Received: 11 December 2019 | Revised: 14 February 2020 | Accepted: 15 February 2020
DOI: 10.1002/cpe.5721

WILEY

SPECIAL ISSUE PAPER

A news image captioning approach based on multimodal pointer-generator network

Jingqiang Chen[1,2,3,4] | **Hai Zhuge**[1,2,3,4]

[1]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

[2]School of Computer Science and Network Engineering, Guangzhou University, Guangzhou, China

[3]Department of Computer Science, School of Engineering and Applied Science, Aston University, Birmingham, UK

[4]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing, China

Correspondence
Hai Zhuge, Department of Computer Science, School of Engineering and Applied Science, Aston University, Birmingham, UK.
Email: haizhuge@gmail.com and h.zhuge@aston.ac.uk

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61876048

Summary

News image captioning aims to generate captions or descriptions for news images automatically, serving as draft captions for creating news image captions manually. News image captions are different from generic captions as news image captions contain more detailed information such as entity names and events. Therefore, both images on news and the accompanying text are the source of generating caption of news image. Pointer-generator network is a neural method defined for text summarization. This article proposes the Multimodal pointer-generation network by incorporating visual information into the original network for news image captioning. The multimodal attention mechanism is proposed by splitting attention into visual attention paid to the image and textual attention paid to the text. The multimodal pointer mechanism is proposed by using both textual attention and visual attention to compute pointer distributions, where visual attention is first transformed into textual attention via the word-image relationships. The multimodal coverage mechanism is defined to reduce repetitions of attentions or repetitions of pointer distributions. Experiments on the *DailyMail* test dataset and the out-of-domain *BBC* test dataset show that the proposed model outperforms the original pointer-generator network, the generic image captioning method, the extractive news image captioning method, and the *LDA*-based method according *BLEU*, *METEOR*, and *ROUGL-L* evaluations. Experiments also show that the proposed multimodal coverage mechanisms can improve the model, and that transforming visual attention to pointer distributions can improve the model.

KEYWORDS
image captioning, multimodal summarization, pointer-generator network, text summarization

1 | INTRODUCTION

News image captioning aims to generate captions or descriptions for news images automatically. News image captioning can provide draft captions for news images which is significant in applications because there are great number of news images on the Internet and to create manual captions is a time- and labor-consuming task. Many online news sites such as *CNN*, *DailyMail*, *BBC*, *Yahoo*! publish images with their stories and even provide photo feeds related to current events. These news sites are a good resource for multimedia files containing information in form of videos, images, and natural language texts, and thus provide captioned news images corpora for supervised machine-learning based captioning approaches.

News image captioning is different from generic image captioning mainly in that news images are related to the texts of news and therefore captions of news images should contain information of the surrounding texts of news images. The caption of a news image often reflects the specific event reported in the news. While generic image captioning generates generic image captions which only contain generic information of images and

	**Angus the rhino with mum Dorothy at Blair Drummond Safari Park.**
	Two rhinos stand before a barrier.
	**The 14-year-old triathlete has become the youngest person ever to complete a marathon on the continent.**
	A person runs beside an iced lake.
	**There was little sign of life at the North Korean embassy in London's Ealing today.**
	A big house with a car in front of it.
	**Two-year-old Archie Watson suffered from a rare genetic disorder and died.**
	A little boy wearing in green clothes smiles.
	**A man who lost his class ring more than 40 years ago was amazed when it was returned by a woman on Facebook.**
	A man's hand wearing a ring in a finger.

FIGURE 1 Five news images taken from *DailyMail*, each with its original caption in the top right and a generic caption in the bottom right

cannot reflect specific information in news. This is mainly because generic image captioning focuses on the image itself which is the only information that can be used for generating the caption. Generic image captioning does not take into consideration the related or surrounding text of the image because not all images have related texts as news images do.

Figure 1 shows the news image captions and generic image captions for five news images taken from *DailyMail* to demonstrate the difference between the two types of captions. One difference is that news image captions usually contain more detail information than generic captions. For example, the generic caption of the first image is "Two rhinos stand before a barrier," while the original news image caption gives the names of the two rhinos, the relationships between them, and the name of the standing place *Safari Park*. The detail information is contained in the text of the news. The other difference is that news image captions often contain information on specific events reported in news. Taking the fourth image, for example, its original caption shows that the boy suffered from a disaster and died while the generic caption does not contain this information. Another example is about the second image. The original caption contains the event that a young person completes a marathon on the continent while the generic caption only shows a person runs beside a lake shown in the image. It is hard to deduce the detailed information and the event information merely from images because they are usually contained in the text of news.

Therefore, to generate news image captions shown in Figure 1, both the image and the text of the news should be taken into consideration. Recent image captioning techniques focus on generic image caption generation and mainly utilize image information to generate captions,[1-5] and do not make full use of the accompanying text of the image because not all images have accompanying texts. On contrary, neural text summarization methods focus on text information to generate summaries,[6] while neglecting image information.

This article proposes the multimodal pointer-generator network for news image captioning by incorporating visual information into the state-of-the-art text summarization model pointer-generator network.[6-8] The proposed network consists of a multimodal generator and a multimodal pointer mechanism. At each decoding step, the model splits the attention into textual attention paid to text and visual attention paid to the image. The generator computes the vocabulary distributions from the text and the image. The pointer mechanism transforms visual attention into textual attention through the Word-VisualPart relationships which are defined between each source word and each visual part of the source image. The pointer distributions are computed by summing up the transformed visual attention and the visual attention. Tow multimodal coverage mechanisms are proposed to alleviate the repetition problems: one is defined over attention and the other is defined over pointer distributions. The model is trained on the corpora constructed by collecting the first images of the news in the original *DailyMail* corpora.[9]

The main contributions of this article are as follows:

(1) The multimodal pointer-generator network is proposed for news image captioning by incorporating visual information into the original pointer-generator network.
(2) The multimodal pointer mechanism is proposed in the network to utilize both textual attention and visual attention to compute pointer distributions by first transforming visual attention into textual attention based on the *Word-VisualPart* relationships modeled by an attention mechanism.
(3) Two multimodal coverage mechanisms are defined in the model by reducing the repetitions of attentions or the repetitions of pointer distributions.

Experiments carried out on the *DailyMail* test dataset and the out-of-domain *BBC* test dataset show that the proposed multimodal pointer-generator network outperforms the original pointer-generator network, the generic image captioning method, the *LDA*-based method, and the neural extractive method according to *BLEU*, *METEOR*, and *ROUGE-L* measures. Experiments also show that incorporating visual attention into the pointer mechanism can improve the proposed model, and that the two multimodal coverage mechanisms can also improve the proposed model.

2 RELATED WORK

News image captioning is tightly related to but is also different from text summarization and generic image captioning. The former generates short summaries from texts, and the latter generates captions from images.

