Multimodal Learning for Automatic Summarization: A Survey

Zhicheng Zhang(✉), Yibo Sun, and Shiyan Su

The University of Queensland, Brisbane, QLD 4072, Australia
zhicheng.zhang3@uqconnect.edu.au

Abstract. With the widespread availability of multiple data sources, such as image, audio-video, and text data, automatic summarization of multimodal data is becoming an important technology in decision support. This paper presents a comprehensive survey and summary of the main articles in the field of multimodal summarization techniques in recent years. Firstly, we define multimodal summarization and briefly describe the development process. Then, we survey existing techniques and their applicability in different domains. Additionally, we provide an analysis of their results and discuss the insights of those approaches, along with the challenges and future research directions. Based on our study, we found that the encoder-decoder approach is currently the best approach for automated summarization. In the future, we believe that the applications of multimodal summarization could develop rapidly in many different fields, particularly in medicine. In our case studies, we demonstrate that multimodal learning is a promising research direction for providing timely and accurate summarizations compared to unimodal approaches.

Keywords: Multimodal Summarization · Feature Engineering · Foundation Models · Attention Mechanism

1 Introduction

In past years, text-based unimodal automatic summarization has been developed and extensively researched [20]. Then, multimodal summarization has begun to receive increasing attention [3,13]. Multimodal automatic summarization can process and correlate information from multiple modalities, such as text, images, audio and video, to produce more coherent and accurate summaries with a high level of information. This approach has shown promising results in improving the quality and effectiveness of automated summaries. It includes a few steps: multimodal input, feature engineering, main model, fine-tuning models, and multimodal summary.

The aim of this article is to provide a comprehensive review of recent approaches in multimodal automatic summarization. We present a comprehensive overview of the main model and application areas of existing methods, categorizing the techniques into different types: methods based on neural networks,

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 362–376, 2023.
https://doi.org/10.1007/978-3-031-46664-9_25

method based on integer linear programming (ILP), method based on submodule optimization, graph- based approaches, method based on LDA Topic Models, and some domain-specific techniques. The application scenarios are considered as universal, news, meetings, movies, sports, medical, and others.

The paper also discusses challenges and future directions of multi-modal automatic summarization, identifies some important datasets, and provides possible directions for improvements of performance and quality with respect to the newly developed technologies.

2 Process of Multimodal Summarization

2.1 Multimodal Summarization

In 2009 Kay L. O'Halloran stated in his article that multimodality generally refers to different properties of the same medium and is a more precise and subdivided concept for representing something through multiple dimensions [40]. It can be expressed as different information properties, data, or representations that describe the same matter or object.

In Mani's book [34], automatic summarization is defined as the process of condensing a group or large amount of information and presenting its most important parts to the user in a short form. Examples include condensing a long report or collection of books into a concise text or presenting the statistics of a season of NBA games in a condensed form as a single image. Therefore, the output of automatic summarization is not limited to text, as numerous studies have shown that even better results can be achieved using images, videos, or multimodality as output. Zhu [55] claims that graphical summaries can increase user satisfaction by an average of 12.4%compared to text-only summaries.

Multimodal summarization can be defined as a computerized method of presenting a large amount of information in many different forms to the user in a streamlined manner. The input to the method must contain multiple forms, and the output can be in any form such as text, images, video or a combination of forms.

2.2 Development of Multimodal Summarization

The concept of multimodality can be traced back as far as the speeches of ancient Greece in BC and is used to express the diversity of behavior [3]. However, with the invention of the computer and the explosion of information flow in the information age, multimodal information has replaced traditional monotypic information in all aspects of life [4].

From the 1980s, the audio-visual speech recognition (AVSR) approach became the beginning of multimodal research [54]. Researchers found that when the demonstrator's lip movements did not match the articulation, the results received by the observed subjects would be affected. When the demonstrator mouthed [ba] and the dubbing was [ga], most subjects would mishear [da], which certainly suggests that multimodality can have a large impact on the results [36].

With the development of neural convolutional networks, multimodality was applied to automatic summarization in conference proceedings in 2003 [35]. By this time the authors had begun to model interactions using Hidden Markov Models and reduced the action error rate on the test set to 5.7%, providing ample evidence of the feasibility and promising future of the project.

