2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) | 978-1-6654-8090-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/AEECA55500.2022.9919012

Research on Multimodal Summarization by Integrating Visual and Text Modal Information

Qiduo Lu*
Academy of Combat Support Rocket Force University of Engineer
Xi'an, China
lqd987273@163.com

Chenhao Zhu
Academy of Combat Support Rocket Force University of Engineer
Xi'an, China
972778371@qq.com

Xia Ye
Academy of Combat Support Rocket Force University of Engineer
Xi'an, China
yex_qing@163.com

***Abstract*—With the rapid development of information technology, Internet data is growing exponentially, which makes it difficult for users to extract key information from massive Internet data. Data compression technology represented by text summary has gradually attracted extensive attention from academia and industry. As an extension of text summarization, multimodal summarization task can effectively reduce users' information burden and improve users' information acquisition speed by integrating visual and auditory modal information and using the mutual supplement and verification of different modal data. It has high research value in the fields of information retrieval, public opinion analysis, content review and so on. This paper combs the related research of multimodal summarization in recent years, summarizes the existing technologies and related data sets for multimodal summarization tasks, and summarizes the development direction of future research in this field.**

Keywords—multimodal summarization, Sequence to Sequence model, attention mechanism

I. INTRODUCTION

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 a continuous stream of information sources. Therefore, data compression technology has gradually become a research hotspot in academia and industry. As a kind of data compression task, automatic summarization technology can effectively reduce users' information burden, improve users' information acquisition speed, free users from cumbersome and redundant information, and save a lot of human and material resources in information retrieval, public opinion analysis Content review and other fields have high research value.

However, the current research on automatic summarization mostly focuses on the field of text summarization, ignoring visual and auditory information. Various studies have shown that using multimodal data as input does help to improve the quality of abstracts [1,2]. Zhu et al. [3] proved that the summary with visual information can improve user satisfaction by an average of 12.4% compared with the plain text summary. Therefore, as an extension of text summarization task, multi-modal summarization (MMS) task makes the application range of summarization task wider by fusing the information of visual and auditory modes and using the mutual supplement and verification of different modal data; At the same time, like any single-modal summarization task, multi-modal summarization may have multiple correct solutions, which makes the task more challenging and interesting.

Multimodal summarization is developed from text summarization. By fusing multimodal information such as text, image and video, more comprehensive and accurate information can be obtained. Srivastava et al. [4] described a deep Boltzmann machine, which is used to learn the data generation model composed of multiple input modes. The model can be used to extract the unified representation of fusion modes. However, due to the uneven development of early computing power and various modal processing methods, less attention is paid to the generation task in the early stage, most of them only discuss the tasks such as classification assisted by multimodal information. At present, the generation of multimodal abstract is mostly based on visual question and answer. Generally, separate coding models are used to encode text information and visual information respectively, and then understand it based on fusion or interaction. Figure 1 shows the basic framework of multimodal summary, which is mainly composed of five parts: multimodal data input, preprocessing, main model, post-processing and final summary output.

According to different classification principles, multimodal summary can be divided into the following categories: a) according to the input mode combination: text image [3,5,6], text video [7,8], audio video [9,10], text image audio video [1,2,11-13]; b) According to the type of input text: single document [3, 6], multi document [1, 2, 11, 12]; c) According to time consistency: time synchronization [9, 10], time asynchrony [1, 2, 11, 12, 14]; d) According to the type of output text summary: extracted text summary [1, 2, 11, 12], abstractive text summary [3, 5, 6, 15]; e) According to the combination of output modes: single mode output [2, 6, 10, 16], multi-mode output [1, 3, 5, 11, 12, 14, 15]; f) According to the way of processing continuous media: extracting information to obtain discrete representation [1, 2, 11, 12], semantic segmentation according to logic technology [10, 14], sliding window [9].

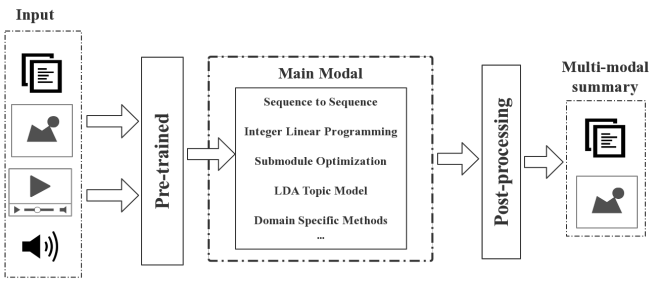

Fig. 1. Basic framework of multimodal summary task

When human beings complete the task of multimodal

summarization, they will use their own prior knowledge and external knowledge, while the computer lacks human perception and knowledge. Therefore, how to make the computer better simulate the human brain to process information promotes the development of multimodal summarization technology. This paper aims to give a brief overview of the current mainstream multimodal summary models and data sets.

