Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 1, No. 1, 2022

Research on Multimodal Summarization Based on Sequence to Sequence Model

Qiduo Lu *, Xia Ye and Chenhao Zhu

Academy of Combat Support Rocket Force University of Engineer, Xi'an, China

Abstract: With the rapid development of information technology in the 21st century, the data on the Internet is growing exponentially, which makes it difficult for users to receive useful information from an endless stream of information sources. Therefore, as a kind of data compression task, automatic summarization technology has gradually become a research hotspot in academia and industry. This paper mainly introduces the multimodal summarization method from sequence to sequence model, then introduces the current multimodal summarization evaluation methods and finally gives some conclusions and thoughts.

Keywords: Multimodal summarization, Sequence to Sequence model, Attention mechanism.

1. Introduction

Multimodal summarization refers to inputting multiple modal information, usually including text, voice, image, video and other information, and outputting a core summary after comprehensively considering multiple modal information. The current summary research usually takes the text as the processing object, and generally does not involve the processing of other modal information. However, various studies have shown that taking multimodal data as input does help to improve the quality of abstracts [1, 2]. Therefore, as an extension of the text summarization task, multimodal summarization task integrates the information of visual and auditory modes, and uses the complementary and verification of different modal data to make the application range of summarization task wider.

According to the definition of Wikipedia, summary refers to the process of shortening a set of data through calculation to create a summary that represents the most important or relevant information in the original content. Mathematically, it can be expressed as the process of obtaining $X_{sum} = f(D)$ that makes $length(X_{sum}) \le length(D)$, in which X_{sum} is the output summary, D is the input data, and $f(\bullet)$ is the summary function.

Multimodal summary task can be defined as "A summary task that takes more than one information representation mode as input and depends on the information sharing between different modes to generate the final summary". Mathematically speaking, when the input data set D can be decomposed into disjoint sets $\{M_1 \bigcup M_2 \bigcup \ldots \bigcup M_n\}$ of different modal information, where $n \ge 2$, the task of obtaining $X_{sum} = f(D)$ is defined as a multimodal summary. If for $X_{sum} = \{M_1^{'} \bigcup M_2^{'} \bigcup \ldots \bigcup M_{n^{'}}^{'}\}$, $n^{'} \ge 2$, the output summary is multimodal, otherwise it is single-mode output.

This paper mainly introduces the current most popular model of multimodal summarization - method based on sequence to sequence, and the summary evaluation methods

2. Method based on sequence to sequence

In recent years, deep learning technology has provided new ideas for the research of automatic summarization. Among them, the research and application of sequence to sequence (seq2seq) model is the most extensive. This model was proposed by Cho et al. and Sutskever et al. for machine translation tasks. The basic idea is to use the global information of the input sequence to infer the corresponding output sequence, which is composed of an encoder and a decoder. The encoder encodes an input variable length sequence $X = (x_1, ..., x_t)$ into a fixed semantic vector. Using recurrent neural network (RNN), the source sequence X is converted into the form of context c by the following formula:

$$h_t = f_{enc}(x_t, h_{t-1}) \tag{2.1}$$

$$c_t = f_c(h_1, ..., h_t) \tag{2.2}$$

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

The decoder extracts semantic information from this vector and outputs another variable length sequence $Y = (y_1, ..., y_{t^{'}})$. Generate y_t from the context vector c_t and the previously generated $\{y_1, ..., y_{t-1}\}$:

$$p(y_t \mid \{y_1, ..., y_{t-1}\}, c_t) = f_{dec}(y_{t-1}, s_t, c_t) \tag{2.3}$$

Where s_t is the hidden state of the decoder, and f_{dec} is a nonlinear activation function used to calculate the probability vector of the output word at time t. The loss function L_t at each time is the negative log likelihood function of y_t:

$$L_t = -\log p(y_t \mid \{y_1, ..., y_{t-1}\}, c_t) \tag{2.4}$$

The disadvantage of this model is that the encoder represents all the information in the source text as a fixed semantic vector, and the decoder refers to this vector when generating each word item, which makes it difficult for neural network to process long text. The experiment of Cho et al. confirmed that the performance of the model decreased rapidly with the increase of text length. In this regard, Bandanas et al. introduced the attention mechanism into the model to make the decoder focus on specific parts of the source text when generating each word item. The attention mechanism estimates the probability distribution of the encoder's hidden state in each decoding step, which is used to calculate the weighted average value of the context vector and the encoder's hidden state as an additional input to the decoder. Experimental results show that the model with attention mechanism achieves better results in machine translation tasks and is more robust to sentence length changes. The addition of attention mechanism makes the seq2seq model more perfect, and a large number of relevant studies have been based on this model. Fig.1 is the schematic diagram of seq2seq model with attention mechanism.