It was not until 2006 that the concept of deep learning was introduced [24]. Since then, CNNs [23] and RNNs [19] have started to develop rapidly, encoder-decoder models, weighted attention mechanisms, and transformers [2] have been proposed, and Multimodal Summarization has started to evolve faster.

3 Methods

3.1 Method Based on Neural Networks

Neural network-based approaches are often preferred by researchers in the generation of multimodal summaries.

Neural network frameworks generally consist of an encoder-decoder, with the addition of a multimodal fusion module to form a complete architecture.

In 2003, McCowan [35] proposed the use of the Hidden Markov Models to model meeting behavior in the meeting domain. Early integration approaches combined the features of all participants in a single HMM and trained them. In 2009, Evangelopoulos [15] applied the spatio-temporal attention mechanism to film summaries, which improved their precision and avoided skimming caused by unimodal or visual-auditory-only modalities. In 2013, Evangelopoulos [14] further improved the method in the same area.

In the general domain, Nallapati [39] started using RNNs to summarize text in 2017, and a year later, Chen [11] used bidirectional RNNs to encode text and sentences, using a convolutional neural network VGGNet [46] to process images. This approach allows for the summarization of documents containing images and outperforms the SummaRuNNer method [39] in ROUGE scoring. Li [26] used VGG19 to extract image features, and Tsai [52] used a Transformer-based model for summarization. Additionally, Khullar [22] proposed a MAST method, which can summarize three modalities of "text-audio-video".

In the field of news summaries, good progress has also been made in multimodal research. Chen [10] used an attentional hierarchical encoder-decoder model to process text-centered information complemented by images, resulting in multimodal summaries. Zhu's MSMO method [55] uses a visual overlay mechanism to select suitable images from the output to supplement the summary. Palaskar [41] used a ResNeXt-101 3D convolutional neural network for video encoding. Another approach used by Chen [12] was to input text and images and use the then state-of-the-art Oxford VGGNet for image vector representation, which greatly improved the processing speed. Zhu [57] improved his MSMO [55] method proposed in 2018 in 2021 (Fig. 1).

In the medical field, Fan [16] proposed the FW-Net method to fuse CT images and MRI images to produce summaries with minimal loss of information, which

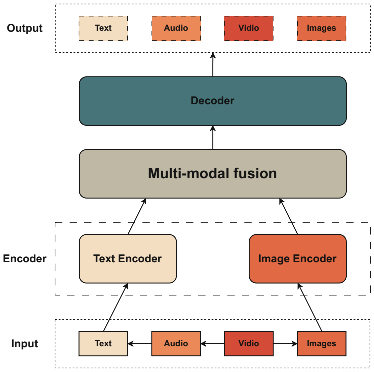

Fig. 1. Framework of Multimodality Automatic Summarization

resulted in good performance. The core of their algorithm is a two-layer U-Net algorithm that follows an encoder-decoder architecture.

Furthermore, Liu [32] also used CNN for the fusion summarization of multimodal medical images, including CT and MRI images. Their algorithm uses CNN to process two images and a weight map through a Siamese network, which uses Gaussian pyramid decomposition. They then perform Laplacian pyramid decomposition on each of the two images and finally perform another Laplacian pyramid decomposition on the resulting fused summary image. Torres [51] also used the DECU framework based on the CNN algorithm to generate an automatic summary of patient activity and determine the patient's health status by collecting other physiological parameters from video acquired by cameras and multiple sensors.

In other domains, Libovický [29] uses the seq2seq method to process instructional videos to generate tutorial summaries. Li [25] employs techniques such as R-CNN, ResNet, encoder-decoder and attention to produce product summaries in the e-commerce domain using a unique dataset. Song [47] utilizes the Swin Transformer and a Generative Pre-trained Language Model (GPLM) to generate product summaries in the e-commerce domain. Gao [17] employs the Sim Net network approach to implement code summaries in programming. Additionally, Ma [33] Gao [18] uses the Transformer architecture for code summarization.

3.2 Method Based on Integer Linear Programming (ILP)

Integer linear programming (ILP) is a method belonging to operations research that requires the decision variables to be integers. Unlike seq2seq, this approach directly intercepts the textual content, avoiding the problem of incoherent statements. This method was first used only for text summaries. Until Boudin [8] proposed an approximation algorithm that solved the NP-hard problem, and showed that it was not limited to text, but could also be applied to multimodal problems.