Text summarization can be used to generate news image captions by summarizing news texts. Recent neural text summarization models are based on the attentional encoder-decoder model which is first proposed in machine translation area to generate text and to align the original text and the translated text.[6-8] The pointer-generator network is proposed to alleviate the out-of-vocabulary (OOV) problem of the encoder-decoder model[6,10] by copying words from source texts. The model is applied to sentence summarization by considering the neural language model and the attention model when generating next words.[11] A neural document summarization model is proposed by extracting sentences and words,[9] where sentences are extracted by computing the probability of sentences belonging to the summary based on an *RNN* model, and word are extracted from the original document based on an attentional decoder. An *RNN*-based extractive summarization named *SummaRuNNer*, which treats summarization as a sentence classification problem and applies a logistic classifier using coverage features and redundancy features computed based on the *RNN* model is proposed.[12] Neural multidocument summarization is also studied. The hierarchical transformer is proposed for multidocument summarization by adding inter-paragraph attention into the transformer.[13,14] The *MMR*[15] is incorporated into the pointer-generator network for multidocument summarization.[16,17] Another category of summarization is based on discovery of features in texts and between texts and images.[18]

Most neural image captioning models are combination of the convolution neural network (CNN) and the recurrent neural network (RNN).[19-28] The *CNN* models such as *VGGNet*,[25] *AlexNet*,[29] *GoogleNet*,[30] *ResNet*,[31] and *SENets*[32] are used to encode images by extracting the last full-connected layers or convolution layers as the vector representations of images. The *RNN* models are used to encode and decode captions. The first deep learning-based image captioning model is proposed in Reference [8] by using a multimodal RNN guided by image information. The encoder-decoder model was further applied to image captioning by encoding image using the *CNN* model which is fed into the *RNN* decoder to generate words one by one.[1] The image encoding is used only once in the decoder as the initial input, and the previously generated word is used as the only input to the next decoding steps to generate the next words. The model is extended in Reference [2] by adding the attention mechanism[7,8] where the image is split into multiple parts which are taken as the initial input of the decoder and are attended to compute the context in each decoding step. The correctness of attention mechanism was further studied in Reference [4] and a supervised attention mechanism is proposed, and the results show that the alignments created by the attention mechanisms are in high accordance with manual alignments. The semantic attention mechanism is proposed in Reference [3] which makes use of image tags as additional information and attends image tags during decoding. A multimodal transformer-based model is proposed in Reference [5] to ingest both entity labels and image features for generic image captioning. Image can also be represented as a collection of objects which are encoded by the *RNN* model, and then the attentional mechanism is applied to the objects in the decoder.[22-24] A hierarchical LSTMs with adaptive attention for visual captioning was proposed.[33] Other advances in image captioning are based on reinforcement learning[25] and the generative adversarial nets.[26]

News image captioning is different from generic image captioning and text summarization. The image information and the text information should be both taken into consideration to generate captions for news images. An early study on news image captioning is based on the probabilistic model.[34] It treats the image and text as a collection of visual words and textual words, and applies the *LDA* model to compute a mixture model of topics and words, based on which the extractive caption generation model and abstractive caption generation model are proposed.[35] A neural extractive news image captioning method is proposed in Reference [36]. The method encoded the ordering embeddings of images and texts with *LSTM* into a context vector to summarize the multimodal document, and uses as the object function the cross entropy between captions and the context vector. The sentences are extracted as captions based on the cosine similarity with the context vector. Other similar work is text-image summarization,[37,38] which creates multimodal summaries with images aligned with sentences.

This work is different from generic image captioning and text summarization in that both the image and the accompanying text are utilized to generate news image captions. The multimodal pointer-generator network is proposed by incorporating image information into the original pointer-generator network.

3 BACKGROUND: NEURAL SUMMARIZATION AND IMAGE CAPTIONING

The encoder-decoder architecture has become the de facto standard for neural abstractive text summarization[11] and neural image captioning.[2] The encoder for text summarization is often a bidirectional *LSTM*[39] or gated recurrent unit (*GRU*)[40] converting the input text to a set of hidden states

$\{h^{eT}{}_i\}$, one for each input word, indexed by *i*. The encoder for image captioning is often a *CNN*-based model pretrained in *ImageNet*[29] converting the image into visual vector representations $\{h^{eV}{}_i\}$ by extracting the last full-connection layer or the last convolution layer. The decoder of both neural text summarization and neural image captioning is a unidirectional *RNN* (*LSTM* or *GRU*) that generates a summary or a caption by predicting one word at a time. The decoder hidden states are represented by $\{h^d{}_i\}$, indexed by *t*. For text summarization, the input text is treated as a sequence of words, and the model is expected to capture the source syntax inherently. For image captioning, the input image is treated as a set of visual parts, and the model is expected to capture relationship between the image and the caption.

$$a_{t,i} = v^T \tanh(W^a[h_t^d \| h_i^e \| \mathrm{cov}_{t,i}] + b^a), \tag{1}$$

$$\alpha_{t,i} = soft\ \max(a_{t,i}), \tag{2}$$

$$\mathrm{cov}_{t,i} = \sum_{t'=0}^{t-1} \alpha_{t,i}. \tag{3}$$

The attention mechanism is proposed to improve the encoder-decoder model. Equations (1) to (3) are the equations for attention calculations in each decoding step used in pointer-generator network.[6] The attention weight $\alpha_{t,i}$ measure the importance the input words or visual parts of the image to generating each output word (Equations (1) and (2)), calculated by measuring the strength of interaction between the decoder hidden stated $h_t{}^d$, the encoder hidden state h^e (the text encoder h^{eT} or the image encoder h^{eV}), and the cumulative attention $\overline{a}_{t,i}$ (Equation (3)). The notation $\mathrm{cov}_{t,\,i}$ denotes the degree of coverage which the *i*th input word of the source text or input visual part of the source image receives for the first decoding step to the *i*−1th step. A large value of $\mathrm{cov}_{t,\,i}$ indicates the *i*th input word or visual part has been used prior to time *t* and it is unlikely to be used again for generating the *t*th output word.

$$c_t = \sum_i \alpha_{t,i} h_i^e, \tag{4}$$

$$P_{\mathrm{vocab}}(w) = soft\ \max(W^y[h_t^d \| c_t] + b^y). \tag{5}$$

The context vector c_t is computed by Equation (4) to summarize the semantic meaning of the input, which is a weighted summation of the encoder hidden states. The vocabulary probability $P^{\mathrm{vocab}}(w)$ which measures the probability of a vocabulary word *w* being selected as the *t*th output word are then computed by using the context vector and the decoder $(h^d{}_t||c_t)$ in Equation (5).

$$P_{\mathrm{gen}} = \mathrm{sigmoid}(w^z[h_t^d \| c_t \| y_{t-1}] + b^z), \tag{6}$$

$$P(w) = p_{\mathrm{gen}} P_{\mathrm{vocab}}(w) + (1 - p_{\mathrm{gen}}) \sum_{i:w_i=w} \alpha_{t,i}. \tag{7}$$

