Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 959–968
DOI : 10.47974/JIOS-1619

Multimodal news document summarization

Hira Javed *
Nadeem Akhtar †
M. M. Sufyan Beg §
Department of Computer Engineering
Aligarh Muslim University
Aligarh
Uttar Pradesh
India

Abstract

With the increase in multimedia content, the domain of multimodal processing is experiencing constant growth. The question of whether combining these modalities is beneficial may come up. In this work, we investigate this by working on multi-modal content for obtaining quality summaries. We have conducted several experiments on the extractive summarization process employing asynchronous text, audio, image,and video. Information present in the multimedia content has been leveraged to bridge the semantic gaps between different modes. Vision Transformers and BERT have been used for the image-matching and similarity-checking tasks. Furthermore, audio transcriptions have been used for incorporating the audio information in the summaries. The obtained news summaries have been evaluated with Rouge Score and a comparative analysis has been done.

Subject Classification: *Primary 68Txx, Secondary 68Uxx.*

Keywords: *Multimodal, Summarization, Machine learning, Transformers.*

1. Introduction

A lot of research has been done in the field of text processing. Text summarization, information retrieval, and question-answering systems have relied heavily on text for decades. However, the growth of large

* *E-mail:* **hira.javed@zhcet.ac.in (Corresponding Author)**
† *E-mail:* **nadeemakhtar@zhcet.ac.in**
§ *E-mail:* **mmsbeg@myamu.ac.in**

© Taru Publications

amounts of multimedia information in the form of images, videos, and audio demands the need for systems that leverage this information. The human comprehension of the text depends on multimodal information, including linguistic and visual cues. In the same way as the use of images, videos, and audio helps humans in a fast and easier grasp of information, multi-modal data must also help machines to understand the task better and represent the information well. The systems in the past were based only on text for the task of summarization. Thus, there is a need for utilizing computer vision and automatic speech recognition (ASR) systems to aid summarization. The process becomes more challenging when the data is asynchronous, i.e. there are no subtitles for videos and images without captions. Works have been done in the past in the areas of movie summarization [1], meeting recording summarization [2], and sports video summarization [3]. However, these researches have mostly been done on synchronous data i.e. the video consisted of subtitles or the images had captions. In this work, we tend to produce a textual summary by utilizing the asynchronous information from video and images as well as the text from news documents. The results have been compared with the Rouge Score metric and the limitations have been discussed. The work utilizes the techniques of CV, NLP, and ASR to find the summaries. The major challenge in this field is to build the semantic gap between text and image, as the multimedia data is complex and heterogeneous. Our contribution is as follows:

1. We have used the visual information to find the most relevant text, from the images embedded in the news documents.
2. We have designed an MMS method that extracts summaries using the audio data from videos.
3. We have summarized the textual documents using BERT embeddings and evaluated the summaries obtained from all 3 modes using the ROUGE score**.**

2. Related Work

This task can support additional applications like multimodal news summarization, image captioning, electronic commerce (e-commerce) product description generation [4], etc. In recent years, machine translation (MT) has extensively investigated multimodal sequence-to-sequence (seq2seq) learning, and their models outperform text-only models in terms

