



# Topic-guided abstractive multimodal summarization with multimodal output

Shaik Rafi<sup>1</sup> · Ranjita Das<sup>2</sup>

Received: 6 September 2022 / Accepted: 28 June 2023

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

## Abstract

Summarization is a technique that produces condensed text from large text documents by using different deep-learning techniques. Over the past few years, abstractive summarization has drawn much attention because of the capability of generating human-like sentences with the help of machines. However, it must improve repetition, redundancy and lexical problems while generating sentences. Previous studies show that incorporating images with text modality in the abstractive summary may reduce redundancy, but the concentration still needs to lay on the semantics of the sentences. This paper considers adding a topic to a multimodal summary to address semantics and linguistics problems. This stress the need to develop a multimodal summarization system with the topic. Multimodal summarization uses two or more modalities to extract the essential features to increase user satisfaction in generating an abstractive summary. However, the paper's primary aim is to explore the generation of user preference summaries of a particular topic by proposing a Hybrid Image Text Topic (HITT) to guide the extracted essential information from text and image modalities with the help of topic that addresses semantics and linguistic problems to generate a topic-guided abstractive multimodal summary. Furthermore, a caption-summary order space technique has been introduced in this proposed work to retrieve the relevant image for the generated summary. Finally, the MSMO dataset compares and validates the results with rouge and image precision scores. Besides, we also calculated the model's loss using sparse categorical cross entropy and showed significant improvement over other state-of-the-art techniques.

**Keywords** Multimodal Abstractive Summary · Topic Modelling · Latent Dirichlet Allocation · Attention Mechanism

## 1 Introduction

As the internet bundles with cluttered data, so the user needs help in identifying the required information. However, searching and reading essential information from numerous documents takes time and effort for the user. This stress the

importance of summarization. The main aim of summarization is to condense a larger piece of text or information into a shorter version while still retaining the most important and relevant information. Summarization provides a quick and concise overview of a longer text. Systems summarize the documents in two types: Automatic summarization and Human summarization. The focus lies on automatic summarization as it contains two types of techniques extractive and abstractive summarization.

Extractive summarization identifies the most important sentences from documents using Term-Frequency Inverse-Document-Frequency (TF-IDF) or Bag-of-Words (BoW) technique that copies the sentences from original text documents' into summarized output, the meaning and semantics of the sentence miss to generate a summary. Conversely, abstractive summarization generates novel sentences from text documents without copying and generates meaningful novel sentences but suffers mainly from the linguistic and syntax of the grammar.

---

Shaik Rafi and Ranjita Das have contributed equally to this work.

---

✉ Ranjita Das  
ranjita.cse@nita.ac.in

Shaik Rafi  
shaik.cse.phd@nitmz.ac.in

<sup>1</sup> Department of Computer Science and Engineering, National Institute of Technology Mizoram, Chaltlang, Aizawl, Mizoram 796012, India

<sup>2</sup> Department of Computer Science and Engineering, National Institute of Technology Agartala, West Jirania, Agartala, Tripura 799046, India

As to performing summarization, deep learning learns data features with the help of neurons. Moreover, it efficiently passes the information from one layer to another by minimizing learning errors to improve the robustness of the model [1]. Furthermore, deep neural networks employ embedding techniques such as word2vec and glove, which help feed data for model training. In addition, these networks use different layers to add non-linearity to the model to learn more complex information, which is cost-effective.

Besides, there is another type of neural network called Spike Neural Networks (SNN), an advanced generation of artificial neural networks evolved from biological neural networks that use signal processing, information passing, data processing and decision-making.

However, artificial neural networks are historically brain-inspired, and the behaviour of these two neural networks differs entirely in structure, computations and learning rules. Spike neural networks are fast and energy efficient for computing spikes in the network as it considers input as spikes and generates output spikes. As spikes exchange information through discrete signals and are not continuous, the network only transmits information at the end of each propagation by assigning a certain threshold. It emits signals to lower-order neurons to share data if it reaches the threshold [2].

Like embedding in ANNs, it transmits data continuously, which leads to high computation costs. Like ANN, SNN also follows three different types of architectures [3]. (1) Feedforward Neural Network (FNN): In FNN, the data transmission takes only a forward direction by processing over many hidden layers. (2) Recurrent Neural Network (RNN): It exhibits the dynamic behaviour of the model. (3) Hybrid Neural Network (HNN) implements FNN and RNN operations.

Currently, there are no such efficient methods for training SNNs, and it does not allow any classical techniques of neural networks. The research community still searches for an optimal approach for training SNNs, which is considered challenging. Still, there exists specific techniques for training SNN's, some of them include Spike-Timing Dependent Plasticity (STDP), Growing Spiking Neural Network (GSNN), Artola Brocher Singer (ABS) rule. However, all the above techniques are considered for Un-Supervised Learning and used for biology rather than machine learning. Still, SNN follows the Leaky Integrate and Fire Model, which is very hard to train because of the availability of few neurons in the network (2100).

Also, simulating SNNs on standard hardware is computationally expensive. Some of the applications areas where spike neural network implementation had done are image and audio processing, also in different applications areas of Robotics, Computer Vision in classification problems like MNIST. However, implementation needs to

be more expensive [4]. Still, more attention requires in respective fields. For example, the ANN-to-SNN conversion method helps to convert ANN to SNNs. It transmits the same information as ANN but decreases the cost of signal transmission and computation. SNNs will turn all input values into binary signals to exchange information with other neurons.

The ANN-to-SNN conversion methods consider a solution to the energy-efficient problem. However, encoding data is the most significant unresolved challenge in SNNs to perform summarization.

Hence, we consider deep learning sequential-to-sequential model for multimodal summarization, as the primary aim of a text summary is to extract features and generate meaningful and novel sentences. With the advancement of deep learning techniques, many researchers [5, 6] introduced the sequential-to-sequential model to generate abstractive text summaries. Later [7, 8] extended their works using RNN model to generate text summaries. However, the traditional abstractive summaries could not catch the most relevant information from the text.

Later over time, different techniques were proposed for text summarization to generate a summary; some of the methods include graph-based [9], cluster-based [10], and template-based [11] approaches to extract essential information from text features and train the model that can generate a summary. The main idea behind these techniques is to identify important information and understand the document's context. Instead, these techniques stepped back in identifying the context of sentences and created learning problems leading to redundancy. Deep learning's encoder-decoder model plays a vital role in overcoming and eliminating learning problems. However, a research gap still exists in abstractive summaries in the form of redundancy, semantics, and linguistic problems.

To address the problems of redundancy, semantics and linguistic issues in the summary generation process, incorporating other modalities like image information with text modality also the addition of a Topic with multimodal features may help to reduce redundant information and helps to identify the context of the sentence to generate a meaningful sentence. Considering two or more modality information can be called a multimodal.

