Multi-modal Sentence Summarization with Modality Attention and Image Filtering

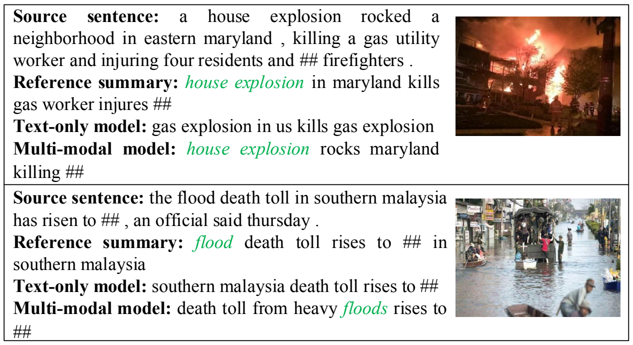
Haoran Li[1,2], **Junnan Zhu**[1,2], **Tianshang Liu**[1,2], **Jiajun Zhang**[1,2] and **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence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

Abstract

In this paper, we introduce a multi-modal sentence summarization task that produces a short summary from a pair of sentence and image. This task is more challenging than sentence summarization. It not only needs to effectively incorporate visual features into standard text summarization framework, but also requires to avoid noise of image. To this end, we propose a modality-based attention mechanism to pay different attention to image patches and text units, and we design image filters to selectively use visual information to enhance the semantics of the input sentence. We construct a multi-modal sentence summarization dataset and extensive experiments on this dataset demonstrate that our models significantly outperform conventional models which only employ text as input. Further analyses suggest that sentence summarization task can benefit from visually grounded representations from a variety of aspects.

1 Introduction

Sentence summarization is a well-studied task that creates a condensed version of a long sentence. Sequence-to-sequence (seq2seq) model that encodes a source sequence into a latent representation and outputs another sequence is the dominating framework for sentence summarization [Rush *et al.*, 2015; Takase *et al.*, 2016; Zhou *et al.*, 2017; Li *et al.*, 2017b]. Intuitively, readers can easier grasp the gist of the event by scanning the image than by reading long sentences, and thus we believe that the image will also reduce the difficulty for machine to understand a news event [Li *et al.*, 2017a]. As shown in Figure 1, the two source sentences contain complex details and multiple event objects, leading to unsatisfactory summaries using the text-only model. However, it is easy to see that the paired images visualize the event highlights which can help to produce better summaries. To explore the effectiveness of images, we introduce in this work a new multi-modal sentence summarization (MMSS) task that generates a short summary based on a pair of sentence and image.



Source sentence: a house explosion rocked a neighborhood in eastern maryland , killing a gas utility worker and injuring four residents and ## firefighters .
Reference summary: *house explosion* in maryland kills gas worker injures ##
Text-only model: gas explosion in us kills gas explosion
Multi-modal model: *house explosion* rocks maryland killing ##

Source sentence: the flood death toll in southern malaysia has risen to ## , an official said thursday .
Reference summary: *flood* death toll rises to ## in southern malaysia
Text-only model: southern malaysia death toll rises to ##
Multi-modal model: death toll from heavy *floods* rises to ##

Figure 1: Example summaries generated by different models. Multi-modal model successfully predicts the main event objects (in green and *italic*).

Multi-modal approaches are explored in machine translation (MT) recently and outperform text-only models [Calixto *et al.*, 2017; Helcl and Libovický, 2017; Caglayan *et al.*,]. The multi-modal MT requires that the paired image should describe the whole information covered by the source sentence and this requirement is hard to meet when the sentence is very long. In contrast to multi-modal MT, images which usually indicate event highlights are more suitable for summarization that aims at abstracting the important information from the original text.

To tackle the MMSS task, we design a novel seq2seq model with hierarchical attention mechanisms. The bottom attention layer focuses on the inner parts of image and sentence. The top attention layer balances the distribution of different modalities. Since some abstract concepts, such as *guilty* and *freedom*, can only be well encoded in textual modality [Paivio, 1990; Hill *et al.*, 2013; Kiela *et al.*, 2014], we design image filter modules to block out visual noises. In addition, we incorporate different visual features into our model to initialize target language decoder and present a visual coverage mechanism to handle the word repetition problem.

