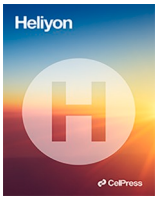
Heliyon 10 (2024) e26162

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Multimodal Abstractive Summarization using bidirectional encoder representations from transformers with attention mechanism☆

Dakshata Argade [a], Vaishali Khairnar [a,*], Deepali Vora [b], Shruti Patil [b,c], Ketan Kotecha [c], Sultan Alfarhood [d,**]

[a] Terna Engineering College, Nerul, Navi Mumbai, 400706, India
[b] Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, 412115, India
[c] Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis Institute of Technology Pune Campus, Symbiosis International (Deemed University) (SIU), Lavale, Pune, 412115, India
[d] Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O.Box 51178, Riyadh, 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Attention mechanism
Bidirectional encoder representations from transformer
Decoder
Encoder
Multimodalities
Multimodal abstractive summarization

ABSTRACT

In recent decades, abstractive text summarization using multimodal input has attracted many researchers due to the capability of gathering information from various sources to create a concise summary. However, the existing methodologies based on multimodal summarization provide only a summary for the short videos and poor results for the lengthy videos. To address the aforementioned issues, this research presented the Multimodal Abstractive Summarization using Bidirectional Encoder Representations from Transformers (MAS-BERT) with an attention mechanism. The purpose of the video summarization is to increase the speed of searching for a large collection of videos so that the users can quickly decide whether the video is relevant or not by reading the summary. Initially, the data is obtained from the publicly available How2 dataset and is encoded using the Bidirectional Gated Recurrent Unit (Bi-GRU) encoder and the Long Short Term Memory (LSTM) encoder. The textual data which is embedded in the embedding layer is encoded using a bidirectional GRU encoder and the features with audio and video data are encoded with LSTM encoder. After this, BERT based attention mechanism is used to combine the modalities and finally, the BI-GRU based decoder is used for summarizing the multimodalities. The results obtained through the experiments that show the proposed MAS-BERT has achieved a better result of 60.2 for Rouge-1 whereas, the existing Decoder-only Multimodal Transformer (D-MmT) and the Factorized Multimodal Transformer based Decoder Only Language model (FLO-RAL) has achieved 49.58 and 56.89 respectively. Our work facilitates users by providing better contextual information and user experience and would help video-sharing platforms for customer retention by allowing users to search for relevant videos by looking at its summary.

☆ This research is funded by the Researchers Supporting Project Number (RSPD2024R890), King Saud University, Riyadh, Saudi Arabia.
* Corresponding author.
** Corresponding author.
E-mail addresses: vaishalikhairnar@ternaengg.ac.in (V. Khairnar), sultanf@ksu.edu.sa (S. Alfarhood).

https://doi.org/10.1016/j.heliyon.2024.e26162
Received 8 August 2023; Received in revised form 28 January 2024; Accepted 8 February 2024
Available online 18 February 2024

1. Introduction

The tremendous use of social media and video sharing platforms creates huge amount of digital data. The high volume of data gets generated with very high velocity in the variety of forms i.e. structured, semi structured and unstructured. The most of the todays data are present in unstructured format i.e. images and videos which is coming from various domains like science, blogs, news, finance, medicine, academics etc. [1,2]. For example, Facebook, Instagram and YouTube has millions of users who posted billons of videos on daily basis. The quantity of user-generated instructional videos posted online has exponentially increased in recent years due to the growing popularity of video-sharing platforms. Due to the amount of videos available online, there is a growing need for the effective methods to find and retrieve relevant videos [3,4]. In order to meet these needs, numerous researchers have employed the concept of video summarization to generate brief summaries from the videos [5,6]. Video summarization is the process of extracting the important information from the input and presenting it into a concise, understandable summary [7,8]. The video summarization makes information extraction easier and efficient. By looking at generated summary user can quickly decide which video relevant or not. These summaries can be used further for auto captioning, annotation purpose which makes search and retrieval process easier. In general, there are two categories for summarization tasks: extractive and abstractive method [9,10]. In extractive summarization key information of input are extracted and combined together to form summary while in abstractive summarization, summary is generated by understanding semantic representation of input [11]. The abstractive method produced more informative summaries than the extractive method when the two approaches were compared. Although extraction-based summaries are syntactically correct, they suffer from coherence issues, ambiguity, and duplication [11–14]. In the case of multi-document summaries, the issue with extractive summary is the imbalance in topic coverage [13,14]. Because of the poor connection between sentences, the extracted summary does not appear to be logically related. Using natural language generation techniques and focusing on a semantic representation of the text, the abstractive approach can address the drawbacks of the extractive approach.

In the abstractive summarization method, the sequence-to-sequence approach has been widely utilized in the architectures of the encoders and decoders using Recurrent Neural Network (RNN). Together, the ability to recognize visual content and generate natural language are necessary for the development of textual summary from sequential videos. The process of converting a video to a text description has several uses in different domains, including automatic subtitles, video retrieval, and summaries [15]. In order to save reading time and maintain key information, abstractive summarizers are crucial [16]. The abstractive method of video summarizing picks out the most important frames from the video and uses pre-trained network architectures to analyze the contents of it [17]. But processing of such data with high dimensional features are challenging. Furthermore, to process the high dimensional data required large computational power as it creates many hidden matrices of input mapping [18]. The proposed research addresses these problems by combining tri-modal features (text, audio, and video) into a BERT-based attention model to summarize the concept in an abstractive multimodal way. In the area of NLP, word representations based on the Transformer like BERT, etc., have shown some incredibly promising outcomes. To improve the performance of summarization, we used BERT, which can pre-train language models on unlabeled data and is quick to train due to parallelization. Our system helps to speed up browsing through large video collections.