As we know, many social sites on the internet, like Facebook, Instagram and Twitter, use different inputs: text, images, audio and video, which give more information to the user. Also, we call it multimodal information. If different types of information exist in documents, we also call those documents multimodal documents [12]. This work consider two types of inputs namely text and image, to process multimodal information.

Multimodal summarization defines extracting multiple features from both modalities and fusing information into a

common multimodal space feature vector to train and generate the multimodal summary. Not only generating novel multimodal sentences but also retrieving the relevant image from the set of images to the summary is called multimodal output. The growth of robust techniques in deep learning has played an important role in extraction of features from text and images and fusing them into a common multimodal space.

However, many research studies have concentrated on unimodal than multimodal summarization. Nowadays, the recent trend has moved from unimodal to multimodal summarization. Adding images to the generated summary will increase the user's understanding level more than the text summary. Hence, a more meaningful summary generates when tagged text with an image to the output.

Many multimodal applications run with deep learning techniques by combining text-image pairs in various fields like Telemedicine and virtually assisting humans like Siri in Apple and Alexa in Google, which can work with different inputs.

Some research works on multimodal summarization [13–16] is also evolved. Studies prove that multimodal summarization has improved the satisfaction of user by 12.4 % compared to text summary, as images give more information to the user than text.

To address the semantics and redundancy problems, we aim to merge the essential information of images with the text. Learning image representation with an image encoder and establishing the contextual relationship between text and image is possible with the help of techniques evolved in deep learning.

Multimodal summarization has concentrated mainly on the extraction of essential features [13–15] of text by applying attention mechanism and local features of images [15, 16]. Extracted features from both modalities are fused to identify the relationship between modalities to generate a meaningful summary. The previous works on multimodal summary have identified promising results compared to text summary. However, the concentration has focused on merging the modality information and generating a summary. However, it has yet to concentrate on generating incomplete sentences, redundant summaries and lexical problems of summaries. However, the research community has not concentrated on topics that guide text and image features to identify the words with a typical relationship between two modalities within the context. So, this paper concentrate on news topics that merge with multimodal features where users can see their interested topic news with a contextual multimodal summary.

So this gives an immense interest in working on multimodal abstractive summarization with the topic as guidance. Several popular advanced techniques are involved in topic modelling to extract topics from documents like

Latent Dirichlet Allocation LDA [17], Latent Semantic Indexing or Latent Semantic Analysis LSA/LSI [18]. These models help retrieve topic distributions of words from documents or sentences and arrange text in the form of topic-related groups. Following this intuition, the proposed work has focused on incorporating a topic with a multimodal summary.

To the best in multimodal summary, this is the first paper to incorporate the topic as other modality information to guide the extracted features of text and image into a topic-related multimodal space and helps to develop a novel architecture by proposing a Hybrid Image Text Topic (HITT) layer that fuses topic with multimodal information. Therefore, topic-related features are identified from both modalities to generate a topic-guided multimodal feature vector.

Figure 1 gives a clear explanation for the understanding of topic-guided multimodal summary. In the example, the text speaks about sportsperson Gabriel Jesus, who plays football and is signing a contract for the Manchester team, as there are some related images of Gabriel Jesus providing more information about the current context. Then, we consider all those related images and text in the document as input, apply some natural language processing techniques to extract topics from the image captions, and use an encoder-decoder to extract both features. Then these features are used to merge and train the model with the help of topics to generate a meaningful multimodal summary based on the topic. Finally, we retrieve a relevant image from the images available to form a multimodal output.

This paper aims to identify topics from image captions and extract essential features from text and images with text and image encoder. These extracted features identify related words with topic words and merge them in a proposed Hybrid Image Text Topics (HITT) layer to form a topic-guided multimodal feature vector. Finally, the model evaluates the MSMO dataset with rouge and image precision scores.

The novel contribution of the papers lies in the following way:

1. Extracting topics from Image Captions using the LDA technique.
2. Proposing a novel technique of caption-summary order space to retrieve the relevant image from a set of images to the generated multimodal summary
3. Proposing a Hybrid Image Text Topic (HITT) for fusing the multimodal features with a topic to address multimodal summaries' redundancy and linguistic and syntactical problems.

The paper frames different sections as follows: Sect. 2 contains related work, Sect. 3 elaborates on the methodology, Sect. 4 explores training and testing, Sect. 5



**Fig. 1** Example for Topic-Guided Multimodal Abstractive Summarization With Multimodal Output

explores the experimental setup Sect. 6 includes results and analysis. Section 7, conclusion and future work, and data availability statement and conflict of interest.

## 2 Related work

Text summarization generates summaries by shortening the most appropriate information from the text document. The sentences or text can be significant when the sentence covers vital information about the document with linguistic and semantical information to generate a summary. Different research ideas have addressed the problems of abstractive summarization. The literature work divides into Spike Neural Networks, Text Summarization, Image Captions and Multimodal Summarization.

The authors proposed a context-dependent with a novel multiple fault-tolerant spike routing scheme [2] One that uses a bypass method for a multicast routing scheme without using virtual channels. Another approach considers digital tolerant context-dependent by the mpfc technique. The study concentrates on the neuromorphic system FCL and works on STDP learning capability. The main intention of FCL is to use bypassing methods that divide the fault nodes and to transfer information when nodes reach to fault area based on specific rules to avoid deadlock. It can also use in various brain-computer interfaces and robotics. The proposed technique combines with 3d mesh NOC topology.

Using gradient-based online meta-learning, the authors have proposed a spike-based minimum error entropy (MeMEE) in [3]. The main intention of the work is to integrate information into the machine and to increase the

capability performance, then testing on agent navigation and memory tasks. Furthermore, they can improve learning performance by changing the firing patterns of neurons, which depends on the threshold value. Finally, for efficient training of the neurons, they use a scaling factor to initialize spikes by enabling self-learning capability to SAM model and the information is encoded with the Gaussian population rate encoding method that helps to find objects in 2d area.

In [4], authors have proposed a Heterogeneous ensemble-based spike-driven few-shot learning mechanism to spike neural networks for learning sequential data. The proposed model uses a leaky integrated fire model for embedding the input. Furthermore, the ensemble entropy technique has fastened the speed of learning to get model accuracy. However, if the ensemble technique results in a higher value using the loss function, then SNN considers it needs to learn more accurately to achieve its accuracy using backpropagation. Therefore, HESFOL improves learning accuracy.

The paper proposes [19], self-adaptive multicompartement (SAM), that uses working memory. Working memory of spike neural network in neuroscience plays a vital role in understanding the learning process. Moreover, working memory retains knowledge about current situations and recent past experiences. The proposed model also uses a spatiotemporal approach for encoding information and achieves a low power consumption on neuroscience hardware platforms.