Our main contributions are as follows:

- We introduce an MMSS task that can generate a textual summary from a pair of sentence and image.
- We propose an inner-modality and inter-modality atten-

tion model to focus on image patches or text units for summary generation.

- We propose image filters to selectively use visual information to benefit summarization.
- We construct a multi-modal sentence summarization corpus that consists of 66,000 summary triples (sentence, image, summary). The experimental results on this dataset demonstrate that our proposed system can take advantage of multi-modal information and outperform other baseline methods.

2 Background: Seq2seq Learning

In this section, we describe the basic seq2seq learning framework. Given a source sequence $\mathbf{x}$, the seq2seq model maximizes the conditional probability of a target sequence $\mathbf{y}$: $p(\mathbf{y}|\mathbf{x})$. The encoder and the decoder are the two basic components in the seq2seq model. Recurrent neural networks (RNN) encoder [Cho *et al.*, 2014] converts $\mathbf{x}$ into a context representation c as follows:

$$h_t = f_{enc}(x_t, h_{t-1}) \tag{1}$$

$$c_t = f_c(h_1, \cdots, h_t) \tag{2}$$

where $h_t \in \mathbb{R}^n$ is a hidden state at time t, and c_t is a context vector generated from the sequence of the hidden states. f_{enc} and f_c are nonlinear activation functions.

The decoder generates word y_t given the context vector c_t and the previously generated words $\{y_1, \cdots, y_{t-1}\}$:

$$p(y_t|\{y_1, \cdots, y_{t-1}\}, c_t) = f_{dec}(y_{t-1}, s_t, c_t) \tag{3}$$

where s_t is the hidden state of the decoder and f_{dec} is a nonlinear activation function that computes the probability vector for output words at time t. The loss function $\mathcal{L}_t$ for each time t is the negative log likelihood of the target word y_t:

$$\mathcal{L}_t = -\log p(y_t|\{y_1, \cdots, y_{t-1}\}, c_t) \tag{4}$$

3 Our Proposed Model

3.1 Overview

We begin by defining the multi-modal sentence summarization task. The input of the task is a pair of sentence and image, and the output is a condensed summary. As shown in Figure 2, our proposed model consists of four modules: sentence encoder, image encoder, summary decoder and image filter. The sentence encoder is a bidirectional GRU (BiGRU) [Chung *et al.*, 2014], and pre-trained CNN (Oxford VGGnet [Simonyan and Zisserman, 2014]) is used to encode images. Our summary decoder is a uni-directional GRU with a hierarchical attention mechanism and a softmax layer over the target vocabulary to generate words. Specifically, beyond the attention over regions of the image and words of the input sentence, we explore the modality-based attention to navigate our model to pay different attention to textual and visual information when generating different words. We design two types of image filters, namely, image attention filter and image context filter, to remove visual noises and enrich the semantic meaning of the corresponding sentence.

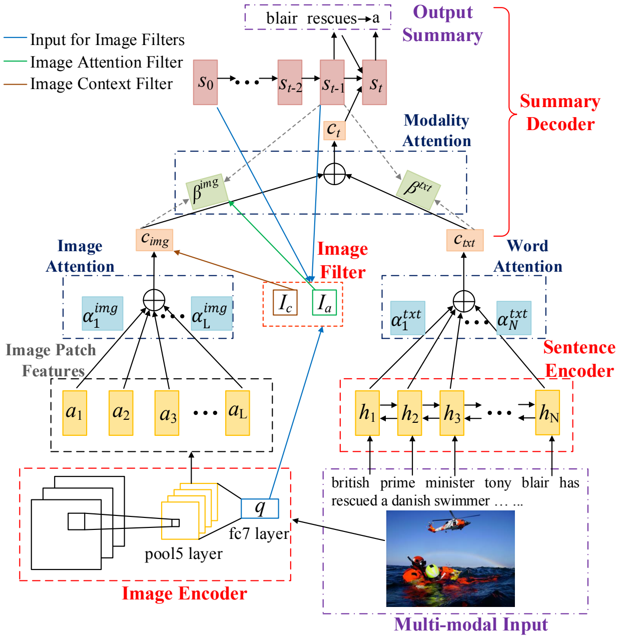


Figure 2: The framework of our model.

