Multimodal Summarization with Guidance of Multimodal Reference

Junnan Zhu,[1,2] **Yu Zhou,**[1,2*] **Jiajun Zhang,**[1,2] **Haoran Li,**[4] **Chengqing Zong,**[1,2,3] **Changliang Li**[5]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
[4]JD AI Research
[5]Kingsoft AI Lab
{junnan.zhu, yzhou, jjzhang, cqzong}@nlpr.ia.ac.cn, lihaoran24@jd.com, lichangliang@kingsoft.com

Abstract

Multimodal summarization with multimodal output (MSMO) is to generate a multimodal summary for a multimodal news report, which has been proven to effectively improve users' satisfaction. The existing MSMO methods are trained by the target of text modality, leading to the modality-bias problem that ignores the quality of model-selected image during training. To alleviate this problem, we propose a multimodal objective function with the guidance of multimodal reference to use the loss from the summary generation and the image selection. Due to the lack of multimodal reference data, we present two strategies, i.e., ROUGE-ranking and Order-ranking, to construct the multimodal reference by extending the text reference. Meanwhile, to better evaluate multimodal outputs, we propose a novel evaluation metric based on joint multimodal representation, projecting the model output and multimodal reference into a joint semantic space during evaluation. Experimental results have shown that our proposed model achieves the new state-of-the-art on both automatic and manual evaluation metrics. Besides, our proposed evaluation method can effectively improve the correlation with human judgments.

1 Introduction

Generally, most existing summarization researches focus on either texts (Wan and Yang 2006; Celikyilmaz et al. 2018) or images (Wang, Jia, and Hua 2011; Sharma et al. 2015) in isolation. Recently, researchers (Chen and Zhuge 2018; Zhu et al. 2018) begin to pay attention to summarizing multimodal news to multimodal outputs, which can be called multimodal summarization with multimodal output (MSMO) (Zhu et al. 2018), to help improve users' satisfaction.

Although great efforts have been made in multimodal summarization, we find that the existing methods have the following problems:

Modality-bias. The current multimodal summarization models are trained by the target of text modality, which causes a modality-bias problem. It means the system tends to only optimize the text summary generation process, while the image quality is ignored during training. We give an example in Figure 1 to illustrate this phenomenon. In the training process, if we only consider the text reference, then Summary B is regarded as better than A; but A will be distinguished as better than B if the multimodal reference is available, which is in line with the ground truth.

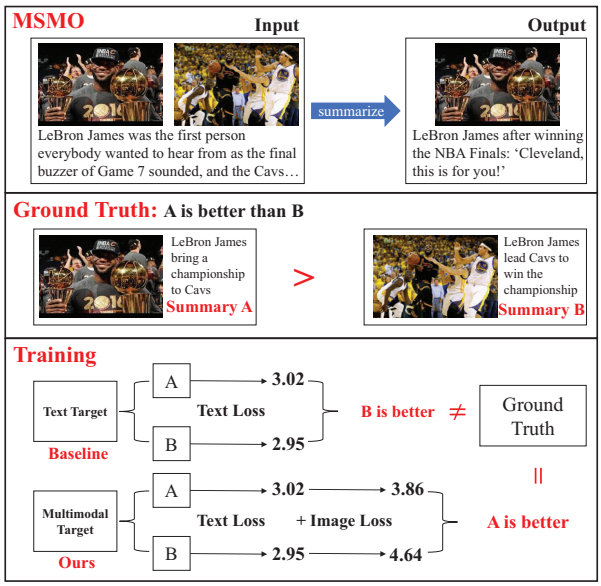

Figure 1: An example of the modality-bias problem. Summary A and B are considered similar when calculating the loss only with the text reference. But with the multimodal reference, we can distinguish A from B.

Lack of good evaluation metric. The existing methods evaluate multimodal summaries from three aspects: (1) the ROUGE value between the texts in the model output and reference, (2) the precision of the images in the model output and reference, and (3) the image-text similarity in the model output. However, all these metrics consider each modality separately. We argue that the multimodal output should be treated as a whole in the evaluation process to maintain the information integrity (See an example in Section 3.3).

Therefore, this paper aims to guide multimodal summa-

* Corresponding author.

rization with the multimodal reference as the target and to evaluate multimodal outputs as a whole. Specifically, we first propose a multimodal objective function which takes into account both the negative log-likelihood loss (NLL) of the summary generation and the cross-entropy loss (CE) of the image selection. In order to extend the text reference to the multimodal reference, we then introduce two strategies: ROUGE-ranking and Order-ranking. **ROUGE-ranking** uses the ROUGE (Lin 2004) value between the corresponding caption and the text reference to sort the images; **Order-ranking** measures the image salience directly according to the order in which they appear in the original news. Finally, to better evaluate the multimodal outputs, we employ the image-caption pairs to train a joint multimodal representation model to help calculate the similarity between two multimodal segments.