The main contributions of this research are mentioned as below:

1. Development of MAS-BERT with the attention mechanism to combine multimodalities such as text, audio and video.
2. The textual data embedding is encoded using a Bi-GRU encoder and features with audio and video data are encoded with the help of the LSTM encoder. The decoding is performed using the Bi-GRU to summarize the modalities as text.

The remaining sections of the manuscript are organized as follows: the related work of this research has been described in Section 2 and the proposed method is explained in Section 3. The results and analysis of this research are described in Section 4 of the manuscript and finally, Section 5 concludes the overall research work.

2. Related works

This section describes some of the related works which are based on multimodal summarization.

Huang et al. [19] have introduced the Hierarchical Multimodal Network (HMNet) to summarize and predict the concepts which are related to educational videos through multimodal data and the analyzing dependency of classes. Initially, a video divider was used for extraction of the essential frames from the video and this sequential video that was categorized into sections with subtitles. After this stage, a multi-modality encoder was used to get the incorporated presentation of the multi-modality. At last, a hierarchical predictor chain was used to predict and summarize the concepts present in the video in a textual manner. The hierarchical predictor chain was highly capable of tackling complex arithmetic tasks, but the computation cost was high at the time of the training and inference process.

Yuan et al. [20] have introduced the Multi-Layer cross-fusion with a Re-constructor (MCR) to create a textual summary from the multimodal video collection. The MCR performs cross-fusion through the layer blocks of cross-model transformers and it results in a cross-modal representation. The feature level re-constructor was used in the process of constraining the complex semantics of every individual modality. Moreover, the constraint separation strategy was utilized in the process of reconstruction with different MCR modalities. However, the specific information loss may be there, as the dataset only consists of the sequences of pre-trained features to imitate the content of the video.

Li et al. [21] have introduced an Inter and Intra modal Contrastive Hybrid (ITCH) framework that uses the automatic alignment of the multimodal information and summarizes it accordingly. ITCH obtains the bi-modal input as text and image to present it in a

patch-oriented encoder and textual encoder to extract the features. Moreover, the module for the fusion of cross-modality was utilized to summarize the semantic features. After this stage, the textual decoder was used in the process of creating the target summary and finally, the ITCH framework introduced an auxiliary objective, which was used in the process of summarizing the multiple modalities. However, the ITCH framework finds difficulty in summarizing the complicated vocabularies.

Liu et al. [22] have introduced a Decoder-only Multimodal Transformer (D-MmT), which is modified from the structure of the decoder by including the in-out multimodal decoder. Moreover, Cascaded Cross-Modal interaction (CXMI) creates the joint fusion representation among the modalities. The suggested D-MmT actively participates in the back-propagation to improve the contextual representation. Additionally, joint in-out loss is used to create summaries for the long transcript text input. However, the suggested D-MmT does not suit multimodal input data.

Atri et al. [23] have introduced a Factorized Multimodal Transformer based Decoder Only Language model (FLORAL) that internments dynamics of inter and intra models with different inputs to provide a summarization of text values. FLORAL uses more self-attention layers, providing better results than traditional encoders and decoders. Moreover, the dataset named AVIATE was introduced to evaluate the performance of abstractive text summarization. The FLORAL approach effectively summarizes the text from the longer videos. However, in some occasions, the suggested approach results in minimized fluency due to automatic speech recognition and optical character recognition.

Dehouche et al. [24] have presented the Text to Image model using the Latent Diffusion technique of deep learning to produce images from the textual description. The image description in the text form is received as input, which is further tokenized using a BERT tokenizer. The tokens are further used to extract topics based on which corresponding images get created.

Palaskar et al. [25] have introduced a multimodal summarization model that creates textual summaries by taking video and audio transcripts as input. The model used the seq-to-seq model and the hierarchical attention to combine information from the different modalities. The new metric for evaluating the performance of the model has been introduced in this work to measure the semantic adequacy of the summarization task.

Khullar et al. [26] presented the MAST model to generate a summary from texts, audio and video modalities. Audio modeling is newly employed by this model. The seq-seq-based model generates summaries while the layer of trimodal hierarchical attention is used to fuse information of all the modalities. They faced challenges for better representation of the audio modality.

Yin et al. [27] proposed the MASCA model, which used text and pictures as input modeling to generate a summary. They provide attention to the information contained in original data by introducing core words fusion attention. They used BART for Text and RESNETl01 for picture modality, respectively to generate the model representation, which was further decoded to obtain a summary. At last, core words and semantic information are combined to form the final summary.

Chen and Zhuge [28] used the Encoder-Decoder model with hierarchical attention to create a summary of the text documents containing images. They extended the Daily mail dataset by gathering its images and corresponding captions from the web. They have not considered other modalities such as video and audio.

Li et al. [29] presented a model to generate a summary from audio and video recordings of the meetings. They joined the topic segments and summarized together to find topics relevant to the data. The word utterance is newly considered in his work.

Mohan [30] et al. proposed a new method to remove redundancy from input video before summarizing video. The domain-independent method is used to reduce the repetition of data. The input video frames are uniformly sampled and then the displacement magnitude of sequential frames is calculated. The frames having a magnitude below the threshold are eliminated.

Overall, the existing approaches faced complexity in the summarization of text due to the longer duration of the time, which hardly consider the multiple modalities and summarization of the complex vocabularies. Although the earlier work for abstractive multimodal text summarization showed promising results, they did not use a Transformer to summarization task to enhance the performance. The audio modality was hardly taken into account in the previous work while creating the summary. Our work addresses this inadequacy by using BERT as a Transformer model with attention, which improves the performance of the summarization task. The Audio modality significance is considered in the generated summary. These existing drawbacks were overwhelmed using the proposed approach by considering the multimodal input values from the How2 dataset.