Especially, to deal with the OOV problem of neural text summarization, a copy mechanism is provided in pointer-generator network by adding a "switch" which is estimated ($p_{\mathrm{gen}} \in [0, 1]$) to indicate whether the system has chosen to select a word from the vocabulary or to copy a word from the source text (Equation (6)). The switch is calculated using a feed forward layer with the sigmoid activation over $[h^d{}_t||c_t||y_{t-1}]$, where y_{t-1} is the embedding of the output word at the *t*−1th decoding step. Equation (7) computes the final probability *P*(*w*) for the word w which is a weighted combination of the vocabulary probability and the copy probability. The attention weights of the word *w* is used to calculate the copy probability $\sum_{i:w_i=w} \alpha_{t,i}$. If the word *w* appears once or more times in the source text, the copy probability is the summation of its occurrences. For image captioning, this type of pointer mechanism is not applicable though it also has *OOV* problems.

$$\mathrm{cov\ loss}_t = \sum_i \min(\alpha_{t,i}, \mathrm{cov}_{t,i}), \tag{8}$$

$$\mathrm{loss}_t = -P(w^*) + \lambda\ \mathrm{cov\ loss}_t. \tag{9}$$

The objective function for training *pointer-generator network* consists of two parts (Equations (8) and (9)): the primary negative log-likelihood loss function ($-P(w^*)$) and the coverage loss ($\sum_i \min(\alpha_{t,\,i}, \mathrm{cov}_{t,\,i})$). The coverage loss is bounded and is always less than 1. The coverage loss is used to penalize repeatedly attending to the same input words of the source text or the same visual parts of the source image, and thus can alleviate the word or phrase repetition problem. This type of coverage loss is initially defined for neural text summarization,[6] and is also applicable for image captioning.

For text summarization, the model can be trained on text summarization data containing a large collection of news articles paired with summaries.[6] For generic image captioning, the model can be trained on image captioning data containing a large collection of images paired with captions.[1,2] The inputs of text summarization and image captioning are single-modal data, either texts or images.

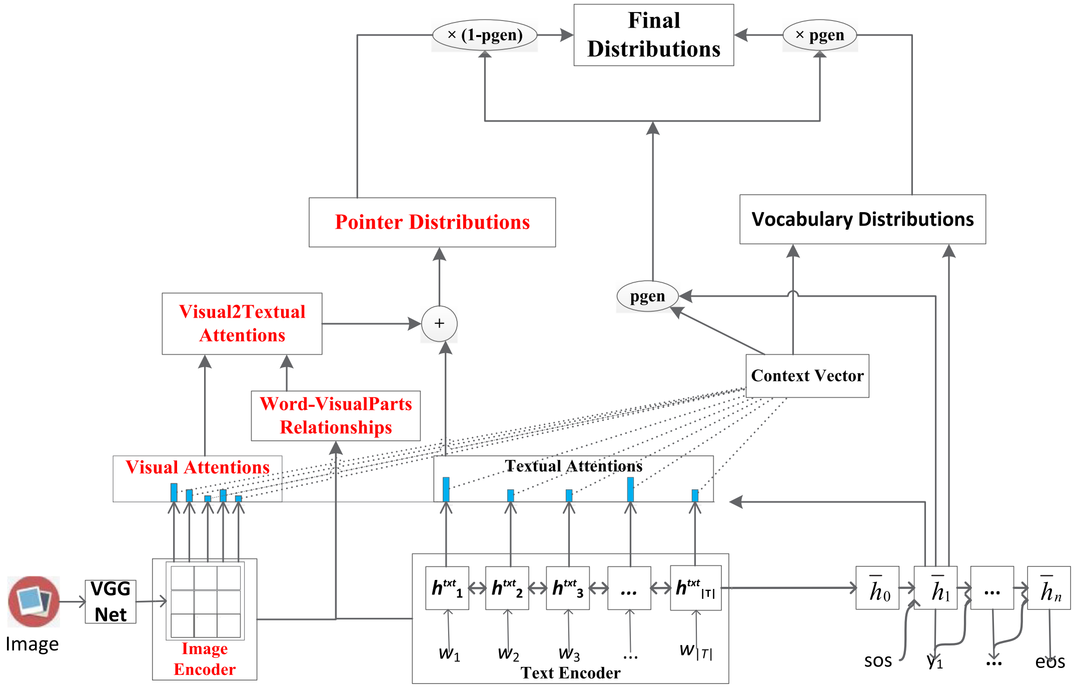


FIGURE 2 The framework of the proposed multimodal pointer-generator network

However, we wish for the model to be applicable in news image captioning, the inputs of which are multimodal data containing both texts and images. This brings up two issues. First, the parameters (in Equations (1) to (8)) of the models are ineffective on modeling the relationships between the source text and the source image of news image captioning. As with captions, the accompanying news texts have tight relationships with the news images and the words in news texts can be well aligned with visual parts or objects in news images. Humans are good at discovering these relationships and alignments, and use them to generate high-quality captions. This inspires us to make the encoder-decoder be able to mine the relations and alignments which will also render the model to generate better captions for news images. Second, the attention mechanism for news image captioning will pay attentions to both words of source texts and visual parts of source images. This will lead to the attention distribution problem and the attention redundancy problem between the two modalities. It needs well decided how much attention the text receives and how much attention the image receives. We conjecture that well attention distribution between the text and the image will also benefit new image caption generation. The attention distribution can be affected by the above mentioned relationships between the text and the image. In the following section, we present our adaptation method of the multimodal pointer-generator network for news image captioning.

4 | MULTIMODAL POINTER-GENERATOR NETWORK

4.1 | Network architecture

Figure 2 shows the framework of the model. The proposed multimodal pointer-generator network is an extension of the original pointer-generator network for news image captioning. The inputs of the model are a news image and its accompanying news text, so the model consists of textual parts and visual parts which are combined in the model to figure out the final caption word distributions for caption generation. The red-colored parts in Figure 2 are the extended new parts in contrast to the original pointer-generator network. As with the original pointer-generator network, the proposed model consists of four parts: the encoders, the decoder, the attention mechanism, and the pointer mechanism.

4.1.1 | Encoders

The textual encoder employs the *RNN* to encode the accompanying new text into vector representations. The visual encoder employs the state-of-the-art CNN *Oxford VGGNet*[28] to extract vector representations for images.

