MSMO: Multimodal Summarization with Multimodal Output

Junnan Zhu[1,2], **Haoran Li**[1,2], **Tianshang Liu**[1,2], **Yu Zhou**[1,2], **Jiajun Zhang**[1,2] and **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{junnan.zhu, yzhou, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Multimodal summarization has drawn much attention due to the rapid growth of multimedia data. The output of the current multimodal summarization systems is usually represented in texts. However, we have found through experiments that multimodal output can significantly improve user satisfaction for informativeness of summaries. In this paper, we propose a novel task, multimodal summarization with multimodal output (MSMO). To handle this task, we first collect a large-scale dataset for MSMO research. We then propose a multimodal attention model to jointly generate text and select the most relevant image from the multimodal input. Finally, to evaluate multimodal outputs, we construct a novel multimodal automatic evaluation (MMAE) method which considers both intramodality salience and intermodality relevance. The experimental results show the effectiveness of MMAE.

1 Introduction

Text summarization is to extract the important information from source documents. With the increase of multimedia data on the internet, some researchers (Li et al., 2016b; Shah et al., 2016; Li et al., 2017) focus on multimodal summarization in recent years. Existing experiments (Li et al., 2017, 2018a) have proven that, compared to text summarization, multimodal summarization can improve the quality of generated summary by using information in visual modality.

However, the output of existing multimodal summarization systems is usually represented in a single modality, such as textual or visual (Li et al., 2017; Evangelopoulos et al., 2013; Mademlis et al., 2016). In this paper, we argue that multimodal output[1] is necessary for the following three reasons: 1) It is much easier and faster for users to get critical information from the images (Li et al., 2017). 2) According to our experiments, the multimodal output (text+image) increases users' satisfaction by 12.4% compared to the single-modality output (text) (more details can be found in Sec. 4.2). 3) Images help users to grasp events while texts provide more details related to the events. Thus the images and text can complement each other, assisting users to gain a more visualized understanding of events (Bian et al., 2013). We give an example in Fig. 1 to illustrate this phenomenon. For the output with only the text summary, user will be confused about the description of "four-legged creatures"; while with a relevant image, user will have a clearer understanding of the text.

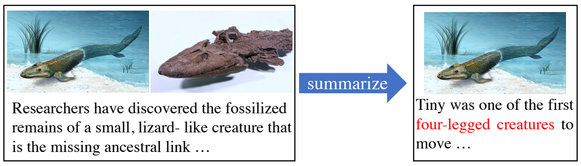
Researchers have discovered the fossilized remains of a small, lizard- like creature that is the missing ancestral link ...
summarize
Tiny was one of the first four-legged creatures to move ...

Figure 1: The illustration of our proposed task – Multimodal Summarization with Multimodal Output (MSMO). The image can help better understand the text in the red font.

In recent years, some researchers(Bian et al., 2013, 2015; Wang et al., 2016) focus on incorporating multimedia contents into the output of summarization which all treat the image-text pair as a basic summarization unit. But in our work, our input comes from a document and a collection of images where there is no alignment between texts and images. So our biggest challenge is how to bridge the semantic gaps between texts and images. Based on the above discussion, in this work, we propose a novel task which we refer to as Multimodal Summarization with Multimodal Output (MSMO). To explore this task, we focus on the

[1] Note that in this work, the multimodal output refers to a pictorial summary which contains one image (for the sake of simplicity, we first consider only one image) and a piece of text. We leave the other multimodal content (like videos) as future work.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4154–4164
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

following three questions: 1) how to acquire the relevant data; 2) how to generate the multimodal output; 3) how to automatically evaluate the quality of the multimodal output in MSMO.

For the first question, similar to Hermann et al. (2015), we collect a large-scale multimodal dataset[2] from *Daily Mail* website and annotate some pictorial summaries. For the second question, we propose a multimodal attention model to jointly generate text and the most relevant image, in which the importance of images is determined by the visual coverage vector. For the last question, we construct a novel multimodal automatic evaluation (MMAE) which jointly considers salience of text, salience of image, and image-text relevance.

Our main contributions are as follows:

- We present a novel multimodal summarization task, which takes the news with images as input, and finally outputs a pictorial summary. We construct a large-scale corpus for MSMO studying.
- We propose an abstractive multimodal summarization model to jointly generate summary and the most relevant image.
- We propose a multimodal automatic evaluation (MMAE) method which mainly considers three aspects: salience of text, salience of image, and relevance between text and image.