3. Multimodal abstractive summarization using BERT attention

The multimodal abstractive summarization using the BERT attention model is a promising approach that combines the video, audio and text features of the input to produce a summarized output by extracting the important concept from the input data. The proposed MAS uses BERT, which can effectively investigate the complexities and works efficiently to summarize the concept. Fig. 1 presents the framework of the proposed methodology in the summarization of the concepts.

3.1. Research methodology

The proposed methodology makes use of the existing How2 dataset, which is a multimodal collection of various videos with english subtitles. The text, audio and video data are taken as input. The text data is further preprocessed by removing extra white spaces and stop words, expanding contractions, and lowercasing all the text. After that word embedding is applied to the processed input, embedding groups together semantically the same input. The data is embedded using an embedded layer using deep learning models like LSTM and GRU. The BERT attention mechanism is used while generating a context vector to fuse all the modalities. The context vector is further provided to the decoder to generate the final summary of the input. The proposed model obtained results were compared with existing models (see Fig. 2).

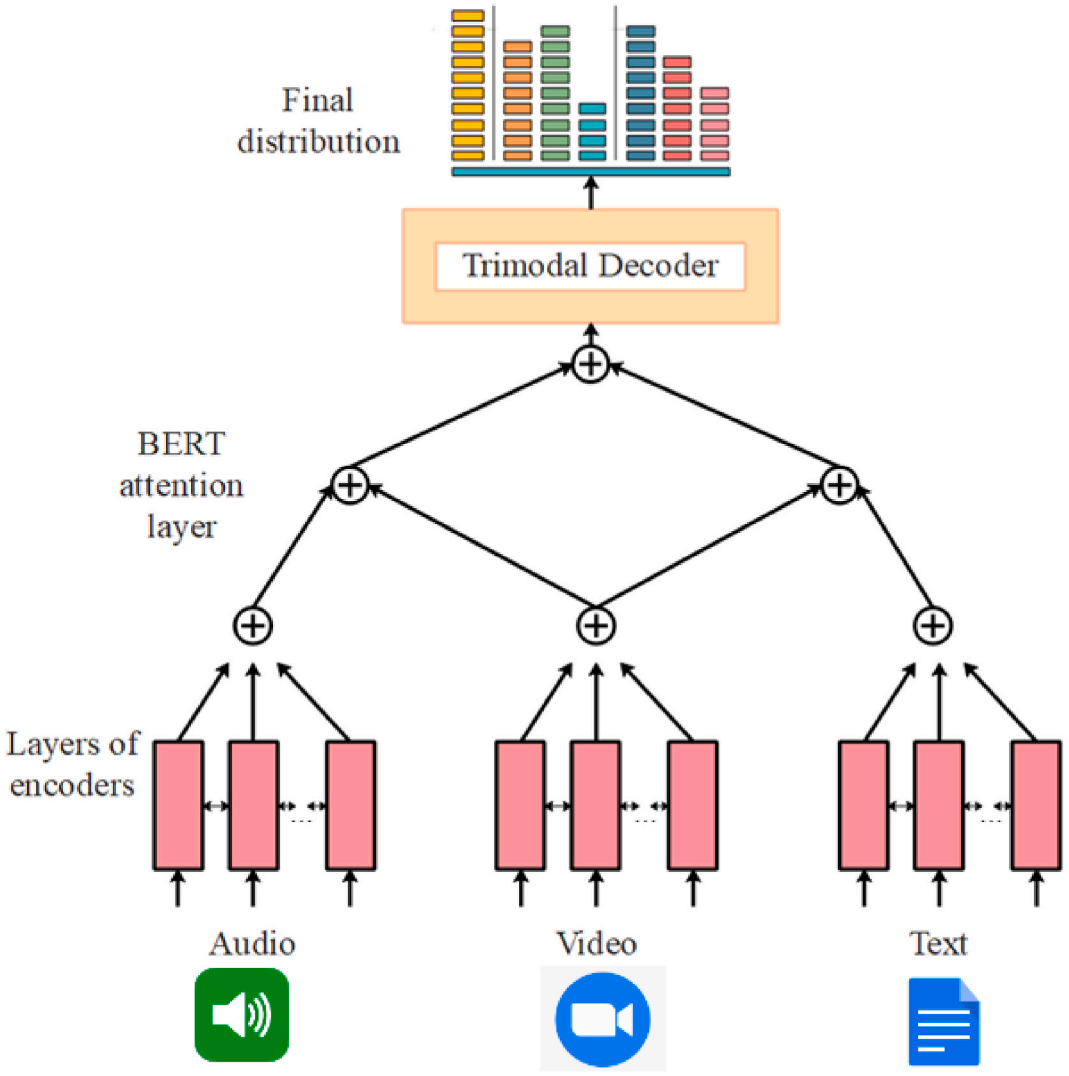


Fig. 1. The overall architecture of the multimodal abstractive summarization using BERT attention: The MAS-BERT model takes Text, Audio and Video as input. The textual data which is embedded in the embedding layer is encoded with the help of a bidirectional GRU encoder and features with audio and video data are encoded with the help of the LSTM encoder. After this, BERT based attention mechanism is used to combine the modalities and finally, BI-GRU based decoder is used in the process of summarizing the multimodalities.

Fig. 2. Research methodology flow diagram.