II. Techniques and Methods Used in Main Model

A. Method Based on Sequence to Sequence

In recent years, deep learning technology provides a new idea for the research of automatic summarization. Among them, the research and application of sequence to sequence (seq2seq) model is the most extensive. The model is proposed by Cho et al. [17] and Sutskever et al. [18] for machine translation tasks, and is composed of encoder and decoder. In order to deal with long text, Bahdanau et al. [19] introduced the attention mechanism into the model. Rush et al. [20] applied this model to generative summary for the first time. Compared with the previous generative method, this model generates summary based on "understanding" text semantics, which is closer to the generation process of manual summary. Therefore, sequence to sequence model is gradually applied to the field of multimodal summarization. Figure 2 shows the basic framework of multi-modal abstract domain sequence to sequence model, which is mainly composed of three modules: encoder, multi-modal fusion module and decoder.

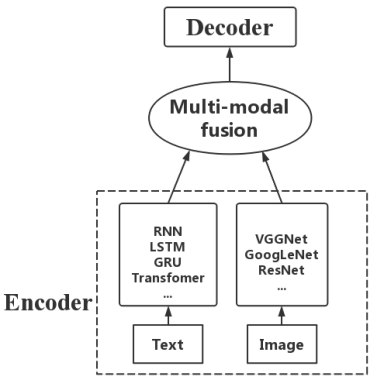

Fig. 2. Basic framework of Seq2Seq model in multimodal summarization

1) Feature Extraction Module (Encoder)

Encoder is a general term, including text encoder and visual encoder. Unlike text, images are discontinuous and have two-dimensional context. Therefore, most current encoders encode each mode separately.

For text data, because multimodal summarization is an extension of text summarization, almost all coding methods of text summarization based on sequence to sequence are suitable for coding text data in multimodal summarization. In addition, the information of our default speech mode can be transformed into text mode through transcription. Next, we will introduce the text coding methods commonly used in multimodal summarization research.

Zhu et al. [3, 15] used bidirectional LSTM for word level coding with reference to the pointer generator network; Khullar et al. [21] used bidirectional GRU to encode text. It is worth noting that Chen and Zhuge [5] adopt hierarchical RNN encoder and use two bidirectional gated recurrent units (BiGRU) to encode sentences and documents respectively. First, encode the sentences to obtain the vector representation of each sentence, then use the vector representation of the sentences as embedding, and then use the BiGRU to encode the documents to finally obtain the vector representation of the documents. Similarly, Fu et al. [7, 22] used hierarchical bidirectional long and short-term memory (BiLSTM) based on word and sentence level for text coding. The first layer of the encoder extracts the fine-grained information of the sentence through word level coding; The second layer of coding obtains the vector representation of the sentence through the weighted average of word level coding. In addition to the composition coding method used by GRU and LSTM, the emergence of RNN variant transfomer has quickly attracted the attention of researchers. Liu et al. [23] used bidirectional transfomer to encode text, in which each layer is normalized by a multi head self attention layer and layer followed by a feedforward sublayer with residual connections.

For video modal information, it can be transformed into picture modal information by extracting key frames and other technologies. Therefore, in this part, we mainly introduce the coding method of picture modal 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 [24], VGGNet [25], GoogLeNet [26]) to train on large image classification data sets such as ImageNet [27]. Li et al. [6] and Zhu et al. [3, 15] used the pre-trained VGG19 to extract the local and global features of the image; Li et al. [28] used ResNet-101 pre trained on ImageNet to extract the global features of the image, and then used Faster R-CNN to extract the local features of the image.