It wasn't until 2020 that Jangra [21] proposed a JILP-Multimodal summarization framework that achieved the task of summarizing multimodalities using ILP, and named the task TIVS, i.e. summarizing text-image-video. The method also simply uses a neural network approach to the output in a pre-processing phase, such as encoding the text using VGG. At its core, it uses the Joint-ILP Framework for core summarization.

Allawadi [1] also uses an ILP based model and has the same inputs and outputs as the previous JILP-Multimodal summarization framework [40]. However, his model is more refined and yields better accuracy and recall on ROGUE.

The decision variable of this method is:

$$M_{txt} = \left[m_{i,j}^{txt}\right]; M_{img} = \left[m_{i,j}^{img}\right]; M_c = \left[m_{ij}^{c}\right] \tag{1}$$

M(txt,img) is a binary square matrix of x*x. Whether txt or img is exemplar. c represents the cross-model, representing the correlation threshold between the image and the sentence. Its core function is:

$$\begin{aligned}
&f(x) = \mathrm{Arg}_{\max} \Big\{ \lambda_1 * m * k_{txt}^2 * \Big(\left[\textstyle\sum_{i=1}^{n} Mtxt_i * SIM_{\text{cosine}}\left(s_i, \quad O_{txt}\right)\right]^{(\alpha)} \\
&+ \left[\textstyle\sum_{i=1}^{n} Mimg_{i,i} * SIM_{\text{cosine}}\left(s_i, O_{img}\right)\right]^{(\beta)} \Big) + \\
&\lambda_2 * \left(k_{txtt} + k_{img}\right) * k_{txt}^2 * \left(\left[\textstyle\sum_{i=1}^{n} \sum_{j=1}^{p} M_{i,i}^{c} * SIM_{\text{cosine}}\left(s_i, \quad I_j\right)\right]^{(\gamma))} \right. \\
&\left. -\lambda_3 * \left(k_{txt} + k_{img}\right) * m * \left(\left[\textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} Mtxt_i * Mtxt_j * SIM_{\text{costne}}\left(s_i, I_j\right)\right]^{(\delta)} \right) \right\}
\end{aligned} \tag{2}$$

In the formula, α and β represent the salience score of the text-set and image-set, respectively. To avoid the problem of modal deviation, the coefficients m and ktxt+kimg are introduced. γ represents the cross-modal correlation score, δ represents the redundant part of the summary.

3.3 Method Based on Submodule Optimization

The submodule function is an aggregation function that provides a more tangible representation of diminishing marginal utility in the economic domain. Similarly to ILP, in 2010, Lin [31] first proposed applying this modified greedy algorithm to text summarization.

Until 2016, in the field of journalism, Modani [38] proposed an approach that uses a five-part submodule function to generate a summary of both text and image modalities. The method innovatively defines the image coverage term and an image diversity reward term for images. Allowing for the generic generation of a bimodal summary of text-image composition. Subsequently, the new method proposed by Li [27] reached new heights by being able to process four modalities: text, image, audio, and video. Chali [9] uses three measures of importance, coverage, and non-redundancy as submodule functions to detect sentence summaries. Tiwari [49] proposed a method for generating final summaries using three measures of coverage, novelty, and significance as submodule functions. The formula is:

$$f_{\text{coverage}}\left(S\right)=\frac{\left|\left\{w\in S|w\in\left(V^{txt}\cup V^{vis}\right)\right\}\right|}{\left|V^{txt}\cup V^{vis}\right|}. \tag{3}$$

$$f_{\text{novelty}}\left(S\right)=\sum\nolimits_w \max\nolimits_{d\in S}\left\{0,\min\nolimits_{d'\in S-\{d\}}\left\{\phi(d,w)-\phi\left(d',w\right)\right\}\right\}. \tag{4}$$

$$f_{\text{sig}}\left(S\right)=\log\left(c_{sig}\right)+\cos\left(d^{\overrightarrow{txt}},v^{\overrightarrow{txt}},\right)+\cos\left(d^{\overrightarrow{vis}},v^{\overrightarrow{vis}}\right). \tag{5}$$

They given a summary S. Coverage is as the fraction of $stxt$ and $svis$ of the vocabulary covered by the summary. Novelty means that the model should give preference to sentences with new information. w is a textual or visual word that appears in the document d of the summary.

They model the vectors $dtxt$ and $dvis$ and calculate their weighted cosine similarity. Capturing the importance of the document to the topic.