The bidirectional *RNN* model is used as the text encoder to encode the accompanying text T. Equations (10) to (12) are the equations of the bidirectional *RNN* model, where x^i represents the word embedding of the ith word of T. The *GRU*[40] is adopted as the *RNN* cell in our method because it is as efficient and effective as *LSTM* while less time-consuming.

$$\overrightarrow{h}^{eT}_{i} = \mathrm{GRU}^{\overrightarrow{eT}}(\overrightarrow{h}^{eT}_{i}, x_i), \tag{10}$$

$$\overleftarrow{h}^{eT}_{i} = \mathrm{GRU}^{\overleftarrow{eT}}(\overleftarrow{h}^{eT}_{i}, x_i), \tag{11}$$

$$h^{eT} = [\overrightarrow{h}^{eT}_{T} \,\|\, \overleftarrow{h}^{eT}_{1}]. \tag{12}$$

The *Oxford VGGNet* is used as the visual encoder to encode the image into vector representations. *VGGNet* is initially used for image classification. It consists of several convolution layers each followed by a pooling layer and the last fully connected layers. The last convolution layer splits the image into 14×14 visual parts denoted as $\{v_1, v_2, \ldots, v_{196}\}$ each of which encoded into a 512-dimensional vector representation, and has been proved suitable for image captioning and for the attention mechanisms to attend.[2] A nonlinear tanh transformation is applied to v_i to get the final encoding of each visual part, Equation (13).

$$h^{eI}_i = \tanh(M^{eI} \cdot v_i + b^{eI}). \tag{13}$$

4.1.2 | Decoder

The *RNN*-based decoder is used to generate words one by one in our method. Equations (14) and (15) are the equations for the decoder. Equation (14) computes the initial hidden state of the decoder from the encoding of the accompanying text. Here only the text encoding is used in initial state computation. Different from that of the original pointer-generator network, the context c_t in Equation (15) is multimodal, which is the summation of the textual encodings and visual encodings weighted by the multimodal attentions. The vocabulary distribution of the words is also computed by Equation (5) as in the original pointer-generator network.

$$h^d_0 = \tanh(W^{d0} \times h^{eT}_{-1} + b^{d0}), \tag{14}$$

$$h^d_t = \mathrm{GRU}^d(h^d_{t-1}, y_{t-1} \| c_t). \tag{15}$$

4.1.3 | Multimodal attentions

The attention is paid to both the image and the accompanying text for news image captioning. The attentions paid to the image are named by *visual attentions*, and the attentions paid to the text are named by *textual attentions*. As in the original pointer-generator network, Equation (1) is used to compute the un-normalized textual attention $a^T{}_{t,i}$ for each word encoding $h^{eT}{}_i$, and to compute the un-normalized visual attention $a^V{}_{t,j}$ for each visual part encoding $h^{eV}{}_j$. The normalized textual attentions and visual attentions are computed by Equations (16) and (17). So the multimodal context c_i is computed by Equation (18).

$$\alpha^T_{t,i} = \frac{\exp(a^T_{t,i})}{\sum_i \exp(a^T_{t,i}) + \sum_j \exp(a^V_{t,j})}, \tag{16}$$

$$\alpha^V_{t,j} = \frac{\exp(a^V_{t,j})}{\sum_i \exp(a^T_{t,i}) + \sum_j \exp(a^V_{t,j})}, \tag{17}$$

$$c_t = \sum_i \alpha^T_{t,i} h^{eT}_i + \sum_j \alpha^V_{t,j} h^{eV}_j. \tag{18}$$

4.1.4 | Final distributions

As with the original pointer-generator network, the final distributions of caption words at the tth decoding step are summation of the vocabulary distributions and the pointer distributions (Equations (6) and (7)). The pointer distributions in the original pointer-generator are only determined by the textual attentions. However, the pointer distributions in the multimodal pointer-generator are determined by both the textual attentions and the visual attentions. The definition of the multimodal pointer mechanism will be discussed in the following subsections.

4.1.5 | Training

The loss function L of our news image captioning model is the negative log likelihood of generating captions over the training set as defined in Equations (19) and (20), where $<I, T, Y>$ is an image-text-caption tuple of the training set, and $Y = [y_1, y_2, \ldots, y_{|Y|}]$ is the word sequences of the caption including the start token *<sos>* and the end token *<eos>*. In Equation (20), $\log(P(y_t|\{y_1, \ldots, y_{t-1}\}, c; \theta))$ is modeled by the proposed news image captioning model. The *Adam*[41] gradient-based optimization method is adopted to optimize the model parameters.

$$L = \sum_{<I,T,Y> \in \mathrm{TrainingSet}} -\log P(Y \mid I, T), \tag{19}$$

$$\log P(Y \mid I, T) = \sum_{t=1}^{|Y|} \log P(y_t \mid \{y_1, \ldots, y_{t-1}\}, c; \theta). \tag{20}$$

4.1.6 | Inferring

For inferring, the beam search algorithm is adopted at test time. The beam width is set as 5 in the experiments. The inferring stops when it generates the end token *<eos>*.

4.2 | Calculations of multimodal pointer distributions

The attention is split to textual attention paid to the source text and visual attention paid to the source image. The textual attention can be straightforwardly used as pointer distributions over the source words as in the original pointer-generator network. While the visual attentions can also influence pointer distributions because an image has relationships with its accompanying text and visual parts of the image can be aligned with the words in the text. Therefore, visual attentions can be transformed into textual attentions and then used to calculate pointer distributions.

The red-colored parts in Figure 2 are especially for calculations of multimodal pointer distributions. The visual attentions are first transformed to pointer distributions through the Word-VisualPart Relationships (Equation (21)), and are then combined with textual attentions to get the whole pointer distributions (Equation (22)).

The Word-VisualPart relationships are calculated in Equations (21) and (22) by applying an attention mechanism between the hidden state ${h^{eT}}_i$ of each source word i and the hidden state ${h^{eI}}_j$ of each visual part j. Here the hidden state ${h^{eI}}_j$ is used as the query and the hidden state ${h^{eT}}_i$ is used as the key. In Equation (19), the matrix $M^{d^{eT} \times d^{eI}}$ is the added parameters to compute the attention between ${h^{eI}}_j$ and ${h^{eT}}_i$, where d^{eT} is the dimension size of h^{eT} and d^{eI} is the dimension size of h^{eI}. In the proposed model, d^{eT} is set as the same with d^{eI}. Equation (22) normalizes the values by using the *softmax* function to make $\sum_i r_{j,i} = 1$.

$$\bar{r}_{j,i} = \tanh(h_j^{eI} M^{d^{eI} \times d^{eT}} (h_i^{eT})^T), \tag{21}$$

$$r_{j,i} = soft\max(r_{j,i}). \tag{22}$$

Next, the visual attentions are transformed into textual attentions (named after Visual2Textual attentions in Figure 2 via the Word-VisualPart relationships. Equation (23) is for calculations of Visual2Textual attentions. At the tth decoding step, the Visual2Textual attention that each source word i receives is the summation of the transformed attentions that each visual part j distributes to the word i. The portion of the transformed attention is determined by the relationship $r_{j,i}$.

$$\alpha_{t,i}^{V2T} = \sum_j \alpha_{t,j} r_{j,i}. \tag{23}$$

Finally, the multimodal pointer distribution of the word w is calculated by summing up the textual attention and the Visual2Textual attention. Equation (24) is the equation for calculations of the pointer distribution. The final distribution is calculated by adding up the vocabulary distribution and the multimodal pointer distribution as in Equation (7).

$$P_{\mathrm{ptr}}(w) = \sum_{i: w_i = w} \alpha_{t,i} + \alpha_{t,i}^{V2T}. \tag{24}$$

4.3 Multimodal coverage mechanism

The repetition problem also exists in the visual attentions as well as in the textual attentions of the original pointer-generator network, so the multimodal coverage mechanism is defined by taking visual attentions into consideration. The aim of the coverage mechanism is to reduce repetitions over attentions as well as repetitions over pointer distributions. However, pointer distributions over source words are not equivalent to textual attentions in the multimodal pointer-generator network. Therefore, two methods for the multimodal coverage mechanism are proposed in the following.