2) Multimodal Fusion Module

At present, the most commonly used multi-modal feature fusion methods are series, element by element multiplication, element by element addition, bilinear pooling and so on. With the proposal of attention mechanism, most tasks based on text image input focus on using multimodal attention mechanism to promote the smooth flow of information between the two modes. Chen et al. [29] fused the extracted multi-layer image features with text features to obtain multiple groups of single-layer text image attention fusion features, and distributed their weights through the attention network. In another work, Fu et al. [7] proposed Bi-hop attention mechanism as an extension of bilinear attention [30], which can effectively fuse text and image information. Li et al. [8] designed a new conditional self-attention module to obtain the local semantic information of video based on the input text information.

3) Summary Generation Module (Decoder)

The decoder is usually designed according to the coding strategy and generation mode. At present, the vast majority of MMS models based on neural network use multimodal data to generate single text mode summary. Most of the decoder structures are based on RNN deformation networks such as ordinary unidirectional RNN or multi-granularity hierarchical RNN. Some researchers found that multimodal summary output can enhance text information and improve user experience. In order to meet the task of multimodal output, some more complex decoders have been developed one after another. Zhu et al. [3] designed a visual coverage mechanism to select the most important image from the input image and realize multimodal summary output. In another work, Zhu et al. [15] took image selection as a classification task, trained the image selector with cross entropy loss, and

incorporated the image selection function into the model.

B. Method Based on Integer Linear Programming

Integer linear programming (ILP) is often used to generate text abstracts [30, 31], especially extracted text abstracts. It regards the automatic text summarization task as a problem to solve the global optimal solution based on 0-1 binary variable. In the early stage, researchers used a relatively simple redundancy removal mechanism, the maximum marginal correlation (MMR) [32] to select the appropriate content to form the summary. Later, McDonald [33] proposed a global optimal method instead of MMR for multi document summarization. One method is to express the multi document summarization problem as an integer linear programming problem and use an efficient branch and bound algorithm to solve the NP hard problem. In order to improve the scalability of ILP, Gillick et al. [34] proposed an extensible global model based on ILP, which operates at the clause or concept level, assuming that the concept is independent, which can be words, named entities and semantic relationships. This work can be more effectively extended to larger problems because it does not need quadratic variables to deal with redundant items. The objective function is:

$$\max \sum_i \omega_i a_i \tag{1}$$

Where, a_i is the indicator variable, indicating whether the concept i exists in the summary, and the weight is ω_i. In addition, Boudin et al. [35] achieved ideal results by using approximate algorithms to eliminate NP hard problems and multiple optimal solution problems caused by pruning. Previous studies have shown that if used reasonably, integer linear programming method can also be used in multimodal summarization tasks.

Jangra et al. [1] used the method of joint integer linear programming in solving the task of text image video summary generation (TIVs) for the first time:

$$\begin{aligned} &\max_{Arg}[\lambda \cdot m \cdot \{Sal(M_{txt}) + Sal(M_{img})\} \\ &+(1-\lambda)\cdot(k_{txt}+k_{img})\cdot MCorr(M_c)] \end{aligned} \tag{2}$$

Among them, $Sal(M_{txt})$ and $Sal(M_{img})$ represent the significance of text and image respectively, and $MCorr(M_c)$ represents multimodal correlation. In order to avoid the problem of modal deviation, coefficients m and $k_{txt}+k_{img}$ are introduced. The model takes the joint embedding of pretrained sentences and images as input. After extracting the most important sentences and images (including video key frames) using ILP framework, the image is separated from the key frames, supplemented with other images with medium similarity, and the threshold and upper limit are determined in advance to avoid noise and redundant information. In this case, the cosine similarity of global image features is used as the basic similarity matrix. The weighted average of language score and visual score is used to determine the video most suitable for multimodal summary. Language score is defined as the information overlap between voice transcription and generated text summary, while visual score is defined as the information overlap between video key frames and generated image summary.

C. Method based on Submodule Optimization

Submodule function is a kind of combinatorial optimization function, which was first used in the research of matroid optimization, Boolean polynomial and warehouse location [36]. Later, the optimization based method has a wide application background in the task of text summary.

Based on the method of sub module function, sentence scoring, abstract sentence selection or abstract sentence compression are transformed into optimization problems under constraints, and a class of set functions with sub modularity are constructed to solve the problems of information redundancy and abstract optimization [37].