The authors use a Markov Random Fields-based similarity measure to compare different descriptions of the same or similar content across different platforms and track events over time to reconstruct the full event. Finally, the final content is selected using a submodule function-based approach. The core function is computed as follows:

$$f(A\cup\{s\})-f(A)\geq f(B\cup\{s\})-f(B) \tag{6}$$

In this formula, $A,B\subseteq s$, $A\subseteq B$, $s\in\frac{S}{B}$. S is a set.

3.4 Method Based on Graphs

Graph-based approaches have been used in the field of automatic summarization for a longer period. In 2004, the Graph based approach was applied to the field of journalism, and the Textrank method proposed by Mihalcea [37] has been able to extract important sentences from large news articles.

Until 2016, the Graph-based approaches was heavily used for multimodal automatic summarization. It was in the above-mentioned work by Modani [38] that a modified graph-based approach and a modification to the submodular approach were used to summarize both text and image modalities. Moreover, the proposed graph-based approach could handle not only images but also documents. The approach sets up images and documents as nodes into the graph, uses the connections between the nodes as weights based on similarity, and sets a reward score, as well as attach a cost. Finally, a greedy algorithm is used to select the most appropriate summary. Schinas [45] also proposed an MGraph framework for the textual, visual, temporal, and social multimodal content in social networking sites for visual summary summarization.

Subsequently, another paper by Li [27], mentioned above, also used the Graph based approach and summarized the four modalities of text-image-audio-video. In this case, the GBA (Graph based approach) is used to calculate the salience score of a text set. The text set here includes text documents, but also a large amount of text that may be incorrectly transcribed from speech. These sentences are treated as nodes to form a graph. The formula for calculating the salience score is computed as follows:

$$\begin{cases} Sa\left(t_i\right) = \mu \sum_j Sa\left(t_j\right) \cdot M_{ji} + \frac{1-\mu}{N} \\ M_{ji} = \operatorname{sim}\left(t_j, t_i\right) \end{cases} \tag{7}$$

In this formula, $\mu = 0.85$, N is total number of the text units; Mji is the relationship between text unit ti and ti; Ti is averaging the embedding of the words in ti. And $Sim(,)$ means cosine similarity between two texts.

Zhu [56] proposed an unsupervised graph-based multimodal summarization model, which does not require the dataset to contain annotations in order to perform summarization. The method classifies modal summarization into modal-mixed and modal non-mixed according to the form of output and can perform either unimodal or multimodal output to suit different application scenarios. Additionally, the method can also measure the similarity between text and images through the model.

Recently, Sun [48] applied the Graph based approach to the field of remote sensing images with Multimodal change detection for remote sensing for Earth observation. The approach performs a regression summary of different modal satellite images for regression summarization.

3.5 Method Based on LDA Topic Model

The Latent Dirichlet allocation (LDA) Topic Model [7] can be utilized for extracting visual words from images through feature extraction and clustering algorithms, thereby facilitating multimodal summarization.

Their approach towards Multimodal summarization has mainly been applied in the field of journalism. Bian [5] proposed the multimodal-LDA method for summarizing social media data in microblogs. The article first detects events, then quickly summarizes the most representative sub-topics, and generates a fluent summary text based on it to restore the entire process of the event quickly. On this basis, different summary focuses are selected based on the type of news to provide a more realistic picture of the event. However, the method may face difficultly in distinguishing the focus for mixed events or events in borderlands. Additionally, the summarization performance is significantly reduced for news with inconsistent text and images. The model Inference formula is as follows:

$$\varphi_k^{TS}(w) = \frac{N^w(Z=k, R=S) + \lambda^{TS}}{\sum_{t \in V^t} \left(N^t(Z=k, R=S) + \lambda^{TS}\right)} \tag{8}$$

$$\varphi_k^{VS}(u) = \frac{N^u(Z=k, Q=S) + \lambda^{VS}}{\sum_{u \in V^v} \left(N^u(Z=k, Q=S) + \lambda^{VS}\right)} \tag{9}$$

In the formula, $\phi_k^{TS}(w)$ represents the probability of w occurring in the kth specific text distribution, while $\phi_k^{VS}(u)$ represents the probability of the visual distribution. Where Vt and Vv denote text words and visual words, respectively.

N^w(Z = k, R = S), N^u(Z = k, Q = S) denote the number of text words after the sampling process.