3.2 Sentence Encoder

Given a source sentence $\mathbf{x} = (x_1, \cdots, x_n)$, we employ a BiGRU to build its hidden representation $(h_1, \cdots, h_n)$.

The BiGRU encodes source sentence forwardly and backwardly to generate two sequences of the hidden states: $(\overrightarrow{h}_1, \cdots, \overrightarrow{h}_n)$ and $(\overleftarrow{h}_1, \cdots, \overleftarrow{h}_n)$, respectively, where:

$$\overrightarrow{h}_i = GRU(E[x_i], \overrightarrow{h}_{i-1}) \tag{5}$$

$$\overleftarrow{h}_i = GRU(E[x_i], \overleftarrow{h}_{i+1}) \tag{6}$$

$E[x_i]$ is the embedding for word x_i.

The final sentence representation h_i is the concatenation of the forward and backward vectors: $h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i]$.

3.3 Image Encoder

Given an image, we apply a pre-trained VGG (VGG19) model to extract visual features. We extract a $7 \times 7 \times 512$ feature map of the last pooling layer (`pool5` layer) of the network as the local feature, and we extract the 4096-dimensional fully-connected layer (`fc7` layer) as the global feature q. We flatten the local patch feature into matrix $\mathbf{A} = (a_1, \cdots, a_L)$ $(L = 49)$ and $a_l \in \mathbb{R}^{512}$ corresponds to a patch of the image.

3.4 Summary Decoder

At each time step t, the state of the decoder s_t is calculated as follows:

$$s_t = GRU(s_{t-1}, y_{t-1}, c_t) \tag{7}$$

In text-only seq2seq model, the decoder's hidden state s_0 is initialized as follow:

$$s_0 = \tanh(\mathbf{W}_{h_1}[\overrightarrow{h}_n; \overleftarrow{h}_1]) \tag{8}$$

Intuitively, s_0 is important because the first few words to be decoded significantly influence the remainder of the generated summary. We incorporate three different visual features into our model to initialize target language decoder: image global features (Equation 9), the average of image local features (Equation 10) and weighted image local features (Equation 11).

$$s_0 = \tanh(\mathbf{W}_{h_2}[\overrightarrow{h}_n; \overleftarrow{h}_1] + \mathbf{W}_q q) \tag{9}$$

$$s_0 = \tanh(\mathbf{W}_{h_3}[\overrightarrow{h}_n; \overleftarrow{h}_1] + \mathbf{W}_v \frac{1}{L}\sum_{i=1}^{L} a_i) \tag{10}$$

$$s_0 = \tanh(\mathbf{W}_{h_4}[\overrightarrow{h}_n; \overleftarrow{h}_1] + \mathbf{W}_w \frac{1}{L}\sum_{i=1}^{L} \alpha_i a_i) \tag{11}$$

where α_i is the weight scalar for a_i:

$$\alpha_i = \sigma(\mathbf{W}_{h_5}[\overrightarrow{h}_n; \overleftarrow{h}_1] + \mathbf{W}_p a_i) \tag{12}$$

where σ denotes sigmoid activation function. The underlying intuition for Equation 11 is that different image patches do not make equal contributions aligning to the source sentence; for instance, background regions often contain trivial information, which would have less contribution to the model than foreground regions. To do so, the image patch and the sentence are matched to determine informative patches.

Context vector c_t in Equation 7 is obtained by a modality-based attention mechanism as follows:

$$c_t = \beta_t^{txt}\mathbf{V}_T c_{txt} + \beta_t^{img}\mathbf{V}_I c_{img} \tag{13}$$

where β_t^{txt} is attention weight for text context vector c_{txt} and β_t^{img} is attention weight for image context vector c_{img}, respectively, calculated as follows:

$$\beta_t^{txt} = \sigma(\mathbf{U}_a s_{t-1} + \mathbf{W}_a c_{txt}) \tag{14}$$

$$\beta_t^{img} = \sigma(\mathbf{U}_b s_{t-1} + \mathbf{W}_b c_{img}) \tag{15}$$

We compute the text context vector c_{txt} as a weighted sum of the source annotations as follows:

$$c_{txt} = \sum_{i=1}^{N} \alpha_{t,i}^{txt} h_i \tag{16}$$

where each vector is weighted by the attention weight $\alpha_{t,i}^{txt}$, as calculated in Equations 17 and 18 as follows:

$$e_{t,i}^{txt} = v_a^T \tanh(\mathbf{U}_c s_{t-1} + \mathbf{W}_c h_i) \tag{17}$$

$$\alpha_{t,i}^{txt} = \frac{\exp(e_{t,i}^{txt})}{\sum_{j=1}^{N}\exp(e_{t,j}^{txt})} \tag{18}$$