Our main contributions are as follows:

- We introduce a multimodal objective function to incorporate the multimodal reference into the training process, in which both the summary generation and the image selection are considered. To the best of our knowledge, this is the first work that guides multimodal summarization with multimodal reference.
- We propose a novel evaluation method to evaluate a multimodal summary by projecting both the multimodal summary and the reference into a joint semantic space.
- The experimental results show that our proposed model outperforms existing methods with both automatic and manual evaluation metrics. Moreover, our proposed evaluation method can effectively improve the correlation with human judgments.

2 Background

For MSMO task, given a multimodal news report $M = \{T, V\}$, where $T = \{t_1, t_2, \cdots, t_m\}$ is a text sequence and $V = \{v_1, v_2, \cdots, v_n\}$ is a collection of images (m denotes the text sequence length and n denotes the image number), the system summarizes M into a multimodal summary $\{Y, v^*\}$, where $Y = \{y_1, y_2, \cdots, y_l\}$ denotes the textual summary limited by length l and v^* is an image extracted from the image collection V.

2.1 Multimodal Attention Model

Zhu et al. (2018) propose a multimodal attention model, in which the news with images is considered as input and a multimodal summary is gained as output. As shown in the left half in Figure 2, the model consists of a text encoder, an image encoder, a multimodal attention layer, and an attentive summary decoder. The text encoder maps the source text to a sequence of hidden states h_i. The image encoder extracts the global `fc7` image feature vectors g for all images and projects g into g^*, the same dimension as h_i. Previous researches (Li et al. 2018a; Zhu et al. 2018) have shown that the global features are more effective than the local features, thus we only consider the global features in this paper.

During decoding, the summary decoder reads the previous predicted word and the multimodal context vector c^t_{mm} to predict the next word. Then, the summary decoder reaches a new decoder state s_t. c^t_{mm} is a weighted sum of the textual context vector c^t_{txt} and the visual context vector c^t_{vis}, which is obtained through the multimodal attention mechanism (Li et al. 2018a). c^t_{txt} and c^t_{vis} are obtained through textual attention layer (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015) and visual attention layer (Li et al. 2018a), respectively. A multimodal coverage mechanism (Li et al. 2018a), which maintains both a textual coverage vector cov^t_{txt} and a visual coverage vector cov^t_{vis}, is introduced to alleviate repeated attention to the source.

The summary generation is based on the pointer-generator network (See, Liu, and Manning 2017), which either generates a word from the vocabulary distribution or copies a word from the source text. The loss for timestep t is the sum of negative log-likelihood (NLL) loss of the target word w_t and the multimodal coverage loss:

$$\begin{aligned} L_t = -\text{logP}(w_t) &+ \sum_i \min(\alpha^t_{txt,i}, \text{cov}^t_{txt,i}) \\ &+ \sum_j \min(\alpha^t_{vis,j}, \text{cov}^t_{vis,j}) \end{aligned} \tag{1}$$

where α^t_{txt} and α^t_{vis} is the attention weight for the text features and the image features, respectively.

The salience of images is measured by the visual coverage vector in the last decoding step, which is the sum of the visual attention on image features over all the decoding steps. The input image with the highest salience score will be selected. The core idea of this model is to sort images by their visual coverage while calculating the NLL loss of text generation, to accomplish the goal of the multimodal summary.

2.2 Multimodal Automatic Evaluation

To evaluate the quality of a multimodal summary, Zhu et al. (2018) propose the multimodal automatic evaluation (MMAE) which is defined to be a linear combination of three metrics: salience of text, salience of image, and image-text relevance. The weight of linear combination is obtained by fitting the human judgment scores.

The salience of text is measured by ROUGE. They define the image precision (**IP**), which represents whether an output image is in the gold summary, to depict the salience of an image. The image-text relevance is indirectly obtained by a cross-modal retrieval model (Faghri et al. 2018), which is trained using the image-caption pairs. Specifically, for images, they directly use the global `fc7` features; for text, a unidirectional GRU with max-over-time pooling (Collobert and Weston 2008) is applied to encode the text to a single vector representation. Next, they employ two feed-forward neural networks to project the text features and the image features into a joint semantic space. The whole network is trained using the max-margin loss:

$$\begin{aligned} L = &\sum_{\hat{c}} \max(\beta - s(i,c) + s(i,\hat{c}), 0) \\ &+ \sum_{\hat{i}} \max(\beta - s(i,c) + s(\hat{i},c), 0) \end{aligned} \tag{2}$$

where i and c denote the paired image and caption, $\hat{i}$ and $\hat{c}$ are the negative ones, $s(\cdot)$ is the cosine similarity between

