhuge, H., 2016. *Multi-Dimensional Summarization in Cyber-Physical Society*. Morgan Kaufmann.

[36] Wang, W. Y., Mehdad, Y., Radev, D. R., et al. A Low-Rank Approximation Approach to Learning Joint Embeddings of News Stories and Images for Timeline Summarization. *NAACL*. 2016:58-68.

[37] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., & Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Computer Science*.

[38] Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

[39] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[40] Mihalcea, R. and Tarau, P., 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[41] Page, L., Brin, S., Motwani, R. and Winograd, T., 1999. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

[42] Sun, X. and Zhuge, H., 2018. Summarization of Scientific Paper Through Reinforcement Ranking on Semantic Link Network. *IEEE Access*, *6*, pp.40611-40625.

[43] Zhuge, H. The Knowledge Grid: Toward Cyber-Physical Society. *World Scientific*, 2012, 2nd Ed.

[44] Chen, J., Zhuge, H. Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems*, 2014, 32(C):246-252.

[45] Chen, J., and Zhuge, H. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience* 31, no. 3 (2019): e4261.