From [2, 3, 12, 19], we conclude that existing learning techniques of deep learning cannot apply directly to SNN. So the authors also concluded that a more efficient technique requires training the snn.

*Text summarization models:* Generation of shorter sentences from long sentences is considered by Mikolov et al. [20] proposing a neural abstractive sentence summarization method based on a neural language model with word distribution for the generation of abstractive summarization with novel words and concentrate on more important words.

Bahadanau et al. developed the Attention mechanism [21] used in summarization to address the repetition problem and for the alignment of the original text. Herman et al. considered encoding sentences with CNN [7] and documents with RNN generates novel words using the probability method to generate an abstractive summary with semantics but suffers mainly from repetition problem. For summarization, the neural language model used a CNN encoder to generate a summary. However, Rush et al. [5] applied a Sequential-to-Sequential deep learning model for the first time to achieve better summarization performance.

Ramesh Nallapati et al. [8] have further used Recurrent Neural Network (RNN) replaced with an encoder to achieve better performance in abstractive summarization. Guleehre et al. [22] proposed a gate control mechanism to copy words from the source document or for a novel word generation from the decoder is considered in abstractive summarization.

See et al. [6] proposed a coverage mechanism technique to address the repetition of words in the abstractive summaries, and he is the first to address the redundancy problem in summarization.

For summarization, Barzilay et al. [23] proposed tree-based approaches that are considered the best technique to extract meaningful information from nodes. However, this technique has a shortcoming of combining salience information among different nodes, which we addressed as syntactic information among the sentences. However, text summarization intends to condense the longer sentences into shorter, more meaningful ones.

In abstractive summarization, semantical and repetition problems seen in Binwahlan et al. [24] to predict shorter sentences. Moreover, the generated summary lacks coherence in the connection between sentences, which produce a weak summary. LSTMs help to remember the long sequences over Recurrent Neural Network (RNN) that uses probabilistic models by Rossiello et al. [25] to generate abstractive summarization. Linking long sentences with semantical meaning for generating summaries will consider the prior knowledge of RNN that overcomes LSTM.

*Image captions:* The task of multimodal summarization is much closer to image captioning. This task generates captions for the images. Different CNN models consider extracting image features, such as AlexNet, GoogleNet, and VGGNet. Decoding can apply to images by splitting visual representation into patches to generate words for

images in Xu et al. [26]. The aligned image set with tags is encoded for recognized objects in images to generate captions using an attention decoder by Liu et al. [27].

The authors, Kiros et al. [28], proposed a multimodal log-bilinear model for generating captions to the images. In addition, they are first to use deep learning-based techniques for image captioning detection of semantic concepts from images by Gan et al. [29]. These concepts retrieve from captions by considering frequencies which aid in generating captions for the images. Sandeep et al. [30] have considered the extraction of topics by LDA proposed from images. The topics have been extracted by pointing to everyday semantic objects from images. This paper's primary concern is extracting and fusing topic vectors with text vectors to generate topic-based image captions by considering a sequential model with an encoder-decoder framework.

*Multimodal summarization:* Multimodal summarization is the subset of Text Summarization, and many researchers have contributed their work in the field of multimodal summary.

The paper [13] proposed a new layer called the Multimodal Attention Layer where text and images are fused to the layer to train two different modalities of text and image separately to generate text and visual context vector predict multimodal summary. The visual coverage vector helps to retrieve images that has highest coverage vector from the available images to include in the predicted text to become a multimodal summary MSMO always suffers from modality bias problem.

A multimodal objective function by Zhu et al. [14] considers negative log-likelihood loss for text and entropy loss for image modality. The model considers text for generating a summary and an image discriminator that selects a set of images to retrieve the relevant image for a multimodal summary by applying attention that generates an attention weight vector to fuse into a new layer called the multimodal attention layer and evaluates with the Multimodal Automatic Evaluation technique for considering the salience of images, text, and text-image relevance.

The paper proposes Multimodal Attentional Hierarchical Encoder-Decoder [15] to summarize both modalities, where text can consider as separate input for both modalities and captions are crawled from the web and maintained according to the order of the text Image summary is performed based on images by assigning ranks to images then visual pair is selected for the summary to become multimodal output.

The authors in [31] proposed MTCA model that uses a two-stream cross attention mechanism to combine information from different modalities such as text, images. The text stream uses a transformer model to encode the textual information. Visual stream uses a cnn for extraction of

visual features from visual information. The two streams are then merged using a cross-attention mechanism, to retrieve relevant information from both streams. This cross-attention mechanism enables the model to capture the relationships between different modalities and generate a more informative multimodal summary.

MMSS by Zhu et al. [16] considers text and visual pairs as input to produce text only as output by proposing inter and intra modalities concentrates on the text and some parts of image patches. The authors have proposed two visual filters, an attention filter and an image context filter, to denoise the images and increase the summary generation semantics. The sequential model considers a hierarchical attention mechanism to generate words from sentences to concentrate more on word vocabulary that enriches the summary's semantics.

The paper [32] proposes a novel approach to multimodal summarization, which involves generating a concise summary of a piece of text and its associated images or videos. The proposed approach uses a dual contrastive loss framework to align and attend to both the textual and visual modalities. The first contrastive loss encourages the alignment between the generated summary and the original text, while the second contrastive loss encourages the alignment between the summary and the visual features of the associated images or videos. The approach also incorporates a self-attention mechanism with transformer-based encoder-decoder to further enhance the performance of the system.

The paper [33] proposes a novel approach of hierarchical, meaning that it generates summaries at multiple levels of granularity, from a high-level overview to more detailed information. The proposed approach uses a combination of image and text features to generate the summary, and incorporates a hierarchical attention mechanism to identify the most relevant information at each level of granularity. The approach also includes a novel loss function that encourages the summary to capture important information from all modalities, while also penalizing redundancy and encouraging diversity in the summary.

The paper [34] essentially proposes a method for summarizing information from multiple sources such as text, images that are all related to a specific topic by proposing a multitask learning which is on-topic and off-topic to identify the topic similarities from the extracted features of both modalities for the generation of topic aware multimodal summary.

The authors Libovicky et al. [35] proposed hierarchical attention on images. However, they have not considered attention to image patches and did not apply image filters to retrieve the relevant image in a multimodal summary. However, calico et al. [36] used separate attention mechanisms on both modalities to pay attention to image patches

and source words to build the relationship between the two modalities to generate a multimodal summary.

From the literature survey, we observe certain shortcomings for improvement in text summary and multimodal summary in the form of redundancy, linguistics, and retrieval of relevant images to the summary is also challenging. The proposed work of the paper incorporates topic modelling that helps to overcome the redundancy problem by selecting maximum words from word vocabulary related to the topics and identifying semantics between sentences with the same meaning to generate a multimodal summary.