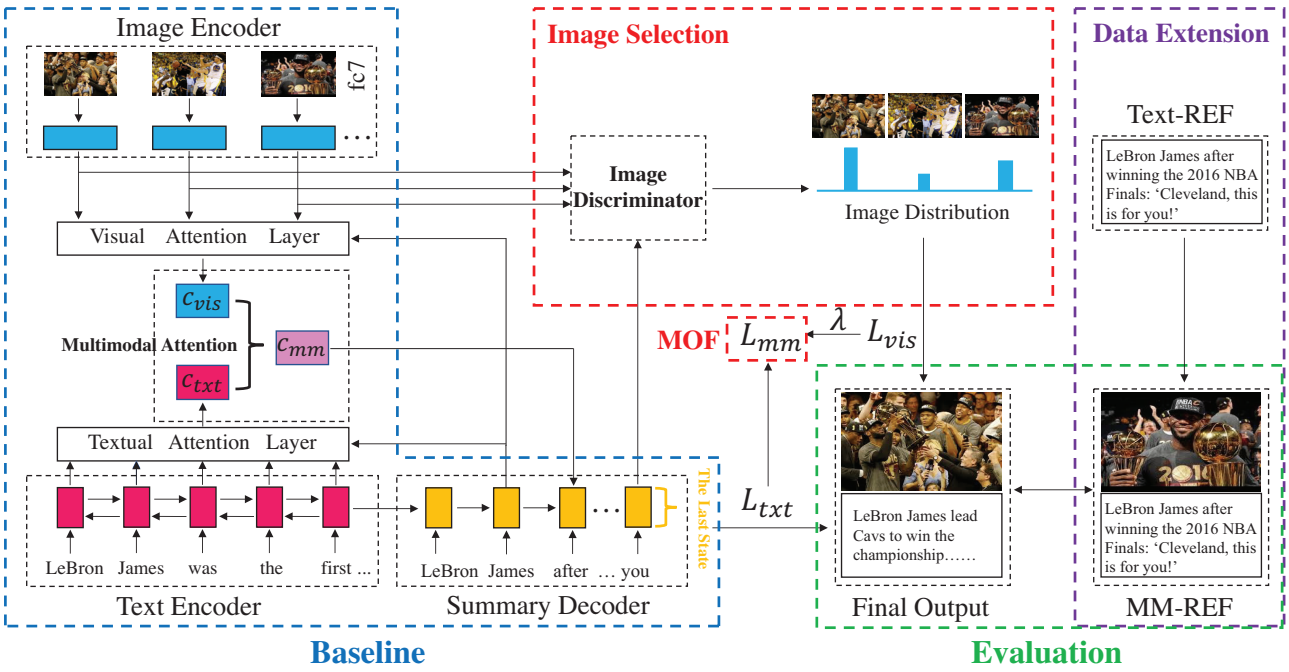

Figure 2: Overview of our work. We divide it into four parts: (1) Baseline (Section 2.1); (2) Image Selection and multimodal objective function (MOF) (Section 3.1); (3) Data Extension (Section 3.2); (4) Evaluation (Section 3.3). We take the model, which uses the last hidden state of decoder, as an example.

the text vector and the image vector, β is the margin. We employ the model to calculate the similarity between an image and a text.

3 Our Model

The current multimodal summarization methods have the following two drawbacks: 1) Due to the lack of multimodal reference, the existing multimodal summarization systems are trained by the target of text modality (Eq.1), which will lead to the modality-bias problem. 2) Existing evaluation metrics consider each modality separately, which ignores the information integrity.

Therefore, in this work, we propose a multimodal objective function, which considers both the text loss and image loss, to improve multimodal summarization with the guidance of multimodal reference. To this end, we introduce an image discriminator based on the multimodal attention model, which is described in Section 3.1 together with our multimodal objective function. Due to the lack of multimodal reference, we explore two strategies to construct the multimodal reference by extending the text reference, which is described in Section 3.2. Finally, we design a multimodal automatic evaluation metric by treating the multimodal outputs as a whole during evaluation, which is described in Section 3.3.

3.1 Multimodal Objective Function

Suppose we have the image reference besides the text reference during model training. To utilize the multimodal reference in training, we propose a multimodal objective function (MOF), which considers the cross-entropy loss of the image selection in addition to the negative log-likelihood loss of text summary. Thus, we decompose the multimodal summarization into two subtasks: summary generation and text-image relation recognition. To achieve that, we propose an image discriminator to guide the image selection. The image discriminator is to determine whether an image is related to the text content. We apply multitask learning (Caruana 1997; Collobert and Weston 2008) to train the two subtasks simultaneously, as illustrated in the right half of Figure 2. In our multitask setup, we share the text encoder and the summary decoder for two subtasks. Since summary generation has been described in Section 2.1, we focus on text-image relation recognition in this section.