2 Our Models

2.1 Overview

We begin by defining the MSMO task. The input of the task is a document and a collection of images and the output is a pictorial summary. As shown in Fig. 2, our proposed model consists of four modules: text encoder, image encoder, multimodal attention layer, and summary decoder. The text encoder is a BiLSTM used to encode text. Our image encoder is VGG19[3] pretrained on ImageNet (Simonyan and Zisserman, 2015) used to extract global or local features. The multimodal attention layer aims to fuse textual and visual information during decoding. Our summary decoder, which is a unidirectional LSTM, makes use of information from two modalities to generate the text summary and select the most relevant image according to visual coverage vector. Our text encoder and summary decoder are based on pointer-generator network which we will describe in Sec. 2.2. We then describe image encoder and multimodal attention layer in our multimodal attention model (Sec. 2.3).

2.2 Pointer-Generator Network

See et al. (2017) propose a pointer-generator network which allows both copying words from the source text and generating words from a fixed vocabulary, achieving the best performance on *CNN/Daily mail* dataset. Their model consists of an encoder (a single-layer bidirectional LSTM) and an attentive decoder (a unidirectional LSTM). The encoder maps the article to a sequence of encoder hidden states h_i. During decoding, the decoder receives the embedding of the previous word and reaches a new decoder state s_t. Then the context vector c_t is computed by the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) as calculated in Eq. 1 and 2. To alleviate the problem of repetition, See et al. (2017) maintain a coverage vector cov^t, which is the sum of attention distributions over all previous decoding timesteps (initialized to zero vector at timestep 0): $cov^t = \sum_{\tilde{t}=0}^{t-1} \alpha^{\tilde{t}}$. The coverage vector is used as an extra input to the attention vector (Eq. 1) and is also used to calculate the coverage loss (Eq. 6). Next, the attention distribution is used to calculate the context vector as follows.

$$e_i^t = v^T \tanh(\mathbf{W}_h h_i + \mathbf{W}_s s_t + \mathbf{W}_c cov^t) \quad (1)$$

$$\alpha^t = \mathrm{softmax}(e^t) \quad (2)$$

$$c_t = \sum_i \alpha_i^t h_i \quad (3)$$

The important part in this model is the calculation of the generation probability p_g. It represents the probability of generating a word from the vocabulary distribution p_v, and $(1 - p_g)$ represents the probability of copying a word from the source by sampling from the attention distribution α^t. p_g is determined by c_t, s_t, and the decoder input x_t in Eq. 4. The final probability distribution over the extended vocabulary, which denotes the union of the vocabulary and all words in the source, is calculated in Eq. 5. Finally, the loss for timestep t is the sum of the negative log likelihood of the target

[2] Our dataset has been released to the public, which can be found in http://www.nlpr.ia.ac.cn/cip/jjzhang.htm.

[3] http://www.robots.ox.ac.uk/~vgg/research/very_deep

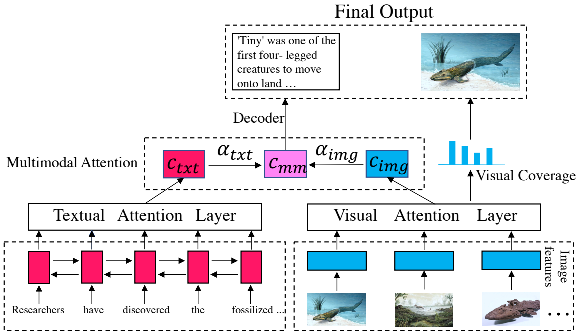

Figure 2: The framework of our model.

word w_t^* and the coverage loss (Eq. 6):

$$p_g = \sigma(\mathbf{W}_h^* c_t + \mathbf{W}_s^* s_t + \mathbf{W}_x x_t) \tag{4}$$

$$p_w = p_g p_v(w) + (1 - p_g) \sum_{w_i = w} \alpha_i^t \tag{5}$$

$$L_t = -\mathrm{log} p_{w_t^*} + \sum_i \min(\alpha_i^t, cov_i^t) \tag{6}$$

2.3 Multimodal Attention Model

We incorporate visual information into the pointer-generator network and propose a novel multimodal attention model. As shown in Fig. 2, there are three main differences between our model and pointer-generator network: 1) We have an extra image encoder and a corresponding visual attention layer; 2) To achieve the fusion of textual and visual information, we introduce a multimodal attention mechanism; 3) We add a visual coverage (Li et al., 2018a) to both alleviate visual repetition and measure the salience of image. More details are as follows.

