LAMS: A Location-aware Approach for Multimodal Summarization (Student Abstract)

Zhengkun Zhang,[1] Jun Wang,[2] Zhe Sun,[3] Zhenglu Yang[4]

[1,4] Tianjin Key Laboratory of Network and Data Security Technology, College of Computer Science, Nankai University, China,
[2]Ludong University, China,
[3]RIKEN, Japan
{zhangzk2017, junwang}@mail.nankai.edu.cn,
zhe.sun.vk@riken.jp, yangzl@nankai.edu.cn

Abstract

Multimodal summarization aims to refine salient information from multiple modalities, among which texts and images are two mostly discussed ones. In recent years, many fantastic works have emerged in this field by modeling image-text interactions; however, they neglect the fact that most of multimodal documents have been elaborately organized by their writers. This means that a critical organized factor has long been short of enough attention, that is, image locations, which may carry illuminating information and imply the key contents of a document. To address this issue, we propose a location-aware approach for multimodal summarization (LAMS) based on Transformer. We investigate image locations for multimodal summarization via a stack of multimodal fusion block, which can formulate the high-order interactions among images and texts. An extensive experimental study on an extended multimodal dataset validates the superior summarization performance of the proposed model.

Introduction

Text summarization aims to generate a brief but refined summary to represent the source document, and it forms the basis for a variety of natural language processing (NLP) tasks, such as text classification, information retrieval, and question answering. Many researches have shown their charming performance in this task by exploring unimodal information (Chen and Bansal 2018; Zhong et al. 2019). While with the development of social media and news sites on the Web, multimodal information, such as texts, images, videos, and audios, becomes considerably available and has received more attention in recent years.

In multimodal summarization, texts and images are two kinds of the most discussed modalities (Zhu et al. 2018; Chen and Zhuge 2018). Existing works try to distill distinctive but complementary information from texts and images to enhance summarization performance. However, they commonly devalue the importance of a critical factor for summarizaiton, that is, analyzing image locations. Image locations are less attentive than textual and visual contents because they are considered hard to be quantitatively evaluated and trivial for summarization. While in multimodal documents, images as well as their locations can give people more intuitive and illuminating information than texts, including pointing out those key sentences, which is crucial for generating excellent summaries.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Methodology

As illustrated in Figure 1, we present LAMS, a location-aware approach for multimodal summarization, which takes multimodal fusion blocks (MFB) to formulate the interactions between different modalities and employ the location-aware mechanism to further exploit the image location information.

Multimodal Fusion Block. Multimodal fusion block (MFB) aims to provide a representation for each multimodal block of the document. We design a unified block for capturing the interaction between text and image. We first encode the text part and the image part and then generate the 2^{nd} order interaction representation through bilinear pooling. What's more, we introduce the maximum influence range of the image and formulate the relative location between sentences and images by the stack of MFBs, which produce the high order interaction representation.

Summary Decoder. For generating multimodal summary, important sentences should be extracted to form the summary. We employ the Pointer Network, which is an auto-regressive decoder with superior performance for summarization generation, as the summary decoder in our model.

Experiments

Dataset. MSMO dataset[1] is proposed by (Zhu et al. 2018) for multimodal summarization. This dataset is formed by collecting articles with images and captions from $DailyMail$ website[2]. However, it lacks any information about how the documents are organized, such as image locations. Therefore, we extend the dataset by collecting the image locations of each document according to the URLs they provide. The documents in our training, validation, and testing are 287467, 10265, and 10261, respectively. Several websites are missing when we crawl the data, and thus, our