Bian [6] proposed a method for removing latent noise images using a spectral filtering model as the core method, which allowed the algorithm to address the above problem well. In another work, Li [28] proposed the hierarchical latent Dirichlet allocation (HLDA) model to analyze the subject structure of news and then used subsequent methods such as crawlers and MST algorithms to process the subject matter. Wadagave [53] proposed the multimodal-LDA (MMLDA) summarization using the TWITTER API, which can also generate visual summaries.

3.6 Domain Specific Techniques

We can see that the above techniques are the dominant approaches to Multimodal summarization, but there are specific times when researchers use their own unique techniques suited to situations and particular data sets. These techniques are often related to relevant characteristics within the domain.

In sports, key sporting moments are often replayed in slow motion, and spectators will remain silent before a serve and then cheer loudly after a goal is scored. These specific phenomena can help the model to better identify key highlights of sports. Tjondronegoro [50] used this idea, together with the Video/Text Alignment Module, Social Media Classification Module and Text Analysis Module to complete automatic summaries of sports matches. Sanabria [44] also uses similar ideas and completes multimodal summaries with methods such as multi-instance learning neural networks. There are also specific features that can indicate the presence of key content in a session to avoid watching meaningless video content from start to finish. Erol [13] suggests that this can be done by analyzing sound direction and audio amplitude, local luminance variations, and term frequency-inverse document frequency measure, or even the video's movements of the characters to identify key sections for summary output.

In the field of e-commerce, Li [25] not only used a method based on neural networks but also adopted an aspect-based reward augmented maximum likelihood (RAML) training method [50], which effectively summarizes the aspects such as “Capacity”, “Control”, “Motor” and so on.

In the field of film summarization, Evangelopoulos [14] also used the concept of saliency to analyze three perspectives - auditory, visual, and textual - to obtain key frames of the film and summarize them. A multimodal fusion technique was eventually used to generate a comprehensive attention model. Finally, a summary is generated by extracting the most important scenes and episodes from the film based on the attention weights.

4 Taxonomy of the Methods

Although there are currently some articles that review similar topic, they are generally published too early or do not describe some areas. In this survey, I browsed through over 200 articles from 2003 to 2022 and selected the most valuable nearly 50 of them for classification statistics. They were classified by the main method into: Method Based on ILP, Method Based on Submodule Optimization, Method Based on Graphs, Method Based on LDA Topic Model, Domain Specific Techniques, and on this basis they are divided chronologically by application area. In the selection of papers, for the early years, we chose articles with high citation numbers. For the less cited articles of the last two years, we chose to use articles from higher quality publications. However, the datasets used for multimodal summarization tasks are not uniform.

Table 1. A list of methods, datasets, input and output patterns and their applications, with T(Text), I(Image), A(Audio), V(video) data.

Paper	Method	Input	Output	Datasets	Application
[35]	neural	A,V	T	60 meeting recordings (30 recordings × 2 participant sets)	Meetings
[15]	neural	T, A, V	A, V	3 movie segments	Movies
[14]	neural	T, A, V	A, V	7 half hour segments of movies	Movies
[26]	neural	T, I	T	66,000 triplets (sentence, image and summary)	News
[22]	neural	T, I, V	T	300 h of short instructional videos spanning different domains	General
[10]	neural	T, I	T, I	219k documents	News
[51]	neural	T, V	T	ICU patient Data set (author created)	Medical
[47]	neural	T, I	T	1.4 million products covering three coarse categories	Other
[17]	neural	T, C	C	10 open source Java projects, 40932 Ethereum Smart Contracts (ESC) code	Other
[21]	ILP	T, I, A, V	T, I, A, V	25 themes (500 documents, 151images, 139 videos)	News
[27]	sub/graph	T, I, A, V	T	66,000 triplets (sentence, image and summary)	News
[56]	graph	T, I	T, I	293,965 document,1,928,356 image	General
[6]	LDA	T, I	T, I	20 topics	News
[50]	specific	T, A, V	T	313k document, 2.0m image, news document, image title pair, sentence summary	Sport

Table 1 shows the basic information on commonly used datasets, the input and output modes of the paper, and the fields and sources of the paper. Method based on neural networks are still the dominant methods in the current methodology and are the focus of research in this survey. Figure 2 shows the current percent of each method.