of performance [5]. Research published in the literature demonstrates that no one modality representation, fusion strategy, or learning algorithm can deliver state-of-the-art performance across all datasets. Two study areas—multi-document summarization and multimodal summarization—serve as inspiration for our work. Various linguistic indicators, including tf*idf [6], and sentence position [7] have been employed by extractive-based models to determine which phrases are the most salient. Commonly used extractive-based MDS models are graph-based approaches [8-14] which are built on the idea that sentences that share a lot of similarities with other phrases in a document set are more significant. Lexrank [13] is a popularly known approach that creates a graph with nodes representing phrases and edges representing the connections between them. After that, each sentence's relevance is calculated using an iterative random walk, which is used in PageRank [15]. Ultimately, the sentences that score highest are chosen to construct summaries.The output of multimodal summarization can be divided into two categories: multimodal output [16] and single-modal output [17]. A lot of work has been done recently to summarize social multimedia, sporting videos, movies, picture narratives, and meeting records. Authors in [17] maximize the salience, non-redundancy, and coverage to produce a textual summary from a collection of documents, photos, audio, and videos. LSTM has been used to develop a special supervised learning technique that automatically extracts features from TVSum50 dataset frames in [18]. Work in [19] offered a technique for condensing a sporting event that involves determining the key points from material gathered from various sources. In [20] a probabilistic two-level topic model has been implemented that infers hierarchical latent themes from sentences by modeling the text at the word and sentence levels. Slab prior and spike have been used in [21] for forming a two-level topic model that identifies concepts in summarization task. [22] and [23] suggested a method, to summarize real-life occurrences for social media summarization, using multimedia content, such as videos from YouTube and images from Flickr.

3. Our Model

3.1 *Problem Description*

Given, a collection of news documents and associated images and videos. We aim to deduce a textual summary that represents the relevant content by utilizing different modes, i.e. textual, visual, and audio.

3.2 *Model description*

We have proposed an extractive summarization method for news documents to extract summaries from different modes. We have performed three experiments on asynchronous news documents using visual audio and text mode.

3.2.1 Visual Mode

We have employed Transformer to obtain the description of images embedded in the text documents. Transformer architecture creates global relationships between input and output using an attention mechanism, unlike the traditional usage of recurrence. An RNN-like encoder-decoder architecture is used by the transformer network [24]. The novelty is that transformers can parallelly process the input sentencesTransformers learn meanings and contexts from sequential data. Images can be converted into a sequence of patches similar to how tokens are changed into embeddings and embeddings are learned to comprehend context. These patches use positional encoding and an n-dimensional embedding is referred to as a patch. Every patch has a weight or attention attributed to it as the model trains and each patch's significance is indicated by its weight. Tokens of the words act as queries to query the patches. Weights are assigned to keys, and they are linked to value V. Attention is represented mathematically as:

$$\text{Attention } (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q and K are the tensors. The dot product increases with the degree of proximity between the vectors q and k. For scaling, the factor d_k is applied. When the query and key vectors are not aligned, attention produces a number that is closer to 0. Various patch representations result in iterative attention computations, which are subsequently added together to get the attention score.The computation of these multiple attention heads is done using various learning Wq, Wk, and Wv matrices.

MultiHead(Q,K,V)=Concat(head_1, head_2,…head_h) W^o

where, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

3.2.2 Audio mode

For finding semantic similarity between sentences we have used Sentence-BERT embeddings. Sentence-BERT is a variant of the BERT

model that creates fix-sized, semantically relevant sentence embeddings by adding a pooling operation to the BERT output and using siamese and triplet network architectures. The closest embeddings from the corpus are determined when the sentences are embedded into the same vector space. Sentences with similar meanings are closer in vector space and this can be computed using cosine similarity between vectors.

3.2.3 Textual Mode

We have used BERT embeddings for text summarization. Vector representations of a particular word are called embeddings. The embeddings are useful in finding the semantic or the context of the word. The BERT model consists of encoder blocks and returns a sequence of hidden states and attention weights. The BERT basic model has 12 levels of transformer encoders, and each token's output from each layer of these encoders can be used to create a word embedding. These embeddings have then been clustered by the k-means algorithm to generate the representative sentences.

4. Experiment

4.1 *Dataset*

We have performed our experiment on the Multi-modal Summarization Dataset v1.0 [25] dataset. The dataset consists of documents from news topics. It comprises multiple news documents and asynchronous images and videos related to those topics. Few of the topics of dataset are Russian Airline crash, the Nepal earthquake, Germanwings Airbus plane crash. There are 3 reference textual summaries available for each topic. We have performed our experiments on 5 news topics and evaluated the generated summaries using ROUGE score with respect to these reference summaries.