The definition of sub module function is as follows: function $F:2^V \to R$ is a sub module function under a given finite set V, if and only if for any subset S, $T \subseteq V$, if there is $s \in V$, $S \subseteq T$ and $s \notin T$, there are:

$$F(S \cup \{s\}) - F(S) \geq F(T \cup \{s\}) - F(T) \tag{3}$$

In summary research, set V is the complete set of sentences in the document, the final summary S is a subset of V, and the score of function $F(\bullet)$ represents the quality of the summary. Obviously, the abstract is more concise than the original text and meets $|S| \leq K$. therefore, the summary research is a maximization sub module function problem under knapsack constraints:

$$\max_{S \subseteq V} \{\mathrm{F(S)}: |S| \leq \mathrm{K}\} \tag{4}$$

Because this is a NP complete problem, usually $F(\bullet)$ can only be approximately solved by a simple greedy selection algorithm.

The sub module function was first introduced into the field of automatic summarization by Lin and Bilmes [38]. They proposed to define automatic summarization as the maximization problem of sub module function of budget constraint, that is, each text unit has a budget [39]. On this basis, Lin et al. [40] designed a class of sub module functions suitable for extraction automatic summarization task. These functions consist of two parts: the first part is used to encourage the summary to contain more information; The second part is used to encourage the diversity of content and reduce redundancy. These functions are monotonic, which means that an efficient and scalable greedy optimization scheme has the guarantee of constant factor optimality. Subsequently, Chali et al. [41] defined three monotonic sub module functions, namely importance, coverage and non redundancy. The objective function is a linear combination of sub module functions. The process of generating summary is formally expressed as the problem of maximizing the objective function under length constraints, and the compressed sentences are extracted through sub module functions. This method first merges the sentences with the same subject but different verb phrases in multiple documents, then compresses the sentences through the dependency tree to generate a more concise and informative new summary

candidate sentence, selects the best sentence from the sentence set to maximize the objective function, and finally uses the greedy algorithm to obtain the approximate optimal summary. Sub module functions have shown excellent performance in text summarization tasks. Therefore, researchers consider applying them to the field of multimodal summarization.

Tiwari et al. [42] used coverage, novelty and importance as sub module functions to extract the most important documents for generating the timeline of social media events in a multimodal environment. In order to link different platforms and topics across time intervals, the author proposes a similarity measure based on Markov random field (MRF), which solves the differences between vocabularies of different platforms, and uses the explicit links between platforms for parameter learning. Because it is difficult to find the best subset of documents related to the optimization goal, the author uses the greedy algorithm to jointly sort the text and visual patterns selected by the documents, and represents the objectives of coverage, importance and novelty as a single tone submodule, which ensures that the greedy solution is at least the best result $1-\frac{1}{e}$ ($\approx 63\%$). In order to carry out multimodal summarization of asynchronous data on the Internet, Li et al. [2] took text saliency, redundancy and visual coverage as sub module functions. The objective function is the linear combination of sub module functions under budget constraints, and the text summary is generated by obtaining the approximate optimal solution at the sentence level.

D. Method based on LDA Topic Model

With the exponential growth of the number of documents on the Internet, probabilistic topic model was initially applied to the field of document summarization. As the most popular and representative probabilistic generation model in topic model, late Dirichlet allocation (LDA)topic model is a document generation model, which originates from the field of natural language processing.

The main idea of LDA topic model is to find hidden topics through text modeling, which is extended by Blei et al. [43] on the basis of pLSA.Too many pLSA parameters will lead to over fitting problem. On this basis, LDA adds super parameters and uses Dirichlet distribution as a priori distribution of document subject and word subject. Although the distribution of topics is specific to each document, the distribution of words related to topics is shared by all documents. Therefore, the topic model can extract meaningful semantic representation from the document by inferring its potential distribution on the topic from the words contained in the document.

In the context of computer vision, LDA transforms the image into a visual word document by extracting the so-called "visual words" from the image, and trains the LDA topic model on the visual word bag. Because multimodal summarization involves many modes such as text, image and video, there are many variants of LDA in order to process multimodal data.

Bian et al. [44] proposed a multimodal probability generation (MMLDA) model, as shown in Figure 3. By exploring the correlation between different media types, we can find sub themes from microblog. Its advantage is to organize the messy microblogs into structured sub themes, then generate high-quality text summaries at the sub theme level, and best represent the text content by selecting the images related to the sub themes. But its disadvantage is that it only focuses on summarizing the content of synchronous multimodality. On this basis, Bian et al. [45] enriched the mmLDA model by designing a noise removal method, that is, estimating the probability of the image related to a given event through the spectrum filtering model, so as to remove the potential noise image.

