SnapSummaries

Sakshi Bohra[1], Sneh Sinha[2], Riddhi Bora[3], Savitri Chougule[4]

[1] Student, School of Engineering, MIT ADT University, Pune, Maharashtra, India
[2] Student, School of Engineering, MIT ADT University, Pune, Maharashtra, India
[3] Student, School of Engineering, MIT ADT University, Pune, Maharashtra, India
[4] Professor, School of Engineering, MIT ADT University, Pune, Maharashtra, India

---***---

Abstract - *Efficient multi-modal summarization (MMS) is crucial in the era of huge digital data due to the increase in multimedia content. In order to automatically compress several content forms—text, images, audio, and videos—related to particular issues, this paper offers an extractive multi-modal summarizing approach. Closing semantic gaps between modalities is the main goal. This approach tries to improve salience, non-redundancy, readability, and coverage in producing textual summaries by selecting transcribing audio, employing neural networks for simultaneous text-image representation, and optimizing submodular functions. The review talks about the progress made in Automatic Text Summarization (ATS), with a focus on Extractive and Abstractive techniques and the function of the Text Rank algorithm. It also investigates video summarization using a viewer-centered computational attention model, providing a substitute for intricate video semantic analysis in the summation of multimodal content. Most notably, ongoing one area of work that is presently being developed is the translation of the summary text into other languages.*

Key Words: **multimedia model, summarization, text, audio, language translation, neural network**

1.INTRODUCTION

With Snap Summaries, an inventive way to speed up information retrieval is presented. It offers concise summaries of multiple documents. Our approach compresses a variety of content into clear, in-depth summaries by using sophisticated algorithms. Snap Summaries seeks to maximize efficiency and accessibility by providing users with a concise and all- encompassing summary of various documents through the use of state-of-the art multi- modal summation techniques.

Anubhav Jangra et al[11] Multi-modal summarization is required to extract important data while eliminating redundant information because the current explosion of multimedia content has made it difficult to extract meaningful information. This method gives a more thorough portrayal by offering a variety of viewpoints and providing more tangible reinforcement for concepts. Nevertheless, this work focuses on text image- video summary generation (TIVS) through a differential evolution-based multi-modal summarizing model (DE-MMS-MOO), whereas previous research largely focused on uni-modal summarization (text or images). Through multiobjective optimization, the model maximizes consistency between modalities and cohesion within them, providing a general framework that may be tailored to different optimization strategies. This innovative model addresses the problem of asynchronous data without alignment among various modalities by taking multimodal input and producing variable-size multimodal output summaries that include text, photos, and videos.

Haoran Li et al[9] Efficient information retrieval is challenged by the exponential rise of multimedia data. Text summaries are provided by Multi-Modal Summarization (MMS), which allows users to quickly understand the main points of multimedia content without having to go through lengthy documents or films. This work presents a novel method for creating text summaries using asynchronous text, images, audio, and video on a given subject. However, bridging the semantic gap between multiple modalities for MMS is a substantial difficulty due to the heterogeneous nature of multimedia data.

MAST, a new model for Multimodal Abstractive Text Summarization, is presented by Aman Khullar et al.[10]. It uses data from a multimodal movie that includes elements of text, audio, and video. A innovative extractive multi- objective optimization based methodology is proposed by Anubhav Jangra et al.[11] to generate a multi- modal summary that includes text, graphics, and videos.