Image Encoder. We apply the VGG19 to extract global and local image feature vectors for all images. The global features g are 4096-dimensional activations of the pre-softmax fully-connected layer `fc7`. The local features l are the $7 \times 7 \times 512$ feature maps of the last pooling layer (`pool5`). We flatten the local feature into a matrix $\mathbf{A} = (a_1, \cdots, a_L)(L = 49)$ where $a_l \in \mathbb{R}^{512}$ corresponds to a patch of an image.

Visual Attention. The attention mechanism is learned to focus on different parts of input text while decoding. Attention mechanisms have also shown to work with other modalities, like images, where they can learn to attend the salient parts of an image (Xu et al., 2015). We then explore to use images with a visual attention to learn text-image alignment. Concretely, we extend attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) to visual attention mechanisms, which attend visual signals. There are three variants of our visual attention mechanisms: 1) attention on global features (ATG), 2) attention on local features (ATL), and 3) hierarchical visual attention on local features (HAN). We take the calculation of ATG as an example. To attend to the salient parts of a collection of images with size M, we flatten the global feature set g into a matrix $g\prime = (g_1, \cdots, g_M)$. In addition to calculating the text context vector in Sec. 2.2, we also obtain a visual context vector. We first project the image feature into the same dimension as the text context vector. The visual attention is calculated as follows:

$$g^* = \mathbf{W}_I^2(\mathbf{W}_I^1 g + b_I^1) + b_I^2 \tag{7}$$

$$e_a^t = v_a^T \tanh(\mathbf{W}_a g_i^* + \mathbf{U}_a s_t + cov_a^t) \tag{8}$$

$$\alpha_a^t = \mathrm{softmax}(e_a^t) \tag{9}$$

where $\mathbf{W}_I^1 \in \mathbb{R}^{4096 \times 4096}$ and $\mathbf{W}_I^2 \in \mathbb{R}^{4096 \times d_h}$ are the image transformation matrices, $b_I^1 \in \mathbb{R}^{4096}$ and $b_I^2 \in \mathbb{R}^{d_h}$ are bias vectors, and cov_a^t denotes the visual coverage vector and is initialized to zero vector in the beginning. Then the visual attention distribution α_a^t is used to obtain the visual context vector c_{img}^t through $c_{img}^t = \sum_i \alpha_{a,i}^t g_i^*$. Similar is the ATL, we flatten the local feature set A into a matrix $A' = (a_1, \cdots, a_{M \times 49})$. The calculation of attention in ATL is the same as in ATG. There is a bit difference in the HAN model, which first attend to the 49 image patches and get an intermediate visual context vector to represent the image, and then attend to the intermediate visual context vectors to get the visual context vector.

Multimodal Attention. To fuse the text and visual context information, we add a multimodal attention layer (Li et al., 2018a), as shown in Fig. 2. And the attention distribution is calculated as follows:

$$e_{txt}^t = v_{txt}^T(\mathbf{W}_{txt} c_{txt}^t + \mathbf{U}_{txt} s_t) \tag{10}$$

$$e_{img}^t = v_{img}^T(\mathbf{W}_{img} c_{img}^t + \mathbf{U}_{img} s_t) \tag{11}$$

$$\alpha_{txt}^t = \mathrm{softmax}(e_{txt}^t) \tag{12}$$

$$\alpha_{img}^t = \mathrm{softmax}(e_{img}^t) \tag{13}$$

$$c_{mm}^t = \alpha_{txt}^t c_{txt}^t + \alpha_{img}^t c_{img}^t \tag{14}$$

where α_{txt}^t is the attention weight for text context vector and α_{img}^t is the attention weight for visual context vector.

Visual Coverage. In addition to the calculation of the text coverage vector as in Sec. 2.2, we also obtain a visual coverage vector cov_{img}^t, which is the sum of visual attention distributions. To help reduce repeated attention to multimodal information, we incorporate a text coverage loss and a visual coverage loss into the loss function. The final loss function is as follows:

$$L_t = -\text{log}p_{w_t^*} + \sum_i \min(\alpha_i^t, cov_i^t) + \sum_j \min(\alpha_j^t, cov_{img,j}^t) \quad (15)$$

The attention mechanism can attend the salient parts of texts or images. Meanwhile, the coverage mechanism sums up all the historical attention distributions. Therefore, we regard the coverage vector as a global salience measure of the source being attended. We then use the visual coverage vector in the last decoding timestep to select the most relevant image. Concretely, we choose the image whose coverage score is the largest. The process is a bit different for the local features. An image corresponds to 49 patches, the coverage scores of these patches are summed up to get the salience score of the image as follows:

$$\mathrm{S}_j = \sum_{patch} cov_{patch,j}^{t^*} \quad (16)$$

where $\mathrm{S_j}$ denotes the salience of the j-th image and $cov_{patch,j}^{t^*}$ denotes the coverage score of each corresponding image patch in the last decoding timestep $t*$. For the HAN, we introduce an extra coverage vector for the image patches attention and calculate coverage loss for it as follows:

$$L_t = -\text{log}p_{w_t^*} + \sum_k \min(\alpha_k^t, cov_{patch,k}^t) + \sum_i \min(\alpha_i^t, cov_i^t) + \sum_j \min(\alpha_j^t, cov_{img,j}^t) \quad (17)$$