3.2. Dataset

How2 dataset is a multimodal collection of various videos with English subtitles. It consists of around 80,000 instructional videos with two versions such as 2000 h and 300 h. Both versions can be used in speech recognition, summarizations and multimodal extensions. The proposed work utilizes 300 h version of the How2 database as the modality based on audio is only present in the 300 h version. The dataset is broken down into three parts i.e. training, testing and validation set. There are 13,168 videos of 298.2 h used for training, 150 videos with 3.2 h duration are used for the validation and for testing 175 videos with 3.7 h duration are utilized. The experimental results are evaluated by considering the 12,798 videos for the training, 520 videos for the validation and 127 videos for the testing. The How2 dataset's brief statistics are displayed in Table 1.

Table 1
Dataset information [31].

Dataset	No of videos	Duration (Hrs)	Content	Clips/Sentences Per Clip Statistics	Clips/Sentences Per Clip Statistics
300 h	Train-13,168	298.2	Open Domain	184, 949	5.8 s & 20 words
	Validation -150	3.2		2022	
	Test- 175	3.7		2305	
2000 h	Train-73,993	1766.6	Open Domain	–	–
	Validation -2965	71.3			
	Test- 2156	51.7			

3.3. Modality encoders

After the stage of the data acquisition from the How2 dataset, the obtained data is encoded using a Bi-GRU encoder and LSTM encoder. The textual data which is embedded in the embedding layer is encoded with the help of a bidirectional GRU encoder and features with audio and video data are encoded with the help of the LSTM encoder.

3.3.1. Bi-GRU neural network

Bi-GRU is a kind of GRU neural network that comprises of two-layered architectures and these layers help to provide contextual information of the obtained input data. It works based on an input sequence, which is distributed through the forward and backward neural network. The forward layer is used to assess each hidden layer's output from front to back, and the backward layer assesses the result of the hidden layer from back to front. Finally, the output layer of the Bi-GRU covers up and standardizes the results of forward and backward layers at each time. The input feature related to *i* th utterance in step *t* gets encoded in two vectors, which is represented in Equation (1).

$$\vec{h}_{i,t} = \vec{G}\,RU\left(X_{i,t}\right), t \in [1, T] \tag{1}$$

where $h_i = \overrightarrow{h_{i,t}} + \overleftarrow{h_{i,t}}$, the feature vector is denoted by X and the terminated time step of each utterance is represented as T. As Bi-GRU is used in the process of embedding the textual information, the dimensionalities of the features are minimized by summing the vectors obtained from forward and backward GRUs.

3.3.2. Long short term memory (LSTM)

The LSTM is an advanced version of a Recurrent Neural Network that has the capability to deal with problems related to long-term dependencies. The LSTM is comprised of the repeated modules in every time step. Fig. 3 shows the architectural diagram of LSTM.

There are three types of gates such as input, output and forget gates which are responsible for regulating each step of the module. The architecture of LSTM was built based on these cell gates, which update the memory of the current cell and the hidden state. The transition function of LSTM is represented in Equations (2)–(6).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2}$$

$$q_t = \tanh\left(W_q.[h_{t-1}, x_t] + b_q\right) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * q_t \tag{4}$$

$$o_t = \sigma(W_0.\,[h_{t-1}, x_t] + b_0) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

where the prior hidden state is denoted as h_{t-1}, the forget gate is represented as f_t, the input gate as i_t and the output gate as o_t. The

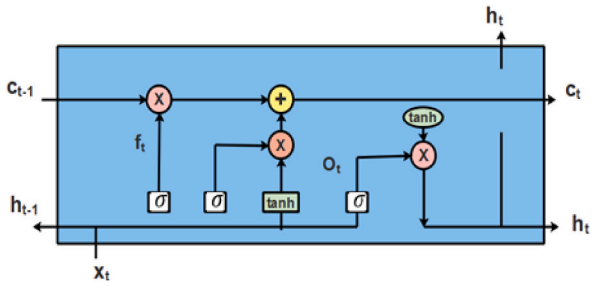


Fig. 3. Architectural diagram of LSTM.

current step time, current hidden state and the current memory cell are represented as x_t, h_t and C_t respectively. The sigmoid function is denoted as σ, which lies in the range of [0,1] and the hyperbolic tangent functionality, which lies in the output range of [-1,1]. The LSTM is well suited for sequential data like videos and images and explicitly high-level features. Thus, the Bi-GRU neural network and LSTM are effectively utilized in the process of modality encoding. The Bi-GRU is one of the active approaches to encode the textual data and the LSTM is used in the process of encoding the features of audio and video.

3.4. BERT-based attention layer

The proposed work combines the various modalities by implementing Bidirectional Encoder Representations from Transformers (BERT) with an attention method. The BERT's attention mechanism operates using the vectors Query (Q), Key (K), and Value (V). These vectors generate a linear transformation while producing weights for different connections and feeds that information into the scaling dot product. Based on the definition of the attention mechanism, the query itself acts as the key, the dimension of the query and key is denoted as d_k. The dot product scaling helps to prohibit the faster growth of the dot product, in case it is not considered, it diminishes the gradient of the softmax function. The feasibility of the attention layer can be improved by evaluating the multi-head attention values in a parallel manner. By assigning the group of randomized parameters to Q, K and V; the value of each head is updated as X_i^Q for query, X_i^K for key and X_i^V for value. The equation of the attention map and the trained head of each attention map is represented in Equations (7) and (8).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

$$Head_i = Attention\left(QX_i^Q, KX_i^K, VX_i^V\right) \tag{8}$$

where the attention of head is denoted as a $Head_i$, the query, key and value of BERT are denoted as Q, K and V. This research make use of multimodal summarization with two-level BERT attention and the context vectors evaluated based on Equations (9)–(11).

$$e_i^{(k)} = v_b^T \tanh\left(W_b s_i + U_b^{(k)} c_i^{(k)}\right) \tag{9}$$

$$\eta_i^{(k)} = softmax\left(e_i^{(k)}\right) \tag{10}$$

$$c_i = \sum_{k\in\{audio,text,videos\}} \eta_i^{(k)} U_c^{(k)} c_i^{(k)} \tag{11}$$

where the distribution of BERT attention in the modalities is denoted as $\eta_i^{(k)}$, the vector related to the k-th modality of the contexts is denoted as $c_i^{(k)}$. The parameters that were shared among the modalities are denoted as W_b and the modalities of the specified projections are represented as $U_c^{(k)}$, $U_b^{(k)}$.