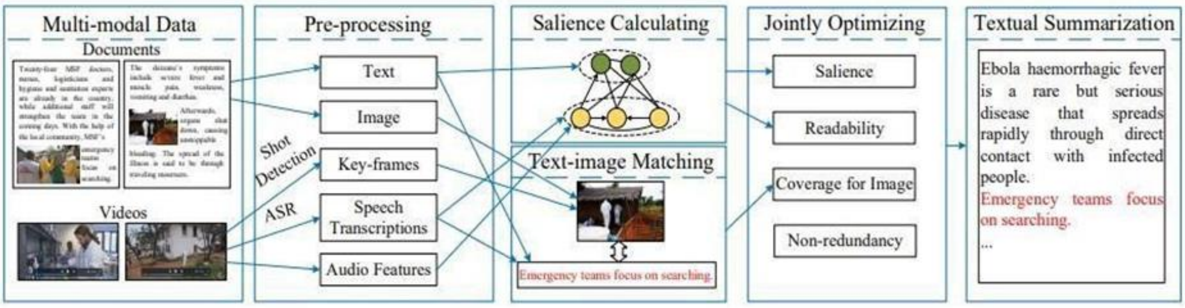


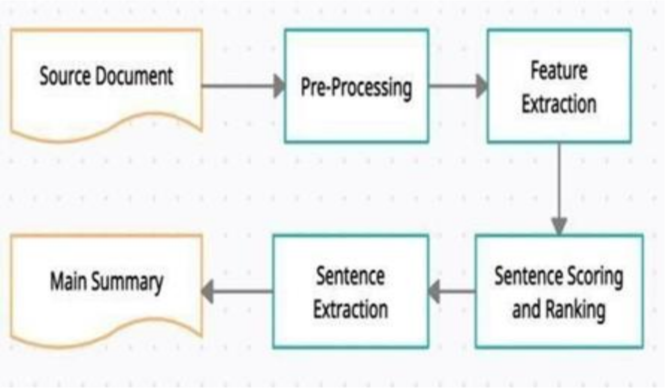
Figure 1: The framework of our MMS model.

2. RELATED WORK

Over time, image processing has attracted a large number of specialists from diverse fields. They have been working on the identification and categorization of different malignant illnesses, such as kidney and brain tumors, and they have suggested a number of cutting-edge methods to get the greatest outcomes. Abualigah L. and others.

[1] Text summarization is becoming increasingly important due to the amount of digital data available. It is crucial to find a quick and efficient method for summarizing lengthy texts without losing any of their important information. The main goal is to keep the important content of the text while condensing it into a more manageable format.

G. Kumar Vijay et al. 2 The primary objective of automatic text summarization is to extract pertinent and concise information from vast amounts of data. The amount of information available on the internet is vast and constantly expanding, making it difficult to gather the most important data from it quickly. Utilizing automatic text summarization facilitates users' ability to extract key information from vast amounts of data.



The method is proposed by Pratibha Devi Hosur et al. [3] and incorporates unsupervised learning into automatic text summarization. The overall picture of text summarization using natural language processing (NLP) is presented in this paper. It involves the following steps: input text document, preprocessing, lesk algorithm, and summary generation. The results, computations, conclusion, and suggested system of the lesk algorithm. Gaikwad, Deepali K. et al. [4] The study examines extractive and abstractive approaches to text summarization, describing the methods employed, how well they work, and the benefits and drawbacks of each approach. It highlights the importance of text summarization in commercial and research contexts. Compared to extractive methods, abstractive summarization produces more appropriate and meaningful summaries, but it is more complex because it requires learning and reasoning. The study draws attention to the paucity of research on abstractive approaches in Indian languages, suggesting significant prospects for additional investigation and better summarization strategies in this field.

Srinivas and colleagues [6] put forth a method for summarizing data using key frames. There are three steps in the process. 1) Determine which frame is the key frame; 2) Score the frames; and 3) Remove duplicate frames. A frame's score is determined by factors such as quality, representativeness, uniformity, and static and dynamic attention. Next, the score is normalized to fall between 0 and 1. The frames are given their weights in the following step. The arrangement of the weights gives the highest weight to the frames with higher scores. The dot product of the score and the weight is then used to determine the final score. The frames are placed in descending order of score, with the highest-scoring frame receiving the top spot. The frame with the highest ranking is designated as the key frame, and the difference between the key frame and the chosen frame is computed. It is chosen as the main frame if the distance is larger. Redundancy is eliminated using a different algorithm after key frame extraction. Images in grayscale are created from the chosen key frames. For those frames, the histograms are calculated, and the Euclidean distance between the frame pair is applied. The key frame is selected when the distance exceeds the threshold.

A systematic review and classification of video abstraction techniques was carried out by Truong et al. [7]. Their work concentrated on offering a methodical classification of approaches to video abstraction and summarizing currently used techniques. This paper is a useful resource for comprehending the landscape of video summarization techniques and is especially beneficial for those seeking an organized overview of the developments in the field.

An approach on real-time surveillance video with alarm facility was proposed by Deepika et al. [8]. After the live video is recorded, it is divided into frames and preprocessed using various methods, such as brightness and contraction. The image is compressed using JPEG image compression before being stored in the system.