We use another image encoder to transform the global image features g to $g\prime$. Then the text information can be presented in two ways: (1) the last hidden state of the text encoder; or (2) the last hidden state of the summary decoder. To project the two vectors into a joint semantic space, we use two multilayer perceptrons with ReLU activation function (Nair and Hinton 2010) to transform the textual vector and the visual vector to I_{txt} and I_{vis}. We then employ the L2-norm to turn I_{txt}, I_{vis} into I^*_{txt}, I^*_{vis}. The degree of relevance between images and text information is calculated as Eq.3.

$$\mathrm{P(img)} = \mathrm{softmax}(\mathrm{I}^*_{txt} \cdot \mathrm{I}^*_{vis}) \tag{3}$$

The images are divided into text-related and non-text-related, which means the text-image relation recognition can be regarded as a classification task. Hence we adopt the

Figure 3: Two multimodal summaries above convey almost identical information: *LeBron James wept with joy for the championship*.

cross-entropy loss here:

$$L_{vis} = \sum -\mathrm{P_A(img)logP(img)} \tag{4}$$

where $\mathrm{P_A(img)}$ denotes the actual probability distribution of images (If we choose the top-k ranked images as the target, then the probability of an image is $1/\mathrm{k}$.). Finally, the cross-entropy loss, weighted by a hyperparameter λ, is added to the loss function of summary generation (Eq.1) to yield a new loss function that takes both the text reference and image reference into consideration:

$$\begin{aligned} L_{txt} &= \sum_t L_t \\ L_{mm} &= L_{txt} + \lambda L_{vis} \end{aligned} \tag{5}$$

3.2 Data Extension

Due to the lack of multimodal reference in existing multimodal summarization dataset, the gold standard is plain text during the training process or validation process. Thus, we consider two methods to sort the images and choose top-k images in order to extend text reference to multimodal reference:

ROUGE-ranking. It sorts the images according to the ROUGE-2 value between the corresponding caption and the text reference since the image is assumed to semantically match with the corresponding caption.

Order-ranking. It sorts the images according to the order in the original news because the core information tends to appear at the front of the news reports.

3.3 Joint Multimodal Representation

A problem with the current multimodal evaluation metric (MMAE) is that it compares the model output with the reference from individual modalities, such as ROUGE and IP. Therefore, we argue that MMAE cannot evaluate the information integrity of the multimodal summary. Consider the example in Figure 3, where two summaries express the same two events: *winning the championship and crying with joy*. Since the text ROUGE value between $\mathrm{T_A}$ and $\mathrm{T_B}$, the image precision between $\mathrm{I_A}$ and $\mathrm{I_B}$, the image-text similarity between $\mathrm{I_A}$ ($\mathrm{I_B}$) and $\mathrm{T_A}$ ($\mathrm{T_B}$) all are very low, A is quite different from B from the perspective of MMAE, which is contradictory to the truth. Thus, it is critical to find a new way to evaluate the overall quality of multimodal summaries.

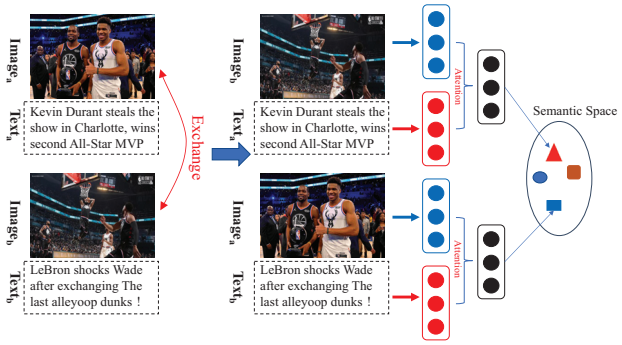

Figure 4: Overview of our proposed joint multimodal representation model.

To the best of our knowledge, no one has studied how to overall evaluate multimodal summaries. Although existing method attempts to measure the cross-modal similarity, it only focuses on the image and text in the modal output, rather than the multimodal output as a whole to compare with the multimodal reference. We extend the basic cross-modal retrieval model (Section 2.2) and propose the joint multimodal representation model in this work. In cross-modal retrieval, the input is a pair of an image and a text. But the input becomes a pair of multimodal segments (m_a, m_b) in our proposed model, where $m_a = (\mathrm{Image}_a, \mathrm{Text}_a)$ and $m_b = (\mathrm{Image}_b, \mathrm{Text}_b)$. The critical problem is how to construct the training data. There are lots of image-caption pairs in MSMO dataset, and each image is assumed to match the corresponding caption. Thus, we exchange the image (or text) of two image-caption pairs to get a matching multimodal segment pair (m_a^*, m_b^*), where $m_a^* = (\mathrm{Image}_b, \mathrm{Text}_a)$ and $m_b^* = (\mathrm{Image}_a, \mathrm{Text}_b)$, as shown in Figure 4. It is worth noting that since Image_a in m_b^* matches Text_a in m_a^* and Image_b in m_a^* matches Text_b in m_b^*, m_a^* semantically matches m_b^*. We encode the image and the text as described in Section 2.2. Then we use the multimodal attention mechanism (Li et al. 2018a) to fuse the text vector and the image features. Finally, our model is trained under a new max-margin loss:

$$L^* = \sum_{\hat{m}} \max(\beta - s(m, m^*) + s(m, \hat{m}), 0) \tag{6}$$

where (m, m^*) is a matching multimodal segment pair, $(m, \hat{m})$ is a negative pair sampled from a batch. We also use the image-caption pairs in MSMO dataset to construct the training data.