The first method is defined over attentions, and uses visual attentions the same way with textual attentions to define the *coverage vector* as in Equation (3) and to define the *coverage loss* as in Equation (8). This is a straightforward method to avoid repetitions over textual attentions and visual attentions, and repetitions over pointer distributions can be indirectly reduced.

The second method is defined over multimodal pointer distributions. This method transforms visual attentions into textual attentions and controls the repetitions in the transformed attentions. Equations (25) and (26) is the equations for calculations of the *coverage vector* and the *coverage loss* for this method, both of which are computed over the multimodal pointer distributions. The repetitions over attentions can be indirectly reduced through controlling repetitions over the multimodal pointer distributions.

$$\mathrm{cov}_{t,i} = \sum_{t'=0}^{t-1} \alpha_{t,i} + \alpha_{t,i}^{V2T}, \tag{25}$$

$$\mathrm{cov}\ loss_t = \sum_{i \in T} \min(\alpha_{t,i}, \mathrm{cov}_{t,i}). \tag{26}$$

Which calculation method for the multimodal coverage mechanism performs better will be discussed in the following experiments.

5 EXPERIMENTS

5.1 Data

Two news image captioning datasets are provided: one large-scale dataset for training and testing, and the other small-scale dataset only for testing.

The large-scale training and testing corpora are constructed from the *DailyMail* news corpora. The standard *DailyMail* corpora are the widely used corpora originally built in Reference [42] by collecting news stories from the *DailyMail* news websites for question answering and document summarization. There are about 210K html-formatted news documents provided in the original *DailyMail* news corpora. Each html-formatted news document contains one or more image-image pairs. To create news image-caption-text dataset, we extract and collect the first image-caption pair and the main text of each news document by parsing the html-formatted documents. The created news image captioning corpora are split into train, dev, and test dataset as in the original *DailyMail* corpora. The split and statistics of the created *DailyMail* news image captioning corpora are shown in Table 1.

The other corpora are originally provided in Reference [34] for probabilistic news image captioning. The corpora are collected from the *BBC* news website, and contain 3361 image-caption-text tuples in all, 240 of which are for testing. Due to such a small size, the corpora are only used to test the model trained with the *DailyMail* corpora introduced above. The statistics of the *BBC* corpora is shown in Table 2.

TABLE 1 The split and statistics of the *DailyMail* news image captioning corpora

Description	Value
Size of training dataset	187 900
Size of dev dataset	11 410
Size of testing dataset	9814
Average number of words in news texts	663.88
Average number of words in news captions	26.78

TABLE 2 The split and statistics of the *BBC* corpora

Description	Value
Size of testing dataset	240
Average number of words in news texts	422.01
Average number of words in news captions	9.59

During training and at the test time, the news texts are truncated to 400 tokens, the ground-truth summaries are truncated to 100 tokens.

5.2 Implementation

The texts of the corpora are preprocessed by tokenizing the text and replacing the digits with the <*NUM*> token. The 40k most frequent words in the corpora are kept and other words are replaced with the <*OOV*> token. The word embeddings are initialized with Google's word2vec tools[43] trained in the whole text of *DailyMail* and *BBC* corpora. The dimension of the word embeddings are set as 128.

The images are encoded by extracting the $14 \times 14 \times 512$ *conv*5_4 layer of the 19-layer *VGGNet*[28] pretrained on ImageNet as the vector representation of images, where 14×14 is the number of visual parts and 512 is the dimension of each visual part.

The proposed model is implemented based on See et al[6] pointer-generator network written with *Tensorflow*. The dimension of the hidden state of the *RNN* decoder is 512. The beam width is set as 5. The parameters of *Adam* are set to those provided in Reference [41]. The batch size is set to 12. Gradient clipping is employed to regularize our models. All models are trained on a *GTX*-1080 TI *GPU* card for 400 000 steps. The best checkpoint is selected based on performance on the validation set and the results on the test set are reported.

5.3 Comparisons with existing methods

The frequently used *BLEU* metric[44] which is the standard in image caption generation research is adopted. *BLEU* without a brevity penalty is reported. There has been, however, criticism of *BLEU*, so another common metric *METEOR*[45] is reported and compared whenever possible. The widely used summarization evaluation metric *ROUGE*[46] is also adopted to evaluate the generated captions.

The proposed method in this article is compared with five existing methods on the *DailyMail* corpora and the *BBC* test dataset. The evaluation results are shown in Tables 3 and 4. The compared methods are listed and described as follows:

- ***MMPtrGen*** is the proposed multimodal pointer-generator network without the coverage mechanism.
- ***MMGen*** is the proposed multimodal pointer-generator network without the pointer mechanism and the coverage mechanism.
- ***PtrGen-T*** is the original pointer-generator network carried out on the news text without the coverage mechanism.
- ***Gen-T*** is the state-of-the-art generative text summarization method based on the attentional encoder-decoder model proposed in Reference [11]. This method treats the news image captioning problem as the text summarization problem and only uses the news text as the input. The method generates the news image caption by treating the news text as a long sentence and summarizing the text.

TABLE 3 Comparisons with the existing methods on the *DailyMail* corpora

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	9.62	19.69	27.70
MMGen	7.24	16.68	26.07
PtrGen-T	9.46	19.49	27.41
Gen-T	6.95	16.49	25.40
Gen-I	0.33	4.20	10.41

TABLE 4 Comparisons with the existing methods on the *BBC* test dataset

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	0.42	10.77	9.26
MMGen	0.36	9.56	9.18
PtrGen-T	0.48	10.58	9.24
Gen-T	0.34	9.39	9.05
Gen-I	0.08	3.00	3.85
NNExtr	0.34	6.77	-
LDA-based	0.30	7.06	-

- ***Gen-I*** is the state-of-the-art generic image captioning method based on the attentional encoder-decoder model proposed in Reference [2]. This method only uses the image as the input, splits the image into 196 visual parts encoded with the *CNN* model, and uses the attentional decoder to generate captions.
- ***NNExtr*** is the neural news image captioning method recently proposed in Reference [36]. This method uses a neural classification model to score the sentences in the news text and extracts the most relevant sentence as the caption. The method first computes the context vector with *LSTM* using the ordering embeddings of images and texts, and then trains a sentence classification model using as the object function the cross entropy between captions and the context vector. The sentences are scored according to the cosine similarity with the context vector.
- ***LDA-based*** is the state-of-the-art probabilistic news image captioning method proposed in Reference [34], and the results on the *BBC* test dataset are reproduced in Reference [36] for comparison. The method works as follows. First, textual dictionaries are synthesized by assigning a unique token *id* to each word present in any of the articles, and visual dictionary is made by clustering *SIFT* descriptors into 2000 different visual words. Second, a *LDA* model is trained with 1000 topics on the *BBC* news dataset containing both text and images. Third, extractive summarization is used for surface realization. It has been shown in Reference [34] that retrieving sentences based on the *Kullback-Leibler divergence* between the topic distribution of a sentence and the topic distribution of a news article gives the best results in terms of human evaluation.