### 3 Methodology

As multimodal summarization consists of different modalities, this paper considers {txt, img} modality information from the MSMO dataset as input. This paper aims to guide multimodal features with topic  $T(txt, img) \in T$  where txt belongs to text documents with a set of words in sentences. Img represents images from the dataset; topic  $T$  guides these multimodal features to identify a typical relationship between topic and features to overcome redundancy and generate lexical and linguistic abstractive multimodal summary with the relevant image as output.

The proposed model considers multimodal information from text documents  $d = (d_1, d_2, \dots, d_d)$  with  $k$ -sentences  $sen = (sen_1, sen_2, \dots, sen_k)$  have words  $w = (w_1, w_2, \dots, w_p)$  as input to build word vocabulary from text and applies attention mechanism on word vocabulary to reduce unnecessary words and targets only important words by training and resulting in a text attention weight vector. As other modality information is also utilized from images  $Img = (img_1, img_2, \dots, img_n)$  to extract features from images with an image encoder and merge them with text-attention weight vector and topics in the Hybrid Image Text Topic (HITT) layer. These merged features with the topic helps to identify domain-specific language relationship and decode it to generate a topic-guided multimodal abstractive summary with the decoder. The proposed architecture of the model is represented in Fig. 2.

Figure 2 uses two different inputs, one that uses topics as input to the hidden state of the LSTM encoder and text attention weight vector as other inputs that process these inputs to result in a topic-based text context vector. In addition, the architecture considers output from the encoder and extracted features of images that use to merge in a proposed layer HITT, and we call these merged features topic-guided multimodal context vectors. Finally, the resultant vector is input to the decoder to generate a topic-guided multimodal summary. From the generated

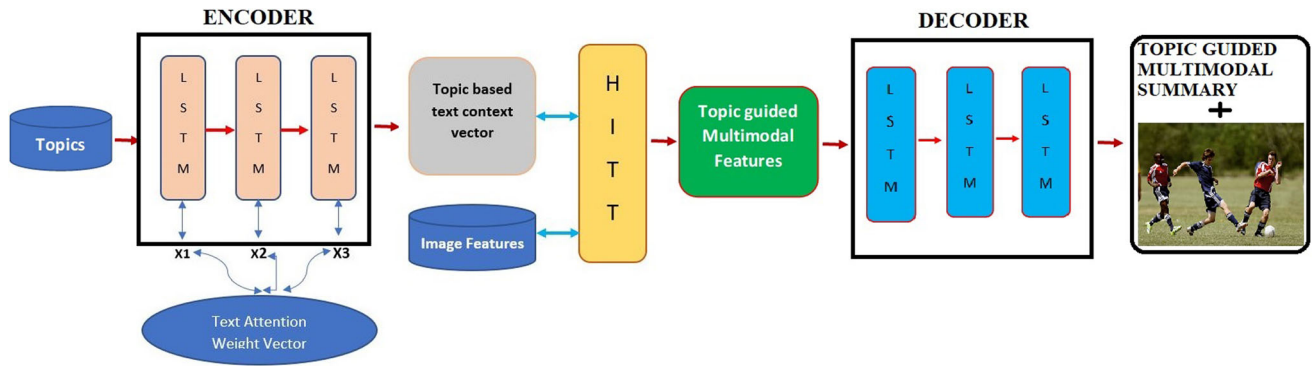


Fig. 2 Proposed Architecture

summary, the user can decide whether the generated summary is of his/her interest based on the topic.

The entire model of the architecture divides into five parts:

- In the first part, extraction of topics from image captions using LDA.
- The second part of the model deals with an encoder that considers topics and text-attention weight vector as inputs.
- The third part deals with the proposed layer HIT where all the modalities of information fused to generate a common multimodal feature vector based on topic.
- The fourth part deals with the decoder generating a topic-guided multimodal summary.
- This part addresses retrieving the relevant image to the generated multimodal summary by proposing a new technique called caption-summary order space.

### 3.1 Pre-Processing

The accuracy of the model lies in the cleaning of raw data. Therefore, we perform different preprocessing techniques individually on text and captions to remove noisy data; we follow different steps for preprocessing, which are as follows:

- Converting the uppercase sentences into lowercase.
- Removal of stop words from sentences
- Removal of numbers from sentences.
- Newline characters and extra spaces are neglected.
- Removal of hyperlinks.
- Words that are less than the length of three are removed from sentences.
- Symbols and punctuation eradicated

These steps help to remove noisy data and to train the model accurately to get better results.

#### 3.1.1 Extraction of Topics

The main intention of the proposed model is to generate user interest in a multimodal summary based on topics. These help the user to understand the concept easily with images when adding a topic to the multimodal summary. The first part of the model extracts topics represented in Fig. 3. Latent Dirichlet Allocation is the most successful generic probability topic modelling for identifying topics. Image captions are used as input to identify topics from images as the images give more meaningful information, which we call context; this context helps extract topics from the images. To extract topics, we assume that each image comprises five or six caption sentences and combine all such captions of the single image to form a single sentence to apply the LDA technique, which follows the process in [30].

From Fig. 3, we would like to use image captions as input and apply the preprocessing on captions and LDA technique to build topic vocabulary and maximum topics are extracted from the text  $d$  to pay more attention towards users interest concept.

Natural Language Toolkit (NLTK) and Scikit contain a variety of packages to preprocess the text. In addition, the Python free open-source library called gensim package helps to extract topics from captions. Table 1 represents some of the topic words extracted from image captions. We aim to extract the top 50 topics that help to guide the text and image features that are semantically related and exhibit the same meaning as we have represented the extracted top 50 topics in the form of learned embedding with a 300-dimensional vector when passed as an initial hidden state of the lstm with input vector, this vector maps topic representation with word representations by training topic-word pairs to learn to associate certain words with specific topics. During training, the LSTM updates the weights of its internal parameters, like the weights of the embedding layers and LSTM layers, to minimize the difference between the predicted output and the actual output. It

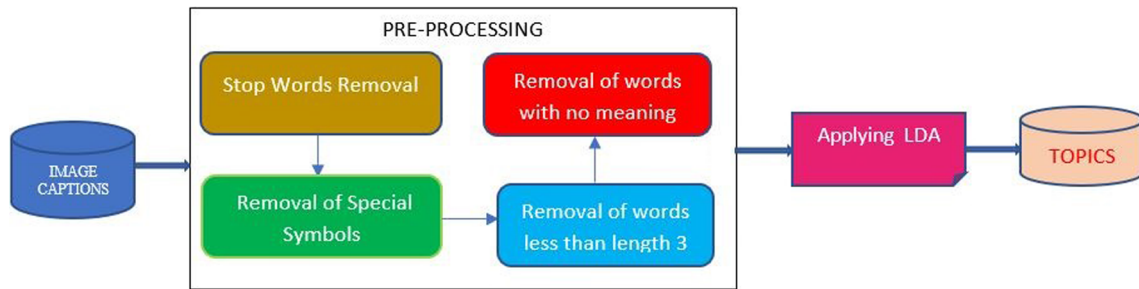


Fig. 3 Extraction of Topics With Image Captions using LDA

Table 1 Topic words

Serial.No	Topic words	Topic No
1	prison, murder, charged, court, dead,defense, temper, charges	Topic 2
2	women, protesters, trump, people, saturday, street, city, crowd	Topic 4
3	fashion, dress, skirt, gown, lady, love, hood, cardigan	Topic 6
4	ivanka, trump, family, year, photos, kamiyah, donald, vacation	Topic16
5	rescue, avalanche, rescuers, farindola, snow, cool, cold, rigopiano	Topic26

allows the model to learn a nonlinear mapping between the topic representation and the word representation, such that words that are semantically related to the topic are more likely to be predicted.

