Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN

Jingqiang Chen[1], **Hai Zhuge**[1,2,3,4]
[1]Nanjing University of Posts and Telecommunications
[2]Aston University; [3]Guangzhou University
[4]Key Laboratory of Intelligent Information Processing, ICT,
University of Chinese Academy of Sciences, Chinese Academy of Sciences
cjq@njupt.edu.cn, haizhuge@gmail.com

Abstract

Rapid growth of multi-modal documents on the Internet makes multi-modal summarization research necessary. Most previous research summarizes texts or images separately. Recent neural summarization research shows the strength of the Encoder-Decoder model in text summarization. This paper proposes an abstractive text-image summarization model using the attentional hierarchical Encoder-Decoder model to summarize a text document and its accompanying images simultaneously, and then to align the sentences and images in summaries. A multi-modal attentional mechanism is proposed to attend original sentences, images, and captions when decoding. The DailyMail dataset is extended by collecting images and captions from the Web. Experiments show our model outperforms the neural abstractive and extractive text summarization methods that do not consider images. In addition, our model can generate informative summaries of images.

1 Introduction

Summarizing multi-modal documents to get multi-modal summaries is becoming an urgent need with rapid growth of multi-modal documents on the Internet. Text-Image summarization is to summarize a document with text and images to generate a summary with text and images. The summarization approach is different from pure text summarization. It is also different from image summarization which summarizes an image set to get a subset of images.

An image worths thousands of words (Rossiter, et al., 2012). Image plays an important role in information transmission. Incorporating images into text to generate text-image

Figure 1: An example of multi-modal news taken from the DailyMail corpora.

Figure 2: The manually generated text-image summary.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4046–4056
Brussels, Belgium, October 31 - November 4, 2018.

summaries can help people better understand, memorize, and express information. Most of recent research focuses on pure text summarization, or image summarization. Little has been done on text-image summarization.

Figure 1 and Figure 2 show an example of text-image summarization. Figure 1 is the original multi-modal news with text and images. The news has 17 sentences (with 322 words) and 4 images each of which has a caption. Figure 2 is the manually generated multi-modal summary. In the summary, the news is distilled to 3 sentences (with 36 words) and 2 images, and each summary sentence is aligned with an image.

To generate such a text-image summary, the following problems should be considered: How to generate the text part? How to measure the importance of images, and extract important images to form the image summary? How to align sentences with images?

In this paper, we propose a neural text-image summarization model based on the attentional hierarchical Encoder-Decoder model to solve the above problems. The attentional Encoder-Decoder model has been successfully used in sequence-to-sequence applications such as machine translation (Luong et al., 2015), text summarization (Cheng and Lapata, 2016; Tan et al., 2017), image captioning (Liu et al., 2017a), and machine reading comprehension (Cui et al., 2016).

At the encoding stage, we use the hierarchical bi-directional RNN to encode the sentences and the text document, use the RNN and the CNN to encode the image set. In the decoding stage, we combine text encoding and image encoding as the initial state, and use the attentional hierarchical decoder which attends original sentences, images and captions to generate the text summary. Each generated sentence is aligned with a sentence, an image, or a caption in the original document. Based on the alignment scores, images are selected and aligned with the generated sentences. In the inference stage, we adopt the multi-modal beam search algorithm which scores beams based on bigram overlaps of the generated sentences and the attended captions.

The main contributions are as follows:

1) We propose the text-image summarization task, and extend the standard DailyMail corpora by collecting images and captions of each news from the Web for the task.
2) We propose an RNN model to encode the ordered image set of the multi-model document as one of the initial states (the other is the text encoding) of the decoder.
3) We propose three multi-modal attentional mechanisms which attend the text and the images simultaneously when decoding.
4) Experiments show that attending images when decoding can improve text summarization, and that our model can generate informative image summaries.

2 Related Work

Recent research on text summarization focuses on neural methods. Attentional Encoder-Decoder model is first proposed in (Bahdanau et al., 2014) and (Luond et al., 2015) to align the original text and the translated text in machine translation. The attention model is applied to sentence summarization by considering the neural language model and the attention model when generating next words (Rush et al., 2015). A selective Encoder-Decoder model that uses a selective gate network to control information from the encoder to the decoder for sentence summarization is proposed (Zhou et al., 2017).

A neural document summarization model by extracting sentences and words is proposed (Cheng and Lapata, 2016). They use a CNN model to encode sentences, and then use a RNN model to encode documents. The model extracts sentences by computing the probability of sentences belonging to the summary based on an RNN model. The model extracts words from the original document based on an attentional decoder. An RNN-based extractive summarization named SummaRuNNer, treating summarization as a sentence classification problem is proposed (Nallapati et al., 2016). A logistic classifier is then applied using features computed based on the RNN model. A hierarchical Encoder-Decoder model, conserving the hierarchical structure of documents is proposed (Li et al., 2015). A graph-based attentional Encoder-Decoder model using a PageRank algorithm to compute the attention is proposed (Tan et al., 2017).