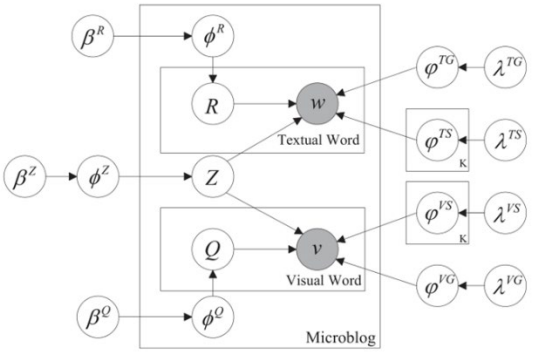

Fig. 3. Graphical representation of MMLDA model

Li et al. [46] introduced the hierarchical late Dirichlet allocation (hLDA) model to the multimedia news summary method of Internet search results, found the hierarchical topic structure from the query of relevant news documents, and proposed a method based on weighted aggregation and maximum pool to identify a representative news article for each topic. At the same time, a representative image is selected for visualization as a supplement to the text information. In addition, the author proposes a time biased MST method, which inserts sub topics into a topic, and gives the news summary of each topic from the perspective of time and space development. Its advantage is that the system can easily present vivid and comprehensive information. Readers can quickly understand the information they need through the multimedia summary in the system, but it is only suitable for news media summary at present.

E. Domain Specific Methods

Most of the models used to solve multimodal summarization problems are general, but there are still models that only apply to some specific fields.

For example, in the field of film, evangelopoulos et al. [10] proposed to detect the key frames in the film based on the significance of individual features represented by hearing, vision and language. Each mode is analyzed independently in a separate significance representation. To obtain the multimodal significance score of each video frame, the features within each mode and the significance between each mode are fused by using linear and nonlinear fusion methods, as well as weighted integration.

In the field of conference, Erol et al. [9] proposed to use text, voice and video features to locate important contents in the conference. In terms of audio, the features are established according to the Transformation Times of sound direction and sound amplitude in unit time. In terms of vision, the features are TF-IDF according to the brightness change of two adjacent frames, Finally, the key content is located by combining the characteristics of the above three modes.

In the field of electronic commerce, Li and others [28] apply multimodal summarization to the field of merchandise summarization, which is more challenging than traditional text summarization tasks. On the one hand, the first impression of goods to customers comes from the appearance of the goods, which has a vital impact on customers' purchase decisions. Therefore, the commodity summary system must be able to fully tap the visual information of commodities and reflect the appearance characteristics of commodities. On the other hand, different products have different selling points. For example, the advantage of a compact refrigerator is to save space, while the advantage of an environmentally friendly refrigerator is to save energy. Therefore, the commodity summary should reflect the most unique aspects of the commodity, so as to maximize the purchase of consumers. In order to effectively integrate the visual and text information of goods to generate the text summary of goods with prominent selling points, fluency and conciseness, the author proposes a multimodal summary model of e-commerce goods. The model is based on the pointer generator network and adopts three strategies to integrate the image information of goods into the model, It includes initializing the encoder with the global features of the commodity picture, initializing the decoder with the global features of the commodity picture, and generating the picture context vector through the attention mechanism to participate in the decoding. Firstly, R-CNN is used to mine valuable local features, such as the panel of refrigerator and the screen of mobile phone, which are integrated into each step of text decoding to make the model describe the parts with selling points of goods; Secondly, the global features of goods are mined through ResNetT, which, together with the global features of text, assigns the initial state of encoder and decoder, so as to enrich the features of goods from more angles, improve the distinguishability of goods and generate more diversified copybooks.

III. DATASET

Due to the flexibility of multimodal summarization task and various input-output combinations, there is no standard data set as a general evaluation benchmark for all methods. The commonly used task sets are summarized in Table I. Table I lists the basic information of 11 common data sets of multimodal summary. As can be seen from table I, among the 11 data sets related to multimodal summary, 7 are related to news, including Sanabria et al's video tutorial data set [47], and 4 data sets in specific fields.