4.2 *Implementation Details*

For our first experiment, we employed the pretrained vit-gpt2-image-captioning [26] model. The vision transformer is used to construct this model. Here, three transformer class pre-trained models have been used. The GPT 2 Tokenizer, the ViT image processor, and the Vision Encoder-Decoder model are the models that are employed. The training time is shortened by these pre-trained models. Here, we have avoided the training overhead of creating our model from scratch. A vision encoder-decoder

model has been used to encode the image. The gpt2 tokenizer utilized here has already been trained to handle token tasks like tokenization, adding new tokens, and the handling of mask tokens, etc. The viT image processor processes the image, rescales and resizes it, and picks the most similar sentences that match the image.In our second experiment, for each topic, we extracted the speech transcriptions from the videos. We have merged the videos belonging to a particular topic When the similarity between a sentence from transcription and the text document exceeds a certain threshold, the sentence from the textual document is included in the summary. Since the speech transcriptions are often ill-formed and consist of many informal cues, the sentence to be included in the summary is taken from the news article. For the last experiment, the text document is tokenized into sentences. These sentences are then passed to the BERT model for inference which results in the output embeddings. These embeddings are then clustered by the k-means algorithm The sentences that are closest to the centroid are selected as the candidate summary sentences. During the tokenization process the sentences that are too short or too long are also removed. The BERT architecture was chosen because it performed better on sentence embedding than other NLP algorithms. The N x E matrix was prepared for clustering where N is the number of sentences and E is the dimension of embedding. The value of k is provided by the user that would output the requested number of sentences in the summary. The final summary consists of sentences close to centroids. We have set the length of the summary to 15.

5. Results and Discussion

Summaries have been assessed using a metric called ROUGE score [27], or Recall-Oriented Understudy for Gisting Evaluation. We have used ROUGE-1 and ROUGE-L for evaluating our summaries. The results obtained from different experimental settings are tabulated in Table 1. We have obtained the best Rouge value for the text only model and Rouge values are the lowest in the vision model. Although the results obtained on news documents by utilizing visual and audio modes are not better than the text-only mode, they are noteworthy. They can supersede this mode, with the development of newer models in the future. Enhancement of summaries can also be achieved by preprocessing the data in these modes, such as eliminating unnecessary or informal content from audio recordings. The text-only mode yields a better outcome because the news documents provide clear and concise information about happenings.

These include accurate and comprehensive information on occurrences and are fairly structured. As a result, summaries produced by extracting sentences from these documents are of higher quality.On the other hand, the audio obtained from the news articles consists of many informal cues too, like greetings, discussion related to the connectivity with the reporter, etc. which is not veryinformative. Nonetheless, the images and videos have good coverage and consist of the highlights of the events. As the model is unable to extract precise and comprehensive information from an image, outcomes employing images are subpar. For example, a picture of a wreckage site does not provide exact details on the incident. For instance, adamaged building may be the consequence of an airstrike or an earthquake. Additionally, in the event of a protest, the visual transformers will only provide a broad description, such as "people holding placards in front of a building," without providing details about the cause of the protests. As a result, the matching sentences that were gleaned from the news stories are not as informative.

Table 1
Results

	Visual Mode		Audio Mode		Text Mode	
Topic Number	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
1	0.3986	0.1854	0.4013	0.1704	0.4071	0.1821
2	0.3504	0.1392	0.4264	0.1497	0.4006	0.1472
3	0.3433	0.1633	0.3306	0.1463	0.4273	0.1882
4	0.3548	0.1488	0.2818	0.1302	0.4261	0.1867
5	0.3478	0.1212	0.4375	0.1713	0.4359	0.1613
Average	0.3589	0.1515	0.37552	0.15358	0.4194	0.1731

6. Conclusion

This paper deals with an asynchronous Multimodal Summarization Task for creating a textual summary using modes of text, audio, images, and video. The asynchronous nature of data has made this task more challenging. We have devised methods to capture the relevant information from multi-modal content associated with the news documents. The techniques of Natural Language Processing, Computer vision, and Automatic Speech Recognition have been used to generate the summaries. Audio, text, and visual modalities have been assessed on this task using saliency of information. Results show that the multi-modal information

can be very beneficial in the process of summary generation and other NLP-related tasks. Visual models need to be developed in the future that can capture the relation between image and text. Different modalities can be combined further to improve the results. Other acoustic features can also be used to improve the results.