Image captioning generates a caption for an image. Text-image summarization is similar to image captioning in that both utilize image information to generate text. Images are encoded with CNN models such as VGGNet (Simonyan

and Zisserman, 2014), AlexNet (Krizhevsky et al., 2012) and GoogleNet (Szegedy et al., 2014) by extracting the last full-connected layers. An attentional model is used in image captioning by splitting an image into multiple parts which is attended in the decoding process (Xu et al., 2015). Image tags was used as additional information, and semantic attention model which attends image tags when decoding was proposed（You et al., 2016). The attention-based alignment of image parts and text is studied（Liu et al., 2017a), and the results show that the alignments is in high accordance with manual alignments. An image to an ordered recognized object set is encoded, and the attentional decoder is applied to generate captions（Liu et al., 2017b).

Multi-modal summarization summarizes text, images, videos, and etc. It is an important branch of automatic summarization. Traditional multi-modal summarization inputs multi-modal documents or pure text documents, and outputs multi-modal documents (Wu, 2011; Greenbacker, 2011; Yan, 2012; Agrawal, 2011; Zhu, 2007; UzZaman, 2011). For example, Yan et al., (2012) generate multi-modal timeline summaries for news sets by constructing a bi-graph between text and images, and apply a heterogeneous reinforcement ranking algorithm. Strategies to summarizing texts with images and the notion of summarization of things are proposed in (Zhuge, 2016). The deep learning related work (Wang et al. 2016) treats text summarization as a sentence recommendation task and applies matrix factorization algorithm. They first retrieve images from Yahoo!, use the CNN to extract image features as the additional information of sentences, use Rouge maximization as the training object function which are trained with SGD. In test time, sentences are extracted based on the model and images are retrieved from the Search Engine.

3 Method

Figure 3 shows the framework, a multi-modal attentional hierarchical encoder-decoder model. The hierarchical encoder-decoder is proposed in (Li et al., 2015) and extended by (Tan et al. , 2017) for document summarization through bringing in the graph-based attentional model.

Our model consists of three parts: a hierarchical RNN to encode the original sentences and the captions, a CNN+RNN encoder to encode the image set, and a multi-modal attentional hierarchical RNN decoder.

The input of our model is a multi-modal document $MD = \{D, PicSet\}$, where D is the main text of the multi-modal document and *PicSet* is the image-caption set ordered by the occurring order of images in the document.

3.1 Main Text Encoder

The main text D consists of sentences, each of which consists of words. Let $D=[s_1, s_2, \ldots, s_{|d|}]$, and $s_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,|s_i|}]$ where $x_{i,j}$ is the word embedding of the j^{th} word in the s_i. We use word2vec (Mikolov et al., 2013) to create word embeddings. GRU is used as the RNN cell (Cho et al., 2014).

We use a hierarchical RNN encoder to encode the main text D to vector representation. The sentence encoder is adopted to encode sentences to vector representations. An <eos> token is appended to the end of each sentence. A bi-directional RNN is used as the sentence encoder:

$$\vec{h}_{i,j} = GRU^{\vec{s}}(\vec{h}_{i,j-1}, x_{i,j}) \quad (1)$$

$$\overleftarrow{h}_{i,j} = GRU^{\overleftarrow{s}}(\overleftarrow{h}_{i,j}, x_{i,j+1}) \quad (2)$$

$$enc^{sent}{}_i = [\vec{h}_{i,1}, \overleftarrow{h}_{i,-1}] \quad (3)$$

where $enc^{sent}{}_i$ denotes the vector representation of s_i. It is the concatenation of $\vec{h}_{i,1}$ and $\overleftarrow{h}_{i,-1}$.

We use $enc^{sent}{}_i$ as inputs to the document encoder to encode the main text to vector representations. A bi-directional RNN is adopted as the document encoder:

$$\vec{h}_i = GRU^{\vec{d}}(\vec{h}_{i-1}, enc^{sent}{}_i) \quad (4)$$

$$\overleftarrow{h}_i = GRU^{\overleftarrow{d}}(\overleftarrow{h}_{i+1}, enc^{sent}{}_i) \quad (5)$$

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

$$enc^{doc}{}_d = [\vec{h}_1, \overleftarrow{h}_{-1}] \quad (7)$$

where enc^{doc} denotes the vector representation of the D, and h_i is the concatenated hidden state of s_i.