TABLE I. MULTIMODAL SUMMARY COMMON DATASETS SUMMARY

No.	Paper	Type	Input Modalities	Output Modalities	Data Statistics	Domin	Source
#1	Li et al.[2]	MISO	T,I,A,V	TE	25 English documents, 25 Chinese documents	News	EMNLP2017
#2	Li et al.[6]	MISO	T,I	TA	66000 summary triples (sentence, image, summary)	News	IJCAI2018
#3	Sanabria et al.[47]	MISO	T,A,V	TA	2000h video	Multiple Domin	NIPS2018
#4	Li et al.[28]	MISO	T,I	TA	1375453 examples of household appliances, clothing, luggage and other categories	E-commerce	AAAI2020
#5	Tjondronegoro et al.[14]	MISO	T,A,V	T	66 hours of video (33 games), 1250 articles on the 2010 Australian Open Tennis Championships	Sports	IEEE2011
#6	Zhu et al.[3]	MIMO	T,I	TA,I	313k document, 2.0m image, news document, image title pair, sentence summary	News	ACL2018
#7	Li et al.[8]	MIMO	T,A,V	TA,I	Each sample contains an article, a video with a text summary, and a cover image	News	EMNLP2020
#8	Fu et al.[22]	MIMO	T,A,V	TA,I	Full modal data sets of English documents, abstracts, pictures, subtitles, videos, audio, text and titles of CNN and the daily mail	News	ACL2021
#9	Jangra et al.[1]	MIMO	T,I,A,V	TE,I,A,V	25 themes (500 documents, 151 images, 139 videos)	News	ECIR2020
#10	Bian et al.[44]	MIMO	T,I	TE,I	10 themes (127k microblogs and 48K pictures)	Social Media	ACM2013
#11	Evangelopoulos et al.[10]	MIMO	T,A,V	A,V	Seven and a half hours of film footage	Movies	IEEE2013

Note: Table I lists the basic information of 11 common data sets of multimodal summary, in which MISO and MIMO represent multimodal input single-mode output and multimodal input multimodal output respectively, T represents text, TA represents generative text summary, TE represents extractive text summary, I represents image, A represents audio and V represents video

In these 11 data sets, four data sets use multimodal input to generate text summary, of which the output of one data set is extracted text summary [2], and the other output results are abstractive text summary [6,28, 47]. The output results of five data sets are text image summary, of which the output results of two data sets are extracted text image summary [1,44], and the rest are abstractive text image summary [3,8,22]. In addition, data set #9[1] is the only data set that contains all modal information of text, image, video and audio in the output, but this data set is very small; Compared with the existing data sets, the data set #8[22] is a large-scale multi-modal data set covering almost all modes such as video, audio, transcribed text, image, caption and title. The data set #3[33] is a video data set, which contains 79114 teaching videos with English subtitles (a total of 2000 hours, with an average length of 90 seconds). In addition, how2 is a multilingual data set

containing English and Portuguese.

For the data sets of four specific fields, the data set #4[28] is a large-scale summary data set of Chinese electronic products, which includes about 14 million images, titles and other product descriptions from household appliances, clothing and luggage categories. The text, video, audio and voice transcripts in the data set #5 [14] come from 33 matches (including men's singles and women's singles) in the 2010 Australian Open, about 66 hours of video records and 1250 articles collected from 278 networks and social media. The data in the data set #10 [44] comes from 127118 microblogs and 48656 pictures under 10 trend topics in Sina Weibo. The data set #11[10] consists of three and a half hours of continuous half-hour clips from seven different theme films in Oscar winning films.

IV. Challenges and Development Trends

The MMS mission is relatively new, and the work done so far has only touched the surface of the services that can be provided in this field[48]. However, on the whole, in recent years, a lot of work has focused on the problem of multi-modal input and single-modal output, and there is less research on data set and evaluation index. In addition, the research on multimodal summarization lacks targeted leapfrog progress, and breakthrough innovation is needed to improve performance in order to more widely adapt to various scenarios. Therefore, the quality and performance of multimodal summarization tasks still face many challenges:

Evaluation criteria are not unique: One of the major problems faced by multimodal summarization is that it is difficult to design evaluation indicators to measure the advantages and disadvantages of the model. Especially in some abstractive tasks, such as image description and annotation, there is often no only correct "standard answer". The mapping process is vulnerable to subjective influence, so that the final result cannot confirm the representation of the same entity between different modes. Although we can also evaluate the mapping quality of the model through manual scoring or pairwise comparison to obtain the quality evaluation closest to human cognition, this kind of manual method is often time-consuming and costly. The labeling results are affected by the gender, age, cultural background and other deviations of the tester, resulting in inaccurate evaluation.