4 Related Work

Multimodal summarization has been proposed to extract the most important information from the multimedia information. The most significant difference between multimodal summarization (Mademlis et al. 2016; Li et al. 2017; 2018b; Zhu et al. 2018) and text summarization (Zhu et al. 2017; Paulus, Xiong, and Socher 2018; Celikyilmaz et al. 2018; Li et al. 2018c; Zhu et al. 2019) lies in whether the input data contains two or more modalities of data. One of the

most significant advantages of the task is that it can use the rich information in multimedia data to improve the quality of the final summary.

Multimodal summarization can be categorized into single-modal output (Evangelopoulos et al. 2013; Li et al. 2017) and multimodal output (Bian et al. 2015; Zhu et al. 2018). Evangelopoulos et al. (2013) detect the keyframes in a movie based on the saliency of individual features for aural, visual and linguistic representations. Li et al. (2017) generate a textual summary from a set of asynchronous documents, images, audios, and videos by maximizing the salience, non-redundancy, and coverage. Bian et al. (2017) propose a multimedia topic model to separately identify the representative textual and visual samples and then produce a comprehensive visualized summary. Zhu et al. (2018) propose a multimodal attention model to generate a multimodal summary from the multimodal input.

However, these researches either generate the text and the image separately or use the single-modal reference. The former ignores the relationship between the texts and the images in the output. The latter may cause a modality-bias problem which means that the system will deviate towards optimizing single-modal output (e.g., text summary). None of the above works focuses on using a multimodal reference to jointly generate the texts and the image as the multimodal output. This is one of the goals in this paper. Another difference is that no one has taken into account the information integrity. In this work, we propose a joint multimodal representation model which maps the multimodal summary and the reference into a joint semantic space during evaluation.

5 Experiments

5.1 Dataset

We use the MSMO dataset (Zhu et al. 2018) which contains online news articles (723 tokens on average) paired with multiple image-caption pairs (6.58 images on average) and multi-sentence summaries (70 tokens on average). It is worth noting that in the definition of MSMO, the input is the text (excluding captions for generalization) and images, and the output is a multimodal summary which is actually a text summary with an image. The dataset includes 293,965 training pairs, 10,355 validation pairs, and 10,261 test pairs. For test data, based on the text reference, at most three images are annotated to produce a multimodal reference.

We design two sets of experiments: (1) To verify the effectiveness of evaluation metric using the joint multimodal representation model (MR), we calculate its correlation with human judgments and compare it with other existing metrics. Then we integrate our proposed MR into the current multimodal automatic evaluation metric (MMAE) to explore whether we can obtain an evaluation model more correlated with human judgments (Section 5.2); (2) We compare our model with existing multimodal summarization methods in the aspect of multiple metrics (including our proposed metrics) and manual evaluation (Section 5.3).

Metric	r	ρ	τ
ROUGE-1	.3006	.2941	.2152
ROUGE-2	.2735	.2742	.2002
ROUGE-L	.3144	.3087	.2272
M_{sim}	.2849	.2749	.2033
Img-Sum	.2380	.2075	.1556
IP	**.6407**	**.6482**	**.5789**
MR_{max}	**.5765**	**.5909**	**.4534**
MR_{avg}	.5328	.5551	.4039
MR-Sum_{max}	.5451	.5625	.4186
MR-Sum_{avg}	.4883	.5013	.3734

Table 1: Correlation with human judgment scores (training set for evaluation), measured with Pearson r, Spearman ρ, and Kendall τ coefficients.

5.2 Analysis of Evaluation metrics

To illustrate the effectiveness of our evaluation metric MR, we conduct an experiment on correlations between MR and human judgment scores. Three graduate students are asked to compare the generated multimodal summary with the reference, and assess each summary from the perspective: *How informative the multimodal summary is?* Each summary is assessed with a score from 1 (worst) to 5 (best), and we take the average value as the final score. We randomly extract 600 multimodal summaries from different systems. These samples are divided into the training set (450 samples to train a new multimodal automatic evaluation model) and test set (150 samples). We compare other existing evaluation metrics with MR in terms of the correlation with human judgments. The correlation is calculated by three widely used metrics, including Pearson correlation coefficient (r), Spearman coefficient (ρ), and Kendall rank coefficient (τ). The existing metrics are as follows:

1) **ROUGE**: It is the standard evaluation metric for text summarization evaluation. We calculate the ROUGE scores between the texts in model summary and reference.