Haoran Li et al. [9] use Automatic Speech Recognition (ASR) to obtain speech transcriptions for the audio data found in videos, then they devise a method to use these transcriptions selectively. Using a neural network to learn the joint representations of texts and images, we can identify the text that is relevant to an image for visual information, such as key frames extracted from videos and images found in documents. This allows for the integration of visual and aural elements into a text summary.

Aman Khullar et al.'s introduction of the audio modality for abstractive multimodal text summarization [10]. analyzing the difficulties in applying audio data and appreciating its role in the produced synopsis.The introduction of a brand- new, cutting-edge model called MAST for the multimodal abstractive text summarization task.

To address the TIVS problem, Anubhav Jangra et al. [11] suggest a differential evolution technique based on multi-objective optimization. The suggested method receives a topic containing several documents, photos, and videos as input and produces an extractive textual summary along with a few key photos and videos.

Sample papragraphDefine abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

3. PROPOSED METHOD

Text Summarization

A class of neural network architectures known as Sequence-to-Sequence (Seq2Seq) model is frequently employed for a variety of natural language processing applications, including text summarization. An encoder and a decoder are the two main parts of the Seq2Seq model. Text summarization involves the encoder processing the input text and the decoder producing the summary. This is a basic overview of text summarization using Seq2Seq:

1. Data preprocessing

The process of text summarization entails compiling an article dataset along with the summaries that go with it. Both the target summaries and the input text (articles) are formatted for the model after the text has been tokenized.

2. Building the Model

• Encoder:

- Word-by-word processing of the input sequence (article) results in the embedding of each word into a high-dimensional vector representation by the encoder. Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM):

- The sequential information of the input text is then captured by feeding these word embedding into an RNN or LSTM.
- Hidden State: The encoded data from the input sequence is represented by the encoder's final hidden state.

• Decoder:

Input Sequence (Summary): After receiving the encoded data from the encoder in its original hidden state, the decoder starts to produce the summary.

Conditional Probability: Based on the context and previously generated words, the decoder predicts the next word in the summary at each time step.

Attention Mechanism (Optional): To enhance performance, it is possible to incorporate an attention mechanism that enables the model to generate each word in the summary by focusing on distinct portions of the input text.

3. Training the Model:

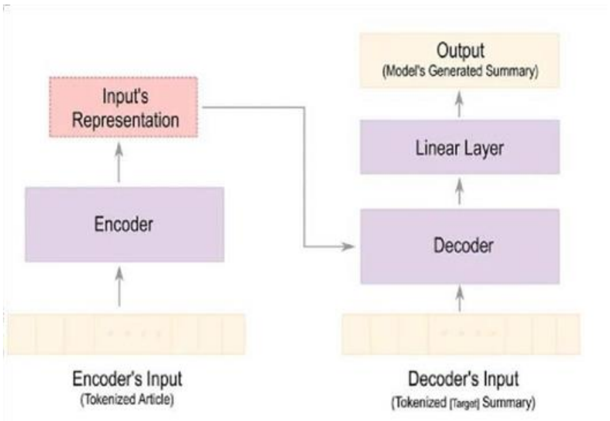
The decoder is trained to predict the next word in the summary at each time step by means of teacher forcing on a dataset of articles and the summaries that correspond to them.

- Using a loss function like cross-entropy, the training goal is to minimize the difference between the target summary and the predicted summary.

4. Inference :

In this stage, summaries of recently published articles are produced using the trained model.

A number of methods, including the use of transformer-based architectures (BERT, GPT, etc.) and reinforcement learning to improve the generated summaries, can be applied to improve Seq2Seq models for summarization.



Video Summarization

Sequence-to-sequence video summarization employs Long ShortTerm Memory (LSTM) networks, where an LSTM model analyzes individual video frames or segments and produces a succinct summary.