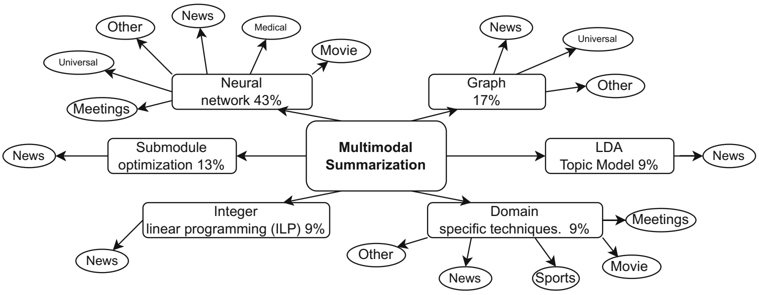

Fig. 2. Framework of Multimodality Automatic Summarization

5 Challenges and Issues

5.1 Challenges

Evaluation Criteria: There is no single correct answer to a multimodal summary. Even for manual assessments, there is no absolute perfect answer, limited by the personal preferences of the reader. And for machine assessment, there are different measures such as Rouge [30] scores. Furthermore, the output of multimodal summaries is also often multimodal, and it is difficult to measure the strengths and weaknesses between different modalities; in many cases, evaluation criteria do not allow for a comparison between methods that output text and images.

Currently, more advanced evaluation criteria [38] use vector function and reward mechanisms and they avoid to use Ground Truth. The methodological equation is:

$$\mu_M = \mu_T + \mu_I + \sigma_{T,I} \tag{10}$$

$$\mu_{(T,I)} = \sum_{w \in T,I} \hat{R}_{v,w} * \max_{x \in S,I} \{\mathrm{Sim}(x,y)\} \tag{11}$$

$$\sigma_{T,I} = \sum_{v \in S} \sum_{w \in I} \left\{ \mathrm{Sim}(v,w) * R_v * \hat{R}_w \right\} \tag{12}$$

where μT and μI are diversity-aware information coverage measures for the text and image parts of the summary, respectively. σT, I denotes the sum of the similarity between sentences and images in the summary across all pairs. $R * \max_{x \in S,I}\{\mathrm{Sim}(x,y)\}$ denotes the maximum similarity between sentence in the document text and any sentence in the summary. R is the reward value. However, this kind of methods lack normalization, and the results are heavily influenced by the length of the content.

No New Image Generated: Many methods in multimodal summarization use multimodal output when outputting, and their output often contains images. However, these methods generally output images by selecting the relevant image in the input video or image, and in the network, for output. The problem is that when there is no suitable image in the input video, the output becomes difficult and the quality of the output becomes lower, even if there is a mismatch between the text and the image.

Poor Quality Data Set: For machine learning, having datasets that perform well across various domains is crucial. However, currently, there are many datasets that are too small [21,27] or lack specific domains, with few datasets available.

Modal Alignment: Most methods have difficulty dealing with asynchronous modal data that is not aligned. Cross-modal alignment requires the resolution of timing asynchronies and scale inconsistencies between modalities.

Multimodal Semantic Understanding: The process of generating summaries requires semantic understanding and analysis of multimodal data. This includes the recognition and understanding of objects, scenes, etc. in images and videos, and the modelling of semantics in text.

5.2 Issues

Application Technologies: New technologies such as chat GPT [42] and DALL-E 2 [43] can solve problems such as no new images being generated, poor output text and difficulties with human-computer interaction. Neural network, with encoder-decoder as the core or use the transformer method are likely to become the mainstream approach to summarisation.

Deeper Applications in Medicine: Current approaches in the medical mainly use cnn for fusion abstraction of images from different modalities [16], while the critical text is neglected. In the future, summarising and outputting text modalities and images acquired by multiple sensors as a reference for doctors' decision making in routine examinations and ICUs will reduce doctors' decision making time.

Multimodal Alignment: Data heterogeneity, modal imbalance and semantic splitting make it difficult for multimodal approaches to achieve alignment. The performance and stability of multimodal alignment can be improved through data pre-processing, feature fusion and migration learning.

Real Time Summarization: Facing the sports domain, multi-modal real-time summarisation can be performed through specific scenarios, broadcast in different languages for different groups of people. Using machine learning algorithms, natural language processing (NLP) techniques, attention mechanisms, combined with text generation models, concise and accurate summaries of the competition can be generated. The generated text summaries are translated into the target language using machine translation models and natural and fluent speech announcements are generated using speech synthesis techniques.