2) **M_{sim}**: It is an image-text relevance metric which calculates the maximum similarity between the image and each sentence in the model summary by cross-modal retrieval model.

3) **Img-Sum**: Similarity between the image and the whole text summary in the model summary.

4) **IP**: The image precision of the model summary with the gold standard as the reference.

We design several metrics based on MR:

5) **MR_{max}**: The maximum similarity between the image-sentence pairs in model summary and the image-sentence pairs in the reference. Similar is **MR_{avg}** (the average value).

6) **MR-Sum_{max}**: The maximum similarity between the image-summary pair in model summary and the image-summary pairs (Since the reference in the original human-labeled test set consists of a text summary and multiple images, and it can be composed of multiple image-summary pairs.) in the reference. Similar is **MR-Sum_{avg}** (the average value).

Our results of correlation test are given in Table 1. We find that **MR_{max}** correlates best with human judgments among the multiple MR metrics. It can be attributed to two reasons: (1) MR model is trained by using the image-caption pairs where the caption is always one sentence. Thus, when the whole text summary is considered into MR, the effect of the MR will be affected, leading to a better performance with MR than MR-Sum; (2) Once one segment in a multimodal summary is found to be similar to another segment in another multimodal summary, people will naturally think of the two as related.

Our proposed MR metrics perform better than most existing metrics, except IP. This is because people can easily tell whether an image appears in the reference and IP is more relevant to human intuition. However, as a discrete metric, the value of IP is binary for a single sample, which is a shortcoming as an evaluation metric. It leads to a phenomenon that given two multimodal summaries, none of their images appears in the reference, and it is impossible to distinguish the quality of them in this way. But MR has this capability, which is an advantage of MR over IP.

We then incorporate $\mathrm{MR}_{\max}$ metric into the MMAE method by the same linear regression method as in Zhu et al. (2018) to explore whether it can further improve the correlation and we note the new method as MMAE++. In MMAE++, the weight for R-L, $\mathrm{M}_{\mathrm{sim}}$, IP, and $\mathrm{MR}_{\max}$ is 1.54, 0.42, 1.25, and 0.98 respectively and the intercept is 1.40. The correlation results over the test evaluation samples are given in Table 2. In addition to the correlation metrics, we compare MMAE with MMAE++ in terms of mean square error and mean absolute error. As shown in Table 2, we can find MMAE++ outperforms MMAE from all the metrics, which further illustrates the effectiveness of MR.

Metric	r	ρ	τ	MSE	MAE
ROUGE-L	.3488	.3554	.2669	-	-
M_{sim}	.2541	.2339	.1773	-	-
IP	**.5982**	**.5966**	**.5485**	-	-
MR_{max}	.4745	.4559	.3523	-	-
MMAE	.6646	.6644	.5265	.2654	.4489
MMAE++	**.6902**	**.6941**	**.5557**	**.2457**	**.4324**

Table 2: Correlation results on test set for evaluation. MSE is the mean square error and MAE is the mean absolute error.

5.3 Multimodal Summarization Methods

To show the effectiveness of our model with the multimodal objective function, we compare our model with the existing multimodal summarization methods (ATG, ATL, HAN, and GR) (Zhu et al. 2018) using multiple metrics including our proposed MR and MMAE++:

1) **ATG**: It refers to the multimodal attention model (Section 2.1). The image salience is measured by the visual attention distribution over the global features.

2) **ATL**: It replaces the global `fc7` features with the local `pool5` image features in multimodal attention and measures the image salience based on the sum of attention distributions over the patches contained in the image. The image features are all extracted by the VGG19 pretrained on ImageNet (Simonyan and Zisserman 2015).

3) **HAN**: Based on **ATL**, a hierarchical attention mechanism is added which first attends to the image patches to get the intermediate vectors to represent images and then attends to these vectors to get the visual context vector. And it calculates the image salience according to the attention distributions over the intermediate vectors.

4) **GR**: It employs LexRank (Erkan and Radev 2004) with guidance strategy where captions recommend the related sentences. And it is an extractive method where the rankings of sentences and captions are obtained by this graph-based method. The salience of an image depends on the ranking of its caption.

5) **MOF**: It is our model using the multimodal objective function (Section 3.1). We incorporate the last hidden states of the text encoder or the summary decoder into the image discriminator and denote it as $\mathbf{MOF}_{\mathrm{enc}}$ and $\mathbf{MOF}_{\mathrm{dec}}$ respectively. There are two kinds of images as the training target: ROUGE-ranking (RR) and Order-ranking (OR) (Section 3.2).