Rely on artificial prior knowledge: Generally speaking, different types of pre training feature extraction models need to be selected in advance for feature extraction. This process depends on strong manual judgment to determine the effective features in advance, which requires certain domain expertise. For example, people usually use transformer and its deformation to extract the features of text classes, and use RNN and its deformations to extract the features of temporal classes.

Shortcomings in multimodal fusion technology: The main goal of multimodal fusion is to reduce the heterogeneity between modes and maintain the integrity of the specific semantics of each mode. On the one hand, the problems of semantic conflict, repetition and noise in deep learning multimodal fusion have not been well solved. Although the attention mechanism can partially deal with these problems, it mainly operates implicitly and is not easy to be actively controlled. Due to the lack of explicit interaction, it can not give full play to the complementary relationship between modal information; On the other hand, the objective function of multimodal deep learning is usually a nonconvex optimization function. The current deep learning training algorithm can not effectively avoid the saddle point, resulting in the failure of the optimization process, which makes researchers unable to know whether the optimization process does not find the optimal solution, resulting in poor prediction results, or the problems in other modal fusion and modal alignment.

Single text output lacks diversity: Existing studies have proved that when the output summary contains multiple modes, it can meet the needs of a wider population. For example, when you are not familiar with the language, you can quickly understand important content through videos and pictures. Although some work has begun to try multimodal input and multimodal output, there is still much room for development.

Modal deviation problem is significant: The current multimodal summarization model is mainly trained by the target of text modality, which often ignores the quality of the image. This means that the system tends to only optimize the text summary generation process, while ignoring the image quality in the training process, which will lead to the problem of modal deviation.

In recent years, due to the rapid development of information technology, the application value of multimodal summary has become more and more prominent, which has attracted the attention of more and more scholars. At the same time, with the rapid development of deep learning, people see the hope that multimodal summarization can be widely used. In the long run, the development of multimodal abstracts has the following trends:

More effective multimodal information fusion: Up to now, almost all work has adopted feature-based fusion, and most of the technologies used focus on pre training and multimodal attention mechanism. Although these methods can capture the semantic overlap between different modes, there is still room for improvement in modeling the common semantic space of multimodal information.

More objective and comprehensive evaluation indicators: Most of the evaluation indexes of multimodal abstracts are ROUGE and image accuracy. This simple statistical index can not measure the semantic information of abstracts. In order to improve the overall satisfaction of users, the focus of the next step is to establish a complete evaluation system that can evaluate text and image abstracts at the same time.

Efficient feature extraction technology: For training depth neural network model, the more parameters of the model, the more complex the model is, and the better the performance is often. However, this is at the cost of sacrificing the calculation time and the sharp increase of calculation, which is obviously not an ideal application model. For the complex multi-modal summarization task, because it involves multi-modal information such as text and image at the same time, there is redundant information or even noise information in the features of different modes, which leads to the large scale of model parameters and brings difficulties to the subsequent information processing. A more efficient feature extraction technology is needed to solve this problem.

Multimodal output summary: Existing work has begun

to try multimodal input and multimodal output. When the output summary contains multiple modes, it can meet the needs of a wider population. For example, when you are not familiar with the language, you can quickly understand important content through videos and pictures. In the future, multimodal summary output will also become an important research focus.

Cross language multimodal summary: Multimodal information has been proved to be useful for multimodal neural machine translation tasks [49, 50]. However, whether language affects visual perception has always been a controversial issue [51]. In fact, this problem is still unresolved, which fully shows that the correct use of multimodal information is useful for multilingual summarization tasks

V. Conclusion

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, multimodal summarization task can effectively reduce users' information burden and improve users' information acquisition speed by integrating visual and auditory modal information and using the mutual supplement and verification of different modal data. It has high research value in the fields of information retrieval, public opinion analysis, content review and so on. This paper reviews the related research of multimodal summarization in recent years, summarizes the existing technologies and related data sets for multimodal summarization tasks, and summarizes the development direction of future research in this field.

Acknowledgments

This work was financially supported by Youth Science Fund Project of National Natural Science Foundation of China (62006240)

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

[4] Srivastava, N. and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. in NIPS. 2012. Citeseer.