3.2 CaptionSet and ImageSet Encoder

The ordered image-caption set *PicSet* consists of an ordered image set and an ordered caption set

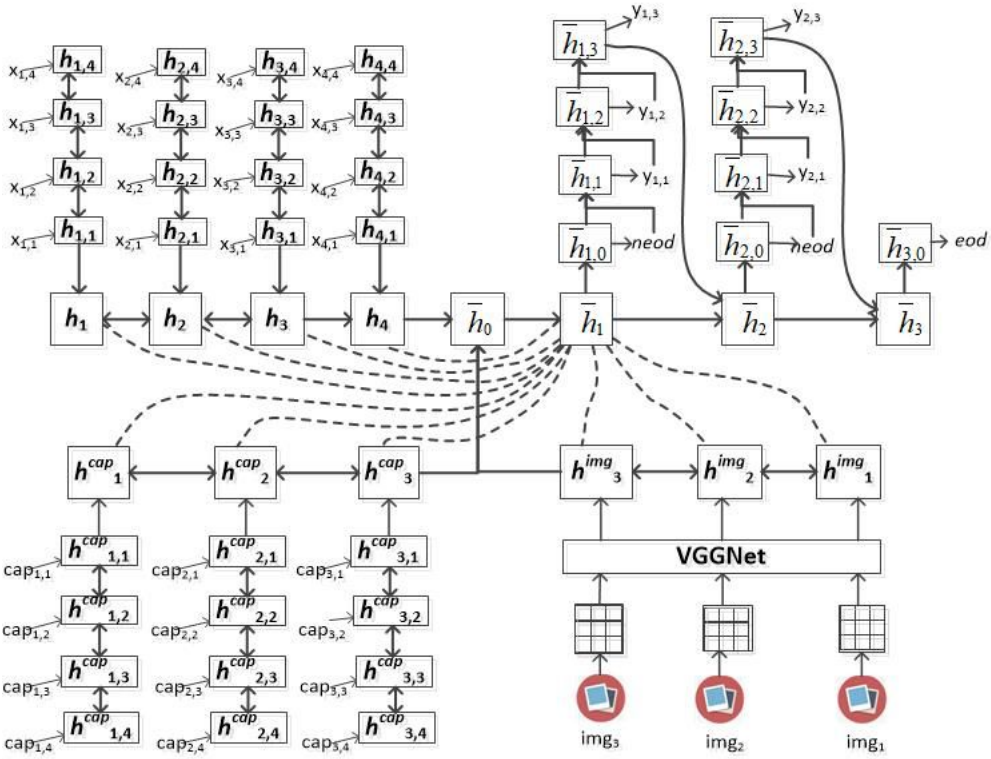

Figure 3: The framework of our neural text-image summarization model.

which are ordered by the occurring order in the multi-modal document. The image occurring order makes sense because images are often put near the most related sentences, and the sentences have strict order in the document.

We treat the ordered caption set as a document, and apply the sentence encoder and the document encoder to the caption document. Then, we get the hidden state $h^{cap}{}_i$ and the vector representation enc^{cap} of the caption document.

We use the CNN model to extract the vector representation of each image, and then use the RNN model to encode the ordered image set to vector representation. The CNN model we adopted is 19-layer VGGNet (Simonyan and Zisserman, 2014). We drop the last dropout layer and keep the last full-connected layer as the image's vector representation, the dimension of which is 4096.

We then use a bi-directional RNN model to encode the ordered image set and the image features are used as inputs of the RNN model.

$$\vec{h}^{img}{}_i = GRU^{\overrightarrow{img}}(\vec{h}^{img}{}_{i-1}, img^{fea}{}_i) \quad (8)$$

$$\overleftarrow{h}^{img}{}_i = GRU^{\overleftarrow{img}}(\overleftarrow{h}^{img}{}_{i+1}, img^{fea}{}_i) \quad (9)$$

$$h^{img}{}_i = [\vec{h}^{img}{}_i, \overleftarrow{h}^{img}{}_i] \quad (10)$$

$$enc^{img} = [\vec{h}^{img}{}_1, \overleftarrow{h}^{img}{}_{-1}] \quad (11)$$

where $img^{fea}{}_i$ is the vector representation of img_i, enc^{img} is the vector representation of the image set, and $h^{img}{}_i$ is the hidden state of img_i when encoding the image set.

To our best knowledge, we are the first to adopt the RNN model to encode the image set.