The sum of the attention layer at every individual participant presents the attention of the token and the attention is provided based on the context and the final decision. The BERT performance depends on the tokens and words present in the context, which get broken into multiple samples. The attention of each word is evaluated using summing the tokens of the word based on Equation (12).

$$attention\ W_{word} = \sum_{token\in word} attention\ W_{token} \tag{12}$$

BERT model needs to be fine-tuned to provide a quality summarization of the text and it can train more sentence pairs at a time. It is comprised of a built-in tokenizer which helps to transform the contextual pairs into tokens and feed them into BERT. The authors have considered top words with higher attention values to specify them as the significant words involved in the process of abstractive summarization. The maximum number of tokens present in the text describes the total count of neurons in the text. The output of the attention layer is combined with the classification layer by using the ReLU activation function to create a relationship between the tokens and the output layers. The attention weight of every individual token is represented as *attention* W_i, which is described using the softmax layer at the output of the attention layer as shown in Equation (13).

$$attentionW_i = \frac{\exp\ (o_i)}{\sum_K \exp\ (o_j)} \tag{13}$$

The architecture of the BERT attention model is generated based on the attention weights of the tokens. From the sequence of attention weights, the token with the highest attention value is considered. When the BERT tokenizer breaks the words into tokens, all the tokens are allotted with varying weights and threshold value of θ, which is evaluated using Equation (14).

$$\theta = p((attentionW_{max} - attentionW_i), n) \tag{14}$$

Multimodal abstractive summarization using BERT attention considers the contextual vectors of textual content based on audio and video. The BERT attention mechanism is utilized in the process of combining audio and video through the second attention layer and

the third BERT attention layer is used to combine contextual vectors. The process of combining the text obtained from audio and video is represented in Equation (15) and Equation (16).

For text obtained from audio,

$$\left.\begin{aligned} e_i^{(k)} &= v_d^T \tanh\left(W_d s_i + U_d^{(k)} c_i^{(k)}\right) \\ \beta_i^{(k)} &= softmax\left(e_i^{(k)}\right) \\ d_{i(1)} &= \sum_{k\in\{audio,text\}} \beta_i^{(k)} U_e^{(k)} c_i^{(k)} \end{aligned}\right\} \tag{15}$$

For text obtained from video,

$$\left.\begin{aligned} e_i^{(k)} &= v_f^T \tanh\left(W_f s_i + U_f^{(k)} c_i^{(k)}\right) \\ \gamma_i^{(k)} &= softmax\left(e_i^{(k)}\right) \\ d_{i(2)} &= \sum_{k\in\{video,text\}} \gamma_i^{(k)} U_e^{(k)} c_i^{(k)} \end{aligned}\right\} \tag{16}$$

where the contextual vector, which is gathered from the combination of modalities is denoted as $d_{i(.)}$, the variable β and γ is the parameter related to the BERT attention mechanism. At last, the text obtained from audio and video is combined using the final attention layer, which is denoted as δ. By using the BERT attention mechanism, the textual modalities are combined with the audio and video modalities in a pair-wise method. Moreover, this strategy helps the model to pay the attention to the text-based modalities at the time of integrating advantageous things from the remaining two modalities, which is presented in Equation (17).

$$\left.\begin{aligned} e_i^{(l)} &= v_h^T \tanh\left(W_g s_i + U_h^{(l)} d_i^{(l)}\right) \\ \delta_i^{(l)} &= softmax\left(e_i^{(l)}\right) \\ C_i^f &= \sum_{l\in\{audio\ \ text,video\ \ text\}} \delta_i^{(l)} U_m^{(l)} d_i^{(l)} \end{aligned}\right\} \tag{17}$$

where the finalized contextual vector at the step of ith decoder is denoted as C_i^f. After obtaining the final text, the decoder is used to create a list of distributed vocabularies at each time step.

3.5. Multimodal decoding

This is the final stage of the multimodal summarization, where the encoded texts are distributed to create the finalized vocabulary at each step time. A conditional GRU (CGRU) decoder is used to decode the texts based on audio and videos, which is obtained from the encoded output. Here, the decoding is performed with two layers included with the attention mechanism. The primary layer of the CGRU is transitional, which is based on the hidden state of the recurrent network and the CGRU's second layer is determined as the primary CGRU layer. Moreover, the second CGRU is evaluated by the hidden states, which are considered as the prior hidden state of the next time step of the CGRU. The contextual text C_t is provided to the attention layer of CGRU as the input and disseminated at each step. At each discrete time step, cumulative data from modalities is obtained by the decoder. The goal of the CGRU decoder is to combine modalities according to the information provided by each modality. Finally, the data obtained from the previous time step is processed through those two linear CGRU layers to create the vocabularies from the next word.