According to Table 3, the proposed multimodal pointer-generator network *MMPtrGen* gets the highest *BLEU* scores, the highest *METEOR* scores, and the highest *ROUGE-L F-measure* scores among the five methods, and the original pointer-generator network gets the second higher scores. The evaluation results of *LDA-based* and *NNExtr* are not shown in Table 3 because they are not reported on the *DailyMail* corpora. *MMPtrGen* outperforms *PtrGen-T*, which implies that incorporating visual information into the pointer-generator network can improves the summarization model, and that by considering both news text and image can generate better captions than only considering text on the *DailyMail* corpora. *PtrGen-T* gets the second higher scores, which implies that news text is more suitable for news image captioning than news images. Both *MMPtrGen* and *PtrGen-T* outperform the corresponding *MMGen* and *Gen-T*, which implies the pointer mechanism plays an important role in the pointer-generator network and can significantly improve the summarization and captioning model. The state-of-the-art generic image captioning method *NNattImg* performs very poorly on the *DailyMail* corpora, though it performs well on the generic image captioning corpora such as *COCO* and *Flickr*. This means that news image captioning is more likely a text summarization problem than an image captioning problem. News image captions contain detail information and event information not contained in images as introduced in Section I. So considering both news text and news images can generates better captions than only considering news images or news text.

Table 4 shows the scores on the *BBC* test dataset. Since our models are trained on the *DailyMail* corpora, the *BBC* test dataset is the out-of-domain dataset for our models. The evaluation results in table of *MMPtrGen*, *MMGen*, *PtrGen*, *Gen-T*, and *Gen-I* are in accordance with the results in Table 3. *MMPtrGen* gets the highest *METEOR* scores and *Rouge-L* scores, and *PtrGen* gets the highest *BLEU* scores. What's new is that the proposed method *MMPtrGen* outperforms *LDA-based* and *NNExtr* which achieve the state-of-the-art performance in the *BBC* dataset. Only the *BLEU* scores and *METEOR* scores of the two baselines are reported in the original article. The proposed *MMPtrGen* achieves new state-of-the-art performance in the *BBC* dataset, though it is trained on the *DailyMail* corpora.

In summary, the proposed method *NNattSim* performs the best and is thus suitable for news image captioning, and incorporating image information into the pointer-generator network does improve the summarization and captioning model.

5.4 | Evaluations of the multimodal coverage mechanism

As previously described, the coverage mechanism aims to alleviate the repetition problem of the encoder-decoder model. In the following, the two proposed multimodal coverage mechanisms are compared with see whether the coverage mechanism will improve the multimodal pointer-generator network. The evaluation results on the *DailyMail* corpora and on the *BBC* test dataset are shown in Tables 5 and 6.

- ***MMPtrGen with COV*1** is the proposed multimodal pointer-generator method with the first multimodal coverage mechanism defined over attentions.
- ***MMPtrGen with COV*2** is the proposed multimodal pointer-generator network with the second multimodal coverage mechanism defined over the pointer distributions.

According to Table 5, *MMPtrGen* with *COV*1 outperforms *MMPtrGen* on the *BLEU* score and the *METEOR* score. *MMPtrGen* with *COV*2 outperforms *MMPtrGen* on all the three scores. This implies that the multimodal coverage mechanism can improve the multimodal pointer-generator network in the *DailyMail* corpora by alleviating the repetition problem. On the other hand, neither of the two multimodal coverage mechanisms outperforms each other on all the metrics. *MMPtrGen* with *COV*1 gets the highest *METEOR* score, and *MMPtrGen* with *COV*2 gets the highest *BLUE* score and *ROUGE-L* score. This implies the coverage mechanism defined over attentions and the coverage mechanism defined over the pointer distributions work differently, and both can improve the proposed pointer-generator network.

TABLE 5 Evaluation results of the multimodal coverage mechanism on the *DailyMail* corpora

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	9.62	19.69	27.70
MMPtrGen with COV1	9.67	20.00	27.66
MMPtrGen with COV2	9.69	19.68	27.83

TABLE 6 Evaluation results of the multimodal coverage mechanism on the *BBC* corpora

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	0.42	10.77	9.26
MMPtrGen with COV1	0.42	9.41	8.82
MMPtrGen with COV2	0.46	9.28	8.72

TABLE 7 Evaluation results of the Word-VisualPart relationships on the *DailyMail* corpora

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	9.62	19.69	27.70
MMPtrGen w./o. WVRela	9.52	19.50	27.64

TABLE 8 Evaluation results of the Word-VisualPart relationships on the *BBC* corpora

Method	BLEU	METEOR	ROUGE-L
MMPtrGen	0.42	10.77	9.26
MMPtrGen w./o. WVRela	0.39	10.75	9.34

Table 6 shows the evaluation results in the out-of-the-domain *BBC* test dataset. The two multimodal coverage mechanisms do not perform as well as in *DailyMail* test dataset. The performance of the coverage mechanism depends on the corpora. Nevertheless, *MMPtrGen* with *COV2* achieves a higher *BLEU* score than *MMPtrGen* does.

In summary, the two multimodal coverage mechanisms can improve the multimodal pointer-generator network.

5.5 | Evaluations of the Word-VisualPart relationships

The Word-VisualPart relationships are the alignments between the source words and the visual parts of the source image, and are used to transform the visual attentions into textual attentions.

In Tables 7 and 8, *MMPtrGen* w./o. *WVRela* is the proposed multimodal pointer-generator network without the *Word-VisualPart* relationships. *MMPtrGen* w./o. *WVRela* does not transform visual attentions into textual attentions and only uses the original visual attentions as the pointer distributions. According to the two tables, *MMPtrGen* outperforms *MMPtrGen* w./o. *WVRela*, which implies using the Word-VisualPart relationships to transform visual attentions into pointer distributions can improve the multimodal pointer-generator network.

Moreover, the *Word-VisualPart* relationships are used to align source words with visual parts as follows: for each visual part j, select the most related source word i to align with such that $r_{i,j}$ is the largest. It is a hard and labor-consuming task to manually create ground-truth alignment of source words and visual parts of images. Therefore, the alignments are evaluated by computing the precision of the aligned word set comparing to the caption word set. The assumption is that the well-aligned words are also the keywords in image captions.

Table 9 shows the average number of the aligned words of the images and the precision on the *DailyMail* corpora. According to the table, there are about eight words aligned by a news image, about 11.3% of which appear in the corresponding captions. This implies that the alignments are focused on several words, some of which are caption words. The alignment is not supervised, and can be improved through supervising methods.