3.3 Decoder

In the decoding state, we adopt the hierarchical RNN decoder to generate text summaries.

$$\begin{aligned} \bar{h}_0 = \tanh(&W^{dec\_doc} \times enc^{doc} \\ &+ V^{dec\_img} \times enc^{img} \\ &+ V^{dec\_cap} \times enc^{cap}) \end{aligned} \quad (12)$$

$$\tilde{h}_i = GRU^{dec\_sent1}(\bar{h}_{i-1}, \bar{h}_{i-1,-1}) \quad (13)$$

$$\bar{h}_i = GRU^{dec\_sent2}(\tilde{h}_i, c_i) \quad (14)$$

$$\bar{h}_{i,j} = GRU^{dec\_word}(\bar{h}_{i,j-1}, y_{i,j-1}) \quad (15)$$

$$y_{i,j} \sim soft\max(W^{soft\max} \bar{h}_{i,j} + b) \quad (16)$$

Equation (12) to (16) are the equations of the hierarchical decoder which consists of a sentence decoder and a word decoder.

Equation (12) computes the initial state $\overline{h}_0$ for the sentence decoder by combining the decoding of the main text information and the decoding of the image information of the multi-modal document. To represent image information, we can use both of the image set decoding and the caption set decoding, or only use one of them, depending on the multi-modal attention mechanism introduced in the next subsection.

The sentence decoder uses a two-level hidden output model (Luong et al., 2015) to generate the representation of the next sentence through equation (13) and equation (14). The two-level hidden output model consistently improves the summarization performance on different datasets (Chen et al., 2016). In equation (14), the two-level model computes $\overline{h}_i$ by capturing a direct interaction between $\tilde{h}_i$ and c_i. $\tilde{h}_i$ is computed by equation (13) using the preceding sentence decoder hidden state $\overline{h}_{i-1}$ and the word decoder hidden state $\overline{h}_{i-1,-1}$ which is the last hidden state of the preceding word decoder. And c_i is the context of the sentence decoder computed based on the multi-modal attention model.

The word decoder uses the sentence representation generated by the sentence decoder as the initial state, and use the <sos> (start of sentence) token as the initial input. Equation (15) and equation (16) generate the next hidden state and the next word. The output of the word decoder in the first step is a switch sign which is either <neod> token or <eod> token. The token <neod> means "not end of document", and the token <eod> means "end of document". If the first output is <eod>, the whole decoding process is finished. If the first output is <neod>, the token is used as the next input of the word decoder. The word decoding process is finished when it generates the <eos> token. The last hidden state of the word decoder is treated as the vector representation of the generated sentence and is used as next input of the sentence decoder.

3.4 Multi-Modal Attention

We propose three multi-modal attention mechanisms to compute the sentence decoding context c_i.

Traditional attention mechanisms for text summarization computes the importance score of the sentence s_j in the original document based on the relationship between the decoding hidden state $\tilde{h}_i$ and the original sentence encoding hidden state h_j. We call the traditional attention model as **Text Attention (*attT* for short)**, which is computed by equation (17), (18) and (19):

$$att^T(\tilde{h}_i, h_j) = v^T \tanh(W^T \tilde{h}_i + U^T h_j) \quad (17)$$

$$\alpha^T(\tilde{h}_i, h_j) = \frac{\exp(att^T(\tilde{h}_i, h_j))}{\sum_{j=1}^{|D|} \exp(att^T(\tilde{h}_i, h_j))} \quad (18)$$

$$c^T(\tilde{h}_i) = \sum_{j=1}^{|D|} \alpha^T(\tilde{h}_i, h_j) h_j \quad (19)$$

where $att^T(\tilde{h}_i, h_j)$ is the attention (Banahama et al., 2014), $\alpha^T(\tilde{h}_i, h_j)$ is the normalized attention, and $c^T(\tilde{h}_i)$ is the context.

The problem is that the multi-modal document has images and captions besides the main text. Therefore, we propose three multi-modal attention mechanisms which take images and captions into consideration.

Text-Caption Attention (*attTC* for short). This attention model uses captions to represent the image information. *attTC* computes the attention score of the caption cap_j based on the relationship between the caption encoding hidden state $h^{cap}{}_i$ and the decoding hidden state $\tilde{h}_j$.