We compute the image context vector c_{img} as a weighted sum of the source image patch representation $\mathbf{A}$ using the "soft" attention strategy [Xu *et al.*, 2015]. This process is very similar to the text attention strategy assuming that decoder needs to attend to different image patches to generate different words.

$$c_{img} = \sum_{i=1}^{L} \alpha_{t,i}^{img} a_l \tag{19}$$

where each visual vector a_l is weighted by the attention weight $\alpha_{t,i}^{img}$, as calculated in Equations 20 and 21 as follows:

$$e_{t,i}^{img} = v_b^T \tanh(\mathbf{U}_d s_{t-1} + \mathbf{W}_d a_l) \tag{20}$$

$$\alpha_{t,i}^{img} = \frac{\exp(e_{t,i}^{img})}{\sum_{j=1}^{N}\exp(e_{t,j}^{img})} \tag{21}$$

The probability for the next target word y_t is computed using the multi-modal hidden state s_t, and the previously emitted word y_{t-1} as follows:

$$p(y_t|\{y_1, \cdots, y_{t-1}\}) \propto \exp(\mathbf{L}_y tanh(\mathbf{L}_s s_t + \mathbf{L}_e E[y_{t-1}]) \tag{22}$$

where $\mathbf{L}_y$, $\mathbf{L}_s$ and $\mathbf{L}_e$ are model parameters.

In addition, to solve the problem of repetition, we adopt the textual coverage mechanism [See *et al.*, 2017] which aims to avoid generating repetitive target words by penalizing repetitive attention to the same position of the source sequence. For coverage model, the attention distributions over all previous decoder time are maintained, which record the attention history for all source words. Then the decoder is not expected to pay current attention to the words which have received enough attention before. For our task, we apply a multi-modal coverage mechanism including textual and visual coverage.

In our multi-modal coverage model, we maintain two coverage vector $c_{t,i}^{txt}$ and $c_{t,i}^{img}$, which is the sum of attention distributions over all previous decoder steps:

$$c_{t,i}^{txt} = \sum_{\tau=1}^{t-1} \alpha_{\tau,i}^{txt} \tag{23}$$

$$c_{t,i}^{img} = \sum_{\tau=1}^{t-1} \alpha_{\tau,i}^{img} \tag{24}$$

Intuitively, $c_{t,i}^{txt}$ and $c_{t,i}^{img}$ denote the degree of coverage that the source words and the image patches have been covered by the target before time t, respectively. Note that $c_{0,i}^{txt}$ and $c_{0,i}^{img}$ are zero vectors, because at that step, none of the source word or the image patch is covered by the target.

The coverage vectors are used to calculate the attention mechanism, changing Equation 17 and 20 to:

$$e_{t,i}^{txt} = v_a^T \tanh(\mathbf{U}_c s_{t-1} + \mathbf{V}_c c_{t,i}^{txt} + \mathbf{W}_c h_i) \tag{25}$$

$$e_{t,i}^{img} = v_b^T \tanh(\mathbf{U}_d s_{t-1} + \mathbf{V}_d c_{t,i}^{img} + \mathbf{W}_d a_l) \tag{26}$$

The coverage losses for text and image at time t are defined as follows respectively:

$$\mathcal{L}_t^{cov_{txt}} = \sum_{i=1}^{N} \min(\alpha_{t,i}^{txt}, c_{t,i}^{txt}) \tag{27}$$

$$\mathcal{L}_t^{cov_{img}} = \sum_{i=1}^{L} \min(\alpha_{t,i}^{img}, c_{t,i}^{img}) \tag{28}$$

Finally, we add the multi-modal coverage loss to the original loss function in Equation 4 to yield new loss functions as follows:

$$\mathcal{L}_t^{txt} = \mathcal{L}_t + \mathcal{L}_t^{cov_{txt}} \tag{29}$$

$$\mathcal{L}_t^{img} = \mathcal{L}_t + \mathcal{L}_t^{cov_{img}} \tag{30}$$

3.5 Image Filter

Compared to text summarization task, the key point for multi-modal summarization is effectively making use of image. Meanwhile, we should avoid introducing noises in the case that (1) the image fails to represent some semantic meanings of words, such as abstract concepts, or (2) the image is insufficiently exact to capture the keypoint of the source sentence. To this end, we introduce an image filtering mechanism and propose two types of image filters, i.e., an image attention filter and an image context filter.