5.6 | Discussion of attention distributions over the text and the image

The above experiments show that the proposed multimodal pointer-generator network outperforms the original pointer-generator network not very significantly, and that the generic image captioning method performs poorly in news image captioning. To interpret the performance, an

TABLE 9 Evaluation of alignments between words and visual parts on the *DailyMail* corpora

Method	Average Number of Words	Precision
MMPtrGen	7.7413	11.33%
MMPtrGen with COV1	7.7945	11.25%
MMPtrGen with COV2	8.8532	11.12%

TABLE 10 Attention distributions on the *DailyMail* corpora

Method	Average Textual Attention	Average Visual Attention
MMPtrGen	98.132%	1.868%
MMPtrGen with COV1	98.449%	1.551%
MMPtrGen with COV2	99.300%	0.700%

additional experiment is carried out to show the average attention distributions on the news texts and the new images as shown in Table 10. Average textual attentions and average visual attentions are computed by making an average over decoding steps and the test datasets. Calculations of the attention distributions in the proposed model are not supervised.

According to the table, more than 98% of the attention is distributed onto the news text, and less than 2% of attention is distributed to the news image in the proposed multimodal pointer-generator network. Note that 100% of the attention is paid to the news text in the original pointer-generator network, and that 100% of the attention is paid to the news image in the generic image captioning model. Most information in the news image caption is contained in the news text and some information is contained in the news image. Textual attentions play a more important role than visual attentions do for news image captioning. This can partly explain the performance of the proposed multimodal pointer-generator network, the original pointer-generator network, and the generic image caption method. Although the precision of the discovered Word-VisualPart relationships is not very high, these relationships can help improve the performance because the visual information can be utilized in the pointer mechanism through these relationships. The supervision method can be used to further improve the attention distributions and the Word-VisualPart relationships.

5.7 A case study

Figure 3 demonstrate the progress and the result of the proposed *MMPtrGen* method. The vertical axis denotes the percentage of the visual pointer distributions averaged over all the decoding steps. For example, the highest visual pointer distribution of example#1 is 0.32%. The total visual pointer distributions of the five examples are 0.81%, 1.62%, 0.62%, 0.75%, and 2.3%, respectively. The summation of textual pointer distributions and visual pointer distribution over all the words is 1. According to Figure 3, for the five examples, the pointer distribution consists of about 98% textual pointer distribution and about 2% visual pointer distributions, and the top-9 words gain most of the visual pointer distribution. Textual attention plays a more important role in news image captioning.

Figure 4 shows the generated captions for the five examples. For each image, the ground-truth caption is provided on the right top, and the generated caption is provided on the right bottom. As described before, The Arabic numbers in the generated captions are replaced with the token *NUM*.

The generated captions have high overlaps with the ground-truth captions. The generated captions are much better than the generic captions shown in Figure 1, because the generated captions contain detailed information provided in news texts. For example, the caption of the second image contains the information of the age of the triathlete, the marathon which are not shown in the image but contained in the news text. Another example is about the third image, the generated caption contains the information of North Korean embassy in London, which is not contained in the image but is contained in news text. These examples demonstrate that the proposed model can generate satisfactory news image captions.

6 CONCLUSIONS

News image captioning task is different from generic image captioning task in that news image captions contain more detailed information such as entity names and events than general image captions do, and detailed information is usually contained in news text but not in news images.

This article proposes a multimodal pointer-generator network for news image captioning by incorporating image information into the original pointer-generator network. The proposed network consists of a multimodal generator which computes vocabulary distributions of the words and

FIGURE 3 The top-9 average visual pointer distributions over all the decoding steps for the source words of the examples

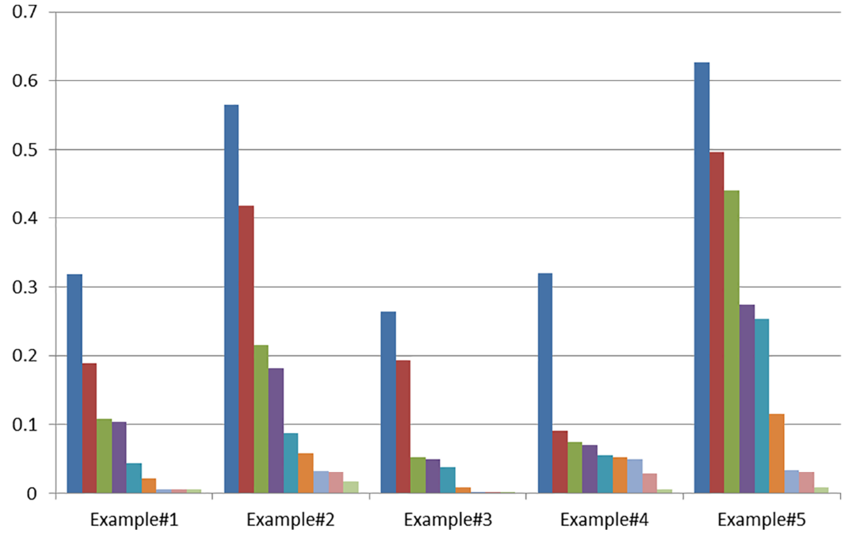
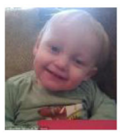


FIGURE 4 The captions generated by the proposed model for the five pictures in Figure 1. The bottom bold captions are the generated ones

	Angus the rhino with mum Dorothy at Blair Drummond Safari Park.
	the scottish government has joined the fight against the illegal trade in rhinoceros horn by setting up a dna database.
	The 14-year-old triathlete has become the youngest person ever to complete a marathon on the continent.
	a triathlete who runs marathons across the globe to honour her cancer victim father has become the youngest person ever to complete a marathon in antarctica -at just NUM .
	There was little sign of life at the North Korean embassy in London's Ealing today.
	the north korean embassy in london was silent today after a sign appeared on a tree outside the property warning that ' loading ' would take place on tuesday
	Two-year-old Archie Watson suffered from a rare genetic disorder and died.
	tragic: two-year-old archie watson died last monday after suffering from tay-sachs, which causes deterioration of nerve cells.
	A man who lost his class ring more than 40 years ago was amazed when it was returned by a woman on Facebook.
	richard hale, NUM, was reunited with a ring he lost more than NUM years ago. he was reunited with a ring he lost more than NUM years ago

a multimodal pointer which computes pointier distributions of the words. The news image is encoded by *VGGNet*, and the news text is encoded by the *RNN* model. The attention is split to textual attention paid to text and visual attention paid to the image to compute the multimodal context vector in each decoding step. In the multimodal pointer, visual attention is first transformed into textual attention through the Word-VisualPart relationships modeled by an attention mechanism, and is then added up with the textual attention to compute pointer distributions. Two multimodal coverage mechanisms are defined by reducing repetitions of multimodal attentions or by reducing repetitions of pointer distributions. The *DailyMail* news image captioning corpora are created for training and testing by collecting news images, captions, and documents through parsing the html-formatted documents. Another small-sized *BBC* dataset is used only for testing. Experiments on the two datasets show that the proposed model outperform the original pointer-generator network, the generic image captioning methods, the *LDA*-based news image captioning method, and the neural extractive new image captioning method, which shows that considering both the news image and the news text for generating news image captions is better than considering only new text or image. It is shown by experiments that the model adding visual attention to compute pointer distributions outperform the one not using visual attention as pointer distributions. It is also shown that more than 98% attention is paid to text and the left is paid to the image, which means that textual attention plays a more important role in the proposed network. Nevertheless, visual attention cannot be neglected because the model considering visual attention outperforms the one without visual attention shown by experiments. Experiments also show the model with the multimodal coverage mechanisms outperforms the ones without the coverage mechanisms. This article is extended from our previous work.[47]

With the development of smart cameras, images will contain more and more physical features such as time, location, temperature, air quality, and weather where they are taken. Those features indicate the content of images and render the content of text, therefore can help identifying the captions of images. The physical features of images that match the content of text are important for increasing the accuracy and reliability of news. In the future, we will incorporate those features into the implementation of the future interconnection environment[48,49] for realizing Cyber-Physical-Society and Cyber-Physical-Socio Intelligence CPSI.[18,50-53]

ACKNOWLEDGMENTS

The research was sponsored by the National Natural Science Foundation of China (No.61806101, No.61876048), and the Open Foundation of Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (IIP2019-2). Professor Hai Zhuge is the corresponding author of this article.