Once trained, the model generates new sequences of words semantically related to a given topic. To do this, we first encode the topic as a fixed-size vector using learned embeddings and then use this vector as the initial hidden state of the LSTM model. We then feed the model a sequence of input words, and the LSTM processes the input words one by one, updating its internal state at each step. Finally, the model produces a sequence of output words predicted to be semantically related to the topic.

$$T = \sum prob(cap_{words} | word^{topics} | d) \cdot prob(word^{topics} | d) \quad (1)$$

where  $T$  is represented as topic vector from learnable embedding  $cap_{words}$  that considers the number of words in caption sentence and  $word^{topics}$  represents the number of topics in particular sentence and  $d$  represents the probability of occurrences of word in document  $d$  to generate topics and more over  $T \in R^{k \times tm}$  where  $k$  represents the number of topics and  $tm$  represents the size of topic vocabulary; in this case, we consider  $k = 50$ .

### 3.1.2 Text Attention Weight Vector

The main aim of the attention mechanism is to focus on the essential words vocabulary built from word2vec  $emb_{txt} = [t_1, t_2, t_3, \dots, t_l]$  as  $emb_{txt}$  represents the text embedding from text articles. To reduce the word vocabulary, we concentrate on the essential information from an extensive vocabulary that gives more meaning to the sentences. The

attention mechanism is such a technique that considers sequence based hidden representation to preserve the context of the whole sentence, which adds semantics to generate text attention weight vector also to preserve the meaning of the sentence. Every row of arbitrary input sequence multiplies its weight with every hidden state of the neuron and aligns weights with a bias of the context vector to maintain the most relevant information and remember the long sequences to shrink down the vocabulary with the essential information to generate text attention weight vector.

Figure 4 explains generating text attention weight vectors from raw data by applying different preprocessing techniques to remove noisy data to build word vocabulary with word2vec. This data helps to apply the attention mechanism to extract essential information from a large vocabulary to generate a text attention weight vector.

$$text_{att} = \tanh(w_{att} \cdot h_{att} + b_{att}) \quad (2)$$

$$e_t = softmax(text_{att}) \quad (3)$$

$$a_t = \sum e_t \cdot w_h \quad (4)$$

from the above equation, we aim to generate text attention weight vector  $\alpha_t$  and  $w_{att}$  is a trainable parameter for each sentence  $h_{att}$  and  $b_{att}$  is a bias used to apply the softmax function by its weights to generate text attention weight vector.

### 3.1.3 Extraction of Image Features

The pictorial representation will give us more concise information to the user than the text model. We aim to



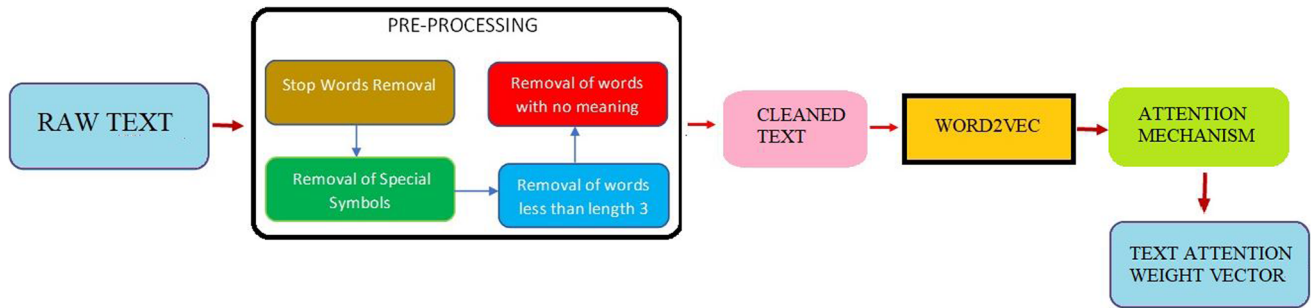


Fig. 4 Text Attention Weight Vector

consider visual modality information from images to identify the semantic relationship with the global features of the images. We consider Google’s inception v3 [1, 37, 38] model to extract the pictorial features of images, as every image is of a different size, and we aim to resize each image to 299\*299 in height and width.

Inception v3 extracts the 2048-dimensional features at the fully connected layer by dropping the softmax layer in the model to provide better performance than other deep learning models like vgg16 and vgg19. Also, Google has announced [37] that it results in 78% accuracy, so usage of inception v3 improves the model’s performance and accuracy in feature learning. Also, to reduce computation costs, we consider the pre-trained layers of  $3 \times 3$ ,  $1 \times 3$ ,  $3 \times 1$  and  $1 \times 1$  to reduce the number of parameters. We can add an extra fully connected layer in inception v3 to reduce 2048-dimensions to 1024 dimensions of image features but to merge the extracted features of images with text and topic at the HITT layer, which can help the model to extract more context from image features, in a later stage we have added two dense layers to scale down the extracted features from 2048 dimensions to 1024 then to a lower-dimensional vector.

Figure 5 shows the extraction of features from the images by applying inception v3. Finally, The extracted pictorial feature helps to feed with the topic and text attention weight vector to generate a multimodal feature vector that guides and considers the topic as supplementary information [39].

$$img_{fet} = \{i_1, i_2, i_3, \dots, i_n\} \tag{5}$$

we generate image features  $img_{fet}$  from different images  $i_1$  to  $i_n$  extracted 2048 dimensional features at a fully connected layer.