3 Multimodal Automatic Evaluation

To evaluate the quality of a pictorial summary, we propose the MMAE method which is defined as $y = f(m_1, m_2, m_3)$. In this definition, m_1, m_2, and m_3 denote scores measured by three metrics which consider salience of text (Sec. 3.1), salience of image (Sec. 3.2), and image-text relevance (Sec. 3.3) respectively, $f(\cdot)$ denotes a mapping function, and y denotes the score of the pictorial summary.

In our experiments, the reference pictorial summary consists of a text summary and a reference image set[4] $\mathrm{ref_{img}}$. In MMAE, m_1 is obtained by comparing the text summary in reference with that in model output, m_2 is obtained by comparing the image set in reference with the image in model output, and m_3 considers the image-text similarity in model output. To learn MMAE, we choose three simple methods to fit y with human judgment scores. These methods include Linear Regression (LR), and two nonlinear methods: Logistic Regression (Logis), and Multilayer Perceptron (MLP).

3.1 Salience of Text

ROUGE (Lin, 2004b) is widely used to automatically assess the quality of text summarization systems. It has been shown that ROUGE correlates well with human judgments (Lin, 2004a; Owczarzak et al., 2012; Over and Yen, 2004). Therefore, we directly apply ROUGE to assess the salience of the text units.

3.2 Salience of Image

We propose a metric, namely, image precision (**IP**), to measure the salience of image. The image precision is defined as follows:

$$\mathrm{IP} = \frac{|\{\mathrm{ref_{img}}\} \cap \{\mathrm{rec_{img}}\}|}{|\{\mathrm{rec_{img}}\}|} \quad (18)$$

where ref_{img}, rec_{img} denote reference images and recommended images by MSMO systems respectively. The reasons for this metric are as follows.

A good summary should have good coverage of the events for both texts and images. The image in the output should be closely related to the events. So we formulate the image selection process as an image recommendation —instead of recommending items to users as in a recommendation system, we recommend the most salient image to an event. It can also be viewed as an image retrieval task, which retrieves the image most relevant to an event. Precision and recall are commonly used to evaluate recommendation systems (Karypis, 2001) and information retrieval task (Zuva and Zuva, 2012). However, we only care about whether the image appears in the reference image set. Thus in our case, we are only interested in calculating precision metric. Therefore, we adapt the precision here as **IP** to measure image salience.

[4] More details can be found in Sec. 4.1

3.3 Image-Text Relevance

A prerequisite for a pictorial summary to help users accurately acquire information is that the image must be related to the text. Therefore, we regard the image-text relevance as one of metrics to measure the quality of the pictorial summary. We consider using visual-semantic embedding (Faghri et al., 2018; Wang et al., 2018) to calculate the cosine similarity between visual feature and textual feature, which we use as image-text relevance. Visual-semantic embedding has been widely used in cross-modal retrieval (Kiros et al., 2014) and image captioning (Karpathy and Fei-Fei, 2015).

We apply VSE0 model of Faghri et al. (2018), which achieves state-of-the-art performance for image-caption retrieval task on the Flickr30K dataset (Young et al., 2014). The difference is that instead of training a CNN model to encode the image, we use the pretrained VGG19 to extract global features. The text is encoded by a unidirectional Gated Recurrent Unit (GRU) to a sequence of vector representations. Then we apply the max-over-time pooling (Collobert et al., 2011) to get a single vector representation. Next, the visual features and text features are projected to a joint semantic space by two feed-forward neural networks. The whole network is trained using a max-margin loss:

$$\begin{aligned} L = \sum_{\hat{c}} \max(\beta - s(i,c) + s(i,\hat{c}), 0) \\ + \sum_{\hat{i}} \max(\beta - s(i,c) + s(\hat{i},c), 0) \end{aligned} \quad (19)$$

The loss comprises two symmetric terms, with i and c being images and captions repectively. The first term is taken over negative captions $\hat{c}$ image i in a batch. The second is over negative images $\hat{i}$ given caption c. If i and c are closer to each other in the joint embedding space than to any other negative pairs, by a margin β, the loss is zero. We choose to use image-caption pairs in our dataset to train the VSE0 model.