Image Attention Filter

The image attention filter is a scalar applied to the image attention β_t^{img} (in Equation 15), which aims to re-balance the attention of the image and sentence based on the following aspects: (1) The correlation between the source image and sentence, which is measured by the image global feature q and the initial state of summary decoder s_0. (2) The relation to the next target word, which relates to the state of the decoder s_{t-1}. The image attention filter $I_a \in [0, 1]$ is calculated as follows:

$$I_a = \sigma(v_s^T s_0 + v_q^T q + v_r^T s_{t-1}) \tag{31}$$

Then, the image attention β_t^{img} in Equation 15 is updated as follows:

$$\beta_t^{img} = I_a \cdot \beta_t^{img} \tag{32}$$

Xu et al. [2015] and Calixto et al. [2017] adopt gating scalars to image context. The difference is that our scalar not only relates to the current decoder state for generating the next target word but also associate different modalities, which aims to evaluate the effect of the image.

Image Context Filter

The image context filter is an element-wise filtering vector applied to the image context c_{img} (in Equation 19) that is expected to filter out the image noises inside the image context. The image context filter is calculated as follows:

$$I_c = \sigma(\mathbf{W}_s s_0 + \mathbf{W}_q q + \mathbf{W}_r s_{t-1}) \tag{33}$$

The image context in Equation 19 is updated as follows:

$$c_{img} = I_c \odot c_{img} \tag{34}$$

Image context filter is partially inspired by gating mechanism which has gained great popularity in neural network models [Hochreiter and Schmidhuber, 1997; Srivastava *et al.*, 2015], and this filter is expected to give fine tune for each dimension, and the information can only flow on the dimensions where the gate is "open". In conclusion, image attention filter is directly applied to change the attention scale between image and text, and image context filter is designed to select the most salient visual features.

4 Related work

4.1 Seq2seq Summarization Models

Rush et al. [2015] are the first to apply the seq2seq model to abstractive sentence summarization. They propose an attentive CNN encoder and a neural network language model [Bengio *et al.*, 2003] decoder. Chopra et al. [2016] use RNN as the decoder and achieve better performance. Nallapati et al. [2016] further replace the encoder with an RNN, forming a full RNN seq2seq model. Gu et al. [2016] and Zeng et al. [2016] incorporate a copying mechanism into seq2seq learning and Gulcehre et al. [2016] propose a switch gate to control whether to copy from the source or generate a word by the decoder. See et al. [See *et al.*, 2017] present a seq2seq model with coverage mechanism to eliminate repetition. Current seq2seq summarization methods only focus on text, while the opportunity to optimize the summary quality with the aid of visually grounded representations is ignored.

4.2 Multi-modal Seq2seq Models

Multi-modal MT has recently attracted much attention from researchers. Libovický and Helcl [2017] propose multi-source seq2seq learning with hierarchical attention combination. Calixto and Liu [2017] use images as words in the source sentences to improve translation quality. They use visual representation to initialize the encoder or decoder. Our work is partially inspired by the models of Libovický and Helcl [2017] and Calixto et al. [2017] with some differences: (1) Libovický et al. [2017] use the hierarchical attention but do not pay attention to image patches. In addition, they do not apply image filters. (2) Calixto et al. [2017] propose two separate attention mechanisms to decode target words, while our model can hierarchically pay attention to source words, image patches and different modalities. (3) Calixto et al. [2017] apply a gating scalar which only relates to decoder state, but our image filters not only relate to the current decoder state for generating the next target word but also consider the correlations between different modalities, which is peculiar for sentence summarization tasks. Our motivation is that (1) Different image patches relate to different words and image filters enable our model to selectively use the image. (2) Hierarchical attention mechanism can decide how much attention the decoder pay to different modalities when generate different words. (3) Our image filters explore association between textual and global visual information, which can measure the extent to which the image can capture the gist of the source sentence.