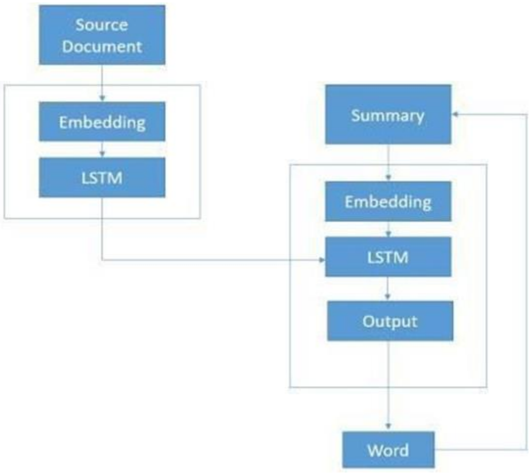
Here's a broad rundown of the process:

1. **Data Preparation**: Assemble a dataset of training videos. Segment and organize videos into frames. These frames or segments will be the input data for the LSTM model.
2. **Sequence Representation**: Convert the frames or segments into a format that is compatible with LSTMs. Presenting each frame as an image or feature sequence could be one way to achieve this. Sequence the segments or frames to produce the input sequences for the LSTM model. Each sequence would represent a segment of the movie.

3. **Model Architecture:** Create an LSTM based architecture. This usually involves an encoder-decoder setup, where the encoder LSTM processes the input sequences and the decoder LSTM uses the learned representations to produce the summary.

4. **Training:** Use the video sequences and their summaries to train the LSTM model. To produce the summary, the model is trained to encode and decode data from video segments.

5. **Evaluation:** Determine how well the generated summaries match the ground truth or human-generated summaries by evaluating the model's performance using metrics like Rouge Score, F1 Score, or human evaluation.



CONCLUSION

Our work presents a novel method for textual summarization from various sources, including audio, video, and text, which outperforms existing state-of-the-art models. Specifically, we aim to produce thorough text-based summaries without the need for explicit image alignment. The approach we suggest lays the groundwork for further research that focuses solely on textual content.

Our developed metrics are a useful tool for assessing the textual output of multimodal sources, even though our current methodology is flexible and based on differential evolution. We achieve the best results in textual summarization from various media sources by tackling an asynchronous Multimodal Summarization task. To put it briefly, our work establishes a method for textual summarization across a variety of sources and creates the foundation for thorough textual summaries. Our framework's flexibility and the useful assessment metrics provide a foundation for further development. The incorporation of a language translation tool will enhance the potential impact and capabilities of our research as it advances in our project.

REFERENCES

[1] Abualigah L., Bashabsheh M. Q., Alabool H., & Shehab M. (2020). Text Summarization: A Brief Review., In Recent Advances in NLP: The Case of Arabic Language (pp. 1-15).

[2] G. Vijay Kumar , Arvind Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, D. Samved Reddy (Year Unavailable).|| Text_Summarizing_Using_NLP||. In M. Rajesh et al. (Eds.), "Recent Trends in Intensive Computing." Publisher.

[3] Pratibha Devihosur, Naseer R. "Automatic Text Summarization Using Natural Language Processing" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 08, Aug-2017.

[4] Deepali K. Gaikwad, C. Namrata Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering||. Vol.5, Issue 3, March 2016.

[5] Aruna Kumara B, Smitha N S , Yashaswini Patil, Shilpa P , Sufiya, Text Summarization Using RankingAlgorithm||, International Journal of Computer Sciences and Engineering||. Vol.-7, Special Issue-14, May 2019.

[6] Srinivas, M., Pai, M. M., & Pai, R. M. (2016). An Improved Algorithm for Video Summarization–A Rank Based Approach. Procedia Computer Science, 89, 812-819.

[7] Truong, B. T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. ACM transactions on multimedia computing, communications, and applications (TOMM), 3(1), 3.

[8] Deepika, T., & Babu, D. P. S. (2007). Motion Detection In Real-Time Video Surveillance with Movement Frame Capture And Auto Record in International Journal of Innovative Research in Science. Engineering and Technology An ISO, 3297

[9] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, Chengqing Zong, et al. 2017. Multimodal summarization for asynchronous collection of text, image, audio and video.

[10] Aman Khullar , Udit Arora ,et al 2010.MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention.

[11] Anubhav Jangra , Sriparna Saha , Adam Jatowt , Mohammad Hasanuzzaman et al 2018.Multi-Modal Summary Generation using Multi-Objective Optimization.

[12] Patel, M., Chokshi, A., Vyas, S., &Maurya, K. (2018). "Machine Learning Approach for Automatic Text Summarization Using Neural Networks". International Journal of Advanced Research in Computer and Communication Engineering.

[13] Falaki, A. A. (2021, December 14). How to Train a Seq2Seq Text Summarization Model with Sample Code (Ft. Huggingface/PyTorch). Towards AI.