4 Experiments

We conduct the following five sets of experiments: 1) To verify our motivation of the multimodal output (pictorial summary), we design an experiment for user satisfaction test (Sec. 4.2); 2) We compare our multimodal summarization with text summarization from both ROUGE score and manual evaluation (Sec. 4.3); 3) To verify the effectiveness of our evaluation metrics, we calculate the correlation between these metrics and human judgments (Sec. 4.4); 4) We conduct two experiments to show the effectiveness of our proposed MMAE and the generalization of MMAE respectively (Sec. 4.5); 5) Finally, we evaluate our multimodal attention model with MMAE (Sec. 4.6).

The hyperparameters in our model are similar to See et al. (2017), except that we set the maximum number of images to 10, 7, and 7 for ATG, ATL, and HAN respectively, because different articles have the image collection of different sizes. The images are sorted in the order of the position in the article.

4.1 Dataset

There is no large-scale benchmark dataset for MSMO. We follow Hermann et al. (2015) to construct a corpus from *Daily Mail* website[5]. Similar to Hermann et al. (2015), we use the manually-written highlights offered by *Daily Mail* as a reference text summary. From *Daily Mail*, we randomly select articles within a week and find that 2,917 out of 2,930 articles contain images. More details are illustrated in Table 1.

	train	valid	test
#Documents	293,965	10,355	10,261
#ImgCaps	1,928,356	68,520	71,509
#AvgTokens(S)	720.87	766.08	730.80
#AvgTokens(R)	70.12	70.02	72.16
#AvgCapTokens	22.07	22.64	22.34
#AvgImgCaps	6.56	6.62	6.97

Table 1: Corpus statistics. Each image on the website is paired with a caption. *#ImgCaps* denotes the number of image-caption pairs. *#AvgTokens(S)*, *#AvgTokens(R)* and *#AvgCapTokens* denote the average number of tokens in articles, highlights, and captions respectively.

To get the pictorial reference, we employ 10 graduate students to select the relevant images from the article for each reference text summary. We allow annotators to select up to three images to reduce the difference between different annotators. If the annotators find that there is no relevant image, they will select none of them. Each article is annotated by at least two students[6]. Since we use the text reference to guide the generation of the pictorial summary, we do not use the reference image during training. Therefore, we only conduct the annotation on the test set.

[5] http://www.dailymail.co.uk

[6] A third annotator will be asked to decide the final annotation for the case of divergence for the first two annotators.

4.2 User Satisfaction Test

We conduct an experiment to investigate whether a pictorial summary can improve the user satisfaction for the informativeness of the summary. For a fair comparison, we propose a novel strategy to compare text summaries and pictorial summaries. We take an example to illustrate our strategy. Given 100 source news pages, we have their corresponding reference text summaries and pictorial summaries. We divide them into two parts of the same size, part 1 and part 2. In part 1, human annotator A evaluates the text summaries according to the input news, and human annotator B evaluates the pictorial summaries. In part 2, annotator A evaluates the pictorial summaries and annotator B evaluates the text summaries. All annotators will give a score of 1 to 5. The input news is the same for annotator A and annotator B.

Format	$\text{Annotator}_\text{A}$	$\text{Annotator}_\text{B}$	Overall
Text	3.67	3.75	3.71
Pictorial	4.14	4.20	4.17

Table 2: User satisfaction test results. In total, we use the strategy mentioned in Section 4.2 to evaluate 400 randomly selected source news pages. *Overall* denotes the average score on these 400 samples.

Table 2 shows our results for user satisfaction test. User ratings of pictorial summaries are 12.4% higher than text summaries. It shows that users prefer this way of presenting information. It also confirms our motivation for MSMO.

4.3 Comparison with Text Summarization

Our user satisfaction test in Sec. 4.2 is done in an ideal situation, comparing the text reference with the pictorial reference. To show the effectiveness of our model, we also compare our model with text summarization from ROUGE and human judgment scores. We compare several abstractive summarization methods with our multimodal summarization methods. **PGC**[7] (See et al., 2017) refers to the pointer-generator network (Sec. 2.2). **AED** (Nallapati et al., 2016) uses an attentional encoder-decoder framework and adds some linguistic features such as POS, named-entities, and TF-IDF into the encoder. We also implement a seq2seq model with attention (**S2S+attn**). To compare the multimodal output with our multimodal model, we propose an extractive method based on GuideRank (**GR**) (Li et al., 2016a, 2018b). GuideRank applies LexRank (Erkan and Radev, 2004) with guidance strategy. In this strategy, captions recommend the sentences related to them. The rankings of sentences and captions are obtained through GR; we extract sentences that satisfy the length limit as a text summary according to the ranking of text. We select an image whose caption ranks the first in the captions. And finally, the pictorial summary is obtained. We evaluate different summarization models with the standard ROUGE metric, reporting the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Our ROUGE results are given in Table 3, and human judgment scores are given in Table 4.