5 Dataset Construction

To the best of our knowledge, there is no benchmark dataset for the MMSS task, and we construct a corpus. Each sample in our corpus is a triple (sentence, image, headline) in which the sentence-headline pair is from the annotated Gigaword corpus [Rush *et al.*, 2015][1] and the image is crawled from Yahoo! Image Search. The Gigaword corpus provides 3.8 million first-sentence-headline pairs. For each first-sentence, we search Yahoo! Image Search, and crawl the top-5 ranked images. Next, we delete the explicit trivial images such as portraits, thumbnails and advertisements.

As a result, we collect 123,839 sentences paired with five images each. Then we employ 10 graduate students to select the best-match image for each sentence. If there is no matching image for the sentence, the annotators label '0'. Each five-image-sentence sample is annotated by two students and

[1] https://github.com/harvardnlp/sent-summary

Model		R-1	R-2	R-L
Lead		33.64	13.40	31.84
Compress [Clarke and Lapata, 2008]		31.56	11.02	28.87
ABS [Rush *et al.*, 2015]		35.95	18.21	31.89
SEASS [Zhou *et al.*, 2017]		44.86	23.03	41.92
Multi-Source [Libovický and Helcl, 2017]		39.67	19.11	38.03
Doubly-Attentive [Calixto *et al.*, 2017]		41.11	21.75	39.92
Our text only model		44.58	22.68	41.91
Multi-modal model without image filter		44.88	23.20	42.11
Multi-modal model with attention filter	Decoder$_{text}$	45.17	23.39	42.20
	Decoder$_{fc}$	45.02	23.06	42.24
	Decoder$_{aconv}$	45.21	23.82	42.50
	Decoder$_{wconv}$	45.78	23.45	43.16
	+text coverage	**47.28**	**24.85**	**44.48**
	+image coverage	47.16	24.43	44.23
Multi-modal model with context filter	Decoder$_{text}$	45.43	23.47	42.67
	Decoder$_{fc}$	45.12	23.29	42.34
	Decoder$_{aconv}$	45.99	23.86	43.19
	Decoder$_{wconv}$	46.08	24.00	43.29
	+text coverage	46.84	24.25	43.76
	+image coverage	46.56	24.34	43.59

Table 1: Main experimental results. **Decoder**$_{text}$, **Decoder**$_{fc}$, **Decoder**$_{aconv}$ and **Decoder**$_{wconv}$ denote different decoder initialization: by only text, by text and global visual feature, by text and average of local visual feature, by text and weighted average of local visual feature (in Equation 8-11), respectively. We report different coverage mechanisms applied to (**MM Decoder**$_{wconv}$) model. Our **MM** models with two kinds of image filters perform significantly better than our **Text-only** model by the 95% confidence interval in the ROUGE script.

the samples with accordance annotations are preserved. Finally, we collect 66,000 samples. Following the settings of the Gigaword summarization corpus, we randomly split our corpus as a training set with 62,000 samples, a test set with 2,000 samples and a development set with 2,000 samples.

6 Experiment

6.1 Comparative Methods

We compare a set of sentence summarization and multimodal MT baselines. **Lead** baseline uses the first 8 words as the summary. **Compress** model [Clarke and Lapata, 2008] produces a compressed result based on the syntactic structure of the original sentence. **ABS** [Rush *et al.*, 2015] uses an attentive CNN encoder and a neural network language model decoder to summarize the sentence. **SEASS** is a state-of-the-art sentence summarization systems, which employs a selective encoding model to control the information flow from the encoder to the decoder. **Multi-Source** [Libovický and Helcl, 2017] and **Doubly-Attentive** [Calixto *et al.*, 2017] are multimodal MT baselines (introduced in Section 4). **Text-only** model is a seq2seq model with attention only using a sentence to generate summary. We report performances of our multi-modal (**MM**) models with different decoder initializations, image attention filter, image context filter, and coverage for different modalities.