$$\alpha^{TC}(\tilde{h}_i, h_j) = \frac{\exp(att^{TC}(\tilde{h}_i, h_j))}{\sum_{j=1}^{|D|} \exp(att^{TC}(\tilde{h}_i, h_j)) + \sum_{j=1}^{|PicSet|} \exp(att^{TC}(\tilde{h}_i, h^{cap}{}_j))} \quad (20)$$

$$\alpha^{TC}(\tilde{h}_i, h^{cap}{}_j) = \frac{\exp(att^{TC}(\tilde{h}_i, h^{cap}{}_j))}{\sum_{j=1}^{|D|} \exp(att^{TC}(\tilde{h}_i, h_j)) + \sum_{j=1}^{|PicSet|} \exp(att^{TC}(\tilde{h}_i, h^{cap}{}_j))} \quad (21)$$

$$c^{TC}(\tilde{h}_i) = \sum_{j=1}^{|D|} \alpha^{TC}(\tilde{h}_i, h_j) h_j + \sum_{j=1}^{|PicSet|} \alpha^{TC}(\tilde{h}_i, h^{cap}{}_j) h^{cap}{}_j \quad (22)$$

Text-Image Attention (*attTI* for short). This attention model only uses images to represent the image information neglecting the captions. *attTI* computes the importance score of the image img_j based on the relationship between the image encoding hidden state $h^{img}{}_i$ and the decoding hidden state $\tilde{h}_j$.

$$\alpha^{TI}(\tilde{h}_i, h^{img}{}_j) = \frac{\exp(att^{TI}(\tilde{h}_i, h^{img}{}_j))}{\sum_{j=1}^{|D|}\exp(att^{TI}(\tilde{h}_i, h_j)) + \sum_{j=1}^{|PicSet|}\exp(att^{TI}(\tilde{h}_i, h^{img}{}_j))} \quad (23)$$

$$c^{TI}(\tilde{h}_i) = \sum_{j=1}^{|D|}\alpha^{TI}(\tilde{h}_i, h_j)h_j + \sum_{j=1}^{|PicSet|}\alpha^{TI}(\tilde{h}_i, h^{img}{}_j)h^{img}{}_j \quad (24)$$

Text-Image-Caption Attention (*attTIC* for short). This attention model uses both captions and images to represent the image information. *attTIC* computes the importance score of the caption cap_j and the importance score of the image img_j simultaneously, and then compute the context of the decoding hidden state $\tilde{h}_i$ using equation (25).

$$c^{TIC}(\tilde{h}_i) = \sum_{j=1}^{|D|}\alpha^{TIC}(\tilde{h}_i, h_j)h_j + \sum_{j=1}^{|PicSet|}\alpha^{TIC}(\tilde{h}_i, h^{cap}{}_j)h^{cap}{}_j + \alpha^{TIC}(\tilde{h}_i, h^{img}{}_j)h^{img}{}_j \quad (25)$$

In the attention mechanisms, $\alpha(\tilde{h}_i, h_j)$ is the normalized attention score of h_j, $\alpha(\tilde{h}_i, h^{cap}{}_j)$ is the normalized attention score of $h^{cap}{}_j$, $\alpha(\tilde{h}_i, h^{img}{}_j)$ is the normalized attention score of $h^{img}{}_j$, and $c(\tilde{h}_i)$ is the context.

The initial state of the decoder is computed by Equation (12) which can be adjusted according to different attention models.

3.5 Model Training

Since there are no existing manual text-image summaries, and most of the existing training and testing data have pure text summaries, we decide to use pure text summaries as training data to train our models. **The sentence-image alignment relationships can be discovered through training the multi-modal attention models.**

The loss function L of our summarization model is the negative log likelihood of generating text summaries over the training multi-modal document set *MDS*.

$$L = \sum_{(D, PicSet, Y)\in MDS} -\log P(Y \mid D, PicSet) \quad (26)$$

where Y=[y_1, y_2, …, $y_{|Y|}$] is the word sequences of the summary corresponding to the main text D and the ordered image set *PicSet*, including the tokens <eos>, <neod> and <eod>.

$$\log P(Y \mid D, PicSet) = \sum_{t=1}^{|Y|}\log P(y_t \mid \{y_1, ..., y_{t-1}\}, c; \theta) \quad (27)$$

where $\log P(y_t \mid \{y_1, ..., y_{t-1}\}, c; \theta)$ is modeled by the multi-modal encoder-decoder model.

We use the Adam (Kingma and Ba, 2014) gradient-based optimization method to optimize the model parameters.

3.6 Multi-Modal Beam Search Algorithm

There are two major problems of the generation of summaries: one is the out-of-vocabulary problem, and the other is the low quality of the generated texts including information incorrectness and repetitions.

For the OOV problem, we use the words in the attended sentences or captions in the original document to replace OOV tokens in the generated summary. Previous research uses the attended words to replace OOV tokens in the flatten encoder-decoder model which attends the words of the original word sequence (Jean, et al., 2015). Our model is hierarchical and multi-modal, and attends sentences, images, and captions when decoding. We use the following algorithm to find the replacement for the j^{th} OOV in a generated sentence:

Step 1: Order the original sentences and captions by the attending scores in descending order.