Post-Joint Representation Approach: This can be addressed using joint representation learning or stepwise fusion strategies. The model considers the relationship between multiple modes at the same time in the training process, rather than just fusion information in the later stage.

Better Evaluation Criteria: Evaluation criteria will become more comprehensive and accurate. We could continue to use the vector function and reward mechanism from the reinforcement Learning. In addition representation learning can be introduced to extract Low-dimensional representations. Convergence modelling using multiple methods

Datasets Expanded: Datasets will cover more areas and a variety of data forms, and many new datasets will be constructed.

6 Conclusions

Multimodal summarization tasks allow people to navigate information from text, images, audio, and video more effectively. In this paper, we defined the problem and analysed the extent to which existing mainstream methods are used in different domains with the datasets provided. We identified a number of papers using new approaches and application areas that have not been summarized before. Through reflection and analysis, we enumerated the challenges currently faced by existing technologies, predicted possible future trends, and described some research issues and directions for future development.

References

1. Allawadi, S., Rana, V., Jain, M., et al.: Multimedia data summarization using joint integer linear programming. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1462–1466. IEEE (2021)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2018)
4. Bateman, J.A.: Multimodality and Genre. Palgrave Macmillan UK, London (2008). https://doi.org/10.1057/9780230582323
5. Bian, J., Yang, Y., Chua, T.S.: Multimedia summarization for trending topics in microblogs. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1807–1812 (2013)
6. Bian, J., Yang, Y., Zhang, H., Chua, T.S.: Multimedia summarization for social events in microblog stream. IEEE Trans. Multimedia **17**(2), 216–228 (2014)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
8. Boudin, F., Mougard, H., Favre, B.: 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)
9. Chali, Y., Tanvee, M., Nayeem, M.T.: 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), pp. 418–424 (2017)

10. Chen, J., Zhuge, H.: Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 4046–4056 (2018)
11. Chen, J., Zhuge, H.: Extractive text-image summarization using multi-modal RNN. In: 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), pp. 245–248. IEEE (2018)
12. Chen, J., Zhuge, H.: News image captioning based on text summarization using image as query. In: 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), pp. 123–126. IEEE (2019)
13. Erol, B., Lee, D.S., Hull, J.: Multimodal summarization of meeting recordings. In: 2003 International Conference on Multimedia and Expo. ICME 2003. Proceedings (Cat. No. 03TH8698), vol. 3, pp. III-25. IEEE (2003)
14. Evangelopoulos, G., et al.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans. Multimedia **15**(7), 1553–1568 (2013)
15. Evangelopoulos, G., et al.: Video event detection and summarization using audio, visual and text saliency. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3553–3556. IEEE (2009)
16. Fan, F., et al.: A semantic-based medical image fusion approach. arXiv preprint arXiv:1906.00225 (2019)
17. Gao, X., Jiang, X., Wu, Q., Wang, X., Lyu, C., Lyu, L.: Gt-SimNet: improving code automatic summarization via multi-modal similarity networks. J. Syst. Softw. **194**, 111495 (2022)
18. Gao, Y., Lyu, C.: M2TS: multi-scale multi-modal approach based on transformer for source code summarization. In: Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension, pp. 24–35 (2022)
19. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Networks Learn. Syst. **28**(10), 2222–2232 (2016)
20. Hahn, U., Mani, I.: The challenges of automatic summarization. Computer **33**(11), 29–36 (2000)
21. Jangra, A., Jatowt, A., Hasanuzzaman, M., Saha, S.: Text-image-video summary generation using joint integer linear programming. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 190–198. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_24
22. Khullar, A., Arora, U.: Mast: multimodal abstractive summarization with trimodal hierarchical attention. arXiv preprint arXiv:2010.08021 (2020)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
24. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
25. Li, H., Yuan, P., Xu, S., Wu, Y., He, X., Zhou, B.: Aspect-aware multimodal summarization for Chinese e-commerce products. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8188–8195 (2020)
26. Li, H., Zhu, J., Liu, T., Zhang, J., Zong, C., et al.: Multi-modal sentence summarization with modality attention and image filtering. In: IJCAI, pp. 4152–4158 (2018)
27. Li, H., Zhu, J., Ma, C., Zhang, J., Zong, C.: 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, pp. 1092–1102 (2017)