6.2 Experimental Settings

We set word embedding size to 300 and GRU hidden state sizes to 512. We use the full source and target vocabularies collected from the training data, which have 36,916 and 26,168 words, respectively. We use dropout [Srivastava *et al.*, 2014] with probability of 0.2. We set the initial learning rate for Adam [Kingma and Ba, 2015] to 5×10^{-4}. At training time, we test ROUGE-2 [Lin, 2004] F1 score on the development set for every 2,000 batches, and we halve the learning rate if model performance drops. Our models typically converge within 50 epochs using an early stopping strategy for our **MM** model without coverage. To obtain our final coverage model, we add the coverage loss to the objective function for further training with the initial learning 5×10^{-5}. At test time, we use beam search with beam size 10 to generate the summary. We report ROUGE F1 score including ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L).

6.3 Experimental Results

Table 1 shows the results of the comparative methods and our proposed methods. The state-of-the-art sentence summarization systems **SEASS** performs better than our **Text-only** model. For our **MM** models, the model without an image filter performs only slightly better than the **text-only** model, and the models with **image filters** outperform the **text-only** model and other baseline models. This proves that image filter plays a significant role in the MMSS task. Initializing decoder with image features is effective for our task. Compared to global and average of local visual features, the models with weighted average of local visual feature achieve greater improvements over the text-only model, which suggests that different image patches do not make equal contributions for our task. Coverage mechanism leads to further improvement, which achieves +2.70% R-1, +2.17% R-2, +2.57% R-L improvement over the **Text-only** model) for our best model. In addition, we have tested an image captioning model [Vinyals *et al.*, 2015] and get poor results (R-1 is lower than 10%), which indicates that image captioning model is not suitable for this task.

6.4 Analysis

In this section, we discuss the effectiveness of visual information for sentence summarization task (results are reported on the test set).

1. Effectiveness of multi-modal initialization of decoder. From Table 1 we observe that incorporating images to initialize hidden state of decoder improves the performance. We conjecture that image features are effective as these features can capture the highlights of the source sentences. Table 2 shows that our image attention filter model with multimodal decoder initialization can predict the first words that appear in the reference summary in much higher accuracies than that with text-only initialization. Example summaries can be found in Figure 1. In part at least, this behavior can be ascribed to better initialization of decoder. Furthermore, we observe the model which predicts the first word correctly achieves 0.235 higher ROUGE-1 score on average than the model which predicts wrong in our experiment.

Decoder initialization	text	fc	aconv	wconv
The first 1 word	42.2	42.6	43.4	44.4
The first 2 words	23.2	23.3	23.5	24.0
The first 3 words	14.2	14.4	14.2	14.6

Table 2: The first words predicting accuracy (%) for the decoders with different initializations (Equation 8-11).

n-gram	1-gram	2-gram	3-gram
Text-only	0.47	0.64	0.46
MM	0.48	0.66	0.48
Reference	2.89	5.01	5.09

Table 3: The average count of correct novel n-grams (i.e., n-grams that don't appear in the source but appear in the reference summary).

2. Are images helpful to produce more abstractive summary? Abstractive summarization can produce novel words that do not appear in source sentence. This phenomenon is common for human-written summaries, but is challenging for a machine. We believe that attention mechanism allows a model to focus on the specific positions of source representation that will help to predict the next word, but it adds a constraint on the context vectors for summarization system, which tends to predict a word obtaining the greatest attention in the source. Table 3 shows that our multi-modal model is more abstractive than text-only model. This is expected because the attention to the image can provide abundant information beyond source text. An example is shown in Figure 3.

3. Which kind of images are more useful? To answer this question, we ask 3 graduate students to annotate the matching scores for 300 sentence-image pairs in the test set. The score ranges from 1 to 3: 1 denotes "partially match", 2 denotes "basically match" and 3 denotes "completely match". The results in Table 4 show that the samples with better matched images achieve higher ROUGE score and also attract higher attention for image, indicating that better matched images are more useful for the MMSS task.

4. Effectiveness of multi-modal coverage. To explicitly demonstrate how the repetition problem is eliminated by coverage mechanism, we calculate the average count of repeated words for each summary with coverage for different modalities. Figure 4 shows the results for (**MM Decoder**$_{wconv}$) model with different image filters, which indicates that textual and visual coverage model can reduce repeated words.

Match score	R-1	R-2	R-L	Image attention
1 (9%)	41.47	21.37	39.88	16.11
2 (15%)	44.42	23.49	41.30	23.32
3 (76%)	45.89	26.03	44.03	35.16

Table 4: ROUGE score and image attention (%) for samples with different text-image matching scores (proportion in 300 annotated samples).