### 3.2 Encoder

The proposed model considers long short-term memory (LSTM) as an encoder to train the model with two different types of information as input. For example, one input to the encoder is the resultant of the text attention weight vector  $\alpha_t$  from Eq. (4) which is the resultant of word2vec vocabulary that captures and builds more meaningful information when merged with the topic vector  $T$  from Eq. (1) The advantage of using LSTM is to handle multiple information simultaneously and processes to generate

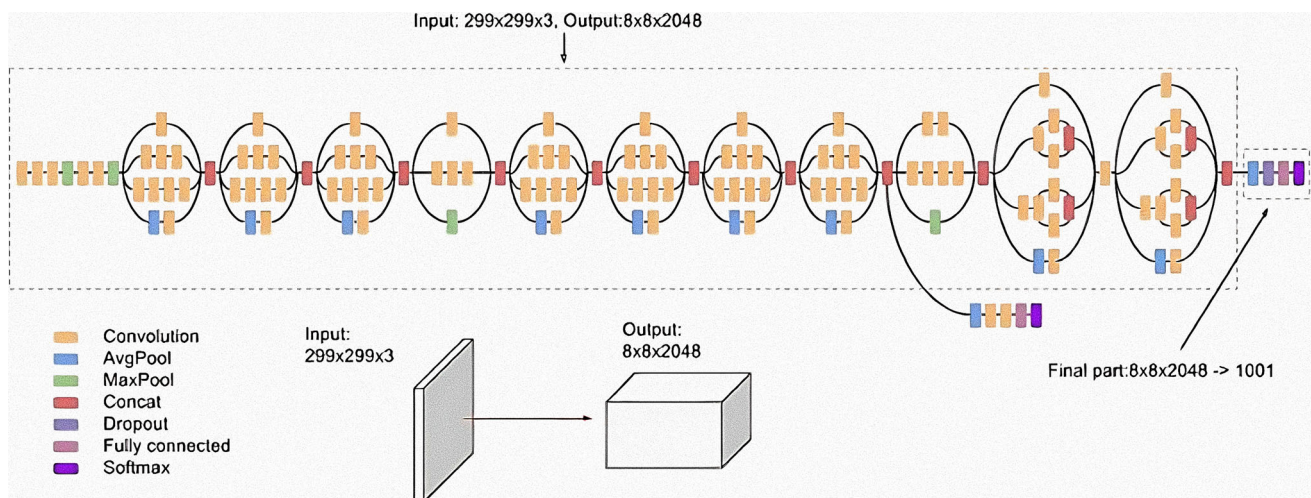


Fig. 5 Inception V3 Architecture to Extract Image Features

topic-guided text context vectors, as this helps to preserve the semantic meaning of the sentence in the generation of topic-based text context vectors, as shown in the proposed architecture in Fig. 6.

$$t_{c_{txt}} = \sum_{i=1}^{50} T * (\alpha_t w_h + w_z h_z + b) \tag{6}$$

in the above equation  $t_{c_{txt}}$  is topic-based text context vector from topic vector  $T$  and  $\alpha_t$  represents text attention weight vector and  $w_h, w_z$  is a learnable parameter and  $h_z$  is hidden state information, and  $b$  represents bias value. Also, the LSTM encoder can handle multiple information by stacking LSTM linearly with structure as inputshape=(topics, text features), where we pass topics as 50 and text features to the LSTM.

### 3.3 Hybrid Image Text Topic (HITT) Layer

The third component of the architecture is the HITT layer. The main aim of the HITT layer is to fuse all information modalities into a single layer. The necessity of a multimodal summarization has gained importance from previous research by fusing [40] two or more modality information into a piece of multimodal information. Also, Jangra [41, 42] stresses the consideration of the fusion technique for improving the multimodal summary. The proposed work considers merging two individual modalities followed by a topic that guides the multimodal information according to semantically related information to solve the redundancy problem. The HITT layer merges vectors of co-occurrences that emit the exact meaning of semantic information. When merged with the features by topics, this information improves the readability of sentence context to generate more fluent sentences of users’ interest to avoid syntactical problems.

Also, we consider two fusion techniques, early fusion [43] for concatenating text attention weight vector and topic vector. Then, the intermediate fusion technique follows for merging topic-guided text features with image

features. Finally, all these features merge into a novel Hybrid Image-Text-Topic layer (HITT) to identify the relationship and address the redundancy problems. This topic vector identifies the semantically related vectors that are similar in meaning and context to generate a meaningful summary with the help of topics.

This layer holds two inputs: topic-based text context vector  $t_{c_{txt}}$  trained from the LSTM encoder and image features  $img_{fet}$  extracted from inception v3 architecture [38]. This layer holds the syntactical relationship between different words due to topics. It can build semantical relationships [2] with visual vectors to learn multimodal relationships when combining topic-based text context vectors with image features. We can generate topic-guided multimodal modal feature vectors from Eqs. (4) and (5).

$$tmm_{txt} = t_{c_{txt}} \cdot w_h + Img_{fet} \cdot w_h \tag{7}$$

as we compute topic-guided multimodal feature vector  $tmm_{txt}$  by merging  $t_{c_{txt}}$  topic-guided text context vector and  $Img_{fet}$  image features and  $w_h$  represents learnable matrix.

### 3.4 Decoder

This module receives topic-guided multimodal features from the HITT layer as shown in Fig. 7 feeds and forward to a stacked LSTM decoder that includes topic information in multimodal features to generate topic-guided multimodal summary.

To calculate the probability of generating a semantic sentence of topic-guided multimodal context vector with a decoder for generating a multimodal linguistic summary with the topic as guidance, we have fixed the decoder summary length to 13 to generate more meaningful sentences.

Based on the information received, the decoder uses <start> as the initial state and generates the next word based on the hidden states. This process continues until all the remembered words from the LSTM decoder generates to form a complete sentence, as <end> end of sentence represents the end of the multimodal summary generation from the decoder.

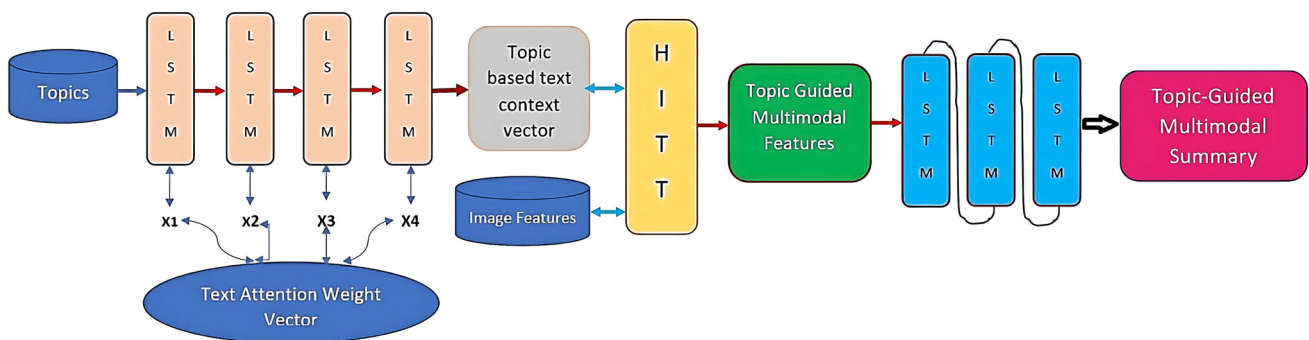
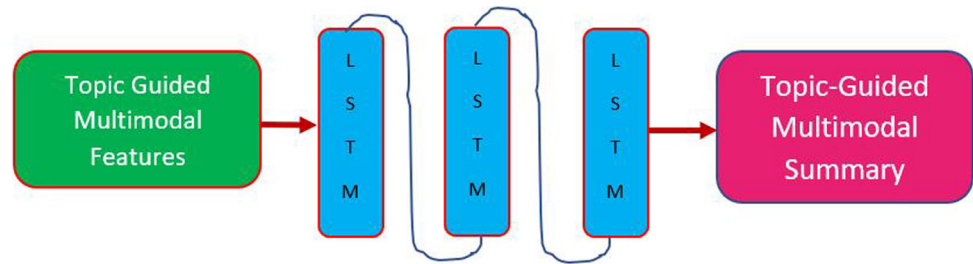