4. Results and analysis

This section presents detailed results and analysis of the proposed MAS-BERT model while comparing it with various approaches based on tri-modal. The proposed MAS-BERT is evaluated using standard performance metrics such as ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL). The metrics such as R-1, R-2 and R-L evaluate the values related to unigram, bi-gram and long sequence among the ground truth values and created summaries. Moreover, an extensive performance analysis is computed based on the human understandability of the results based on its fluency and summary. The results part is divided into two subsections: performance analysis and comparative analysis. In performance analysis, the three modalities are used to measure efficiency. In comparison, the efficiency of the MAS-BERT model is compared with its existing methodologies described in related works. The proposed MAS-BERT approach is evaluated and implemented in the Python software and the system with specifications of intel i7 processor, 8 GB random access memory and windows 10 operating system.

4.1. Performance analysis

In this sub-section, the performance of the proposed approach is assessed based on different modalities. Table 2 presents the results

Table 2
Performance analysis based on different modalities.

Modalities	Models	ROUGE-1	ROUGE-2	ROUGE-L
Tri modal	TrimodalH2	47.85	28.46	42.17
	MAST-Binned	46.22	25.94	40.34
	MAST	48.85	29.51	43.23
	MAS-BERT	**60.02**	**30.30**	**59.9**

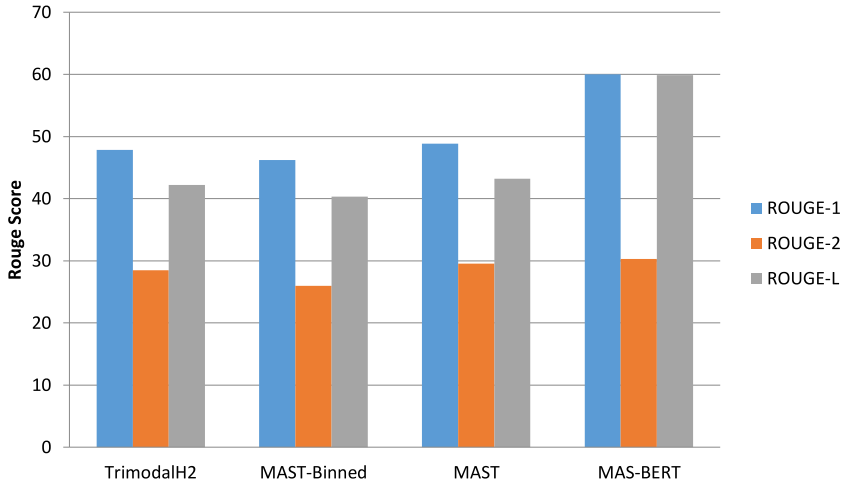


Fig. 4. Graphical representation for performance evaluation of tri-modals.

Table 3
Performance analysis with different methodologies while summarizing three modalities of text, audio and video.

Modality Models	How2		
	R-1	R-2	R-L
Multimodal HA	55.87	26.32	54.9
Multimodal Systems based encoder and decoder	55.89	26.79	55.1
Factorized Multimodal Transformer based encoder and decoder	55.98	26.83	55.4
Multi-Language model	56.13	26.89	55.41
MAS-BERT	**60.02**	**30.30**	**59.9**

of the proposed methodology.

According to findings of Table 2, the proposed approach has produced better results in overall metrics like R-1, R-2, and R-L. For instance, the result obtained by the proposed MAS-BERT model while summarizing the three modalities such as text, audio and video for R-1 metric is 60.02, whereas the existing Multimodal Abstractive Summarization of Tri-modal (MAST), Binned MAST and Tri-modal H2 is 48.85, 46.22 and 47.85 respectively. The increased rouge score of the proposed method helps to enhance the summarization capability and helps to achieve better results. Fig. 4 shows a graphical representation of performance evaluation of the tri-modal system's.

Secondly, the results are evaluated based on various methodologies used in tri-modal summarization.

The results obtained from Table 3 are evaluated for various tri-modal methodologies that used How2 dataset to acquire the input data. Moreover, the results are evaluated using standard metrics such as R1, R2 and R-L. The rouge score is calculated as per Equation (18).

$$Rouge_N = \frac{\sum S \in \{\text{Reference Summaries}\} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum S \in \{\text{Reference Summaries}\} \sum_{gram_n \in S} Count(gram_n)} \tag{18}$$

where {Reference Summaries} indicates the reference summaries, $Count_{match}(gram_n)$ represents number of n-grams matched in the reference summary and the generated summary, and $Count(gram_n)$ denotes the number of n-grams in the reference summary.

The findings of Table 3 show that the proposed MAS-BERT model has attained a better result of 60.02 for the R-1 metric whereas the existing multimodal HA, MulT En-De, FMT En-De and MulT-LM have achieved 55.87, 55.89, 55.98 and 56.13 for R-1 metric. Fig. 5 shows the graphical representation of various modalities vs summarization values.

Finally, the results are computed by considering the results obtained based on Informative (INF), Relevance (REL), Coherence

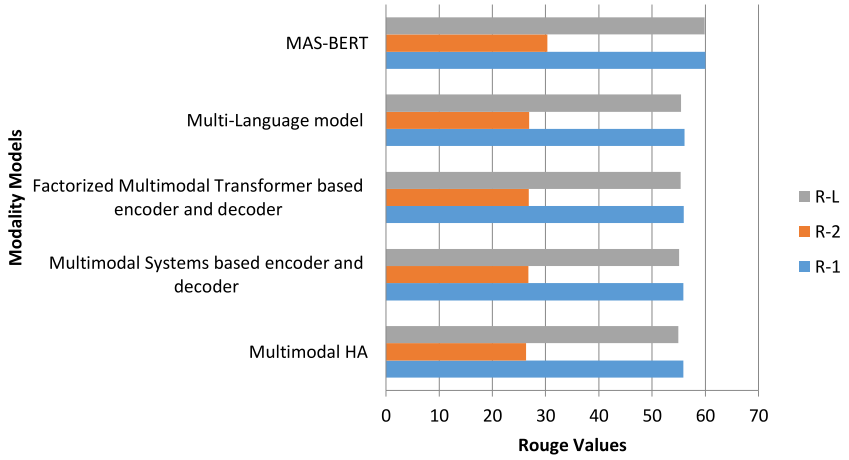


Fig. 5. Graphical representation of various modalities vs. summarization values.