	Model	ROUGE-1	ROUGE-2	ROUGE-L
Base	S2S+attn	32.32	12.44	29.65
	AED	34.78	13.10	32.24
	PGC	**41.11**	**18.31**	**37.74**
MM	ATG	40.63	18.12	37.53
	ATL	**40.86**	18.27	**37.75**
	HAN	40.82	**18.30**	37.70
	GR	37.13	15.03	30.21

Table 3: ROUGE F1 scores on our test set. All our ROUGE scores are reported by official ROUGE script.

Model	PGC	ATG	ATL	HAN
HS	3.07	3.30	3.22	3.20

Table 4: Human judgment scores for our multimodal model and PGC. We randomly select 400 articles and use the same strategy as Sec. 4.2. HS denotes the average human judgment scores.

From Table 3, all multimodal models lead to a decrease in ROUGE scores which can attribute to the following reasons. There are 6.56 images on average in each article and not every image is closely related to the event of the article. In other words, some images are noise. On the other hand, our text input is long text, and it contains enough information for text generation. In Table 4, multimodal models are better than text model in human judgments. It further illustrates our motivation, and also proves the effectiveness of our models.

4.4 Correlation Test

To illustrate the effectiveness of our evaluation metrics, we conduct an experiment on correlations between these metrics and human judgment scores. Human annotators give a score which ranges from 1 to 5 to a pictorial summary accord-

[7] https://github.com/abisee/pointer-generator

ing to the reference[8]. The reference consists of a text summary and up to three relevant images selected by humans. We randomly extract the pictorial summaries from the output of different systems. In response to the three aspects we proposed in Section 3, we propose some related metrics respectively. For text salience, we apply **ROUGE-1**, **ROUGE-2**, **ROUGE-L**, and **BLEU**. For image-text relevance of candidate pictorial summaries, we propose two ways. One is to calculate the similarity (**Img-Sum**) between the image and the whole text summary. The other is to calculate the similarities between the image and each sentence in the text summary. Then we take the maximum and average values as two metrics: **MAX**$_{\textbf{sim}}$ and **AVG**$_{\textbf{sim}}$. For image salience, in addition to the **IP** metric mentioned in Section 3.2, we try to calculate the similarity between the candidate image and each reference image in three ways: 1) **I-I**: similarities between the global `fc7` features, 2) **Hist**: Bhattacharyya distance[9] for histogram comparison (Bhattacharyya, 1943), and 3) **Temp**: Fourier Analysis template matching (Briechle and Hanebeck, 2001).

We employ annotators to evaluate 600 samples (randomly selected from the outputs of each model on the validation set). Each sample is scored by two persons and we take the average score as the final score. We use 450 of them as training set to train the MMAE model in Sec. 4.5, the rest is used as test set. The scores calculated by each evaluation metric are then tested on the training set to see how well they correlate with human judgments. The correlation is evaluated with three metrics, including 1) Pearson correlation coefficient (r), 2) Spearman rank coefficient (ρ), and 3) Kendall rank coefficient (τ). Our results of correlation test are given in Table 5.

As shown in Table 5, **IP** (Image Precision) correlates best with human assessments according to the three correlation coefficients. It illustrates that people pay more attention to images when assessing pictorial summaries. If we choose the right image for the summary, people are more likely to assign a high score. We also note that the correlation score of **IP** is significantly higher than four text metrics. Because it is easy for a person to judge the importance of images based on reference, such as to see whether the image appears in reference. However, measuring the semantic similarity of two texts is difficult. The four metrics all measure the degree of n-gram overlap which cannot accurately measure semantic similarity.

	Metric	r	ρ	τ
Text	BLEU	.1949	.1542	.1198
	ROUGE-1	.3006	.2941	.2152
	ROUGE-2	.2735	.2742	.2002
	ROUGE-L	**.3144**	**.3087**	**.2272**
Image-Text	AVG$_{\text{sim}}$	.2662	.2388	.1774
	MAX$_{\text{sim}}$	**.2849**	**.2749**	**.2033**
	Img-Sum	.2380	.2075	.1556
Image	I-I (max)	.0169	.0258	.0196
	I-I (avg)	-.0262	-.0140	-.0113
	Hist$_{\text{avg}}$	.4688	.5077	.3725
	Hist$_{\text{max}}$	.5974	.6388	.5149
	Temp$_{\text{avg}}$	.4913	.4944	.3631
	Temp$_{\text{max}}$	.5967	.6435	.5080
	IP	**.6407**	**.6482**	**.5789**

Table 5: Correlation with human judgment scores (training set), measured with Pearson r, Spearman ρ, and Kendall τ coefficients. The max and avg denote the maximum and average value of the scores.