Step 2: Return the j^{th} OOV word in the ordered sentences and captions as the replacement.

For the *attTI* mechanism that attends images neglecting captions, we use captions instead of the attended images in the algorithm.

For the low-quality generated text problem, we adopt the hierarchical beam search algorithm (Tan el al., 2017). We extend the algorithm by adding caption-level and image-level beam search. The multi-modal hierarchical beam search algorithm comprises *K*-best word-level beam search and *N*-best sentence-caption-level beam search. In particular, we use the corresponding captions instead of images in beam search algorithm for the *attTI* mechanism which attends images.

$$score(y_t) = p(y_t) + \gamma(ref(Y_{t-1} + y_t, s_*) - ref(Y_{t-1}, s_*)) \quad (28)$$

At the word-level search algorithm, we compute the score of generating word y_t using equation (28) where *ref* is a function calculating the ratio of bigram overlap between two texts, s_* is the attended sentence or caption, and γ is the weighting factor. The added term aims to increase the overlap of the generated summary and the original text.

At the sentence level and the caption level, we set the sentence beam width as *N*, and keep *N*-best previously un-referred sentences or captions which have highest attending scores. For each sentence beam, we try *M* sentences or captions and keep the one achieving best word-level scores.

3.7 Image Selection and Alignment

We rank the images, select several most important images as the image summary, and align each sentence with an image in the image summary. The score of images is computed by equation (29).

$$score(img_j) = \sum_{i=1}^{|TextSum|} \alpha_{i,j} \quad (29)$$

where $\alpha_{i,j}$ is the attention score of the j^{th} image when generating the i^{th} sentence of the text summary, and $|TextSum|$ is the number of summary sentences.

The images are ranked by the scores in descending order, and the top K images are selected to form the image summary *ImgSum*. We align each sentence i in *TextSum* to the image j in *ImgSum* such that $\alpha_{i,j}$ is the biggest.

4 Experiments

4.1 Data preparation

We extend the standard DailyMail corpora through extracting the images and the captions from the html-formatted documents. We call the corpora as E-DailyMail. The standard DailyMail and CNN datasets are two widely used datasets for neural document summarization, which are originally built in (Hermann et al., 2015) by collecting human generated highlights and news stories from the news websites. We only extend the DailyMail dataset because it has more images and is easier to collect than the CNN dataset does. We find that the text documents provided by the original DailyMail corpora contain captions. This is due to that all related texts are extracted from the html-formatted news when the corpora are created. We keep the original text documents unchanged in E-DailyMail. The split and statistics of E-DailyMail are shown in Table 1.

4.2 Implementation

We preprocess the text of the E-DailyMail corpora by tokenizing the text and replacing the digits with the <NUM> token. The 40k most frequent words in the corpora are kept and other words are replaced with OOV.

Our model is implemented by using Google's open-source seq2seq-master project written with Tensorflow. We use one layer of the GRU cell. The dimension of the hidden state of the RNN decoder is 512. The dimension of the word embedding vector is 128. The dimension of the hidden state of the bi-directional RNN encoder is 256. We initialize the word embeddings with Google's word2vec tools (Mikolov et al., 2013) trained in the whole text of DailyMail/CNN corpora. We extract the 4096-dimension full-connected layer of 19-layer VGGNet (Simonyan and Zisserman, 2014) as the vector representation of images. We set the parameters of Adam to those provided in (Kingma and Ba, 2014). The batch size is set to 5. Convergence is reached within 800k training steps. It takes about one day for training 40k ~ 50k steps depending on the models on a GTX-1080 TI GPU card. The sentence beam width and the word beam width are set as 2 and 5 respectively. M is set as 3. The parameter γ is set as 3 or 300 tuned on the validation set.

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

Table 1: The split and 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 number of word in the sentence and the caption respectively.

To train the multi-modal attention mechanism such as *attTIC*, we concatenate the matrix of text representations, image representations, and caption representations to one matrix $M = [h_1, h_2, ... h_{|D|}, h^{cap}_1, h^{cap}_2, \ldots, h^{cap}_{|PicSet|}, h^{img}_1, h^{img}_2, \ldots, h^{img}_{|PicSet|}]$. The parameters of the attention mechanisms are trained simultaneously. This way the model training can converge faster.

4.3 Evaluation of Text Summarization

The widely used ROUGE (Lin, 2004) is adopted to evaluate text summaries.