28. Li, Z., Tang, J., Wang, X., Liu, J., Lu, H.: Multimedia news summarization in search. ACM Trans. Intell. Syst. Technol. **7**(3), 1–20 (2016)
29. Libovickỳ, J., Palaskar, S., Gella, S., Metze, F.: Multimodal abstractive summarization for open-domain videos. In: Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NIPS (2018)
30. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
31. Lin, H., Bilmes, J.: 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, pp. 912–920 (2010)
32. Liu, Y., Chen, X., Cheng, J., Peng, H.: A medical image fusion method based on convolutional neural networks. In: 2017 20th International Conference on Information Fusion (Fusion), pp. 1–7. IEEE (2017)
33. Ma, Z., Gao, Y., Lyu, L., Lyu, C.: MMF3: neural code summarization based on multi-modal fine-grained feature fusion. In: Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 171–182 (2022)
34. Mani, I.: Automatic Summarization, vol. 3. John Benjamins Publishing (2001)
35. McCowan, I., et al.: Modeling human interaction in meetings. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 2003), vol. 4, pp. IV-748. IEEE (2003)
36. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (1976)
37. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
38. Modani, N., et al.: Summarizing multimedia content. In: Cellary, W., Mokbel, M., Wang, J., Wang, H., Zhou, R., Zhang, Y. (eds.) Web Information Systems Engineering-WISE 2016: 17th International Conference, Shanghai, China, November 8–10, 2016, Proceedings, LNCS, Part II 17, pp. 340–348. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48743-4_27
39. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
40. O'Halloran, K.L.: Interdependence, interaction and metaphor in multisemiotic texts. Soc. Semiot. **9**(3), 317–354 (1999)
41. Palaskar, S., Libovický, J., Gella, S., Metze, F.: Multimodal abstractive summarization for how2 videos. arXiv preprint arXiv:1906.07901 (2019)
42. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
43. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
44. Sanabria, M., Sherly, Precioso, F., Menguy, T.: A deep architecture for multimodal summarization of soccer games. In: Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, pp. 16–24 (2019)
45. Schinas, M., Papadopoulos, S., Kompatsiaris, Y., Mitkas, P.A.: MGraph: multimodal event summarization in social media using topic models and graph-based ranking. Int. J. Multimedia Inf. Retrieval **5**, 51–69 (2016)

46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
47. Song, X., Jing, L., Lin, D., Zhao, Z., Chen, H., Nie, L.: V2P: vision-to-prompt based multi-modal product summary generation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 992–1001 (2022)
48. Sun, Y., Lei, L., Tan, X., Guan, D., Wu, J., Kuang, G.: Structured graph based image regression for unsupervised multimodal change detection. ISPRS J. Photogramm. Remote. Sens. **185**, 16–31 (2022)
49. Tiwari, A., Weth, C.V.D., Kankanhalli, M.S.: Multimodal multiplatform social media event summarization. ACM Trans. Multimedia Comput. Commun. Appl. **14**(2s), 1–23 (2018)
50. Tjondronegoro, D., Tao, X., Sasongko, J., Lau, C.H.: Multi-modal summarization of key events and top players in sports tournament videos. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 471–478. IEEE (2011)
51. Torres, C., Rose, K., Fried, J.C., Manjunath, B.: Summarization of ICU patient motion from multimodal multiview videos. arXiv preprint arXiv:1706.09430 (2017)
52. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, p. 6558. NIH Public Access (2019)
53. Wadagave, P., Garg, B.: A heterogeneous data summarization system to generate automatic summary of data using twitter API using tweets, images etc. Harbin Gongye Daxue Xuebao J. Harbin Inst. Technol. **54**(6), 167–172 (2022)
54. Yuhas, B.P., Goldstein, M.H., Sejnowski, T.J.: Integration of acoustic and visual speech signals using neural networks. IEEE Commun. Mag. **27**(11), 65–71 (1989)
55. Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., Zong, C.: MSMO: multimodal summarization with multimodal output. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 4154–4164 (2018)
56. Zhu, J., Xiang, L., Zhou, Y., Zhang, J., Zong, C.: Graph-based multimodal ranking models for multimodal summarization. Trans. Asian Low Resour. Lang. Inf. Process. **20**(4), 1–21 (2021)
57. Zhu, J., Zhou, Y., Zhang, J., Li, H., Zong, C., Li, C.: Multimodal summarization with guidance of multimodal reference. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9749–9756 (2020)