Table 4
Performance evaluation based on human evaluation.

Modality	Model	INF	REL	COH	FLU
Extractive systems	KLSumm	2.82	2.54	2.98	3.14
(Text only)	Text Rank	2.92	2.73	2.82	3.12
Abstractive systems	PG	3.45	3.17	3.12	3.32
(Text only)	CopyTrans.	3.46	3.18	3.18	3.36
Abstractive systems (Video only)	Action Ft.+RNN	3.54	3.2	3.21	3.4
Multimodal systems	FMT En-De	3.61	3.39	3.37	3.67
(Video + Audio + Text)	MulT LM	3.57	3.38	3.34	3.68
MAS-BERT (Video + Audio + Text)	BERT attention	**4.02**	**3.88**	**3.87**	**3.83**

(COH) and Fluency (FLU) for various system modalities such as extractive-based, abstractive and multimodal-based. Coherence is the ability to have seamless transition between different summary statements so that no two sentences are entirely the same or unrelated. Consistency in the summaries is factual accuracy compared to the source document. Fluency deals with grammatical accuracy and sentence readability. Relevance refers to a summary's capacity to draw significant details from the source text. Table 4 presented the obtained results based on human-evaluated metrics.

The results obtained from Table 4 show that the MAS-BERT has achieved better values in all metrics such as informative, relevance, coherence and fluency. The BERT attention model has achieved a better fluency of 3.83 which is comparatively higher than the existing models. Moreover, improvement in the performance is due to the BERT attention layer which effectively enhances the overall efficiency of the summarization model. The BERT attention is updated frequently and the obtained metrics can be fine-tuned and utilized.

4.2. *Comparative analysis*

In this section, the efficiency of the proposed MAS-BERT model is assessed with the existing methodologies such as D-MmT [22] and FLORAL [23] based on performance metrics such as R-1, R-2 and R-L. Table 5 presented the comparative results of the proposed method compared to the existing techniques, including D-MmT and FLORAL.

The results from Table 5 show that the proposed MAS-BERT acquired comparatively higher results when compared with D-MmT and FLORAL. For the metric Rouge -1, MAS-BERT has achieved the value of 60.2 whereas the existing D-MmT and FLORAL have achieved 49.58 and 56.89 respectively. Moreover, the better result in multimodal summarization is due to the presence of the BERT attention layer which generates a linear transformation while producing the weights for different connections and providing them to the decoder.

Table 5
Comparison table to evaluate the performance of different modalities.

Methodologies	Rouge-1	Rouge-2	Rouge-L
D-MmT [22]	49.58	30.3	44.56
FLORAL [23]	56.89	26.93	56.8
MAS-BERT	**60.02**	**30.30**	**59.9**

5. Conclusion

This research symbolized the effectiveness of the multimodal abstractive text summarization using the MAS-BERT. Although earlier work for abstractive multimodal text summarization showed promising results, they did not use transformer to summarization tasks for enhancing the performance. Our work addresses this inadequacy by using BERT as a transformer model with attention for summarization task. To create a summary, the authors have taken into account, the importance of the audio modality, which has rarely been done before. The proposed approach outperforms well in the overall metrics when compared with the existing approaches such as D-MmT and FLORAL. The MAS-BERT obtained R-1 of 60.02 whereas D-MmT and FLORAL obtained 49.58 and 56.89 respectively. The increased rouge score of the proposed method helps to enhance the summarization capability and helps to achieve better results.

In the future, a better pre-trained model can be used to improve the process of multimodal summarization. In the future, we want to extend this work to adapt unsupervised deep learning methods to train models well without any requirement of large amounts of the ground truth data that have been annotated by humans. The standard performance metric needs to be evolved, which evaluates summary based on the original data.

Data availability

Data will be made available on the request.

CRediT authorship contribution statement

Dakshata Argade: Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Vaishali Khairnar:** Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Deepali Vora:** Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Shruti Patil:** Writing – review & editing. **Ketan Kotecha:** Writing – review & editing, Project administration, Methodology, Conceptualization. **Sultan Alfarhood:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr. Vaishali Khairnar reports financial support was provided by King Saud University, Riyadh, Saudi Arabia.

Acknowledgements

The authors extend their appreciation to the King Saud University for funding this research through Researchers Supporting Project Number (RSPD2024R890), King Saud University, Riyadh, Saudi Arabia.