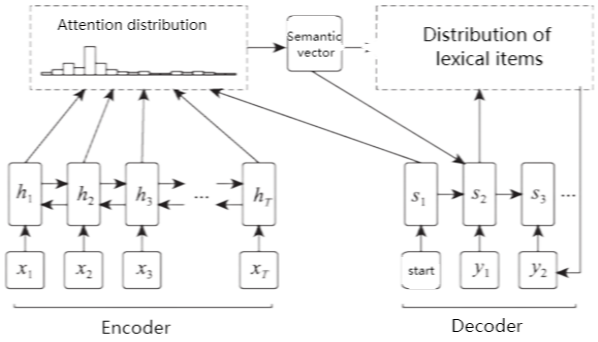

Fig.1 Schematic diagram of seq2seq model with attention mechanism

Rush et al. applied this model to generative summarization for the first time. Compared with the previous generative method, this model generates abstracts based on "understanding" the semantics of text, which is closer to the generation process of artificial abstracts. Subsequently, a series of generative summarization models based on seq2seq have been proposed, and fruitful research work has been carried out on encoders, decoders and training methods. The abstracts generated based on this model let the academic community see the hope of the practicality of automatic abstracts in terms of language fluency and coherence . Compared with text summarization, multimodal summarization extracts higher quality summarization by adding modal information such as image, video, audio, etc. Therefore, sequence to sequence model is gradually applied to the field of multimodal summarization. The multimodal summarization model based on sequence to sequence is mainly composed of three modules: encoder, decoder and multimodal fusion module.

2.1. Encoder

Due to the significant progress in feature extraction, various encoders have been explored to encode context information in the form of text. Due to the long-term dependence of RNN, it is difficult to learn information far away, so most models use long short-term memory (LSTM) or gated recurrent unit (GRU) to replace it. LSTM introduces the concept of gating, which can learn the memory / forgetting, input and output of memory units, so that the machine knows when to remember or discard some information. GRU uses gating network to learn the combination of new input and previous memory and the retention degree of previous memory. Although the principles of LSTM and GRU are different, they can learn long-term dependent information and perform better than RNN on many tasks. Compared with LSTM, GRU parameters are less and easier to train, but when there are more training data, LSTM with stronger expression ability will perform better. Zhu et al. [9, 10] used bidirectional LSTM for word level coding with reference to the pointer generator network; Khullar et al. used bidirectional GRU to encode text. Except that GRU and LSTM are encoded in writing, the emergence of transformer, a variant of RNN, has quickly attracted researchers' attention. Table 1 shows the advantages and disadvantages of different text coding methods.

Table 1. Text coding method and its advantages and disadvantages

Text Encoding Method	Advantages	Disadvantages
Traditional RNN	It has simple internal structure, low requirements for computing resources, and excellent performance on short sequence tasks	Gradient disappearance or explosion is easy to occur when processing long sequences, and parallel computing is impossible
LSTM [9, 10]	With long-term memory function, it can alleviate the gradient disappearance or explosion of traditional RNN, and is suitable for processing long sequences	The internal structure is complex, and the training efficiency is far lower than that of traditional RNN under the same computing power, so parallel computing is impossible
GRU	With long-term memory function, it can alleviate the gradient disappearance or explosion of traditional RNN. It is suitable for processing long sequences, and the computational complexity is smaller than LSTM	It cannot completely solve the problem of gradient disappearance and parallel computing
Transformer	It can be calculated in parallel and has infinite memory function	Weak local information acquisition ability, unable to characterize location information

Most visual encoders do not train parameter weights from scratch, but prefer to use CNN based pre-training embedding. Most existing work uses pre-trained networks (such as ResNet , VGGNet , GoogLeNet) to train on large image classification datasets such as ImageNet . Table 2 lists the comparison, advantages and disadvantages of visual coding methods.