We evaluate different multimodal summarization models with the standard ROUGE metric, reporting the F1 scores for ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L). Besides, we use image precision (**IP**), $\mathbf{M_{sim}}$, **MR**, **MMAE**, and **MMAE++** to measure the performances of different models. Note that when calculating IP and MR metrics, we extend the dataset in the RR and OR manner, where the test set and training set are ensured to be independent and identically distributed. The main results are shown in Table 3.

Compared with the baselines, MOF models achieve a slightly higher ROUGE value. It demonstrates that the multimodal objective function can improve the quality of generated text summary. From the IP metric, the multimodal reference we construct can help select more salient images, which indirectly leads to the improvement of image-text relevance. From IPR and IPO, our model significantly outperforms the baselines, which indicates that our model can effectively improve the visual informativeness if a real large-scale dataset with multimodal reference is available. Notice that, with the multimodal reference obtained by the two strategies (RR or OR), the model (take $\mathrm{MOF}_{\mathrm{dec}}$ as an example) is trained and applied to the human-labeled dataset, of which the OR score (from 71.78 to 64.00) decreases more obviously than the RR score (from 68.62 to 65.45). This shows that the real distribution of images in the manual annotation is closer to RR, which reveals that people pay more attention to the semantic matching between images and texts during annotation and often ignore the image order. Although people always follow the characteristics of sequence during reading, it is easy to find a salient image quickly due to the intuitiveness of looking through images and get rid of the sequence or space constraints. From the MR metric, the multimodal objective function can still help improve the quality of information integrity, both on the human-labeled dataset (MR) and automatically constructed test set (MRR and MRO). It further illustrates the effectiveness of our model. Comparing $\mathrm{MOF}_{\mathrm{enc}}$ with $\mathrm{MOF}_{\mathrm{dec}}$, we

	Model	R-1	R-2	R-L	M_{sim}	IP	IPR	IPO	MR	MRR	MRO	AE	AE++
Base	ATG	40.63	18.12	37.53	25.82	59.28	59.42	64.04	**56.54**	**57.32**	57.82	**65.88**	**67.63**
	ATL	**40.86**	18.27	**37.75**	13.26	**62.44**	**62.77**	**67.04**	55.67	55.79	57.34	64.26	67.26
	HAN	40.82	**18.30**	37.70	12.22	61.83	60.14	64.24	55.29	54.83	56.36	63.96	66.93
	GR	37.13	15.03	30.21	**26.60**	61.70	60.45	65.54	55.81	56.60	**58.33**	63.94	65.90
Ours	$\mathrm{MOF}_{\mathrm{enc}}^{\mathrm{RR}}$	41.05	18.29	37.74	26.23	62.63	67.85	-	57.13	59.26	-	66.52	68.68
	$\mathrm{MOF}_{\mathrm{dec}}^{\mathrm{RR}}$	**41.20**	18.33	37.80	**26.38**	**65.45**	**68.62**	-	**58.38**	**59.58**	-	**67.02**	**69.66**
	$\mathrm{MOF}_{\mathrm{enc}}^{\mathrm{OR}}$	41.16	**18.35**	**37.85**	26.15	63.55	-	68.76	57.66	-	59.55	66.69	69.04
	$\mathrm{MOF}_{\mathrm{dec}}^{\mathrm{OR}}$	40.95	18.12	37.75	26.30	64.00	-	**71.78**	58.16	-	**60.58**	66.76	69.24

Table 3: Results of different metrics on the test set. $\mathrm{MOF}_{\mathrm{enc}}^{\mathrm{RR}}$ means using the ROUGE-ranking (RR) images to train the network, while Order-ranking (OR) in $\mathrm{MOF}_{\mathrm{enc}}^{\mathrm{OR}}$. We set λ to 1.0 and the image number K (the target when calculating the cross-entropy loss) to 3 here. IPR (IPO) denotes the image precision in the RR (OR) manner, where the top-3 ranked images are considered as the reference. MR is the $\mathrm{MR}_{\mathrm{max}}$ metric (Section 5.2) calculated by our proposed joint multimodal representation model, and MRR (MRO) is the MR score in the RR (OR) manner. AE (%) denotes MMAE score, and AE++ (%) denotes MMAE++ score.

find $\mathrm{MOF}_{\mathrm{dec}}$ performs better, which can be attributed to the fact that the decoder contains the summary information while the encoder contains information of the original text.

λ	R-L	M_{sim}	IP	AE
0.5	37.36	**26.58**	64.48	66.76
1.0	**37.80**	26.38	**65.45**	**67.02**
1.5	37.77	26.33	64.35	66.83
2.0	37.68	26.46	63.44	66.67

Table 4: Results of $\mathrm{MOM}_{\mathrm{dec}}^{\mathrm{RR}}$ model under different hyperparameters, where λ is the balance weight of NLL loss and CE loss. The image number is set to 3 here.

Discussion on λ (See Table 4). To study the impact of λ, we conduct an experiment on how the model performance changes when λ varies from 0.5 to 2.0. When λ is 1.0, the model achieves the best performance. When λ is small, the image discriminator is not optimized enough, if otherwise, it may lead to over-fitting.