Fig. 6 Topic-guided multimodal summary generation

**Fig. 7** Generating Topic-Guided Multimodal Summary With Decoder




---

**Algorithm:** Topic-Guided Multimodal Abstractive Summarization with Multi-modal output.

---

**Input:** Three different inputs are considered Text Articles, a set of Images, and topics extracted from Captions.

**Output:** Topic-Guided Multimodal Summary with relevant Image as output.

---

**Step1:** A preprocessing step is applied on text articles and image captions to remove noisy data from MSMO dataset

**do:**

- Removing Stop Words
- Removing special symbols.
- Removing words that are less than length three.
- Removing numbers and extra spaces.

**Step2:-** Building word vocabulary with word2vec from text and applying attention mechanism to reduced vocabulary.

**Step3:-** Applying LDA technique to extract topics from image captions by using step 1.

**Step4:-** Keywords are used to generate a topic vector from topic vocabulary using LDA.

**Step5:-** Inception v3 helps to extract features of images .

**Step6:-** Training the topic vector with LSTM guides text vocabulary for semantically related words.

**Step7:-** HITT layer helps to fuse topic-guided text context vector with image features.

**Step8:-** Feeding the features of the model to the encoder.

**Step9:-** The model can learn the features with different weights and feed weight vectors to the decoder to generate a multimodal summary.

**Step10:-** Caption-Summary Order Space is used to retrieve the relevant image for the generated multimodal summary.

---

### 3.5 Image Retrieval

Retrieving relevant images for multimodal summaries is a tedious task. We consider image captions to retrieve images for a multimodal summary. We propose a new model for retrieving images called caption-summary order space.

This technique uses the generated multimodal summary and image captions from the MSMO dataset and checks which caption matches the highest possible words in the generated multimodal summary and with the topic word. First, the captions with the highest possible matching will be retrieved. Sometimes two or more captions will also be retrieved that match the summary. On all such captions, we perform intersection operations that exactly match with the summary or the caption that has the highest match in the summary is considered, and all remaining such captions remove from consideration. Then the caption-related image will be retrieved and added to the multimodal summary to call it a topic-guided multimodal summary with multimodal output, as shown in Fig. 8.

## 4 Training and Testing

The objective of the training is to map input information to output for a given training dataset that finds a set of weights and progresses with minor updates that change the performance of the model with each iteration and solve optimization problems in evaluation with the seq-to-seq model. As the MSMO dataset contains 2,93,965 training articles from these, we consider 44,007 sentences to train the model. During training, the main objective is to minimize the loss and to make the model learn efficiently. The loss function mainly helps to know how well the model has organized or learned the data during training without losing the given information. The model adopts sparse categorical cross-entropy to measure the loss during training. As the loss function handles numerical values because the embedding technique results in numerical values. Therefore, we consider sparse categorical cross-entropy as a loss function to measure the loss and is speedy compared to other loss functions and can be evaluated as

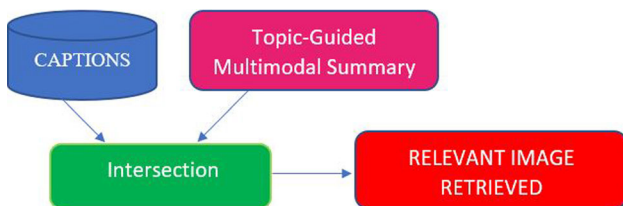


Fig. 8 Image Retrieval Process

$$L_{mm} = -1/n \sum_{k=1}^n [y_i \cdot \log(y^i) \cdot w_i + (1 - y_i) \cdot \log(1 - y^i)] \tag{8}$$

where  $L_{mm}$  represents loss function for multimodal summary and  $y_i$  represents true labels from training data and  $Y^i$  reflects the vector numbers and  $w_i$  are model parameters of the network and  $n$  represents number of samples.

We have tested the model with 10,261 articles and adopted Root Mean Squared Propagation (RMSProp) optimizer for learning the model parameters in each step by dividing the initial step size of 0.001 hyperparameters by square root RMSProp can quickly learn the parameters, so it reduces the computation time compared to another optimizer. Also, to reduce the over-fitting problems, we use a dropout ratio of 0.3.

Figure 9 resembles the data learning procedure in the model by observing the training and validation graph we can observe that accuracy gets increased by each epoch and over-fit problems gets decreased.

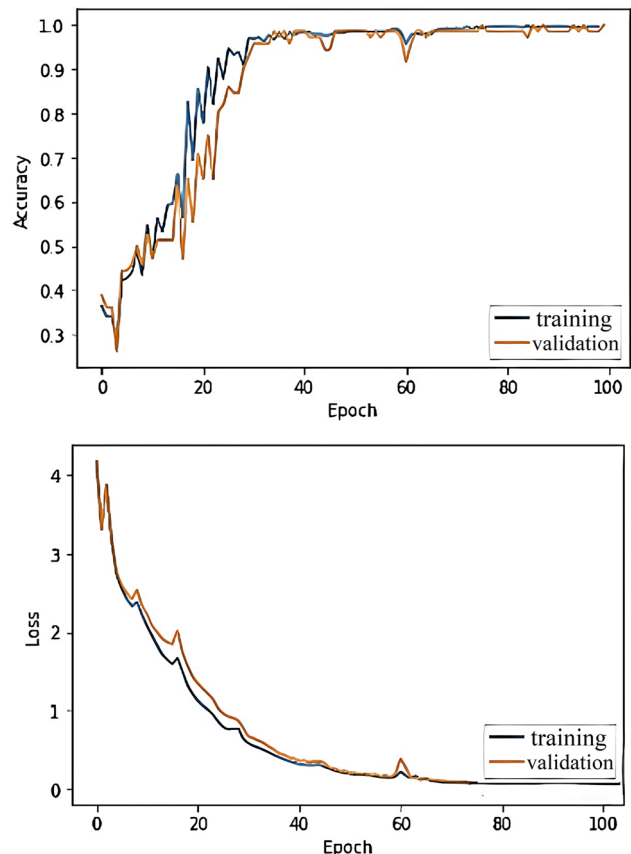


Fig. 9 Training and Loss Graph of Learning Multimodal Features

## 5 Experimental Setup

### 5.1 Dataset

The proposed model uses the MSMO [14] dataset and mostly news articles covering different topics. The dataset consists of three parts: Text articles, Images and Captions for images. This dataset has overcome the modality bias problem where a relationship exists between text and images tagged with captions. The dataset consists of 723 text article tokens and 6.58 image-captions pairs on the average scale to match images in output. In addition, the dataset consists of 10,355 validation pairs and 10,261 test pairs.