[1]http://www.nlpr.ia.ac.cn/cip/jjzhang.htm
[2]http://www.dailymail.co.uk

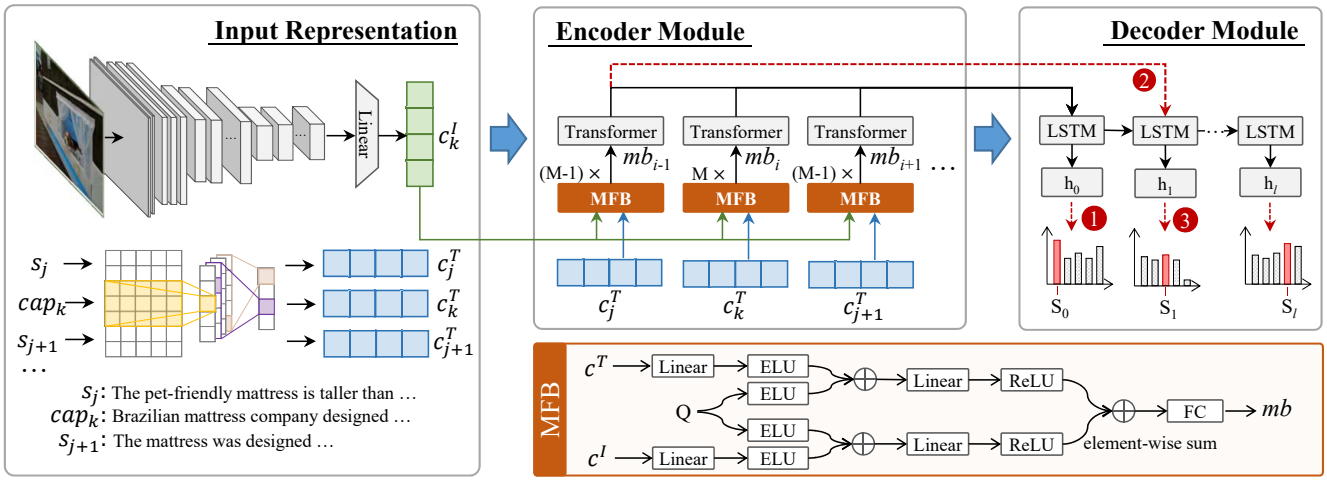

Figure 1: Illustration of the proposed location-aware approach for multimodal summarization. At first, VGG19 and BERT with a convolutional layer are firstly used to encode the text and image part of the document. Then, a stack of multimodal fusion blocks are employed to encode the high order interactions between the two parts. The stack folds are determined via the interaction range of the image. Pointer Network is further utilized in summary decoder to perform multimodal summarization.

Model	R-1	R-2	R-L
(Nallapati et al. 2016)*	34.78	13.10	32.24
(See, Liu, and Manning 2017)*	41.11	18.31	37.74
(Zhong et al. 2019)	42.69	19.92	38.89
ATG*	40.63	18.12	37.53
ATL*	40.86	18.27	37.75
HAN*	40.82	18.30	37.70
ATL+PN	42.48	19.75	38.78
LAMS	**43.07**	**20.28**	**39.34**

Table 1: Summarization performance the compared methods on the extended MSMO dataset: Results with * marks are taken from (Zhu et al. 2018). All the ROUGE scores have 95% confidence intervals of at most ±0.22 as reported by the official ROUGE script.

training and validation datasets are slightly different from the original MSMO dataset.

Summarization Performance Comparison. The first three models in Table 1 are text summarization approaches, including abstractive and extractive ones. Compared with the abstractive methods (Nallapati et al. 2016) and (See, Liu, and Manning 2017), the extractive model (Zhong et al. 2019) can achieve generally better performance, as shown in the table. As to the multimodal summarization methods in the middle of Table 1, ATG, ATL, and HAN are all constructed based on the abstractive mode. As evaluated by the 95% confidence interval in the official ROUGE script, our proposed model LAMS achieve statistically significant improvements over all of the baseline models.

Conclusions and Future Work

In this paper, we discuss the importance of image locations and the interactions between images and texts. To address this issue, we propose a location-aware approach for multimodal summarization (LAMS) based on Transformer. We employ the multimodal fusion block to formulate the interactions. Then, we introduce the location-aware mechanism to further leverage the image location information. For evaluating our method, we supplement the information about the image locations to the previous multimodal summarization dataset, and the experiments on this dataset validate the promising performance of our model in the multimodal summarization task. In the future, it will be interesting to study the abstractive approach in multimodal summarization.

References

Chen, J.; and Zhuge, H. 2018. Abstractive Text-Image Summarization using Multi-modal Attentional Hierarchical Rnn. In *EMNLP 2018*, 4046–4056.

Chen, Y.-C.; and Bansal, M. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *ACL 2018*, 675–686.

Nallapati, R.; Zhou, B.; dos Santos, C.; Gulcehre, C.; and Xiang, B. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL 2016*, 280–290.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL 2017*, 1073–1083.

Zhong, M.; Liu, P.; Wang, D.; Qiu, X.; and Huang, X.-J. 2019. Searching for Effective Neural Extractive Summarization: What Works and What's Next. In *ACL 2019*, 1049–1058.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *EMNLP 2018*, 4154–4164.