References

[1] Anubhav Jangra, Sourajit Mukherjee, Jatowt Adam, Sriparna Saha, Mohammad Hasanuzzaman, A survey on multimodal summarization, ACM Comput. Surv. 55 (13s) (2023) 1–36.
[2] Blaž Škrlj, Matej Bevec, Nada Lavrač, Multimodal AutoML via representation evolution, Mach. Learn. Knowl. Extract. 5 (1) (2023) 1–13.
[3] Theodoros Psallidas, Panagiotis Koromilas, Theodoros Giannakopoulos, Evaggelos Spyrou, Multimodal summarization of user-generated videos, Appl. Sci. 11 (11) (2021) 5260.
[4] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, Bouridane Ahmed, Azeddine Baghdadi, A combined multiple action recognition and summarization for surveillance video sequences, Appl. Intell. 51 (2021) 690–712.
[5] Chengzhe Yuan, Zhifeng Bao, Mark Sanderson, Yong Tang, Incorporating word attention with convolutional neural networks for abstractive summarization, World Wide Web 23 (2020) 267–287.
[6] P. Kadam, et al., Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms, vol. 10, IEEE Access, 2022, pp. 122762–122785, https://doi.org/10.1109/ACCESS.2022.3223379.
[7] Liqiang Jing, Yiren Li, Junhao Xu, Yongcan Yu, Pei Shen, Xuemeng Song, Vision enhanced generative pre-trained Language model for multimodal sentence summarization, Mach. Intellig. Res. (2023) 1–10.
[8] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, Bouridane Ahmed, Azeddine Baghdadi, A combined multiple action recognition and summarization for surveillance video sequences, Appl. Intell. 51 (2021) 690–712.
[9] A. Dilawari, M.U.G. Khan, ASoVS: abstractive summarization of video sequences, IEEE Access 7 (2019) 29253–29263.
[10] S.P. Patil, L. Chavan, J. Mukane, D.R. Vora, V. Chitre, State-of-the-art approach to e-learning with cutting edge NLP transformers: Implementing text summarization, question and distractor generation, question answering, Int. J. Adv. Comput. Sci. Appl. (2022).
[11] Ming-Hsiang Su, Chung-Hsien Wu, Hao-Tse Cheng, A two-stage transformer-based approach for variable-length abstractive summarization, IEEE/ACM Trans. Audio Speech Language Process. 28 (2020) 2061–2072.
[12] Yangbin Chen, Yun Ma, Xudong Mao, Qing Li, Multi-task learning for abstractive and extractive summarization, Data Sci. Eng. 4 (2019) 14–23.
[13] A. Qaroush, I.A. Farha, W. Ghanem, M. Washaha, E. Maali, An efficient single document Arabic text summarization using a combination of statistical and semantic features, J. King Saud Univ.-Comp. Inform. Sci. 33 (6) (2021) 677–692.
[14] D. Argade, V. Khairnar, in: J.S. Raj, Y. Shi, D. Pelusi, V.E. Balas (Eds.), Intelligent Sustainable Systems. Lecture Notes in Networks and Systems, 458, Springer, Singapore, 2022. https://doi.org/10.1007/978-981-19-2894-9_39.
[15] Laura Perez-Beltrachini, Mirella Lapata, Multi-document summarization with determinantal point process attention, J. Artif. Intell. Res. 71 (2021) 371–399.
[16] Mateusz Krubiński, Pecina Pavel, MLASK: multimodal summarization of video-based news articles, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 880–894.
[17] TamiresTessarolli Barbieri de Souza, Rudinei Goularte, Content selection criteria for news multi-video summarization based on human strategies, Int. J. Digit. Libr. 22 (2021) 1–14.

[18] Bin Zhao, Xuelong Li, Xiaoqiang Lu, TTH-RNN: tensor-train hierarchical recurrent neural network for video summarization, IEEE Trans. Ind. Electron. 68 (4) (2020) 3629–3637.
[19] Wei Huang, Xiao Tong, Liu Qi, Zhenya Huang, Jianhui Ma, Enhong Chen, HMNet: a hierarchical Multimodal network for educational video concept prediction, Int. J. Mach. Learning Cybern. (2023) 1–12.
[20] Jingshu Yuan, Jing Yun, Bofei Zheng, Lei Jiao, Limin Liu, MCR: multilayer cross-fusion with reconstructor for multimodal abstractive summarization, IET Comput. Vis. (2023).
[21] Jiangfeng Li, Zijian Zhang, Bowen Wang, Qinpei Zhao, Chenxi Zhang, Inter-and intra-modal contrastive Hybrid learning framework for multimodal abstractive summarization, Entropy 24 (6) (2022) 764.
[22] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, Guangluan Xu, D-MmT: a concise decoder-only Multimodal transformer for abstractive summarization in videos, Neurocomputing 456 (2021) 179–189.
[23] Yash Kumar Atri, ShramanPranick, Vikram Goyal, Tanmoy Chakraborty, See, hear, read: leveraging multimodality with guided attention for abstractive text summarization, Knowl. Base Syst. 227 (2021) 107152.
[24] Nassim Dehouche, Dehouche Kullathida, What's in a text-to-image prompt? The potential of stable diffusion in visual arts education, Heliyon (2023) e16757.
[25] Shruti Palaskar, Jindrich Libovický, Spandana Gella, Florian Metze, Multimodal Abstractive Summarization for How2 Videos, 2019 arXiv preprint arXiv: 1906.07901.
[26] Aman Khullar, Udit Arora, Mast: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention, 2020 arXiv preprint arXiv:2010.08021.
[27] X. Yin, L. Sun, J. Wu, Y. Gao, X. Wu, L. Qiu, MASCA: a multimodal abstractive summarization model based on core words fusion attention, in: 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), Guilin, China, 2022, pp. 455–461, https://doi.org/10.1109/DSC55868.2022.00069.
[28] Jingqiang Chen, Hai Zhuge, Abstractive text-image summarization using multimodal attentional hierarchical rnn, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4046–4056.
[29] Manling Li, Lingyu Zhang, Heng Ji, Richard J. Radke, Keep meeting summaries on topic: abstractive multimodal meeting summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2190–2196.
[30] Jesna Mohan, Madhu S. Nair, Domain independent redundancy elimination based on flow vectors for static video summarization, Heliyon 5 (10) (2019) e02699.
[31] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, Florian Metze, How2: a Large-Scale Dataset for Multimodal Language Understanding, 2018 arXiv preprint arXiv:1811.00347.