We compare four attention models. *HNNattTC*-3, *HNNattTIC*-3, *HNNattTI*-3, and *HNNattT*-3 are our hierarchical RNN summarization models with the *attTC*, *attIC*, *attTI*, and *attT* attention mechanisms respectively, and 3 is the γ value. *HNNattT* is similar to the model introduced in (Tan et al., 2017) without the graph-based attention. We compare our models with *HNNattT* to show the influence of multi-modal attentions. The first 4 lines in Table 2 are the results with summary length of 75 bytes. The results show that *HNNattTI* has considerable improvement over *HNNattT*. An interesting observation is that *HNNattTC* and *HNNattTIC* are not better than *HNNattT*. One of the

Method	Rouge-1	Rouge-2	Rouge-L
***HNNattTI*-3**	**24.84**	8.7	**16.99**
HNNattTC-3	18.61	6.7	13.44
HNNTattTIC-3	21.17	8.1	15.24
HNNattT-3	22.09	7.9	15.97
Lead	21.9	7.2	11.6
NN-SE	22.7	8.5	12.5
SummaRuNNer-abs	23.8	9.6	13.3
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 2: Comparison results on the DailyMail test set using Rouge recall at 75 bytes.

Method	Rouge-1	Rouge-2	Rouge-L
***HNNattTI*-300**	**32.64**	**12.02**	**23.88**
HNNattTC-300	26.75	10.12	19.42
HNNattTIC-300	30.52	11.04	21.81
HNNattT-300	31.34	11.81	22.93

Table 3: Comparison results on the DailyMail test set using full-length F1metric.

reasons is that the text documents provided by the DailyMail corpora contain captions. Captions are already parts of the text documents. The other reason is that captions distract attentions and cannot attract sufficient attentions from the original sentences, which will be discussed in the next subsection.

We compare our methods with state-of-the-art neural summarization methods reported in recent papers on the DailyMail corpora. Extractive models include Lead which is a strong baseline using the leading 3 sentences as the summary, NN-SE (Cheng and Lapata, 2016), and *SummaRuNNer-abs* (Nallapati et al., 2017) which is trained on the abstractive summaries. Abstractive models include *NN-ABS*, *NN-WE*, *LREG*, though they are tested on 500 samples of the test set. *LREG* is a feature-based method using linear regression. *NN-ABS* is a simple hierarchical extension of (Rush et al., 2015). *NN-WE* is the abstractive model restricting the generation of words from the original document. The results are shown in the last 6 rows in Table 2. Our method *HNNattTI* outperforms the three extractive models and the three abstractive models.

We compare our models under the full-length F1 metric by setting the γ value as 300. According to (Tan et al., 2017), a large γ makes the generated summary has more overlaps with the attended texts, and thus partly overcome the repeated sentences problem in the generated summary. We do not incorporate the attention distraction mechanism (Chen et al., 2016) into our model, because we want to focus on our own model to see whether considering images improves text summarization. Results in Table 3 also show that *HNNattTI* performs better than *HNNattT*, *HNNattTC*, and *HNNattTIC*.

Method	Rouge-1	Rouge-2	Rouge-L
***HNNattTI*-3-OOV**	24.03	8.2	16.52
HNNattTC-3-OOV	18.18	6.53	12.87
HNNTattTIC-3-OOV	20.50	7.67	14.36
HNNattT-3-OOV	21.60	7.82	15.05

Table 4: Comparison results using Rouge recall at 75 bytes without OOV replacement. *HNNattTI*-3-OOV is the version of *HNNattTI*-3 without the OOV replacement mechanism.

Method	Rouge-1	Rouge-2	Rouge-L
***HNNattTI*-300-OOV**	32.03	11.52	22.67
HNNattTC-300-OOV	26.13	9.87	19.03
HNNattTIC-300-OOV	30.11	10.87	21.12
HNNattT-300-OOV	30.74	11.21	22.28

Table 5: Comparison results using full-length F1 metric without OOV replacement. *HNNattTI*-300-OOV is the version of *HNNattTI*-300 without the OOV replacement mechanism.

To show the influence of our OOV replacement mechanism, we eliminate the mechanism from our models, and show the evaluation results in Table 4 and Table 5. We can see from the two tables that the scores are lower than the corresponding scores in Table 2 and Table 3. Our OOV replacement mechanism improves the summarization models, though the mechanism is relatively simple.

In short, combining and attending images in the neural summarization model improves document summarization.

4.4 Evaluation of Image Summarization

To evaluation the image summarization, the gold standard image summary is generated based on 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 ground truth text summary and

num	*HNNattTI*	*HNNattTC*	*HNNattTIC*	*Random*
1	**0.4978**	0. 4137	0. 4362	0. 4721
2	**0.4783**	0. 3998	0. 4230	0. 4517

Table 6: Image summarization using the recall metric for the 1-image or 2-images summary. γ is set as 300.

cap_k is the caption of img_k. The average number of images in summaries is 2.15. The average Rouge-1, Rouge-2, and Rouge-L scores of the caption summaries with respect to the ground truth summaries are 43.85, 19.70, and 36.30 respectively.