ORCID

Jingqiang Chen https://orcid.org/0000-0001-7242-0141
Hai Zhuge https://orcid.org/0000-0001-8250-6408

REFERENCES

1. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(4):652-663.
2. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. Paper presented at: International Conference on Machine Learning; June 1, 2015;2048-2057.
3. Liu C, Mao J, Sha F, Yuille AF. Attention correctness in neural image captioning. Paper presented at: AAAI; February 4, 2017a;4176-4182.
4. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016;4651-4659.
5. Zhao S, Sharma P, Levinboim T, Soricut R. Informative image captioning with external sources of information. arXiv preprint arXiv:1906.08876; 2019.
6. See A, Liu PJ, Manning CD. Get to the point: summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368; 2017.
7. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473; 2014.
8. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025; 2015.
9. Cheng J, Lapata M. Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252; 2016.
10. Zeng W, Luo W, Fidler S, Urtasun R. Efficient summarization with read-again and copy mechanism. arXiv preprint arXiv:1611.03382; 2016.
11. Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685; 2015.
12. Nallapati R, Zhai F, Zhou B. Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. Paper presented at:Thirty-First AAAI Conference on Artificial Intelligence; February 12, 2017.
13. Vaswani A, Shazeer N, Parmar N. Attention is all you need. Paper presented at: Advances in Neural Information Processing Systems; 2017;5998-6008.
14. Liu Y, Lapata M. Hierarchical transformers for multi-document summarization. arXiv preprint arXiv:1905.13164; 2019.
15. Carbonell JG, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Paper presented at: SIGIR, Vol. 98; 1998;335-336.
16. Fabbri AR, Li I, She T, Li S, Radev DR. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. arXiv preprint arXiv:1906.01749; 2019.
17. Lebanoff L, Song K, Liu F. Adapting the neural encoder-decoder framework from single to multi-document summarization. arXiv preprint arXiv:1808.06218; 2018.
18. Zhuge H. *Multi-Dimensional Summarization in Cyber-Physical Society*. Amsterdam, NL: Morgan Kaufmann; 2016.
19. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632; 2014.
20. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. Paper presented at: Proceedings of the IEEE International Conference on Computer Vision; 2015;2407-2415.
21. Wang C, Yang H, Bartz C, Meinel C. Image captioning with deep bidirectional LSTMs. Paper presented at: Proceedings of the 2016 ACM on Multimedia Conference; 2016;988-997.
22. Liu C, Sun F, Wang C, Wang F, Yuille A. MAT: a multimodal attentive translator for image captioning. arXiv preprint arXiv:1702.05658; 2017b.
23. Wang J, Madhyastha P, Specia L. Object counts! Bringing explicit detections back into image captioning. arXiv preprint arXiv:1805.00314; 2018.
24. Venugopalan S, Anne HL, Rohrbach M, Mooney R, Darrell T, Saenko K. Captioning images with diverse objects. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017;5753-5761.
25. Ren Z, Wang X, Zhang N, Lv X, Li LJ. Deep reinforcement learning-based image captioning with embedding reward. arXiv preprint arXiv:1704.03899; 2017.
26. Chen C, Mu S, Xiao W, Ye Z, Wu L, Ma F, Ju Q. Improving image captioning with conditional generative adversarial nets. arXiv preprint arXiv:1805.07112; 2018.
27. Wang Q, Chan AB. CNN + CNN: convolutional decoders for image captioning. arXiv preprint arXiv:1805.09019; 2018.
28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556; 2014.

29. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in Neural Information Processing Systems; 2012;1097-1105.
30. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015;1-9.
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016;770-778.
32. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018;7132-7141.
33. Gao L, Fan K, Song J, Liu X, Xu X, Shen HT. Deliberate attention networks for image captioning. Paper presented at: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019;8320-8327.
34. Feng Y, Lapata M. Automatic caption generation for news images. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(4):797-812.
35. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
36. Batra V, He Y, Vogiatzis G. Neural caption generation for news images. Paper presented at: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018); 2018.
37. Chen J, Zhuge H. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. Paper presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018;4046-4056.
38. Chen J, Zhuge H. Extractive summarization of documents with images based on multi-modal RNN. *Future Gener Comput Syst*. 2019a;99:186-196.
39. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. Paper presented at: Thirteenth Annual Conference of the International Speech Communication Association; 2012.
40. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259; 2014.
41. Kingma D, Ba J. Adam: a method for stochastic optimization. *Comput Sci*. 2014;arXiv: 1412.6980.
42. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. Paper presented at: Advances in Neural Information Processing Systems; 2015;1693-1701.
43. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013;26:3111-3119.
44. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. Paper presented at: Proceedings of the Meeting of the Association for Computational Linguistics (ACL); 2002.
45. Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. Paper presented at: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation; 2014.
46. Lin CY. Rouge: a package for automatic evaluation of summaries. Paper presented at: Text Summarization Branches Out, Association for Computational Linguistics, https://www.aclweb.org/anthology/W04-1013; 2004.
47. Chen J, Zhuge H. News image captioning based on text-summarization using image as query. Paper presented at: The 15th International Conference on Semantics, Knowledge and Grids (SKG2019), Guangzhou, China; 2019b.
48. Zhuge H. The future interconnection environment. *IEEE Comput*. 2005;38(4):27-33.
49. Zhuge H. *The Web Resource Space Model*. Berlin, Germany: Springer; 2008.
50. Zhuge H. Semantic linking through spaces for cyber-physical-socio intelligence: a methodology. *Artif Intell*. 2011;175:988-1019.
51. Zhuge, H and Xing, Y. Probabilistic resource space model for managing resources in cyber-physical society. *IEEE Trans Serv Comput*. 2012;5(3):404-421.
52. Zhuge H. *The Knowledge Grid: Toward Cyber-Physical Society*. Singapore: World Scientific Co.; 2012.
53. Zhuge H. *Cyber-Physical-Social Intelligence on Human-Machine-Nature Symbiosis*. Berlin, Germany: Springer; 2020.

How to cite this article: Chen J, Zhuge H. A news image captioning approach based on multimodal pointer-generator network. *Concurrency Computat Pract Exper*. 2020;e5721. https://doi.org/10.1002/cpe.5721