References

[1] Evangelopoulos, Georgios, et al. "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention." *IEEE Transactions on Multimedia* 15.7 : 1553-1568 (2013).

[2] Erol, Berna, D-S. Lee, and Jonathan Hull. "Multimodal summarization of meeting recordings." 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). Vol. 3. IEEE (2003).

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

[4] Zhang, Tao, et al. "Automatic generation of pattern-controlled product description in e-commerce." The World Wide Web Conference (2019).

[5] Calixto, Iacer, Qun Liu, and Nick Campbell. "Incorporating global visual features into attention-based neural machine translation." arXiv preprint arXiv:1701.06521 (2017).[6]D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based´ summarization of multiple documents," *Information Processing Management*, vol. 40, no. 6, pp. 919–938 (2004).

[7] Katragadda, Rahul, Prasad Pingali, and Vasudeva Varma. "Sentence Position revisited: A robust light-weight Update Summarization 'baseline' Algorithm." Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3) (2009).

[8] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing (2004).

[9] Li, Haoran, et al. "Guiderank: A guided ranking graph model for multilingual multi-document summarization." Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016,

and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24. Springer International Publishing (2016).

[10] Goyal, Pawan, Laxmidhar Behera, and Thomas Martin McGinnity. "A context-based word indexing model for document summarization." *IEEE Transactions on Knowledge and Data Engineering* 25.8 : 1693-1705 (2012).

[11] X. Li, L. Du, and Y. D. Shen, "Update summarization via graph_ based sentence ranking," *IEEE Transactions on Knowledge & Data Engineering,* vol. 25, no. 5, pp. 1162–1174 (2013).

[12] Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao. "CMiner: Opinion extraction and summarization for chinese microblogs." *IEEE Transactions on Knowledge and Data Engineering* 28.7 : 1650-1663 (2016).

[13] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of Artificial Intelligence Research* 22 : 457-479 (2004).

[14] Wan, Xiaojun, and Jianwu Yang. "Improved affinity graph-based multi-document summarization." Proceedings of the human language technology conference of the NAACL, Companion volume: Short papers (2006).

[15] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 : 107-117 (1998).

[16] Bhatnagar, Vaibhav, et al. "Descriptive analysis of COVID-19 patients in the context of India." *Journal of Interdisciplinary Mathematics* 24.3 : 489-504 (2021).

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

[18] Kumari, Rajani, et al. "Analysis and predictions of spread, recovery, and death caused by COVID-19 in India." *Big Data Mining and Analytics* 4.2 : 65-75 (2021).

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

[20] Akhtar, Nadeem, et al. "Tiered sentence based topic model for multi-document summarization." *Journal of Information and Optimization Sciences* 43.8 : 2131-2141 (2022).

[21] Akhtar, Nadeem, M. M. Sufyan Beg, and Md Muzakkir Hussain. "Sparse two level topic model for extraction of general summary words." *Journal of Interdisciplinary Mathematics* 23.1 : 303-310 (2020).

[22] M. Del Fabro, A. Sobe, and L. Boszormenyi, "Summarization of ¨ real-life events based on community-contributed content," in The Fourth International Conferences on Advances in Multimedia : 119–126 (2012).

[23] Schinas, Manos, et al. "Multimodal graph-based event detection and summarization in social media streams." Proceedings of the 23rd ACM International Conference on Multimedia (2015).

[24] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

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

[26] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning Accessed on 20/2/24

[27] Lin, Chin-Yew, and Eduard Hovy. "Automatic evaluation of summaries using n-gram co-occurrence statistics." Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics (2003).