Table 2. Visual coding method and its advantages and disadvantages

Visual encoding method	VGGNet	GoogLeNet	ResNet
First appearance	2014	2014	2015
Number of layers	19	22	152
Top-5 error	7.3%	6.7%	3.57%
Convolution layers	16	21	151
Convolution kernel size	3	7,1,3,5	7,1,3,5
Number of fully connected layers	3	1	1
Full connection layer size	4096,4096,1000	1000	1000
advantages	Good generalization performance	With small parameters and good performance, it can make more efficient use of computing resources and extract more features under the same amount of computing	Deeper level and higher accuracy
disadvantages	Large number of parameters, easy over fitting and low efficiency		Efficiency depends on Model

Although most works use the information of the target data set to train their own shared embedding space for multiple modes , a considerable number of works [1, 2] also tend to use the pre training neural network model trained on the image caption dataset [17,18], such as pascal1k, flickr8k, flick30k, so as to take advantage of the information overlap between different modes. This is necessary for small datasets that are mainly used to extract summaries. However, even these pre-trained models cannot process the original data. Therefore, text and image input are first preprocessed into the required embedded format, and then fed to these models with pre-trained weights. For example, Wang et al. obtained a 6000-dimensional sentence vector by applying principal component analysis on the 18000-dimensional vector obtained by the Gaussian Laplace mixture model, and a 4096-dimensional image vector by extracting weights from the full connection layer of VGGNet.

2.2. Multimodal Fusion Module

With the introduction of attention mechanism, most tasks based on text image input focus on the use of multimodal attention mechanism. Attention strategies have proved to be a very useful technique to eliminate noise and focus on relevant information . It has been applied to specific modal information and information sharing steps in the form of multimodal attention to determine the participation of specific modes of each input. Li et al. first proposed hierarchical multimodal attention to solve the multimodal summary task of long sentences. The attention module includes a separate text attention layer and image attention layer, and then multimodal attention layer. Although multimodal attention shows great potential in the task of text image summarization, it is not enough to complete the task of text video audio summarization by itself. Therefore, in order to overcome this weakness, Fu et al. proposed the double jump attention mechanism as an extension of bilinear attention , while Li et al. developed a new conditional self-attention mechanism module to capture video local semantic information based on input text information.

2.3. Decoder

According to the coding strategy used, the text decoder is also different, from the ordinary one-way RNN that generates one word at a time to the hierarchical RNN decoder that performs this step at multiple granularity levels. Although most neural models only focus on using multimodal information as input to generate text summaries, some work also outputs images as a supplement to the generated summaries [9, 10] to enhance text information and improve user satisfaction.

There are mainly two methods to select the output image. One is to use the multimodal attention mechanism to determine the correlation of different modes, so as to select the most relevant image . More precisely, the visual coverage score after the last decoding step is used to select the most relevant image, such as the sum of attention values when generating text summaries. The second method is to classify the image selection task into the model . For example, Fu et al. used unsupervised learning technology using reinforcement learning (RL) method to select images, which takes representativeness and diversity as two reward functions of RL learning. Li et al. proposed a cover box selector, which is based on CNN-RNN layered video coding, and uses the conditional self-attention module to select an image based on the semantics of the article. The three frameworks with implicit text image summarization generation features cited above adjust the final loss together with the image selection loss to the weighted average of the text generation loss.

3. Evaluation method

In order to evaluate multimodal abstracts, Zhu et al. proposed a multimodal automatic evaluation (MMAE) technology, which comprehensively considers the unimodal significance and cross modal correlation. The final summary is composed of text and image, and the final objective function is composed of three objective functions: 1) text significance, 2) image significance, and 3) text image correlation. Monitoring techniques (linear regression, logistic regression and multilayer perceptron) are used to learn the mapping function to minimize the training loss of human judgment scores. Although this indicator seems promising, many conditions must be met for evaluation.

Because the MMAE index takes the unimodal significance score as a feature of the overall judgment process, resulting in cognitive bias, it cannot effectively evaluate the information integrity of multimodal summaries. Subsequently, Zhu et al. improved this and proposed an evaluation index to summarize the generated summary and basic facts into a joint multimodal

representation of a common semantic space. Compared with other multimodal evaluation indicators, we try to treat the multimodal summary as a whole, rather than a combination of segmented important elements. The model based on neural network is used to train this joint representation. In order to automatically obtain the training data of joint representation, the images of two image title pairs are exchanged to obtain two image text pairs with similar semantics. Then, the evaluation model is trained using the multimodal attention mechanism to fuse text and image vectors, and the maximum edge loss is taken as the loss function. Table 3 shows the evaluation methods and advantages and disadvantage.