We use the 1-image and 2-image random selected image summaries as the baselines which we compare our models with. The top 1 or 2 images ranked by our model are selected out to form the summaries. Results in Table 4 show that *HNNattTI* outperforms the random baseline, while *HNNattTC* and *HNNattTIC* perform worse. This implies that attending images can generate better sentence-image alignment in the multi-modal summaries than the model attending captions does. And this can also partly explain why our summarization model attending images when decoding can generate better text summaries than the one attending captions does.

4.5 Instance

Figure 4 shows the text-image summary of the example demonstrated in Figure 1 generated by the *HNNattTI* model. In the summary, there are 2 images and 3 generated sentences, and each sentence is aligned with an image. The image summary has one common image with Figure 2.

Figure 4: The generated text-image summary of the example in Figure 1.

	IMG1	IMG2	IMG3	IMG4
S1	0.0947	0.1089	0.1157	0.1194
S2	0.0893	0.1020	0.1070	0.1052
S3	0.0853	0.0769	0.0946	0.0969

Table 7: The sentence-image alignment scores of the generated summary for the news in Figure 1.

The sentences are named by S1, S2, and S3 respectively.

Table 7 shows the sentence-image alignment scores. The four images in the original document are numbered from top to bottom and left to right by IMG1, IMG2, IMG3, and IMG4. The summation of alignment scores for a summary sentence is less than 1, because the sentence is also aligned with the sentences in the original document.

5 Conclusions

This paper proposes the text-image summarization task to summarize and align texts and images simultaneously. Most previous research summarizes texts and images separately, and few has been done on text-image summarization. We propose the multi-modal attentional mechanism which attends original sentences, images, captions simultaneously in the hierarchical encoder-decoder model, use the RNN model to encode the ordered image set as the initial state of the decoder, and propose the multi-modal beam search algorithm which scores beams using the bigram overlaps of the generated sentences and the captions. The model is trained by using abstractive text summaries as the targets, and the attention scores of images are used to score images. The original DailyMail dataset is extended by collecting images and captions from the Web. Experiments show that our model attending images outperforms the models not attending images, three existing neural abstractive models and three existing extractive models. Experiments also show our model can generate informative summaries of images.

Acknowledgments

The research was sponsored by the National Natural Science Foundation of China (No.61806101, No.61876048) and the Natural Science Foundation of Jiangsu Province (BK20150862). We thank the anonymous reviewers for helpful comments. Professor Hai Zhuge is the corresponding author.

References

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2011). Enriching textbooks with images. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1847-1856). ACM.

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

Chen, Q., Zhu, X., Ling, Z., Wei, S., and Jiang, H. (2016, July). Distraction-based neural networks for modeling documents. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 2754-2760). AAAI Press.

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

Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T. and Hu, G., 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). IEEE.

Gu, J., Lu, Z., Li, H. and Li, V.O., 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1631-1640).

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

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).

Jean, S., Cho, K., Memisevic, R. and Bengio, Y., 2014. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007*.

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

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

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

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

Liu C., Mao J., Sha F., Yuille AL. 2017a. Attention Correctness in Neural Image Captioning. InAAAI 2017 Feb 4 (pp. 4176-4182).

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

Luong, M. T., Hieu P., and Christopher D. M.. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

Mikolov T., Sutskever I., Chen K., et al. Distributed Representations of Words and Phrases and their Compositionality. 2013, 26:3111-3119.

Nallapati R., Zhai F., Zhou B. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. 2016.

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.

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

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

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

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).

UzZaman, N., Bigham, J. P., & Allen, J. F. (2011). Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces* (pp. 43-52). ACM.

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.

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

Xu K., Ba J., Kiros R., et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Computer Science, 2015:2048-2057.

Yan R., Wan X., et al. Visualizing timelines: evolutionary summarization via iterative reinforcement between text and image streams. CIKM2012. ACM, 2012:275-284.

You Q., Jin H., Wang Z., et al. Image Captioning with Semantic Attention. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:4651-4659.

Zhou Q., Yang N., Wei F., et al. Selective Encoding for Abstractive Sentence Summarization. Meeting of the Association for Computational Linguistics. 2017:1095-1104.

Zhu, X., Goldberg, A. B., et al. (2007). A text-to-picture synthesis system for augmenting communication. In *AAAI* (Vol. 7, pp. 1590-1595).

Zhuge, H. Multi-Dimensional Summarization in Cyber-Physical Society, Morgan Kaufmann, 2016.