For the image-text relevance, **MAX**$_{\textbf{sim}}$ performs best and is comparable to the several ROUGE metrics. It shows that in a good pictorial summary, the image and text should be relevant. In some cases, even though the generated text is not so important, the image is closely related to the text. At this time, people can also be satisfied. On the other hand, our VSE0 (Sec. 3.3) model can capture some fluency of sentences by adopting GRU. Compare **MAX**$_{\textbf{sim}}$, **AVG**$_{\textbf{sim}}$, and **Img-Sum**, this is very intuitive. Once people find a sentence (or a part) relevant to the image, they will think the image is related to the text. Besides, the worst performance of **Img-Sum** metric is probably because the average length of captions used to train VSE0 model is about 22, far less than the length of the summary. We find the **I-I (max)** and **I-I (avg)** nearly do not correlate with human assessments. It shows that the visual features extracted from VGG19 are not suitable for calculating the similarity between news images. The analysis of **Hist (Temp)**$_{\textbf{avg}}$ and **Hist (Temp)**$_{\textbf{max}}$ is similar to the analysis of **MAX**$_{\textbf{sim}}$ and **AVG**$_{\textbf{sim}}$ above.

4.5 Effectiveness and Generalization of MMAE

We then select the best-performing metrics separately from the three sets of metrics, namely

[8] Some articles are annotated with no relevant images (about 3.9%), we directly skipped these articles without manual scoring

[9] In statistics, the Bhattacharyya distance measures the similarity of two discrete or continuous probability distributions. For a distance d, we take $(1 - d)$ as the similarity.

ROUGE-L, $\textbf{MAX}_{\textbf{sim}}$, and **IP**. We apply LR, MLP, and Logis to learn our MMAE model that combines the three metrics. We calculate the three coefficients for the three metrics on the test set as a comparison. The correlation results are given in Table 6.

Metric	r	ρ	τ
ROUGE-L	.3488	.3554	.2669
MAX_{sim}	.2541	.2339	.1773
IP	**.5982**	**.5966**	**.5485**
MMAE_{LR}	**.6646**	.6644	.5265
MMAE_{MLP}	.6632	.6646	.5265
$\text{MMAE}_{\text{Logis}}$	.6630	**.6653**	**.5277**

Table 6: Correlation with human judgment scores (test set).

As shown in Table 6, the MMAE learned by three methods correlates better with human judgments. Although $\text{MMAE}_{\text{Logis}}$ gets a slightly higher correlation score according to Spearman and Kendall coefficients, we choose the MMAE_{LR}[10] as our final MMAE model due to Occam's Razor[11].

It is crucial that MMAE can generalize for a previously unseen system. To test the generalization of MMAE, we use MMAE to evaluate a new system and calculate the correlation with human judgment scores. The new system is a naive model which applies LexRank to extract sentences and randomly select an image from source. We can observe that MMAE still correlates well with human judgment scores, as shown in Table 7. It illustrates that MMAE generalize well for a new model. We give some examples of MMAE in supplementary material.

Metric	r	ρ	τ
ROUGE-L	.3223	.3514	.2615
MMAE	.6352	.6318	.4728

Table 7: Correlation results for the new model on the same 150 test samples as in Sec. 4.4.

4.6 Model Performances

According to our analyses above, we have proved MMAE can evaluate multimodal output. In this section, we report the MMAE scores for our proposed multimodal attention model, as shown in Table 8.

Model	ROUGE-L	MAX_{sim}	IP	MMAE
ATG	40.76	25.82	59.28	**3.35**
ATL	40.80	13.26	**62.44**	3.26
HAN	**40.82**	12.22	61.83	3.25
GR	30.20	**26.60**	61.70	3.20

Table 8: Results evaluated by our MMAE method. We skipped the articles that are labeled as no relevant images. Finally, only 9,851 of the 10,261 articles are left.