[5] Chen, J. and H. Zhuge. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

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

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

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

[9] Erol, B., D.-S. Lee, and J. Hull. Multimodal summarization of meeting recordings. in 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). 2003. IEEE.

[10] Evangelopoulos, G., et al., Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia, 2013. 15(7): p. 1553-1568.

[11] Jangra, A., et al. Multi-modal summary generation using multi-objective optimization. in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.

[12] Jangra, A., et al. Multi-Modal Supplementary-Complementary Summarization using Multi-Objective Optimization. in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021.

[13] UzZaman, N., J.P. Bigham, and J.F. Allen. Multimodal summarization of complex sentences. in Proceedings of the 16th international conference on Intelligent user interfaces. 2011.

[14] Tjondronegoro, D., et al. Multi-modal summarization of key events and top players in sports tournament videos. in 2011 IEEE Workshop on Applications of Computer Vision (WACV). 2011. IEEE.

[15] 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.

[16] Libovický, J., et al. Multimodal abstractive summarization for open-domain videos. in Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NIPS. 2018

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

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

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

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

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

[22] Fu, X., J. Wang, and Z. Yang. MM-AVS: A Full-Scale Dataset for Multi-modal Summarization. in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

[23] 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.

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

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

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

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

[28] Li, H., et al. Aspect-aware multimodal summarization for chinese e-commerce products. in Proceedings of the AAAI Conference on Artificial Intelligence. 2020.

[29] Chen, Sun., et al., Graphic emotion analysis based on multi-layer cross modal attention fusion. Journal of Zhejiang University of Technology (Natural Science Edition), 2021: p. 1-10.

[30] Duan, Y. and A. Jatowt. Across-time comparative sum-marization of news articles. in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019.

[31] Alguliev, R., R. Aliguliyev, and M. Hajirahimova, Multi-document summarization model based on integer linear programming. Intelligent Control and Automation, 2010. 1(02): p. 105.

[32] Carbonell, J. and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998.

[33] McDonald, R. A study of global inference algorithms in multi-document summarization. in European Conference on Information Retrieval. 2007. Springer.

[34] Gillick, D. and B. Favre. A scalable global model for summarization.

in Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing. 2009.

[35] Boudin, F., H. Mougard, and B. Favre. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. in Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015. 2015.

[36] Wang, et al., Approximate algorithm for priority facility location problem with submodular penalty. Operations Research Transactions, 2015. 19(2): p. 1-14.

[37] Mao,et al, Research on Automatic Summarization Based on submodular function. 2017, Central China Normal University.

[38] Lin, H. and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010.

[39] Li, et al., A review of automatic text summarization. Computer Research and Development, 2021. 58(1): p. 1.

[40] Lin, H. and J. Bilmes. A class of submodular functions for document summarization. in Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011.

[41] Chali, Y., M. Tanvee, and M.T. Nayeem. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2017.

[42] Tiwari, A., C.V.D. Weth, and M.S. Kankanhalli, Multimodal Multiplatform Social Media Event Summarization. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018. 14(2s): p. 1-23.

[43] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. the Journal of machine Learning research, 2003. 3: p. 993-1022.

[44] Bian, J., Y. Yang, and T.-S. Chua. Multimedia summarization for trending topics in microblogs. in Proceedings of the 22nd ACM international Conference on information & knowledge management. 2013.

[45] Bian, J., et al., Multimedia summarization for social events in microblog stream. IEEE Transactions on multimedia, 2014. 17(2): p. 216-228.

[46] Li, Z., et al., Multimedia news summarization in search. ACM Transactions on Intelligent Systems and Technology (TIST), 2016. 7(3): p. 1-20.

[47] Sanabria, R., et al., How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347, 2018.

[48] Jangra A , Jatowt A , Saha S , et al. A Survey on Multi-modal Summarization[J]. 2021.

[49] Qian, X., Z. Zhong, and J. Zhou, Multimodal machine translation with reinforcement learning. arXiv preprint arXiv:1805.02356, 2018.

[50] Specia, L., Multi-modal Context Modelling for Machine Translation. 2018.

[51] Vulchanova, M., et al., Language and perception: introduction to the special issue “Speakers and listeners in the visual world”. Journal of Cultural Cognitive Science, 2019. 3(2): p. 103-112.