Source sentence: at least ## people were killed and ## injured when a passenger bus plunged into a deep ravine in north ethiopia , police said wednesday .
Reference summary: *bus accident* kills at least ## in north ethiopia
Text-only model: ## killed as bus plunges into ravine
Multi-modal model: ## killed in ethiopian *bus accident*

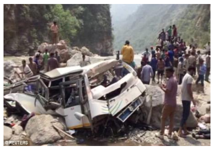


Figure 3: Multi-modal model successfully predicts the novel words (in blue and *italic*).

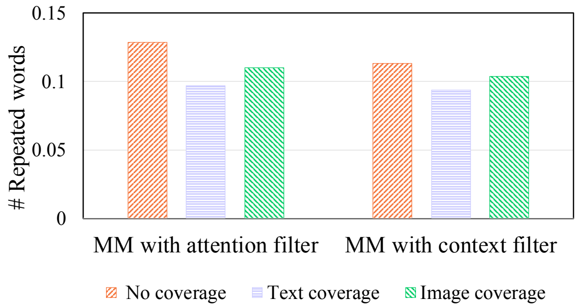


Figure 4: The count of repeated words for different models.

7 Conclusions

This paper addresses a multi-modal sentence summarization task, namely, how to transform a sentence-image pair into a short-length summary. Our proposed model can simultaneously focus on image patches and text units to generate summaries with a modality-based attention mechanism. We design image filters to selectively use visual information to enrich the semantics for the source sentence. We introduce a visual coverage mechanism which prove that multi-modal coverage is effective for our task. We provide a publicly available multi-modal sentence summarization corpus and the experiments on this dataset show our proposed model significantly outperforms the baseline models. Our dataset is released to the public[2].

Acknowledgments

The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303 and the Natural Science Foundation of China under Grant No. 61333018.

References

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *JMLR*, pages 1137–1155, 2003.

[Caglayan *et al.*,] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. Lium-cvc submissions for wmt17 multimodal

[2] http://www.nlpr.ia.ac.cn/cip/jjzhang.htm

translation task. In *Proceedings of the Second Conference on Machine Translation*.

[Calixto and Liu, 2017] Iacer Calixto and Qun Liu. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*, pages 992–1003, 2017.

[Calixto *et al.*, 2017] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pages 1913–1924, 2017.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Chopra *et al.*, 2016] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL*, pages 93–98, 2016.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014.

[Clarke and Lapata, 2008] James Clarke and Mirella Lapata. Global inference for sentence compression an integer linear programming approach. *AI Access Foundation*, 2008.

[Gu *et al.*, 2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, pages 1631–1640, 2016.

[Gulcehre *et al.*, 2016] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *ACL*, pages 140–149, 2016.

[Helcl and Libovický, 2017] Jindřich Helcl and Jindřich Libovický. Cuni system for the wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 450–457, 2017.

[Hill *et al.*, 2013] Felix Hill, Douwe Kiela, and Anna Korhonen. Concreteness and corpora: A theoretical and practical analysis. In *Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–83, 2013.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, pages 1735–1780, 1997.

[Kiela *et al.*, 2014] Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*, pages 835–841, 2014.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

[Li *et al.*, 2017a] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *EMNLP*, pages 1092–1102, 2017.

[Li *et al.*, 2017b] Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. Deep recurrent generative decoder for abstractive text summarization. In *EMNLP*, pages 2091–2100, 2017.

[Libovický and Helcl, 2017] Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *ACL*, pages 196–202, 2017.

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

[Nallapati *et al.*, 2016] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*, pages 280–290, 2016.

[Paivio, 1990] Allan Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1990.

[Rush *et al.*, 2015] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, pages 1073–1083, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[Srivastava *et al.*, 2015] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv:1505.00387*, 2015.

[Takase *et al.*, 2016] Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. Neural headline generation on abstract meaning representation. In *EMNLP*, pages 1054–1059, 2016.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[Zeng *et al.*, 2016] Wenyuan Zeng, Wenjie Luo, Sanja Fidler, and Raquel Urtasun. Efficient summarization with read-again and copy mechanism. *arXiv:1611.03382*, 2016.

[Zhou *et al.*, 2017] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. In *ACL*, pages 1095–1104, 2017.