Surprisingly, the model ATG achieves the highest MMAE score despite the mediocre performance in three individual metrics. The $\textbf{MAX}_{\textbf{sim}}$ score of ATG is much higher than ATL and HAN. It shows the global features can help to learn better image-text alignments. Since GR itself makes use of the image-caption pairs, it is natural to get a high image-text relevance score. Our proposed multimodal attention models all achieves higher performance than the extractive baseline GR, which further indicate the effectiveness of our models.

5 Related Work

Different from text summarization (Wan and Yang, 2006; Rush et al., 2015; Zhu et al., 2017; See et al., 2017; Celikyilmaz et al., 2018; Paulus et al., 2018), Multimodal Summarization is a task to generate a condensed text summary or a few keyframes to help acquire the gist of multimedia data. One of the most significant advantages of the task is that it does not rely solely on text information, but it can also utilize the rich visual content from the images.

In recent years, much work has focused on multimodal summarization. Evangelopoulos et al. (2013) detect salient events in a movie based on the saliency of individual features for aural, visual, and linguistic representations. Li et al. (2017) generate the text summary from an asynchronous collection of text, image, audio, and video. There has also been some work (Bian et al., 2013, 2015; Wang et al., 2016; Qian et al., 2016) focused on producing multimodal output for summarization. Bian et al. (2013, 2015) aim to produce a visualized summary for microblogs. Wang et al. (2016) generate a pictorial storyline for summarization. Qian et al. (2016) generate the multimedia topics for social events. But these researches all treat image-text pairs, in which texts and images are aligned, as a basic summarization unit. For example, the images are aligned with the text in a mi-

[10] The weight for ROUGE-L, MAX_{sim}, and IP is 1.641, 0.854, 0.806 respectively and the intercept is 1.978.

[11] https://en.wikipedia.org/wiki/Occam%27s_razor

croblog post; Wang et al. (2016) obtain the image-text pairs by using image search engine. None of the above works focuses on generating multimodal output from a collection of texts and images that are not explicitly aligned. This is one of the goals in this paper. Another difference is that they separately evaluate texts and images when evaluating the final results. In our work, we propose a new automatic evaluation which jointly considers two aspects of textual and visual modalities.

6 Conclusion

In this paper, we focus on a novel task which aims to automatically generate a multimodal summary from multimodal news, where the images and the texts are not explicitly aligned. We provide a multimodal summarization method to jointly generate text and the most relevant image, which can be referred as the baseline for further study. Our proposed metrics have been proved to be effective in evaluating the multimodal output. Moreover, the idea of constructing our MMAE can be easily extended to other modalities. That is, we both consider the intramodality salience and the intermodality relevance.

7 Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303. We thank the three anonymous reviewers for their careful reading of our paper and their many insightful comments and suggestions. We also thank Shaonan Wang, Xiaomian Kang, and Yang Zhao for helpful discussions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Anil Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109.

Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1807–1812.

Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on multimedia*, 17(2):216–228.

Kai Briechle and Uwe D Hanebeck. 2001. Template matching using fast normalized cross correlation. In *Optical Pattern Recognition XII*, volume 4387, pages 95–103.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1662–1675.

Ronan Collobert, Jason Weston, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12(1):2493–2537.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1693–1701.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3128–3137.

George Karypis. 2001. Evaluation of item-based top-n recommendation algorithms. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 247–254.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Haoran Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016a. Guiderank: A guided ranking graph model for multilingual multi-document summarization. In *The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (NLPCC-ICCPOL)*, pages 608–620.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1092–1102.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018b. Read, watch, listen and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Zechao Li, Jinhui Tang, Xueming Wang, Jing Liu, and Hanqing Lu. 2016b. Multimedia news summarization in search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):33.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In *Proceedings of NII Testbeds and Community for information access Research (NTCIR)*.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. 2016. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing*, 25(12):5828–5840.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Paul Over and James Yen. 2004. An introduction to duc 2004 intrinsic evaluation of generic new text summarization systems. In *Proceedings of the Document Understanding Conference (DUC)*.

Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal multi-view topic-opinion mining for social event analysis. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 2–11.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083.

Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, and Roger Zimmermann. 2016. Leveraging multimodal information for event summarization and concept-level sentiment analysis. *Knowledge-Based Systems*, 108:102–109.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xiaojun Wan and Jianwu Yang. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 181–184.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

William Yang Wang, Yashar Mehdad, Dragomir R Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 58–68.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Junnan Zhu, Long Zhou, Haoran Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2017. Augmenting neural sentence summarization through extractive summarization. In *Proceedings of the 6th Conference on Natural Language Processing and Chinese Computing (NLPCC)*, pages 16–28.

Keneilwe Zuva and Tranos Zuva. 2012. Evaluation of information retrieval systems. *International Journal of Computer Science & Information Technology*, 4(3):35.