Table 3. Evaluation methods and advantages and disadvantages

Evaluation method	advantages	disadvantages
MMAE	It is highly similar to the score of manual discrimination	Datasets that require a large number of manual annotations; Abstracts from other areas may be ambiguous when evaluated
MMAE++	The joint multimodal representation of sentence image pairs is used to improve the correlation score of MMAE	Datasets that require a large number of manual annotations; Abstracts from other areas may be ambiguous when evaluated

To sum up, only a few works focus on the evaluation of multimodal abstracts, and the proposed evaluation indicators also have many shortcomings. The evaluation index proposed by Zhu et al. and Zhu et al. requires a large amount of training data based on manual evaluation scores to learn parameter weights. In addition, the evaluation indicators depend on the field of training data, which limits the scope of application of these indicators.

4. Conclusion

Although multimodal summarization has made some progress, there are still several key points worthy of serious consideration:

(1) The existing model is simple in structure. The existing model architecture is basically sequence to sequence model combined with hierarchical attention mechanism. Different tasks will be improved according to the characteristics of tasks. In order to integrate multimodal information more effectively and give full play to the interaction and complementarity of modal information, we should consider a more appropriate architecture based on the current architecture.

(2) There is less information interaction between different modes. The core of the existing work mode fusion lies in the hierarchical attention mechanism. In addition, different modal information lacks an explicit interaction mode, which cannot give full play to the complementary relationship between modal information.

(3) Little consideration is given to data privacy. Multimodal data not only provides more information, but also brings some challenges to data confidentiality. For example, in multimodal conference data, voiceprint features and facial features are very important personal privacy information. Therefore, data privacy needs to be fully considered in the actual implementation.

In general, under the background of the hot development of multimodality, multimodal summarization, as a branch of multimodal learning, has both the common problems of multimodal learning and the personality problems of the summarization task itself. This task has begun to flourish in recent years and will become an important research direction in the future.

References

[1] Jangra, A., et al., Text-Image-Video Summary Generation Using Joint Integer Linear Programming, in Advances in Information Retrieval. 2020. p. 190-198.

[2] Li, H., et al. Multi-modal summarization for asynchronous collection of text, image, audio and video. in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[3] [Cho, K., et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[4] Sutskever, I., O. Vinyals, and Q.V. Le. Sequence to sequence learning with neural networks. in Advances in neural information processing systems. 2014.

[5] Bahdanau, D., K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[6] Rush, A.M., S. Chopra, and J. Weston, A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685, 2015.

[7] Shi, T., et al., Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science, 2021. 2(1): p. 1-37

[8] Hochreiter, S. and J. Schmidhuber, Long short-term memory. Neural computation, 1997. 9(8): p. 1735-1780.

[9] Zhu, J., et al. MSMO: Multimodal summarization with multimodal output. in Proceedings of the 2018 conference on empirical methods in natural language processing. 2018.

[10] Zhu, J., et al., Multimodal Summarization with Guidance of Multimodal Reference. Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 34(05): p. 9749-9756.

[11] Khullar, A. and U. Arora, MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention. arXiv preprint arXiv:2010.08021, 2020.

[12] Liu, N., et al. Multistage Fusion with Forget Gate for Multimodal Summarization in Open-Domain Videos. in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[13] He, K., et al. Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[14] Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[16] Deng, J., et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. IEEE.

[17] Karpathy, A., A. Joulin, and L. Fei-Fei, Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679, 2014.

[18] Wang, L., Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[19] Rashtchian, C., et al. Collecting image annotations using amazon's mechanical turk. in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010.

[20] Hodosh, M., P. Young, and J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013. 47: p. 853-899.

[21] Young, P., et al., From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2014. 2: p. 67-78.

[22] Wang, L., Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Vaswani, A., et al. Attention is all you need. in Advances in neural information processing systems. 2017.

[24] Li, H., et al. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. in IJCAI. 2018

[25] Fu, X., J. Wang, and Z. Yang, Multi-modal Summarization for Video-containing Documents. arXiv preprint arXiv:2009.08018, 2020

[26] Kim, J.-H., et al., Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325, 2016.

[27] Li, M., et al., VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles. arXiv preprint arXiv:2010.05406, 2020