K	R-L	M_{sim}	IP	AE
1	37.56	**26.44**	63.32	66.61
2	37.48	26.35	62.92	66.50
3	**37.80**	26.38	**65.45**	**67.02**
4	37.44	26.23	63.55	66.57

Table 5: Results of $\mathrm{MOM}_{\mathrm{dec}}^{\mathrm{RR}}$ under different hyperparameters, where K is the image number. λ is set to 1.0.

Discussion on K (See Table 5). Table 5 depicts the experimental results of the model performance varying with K (the image number at target). Since the IP is calculated based on the top-3 images on the test set, the consistency between training and test makes the model perform best when K is 3.

According to our analysis in Section 5.2, our MMAE++ can better evaluate multimodal summary, thus we report the MMAE++ scores for our proposed models in Table 3. Besides, we conduct a manual evaluation to further compare the performance of different models, as shown in Table 6. Specifically, we select 200 multimodal summaries (randomly shuffled) from each system output, where the inputs are the same, for manual evaluation the same as described in Section 5.2. Our proposed MOF models all outperform the baselines in terms of manual evaluation or automatic evaluation, which further indicates the effectiveness of our model. $\mathrm{MOF}_{\mathrm{dec}}^{\mathrm{RR}}$ achieves both the highest MMAE++ score and the highest human judgment score, hence RR strategy is better when multimodal reference is unavailable.

Model	HS	Model	HS
ATG	**3.45**	$\mathrm{MOM}_{\mathrm{dec}}^{\mathrm{RR}}$	**3.67**
ATL	3.39	$\mathrm{MOM}_{\mathrm{enc}}^{\mathrm{RR}}$	3.52
HAN	3.35	$\mathrm{MOM}_{\mathrm{dec}}^{\mathrm{OR}}$	3.62
GR	3.30	$\mathrm{MOM}_{\mathrm{enc}}^{\mathrm{OR}}$	3.56

Table 6: Results evaluated by human annotators. HS denotes human judgment scores. Each summary is scored by two persons, and we take the average value.

6 Conclusion

In this paper, we focus on improving multimodal summarization by proposing a multimodal objective function which considers both the negative log-likelihood loss of the text summary generation and the cross-entropy loss of the image selection. Experiments show that our model can improve the quality of multimodal output on both real human-labeled test set and automatically constructed test set. Besides, we are the first to evaluate the multimodal summaries from the aspect of the information integrity which learns the joint multimodal representation for the model summary and the reference summary. We find the evaluation metric containing information integrity correlates much better with human judgments.

7 Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Pro-

gram of China under Grant No. 2016QY02D0303.

References

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

Bian, J.; Yang, Y.; Zhang, H.; and Chua, T.-S. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on multimedia (TMM)* 17(2):216–228.

Caruana, R. 1997. Multitask learning. *Machine learning* 28(1):41–75.

Celikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 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)*, 1662–1675.

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 (EMNLP)*, 4046–4056.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the international conference on Machine learning (ICML)*, 160–167.

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

Evangelopoulos, G.; Zlatintsi, A.; Potamianos, A.; Maragos, P.; Rapantzikos, K.; Skoumas, G.; and Avrithis, Y. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia (TMM)* 15(7):1553–1568.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2017. Multimodal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1092–1102.

Li, H.; Zhu, J.; Liu, T.; Zhang, J.; and Zong, C. 2018a. Multimodal sentence summarization with modality attention and image filtering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4152–4158.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2018b. Read, watch, listen and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31(5):996–1009.

Li, H.; Zhu, J.; Zhang, J.; and Zong, C. 2018c. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 1430–1441.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81.

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

Mademlis, I.; Tefas, A.; Nikolaidis, N.; and Pitas, I. 2016. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing (TIP)* 25(12):5828–5840.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML)*, 807–814.

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

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1073–1083.

Sharma, V.; Kumar, A.; Agrawal, N.; Singh, P.; and Kulshreshtha, R. 2015. Image summarization using topic modelling. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 226–231.

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

Wan, X., and Yang, J. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL (NAACL-HLT)*, 181–184.

Wang, J.; Jia, L.; and Hua, X.-S. 2011. Interactive browsing via diversified visual summarization for image search results. *Multimedia systems* 17(5):379–391.

Zhu, J.; Zhou, L.; Li, H.; Zhang, J.; Zhou, Y.; and Zong, C. 2017. Augmenting neural sentence summarization through extractive summarization. In *Proceedings of the 6th Conference on Natural Language Processing and Chinese Computing (NLPCC)*, 16–28.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4154–4164.

Zhu, J.; Wang, Q.; Wang, Y.; Zhou, Y.; Zhang, J.; Wang, S.; and Zong, C. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3045–3055.