Evaluating a multimodal summary is also a significant task, as the MSMO dataset consists of multimodal reference summaries, which we call the gold summary. The paper uses rouge and image precision to evaluate generated multimodal summaries and compares them with gold summaries for evaluation with metrics. In addition, it shows user satisfaction when the image tags with the summary, which we call a multimodal summary with multimodal output.

### 5.2 Implementation Details

The proposed work implements deep learning models and techniques with a sequential-to-sequential model with an encoder and decoder framework. First, we consider text articles about building a word2vec vocabulary, and then essential vocabulary uses the attention technique to generate meaningful sentences. The resultant is a 300-dimensional attention weight vector. Finally, these inputs with 300-dimensional vectors are passed to the encoder to train the model with the RNN [44] LSTM [45–47] and layers structured in the form of a linear stack. For LSTM, we pass parameters as units = 300 because we use 300-dimensional features, and we set return sequences = True because the output of every LSTM will be passed as input to the next LSTM cell. Finally, we add a dropout layer with 0.3, and we flatten the features into a one-dimensional topic based text context vector. The entire model runs using a 16GB ram on the tpu machine in colab with an intel core i7 processor with 100 epochs for training and testing.

#### 5.2.1 Extraction of Topics and Image Features

An unsupervised technique, LDA is used to extract topics from image captions. First, we extract the top 50 topics from the captions using the natural language toolkit and scikit packages from Python to build the topic vectors. The topic vector guides the text with the related words to

generate a 300-dimensional topic-guided context vector from the encoder. In the second part, we resize every image with height and width as 299\*299. Inception v3 model extracts the image features of 2048 dimensions at a fully connected layer by dropping the softmax layer. Furthermore, we adopt two dense layers to resize the feature vector, merge both vectors in the HITT layer, and generate a 300-dimension vector. Finally, the decoder decodes a 300-dimension vector to generate a topic-guided multimodal summary as output.

## 6 Results and Analysis

Table 2 explores and compares with other state-of-art results of image precision scores. We also evaluated the proposed model performance with Recall-Oriented Understudy for Gisting Evaluation Package [50] in Table 3 for automatic system summarization that generates multimodal summaries. Table 4 shows some of the output screens generated by the model. The model has also learned the topic's multimodal features that guide them to a common multimodal space with a typical relationship between modalities to generate a meaningful summary. At the same time, it considerably reduces the loss during multimodal feature learning.

Rouge 1: Represented as R-1, It considers 1(single) word overlap in reference and system-generated summaries.

Rouge 2: Represented as R-2, It provides 2(bi) word overlap between reference and system-generated summaries.

Rouge L: Represented as R-L, It shows the longest common sequence in reference and system-generated summaries.

However, the model sometimes needed to learn semantical relationships because of the failure of the non-availability of related words with a related topic. As a result, it leads the model to generate 'NaN' as a summary; when generating a multimodal summary further, we like to





**Table 2** Image precision score with state-of-art-results

Sno	Method	IP score
1	TSC MSMO-IS [34]	68.57
2	TSC MSMO SIMPAD [34]	68.05
3	HNNattT1 [15]	49.78
4	MSMO [13]	64.82
5	MSMR [14]	68.62
6	Proposed model	69.71

**Table 3** State of art results compared with our model

S. No	Method	R-1	R-2	R-L
1	A2Summ [32]	30.8	11.4	27.4
2	SummaRunner [48]	26.2	11.1	14.5
3	Topic Augmented Generator [49]	28.36	9.05	27.48
4	Abstractive Text-Image Summarization using Multi-Modal RNN [15]	32.64	12.08	23.88
5	MSMO [14]	41.11	18.31	37.74
6	MSMR [13]	41.20	18.35	37.85
7	Text summarization using topic based vector space model and semantic measure [49]	42.3	20.3	39.5
8	Topic aware multimodal summarization [34]	37.4	15.4	34.5
9	Proposed Model	46.34	33.66	44.3

**Table 4** Output of Topic-Guided Multimodal Summary With Multimodal Output

Topic name	Reference summary	Generated multimodal summary	Retrieved image for multimodal summary
politics (topic-1)	within minutes account accumulated million followers	start within minutes accumulated million followers end	
crime(topic-2)	device brought pub pontardaw swansea	start device brought pub pontardaw pontardaw swan sea sea end	
fashion (topic-7)	riley alessandra found perfect fashion formula rock winter long	start riley alessandra perfect fashion rock winter winter end	
military (topic-10)	fighter jets reportedly scrambled escort plane travels uk	start fighter jets reportedly escort escort plane travels end	

concentrate on this type of problem. Also, it helps solve this problem by increasing topic vocabulary in the future.

We also evaluated the image precision scores of the Model in Table 2. Also, it highlights how well the image has aligned with the multimodal summary and compares them with other state-of-art methods for retrieving images. The proposed method has improved its performance.

## 7 Conclusion and Future work

Multimodal problem considers two different inputs like text and image. An unsupervised technique called LDA is used extensively to identify topics from captions. Individual extraction of both modality features is possible. Adding a topic to multimodal features in the HITT layer helps the model to guide multimodal features according to topics. The topic-guided multimodal summary generates better results when incorporating an intermediate fusion for

semantic-related words. When the Topic is added as a solution to address the semantics and linguistic problems of multimodal summary, the model generates an efficient summary, but sometimes the model generates 'NaN' values as this fails to learn topic-related words at the time of summary generation because of a lack of more topic words. As this problem needs more concentration in future by increasing topic vocabulary. Also, the model retrieves relevant images to the generated summary by proposing a caption-summary order space technique to the abstractive multimodal summary.

However, the problem can enhance by adding audio or video modalities to make a complex problem such as text, image and video or audio by extracting features from all modalities and merging them with topic modelling. Then, the related and identical features of the topic are extracted with the same context and form a cluster by proposing a new technique that helps address the linguistics and

semantics of multimodal summary and guides the multimodal features to generate a meaningful summary.

**Data availability** Sharing data does not apply to this article, as no data set was generated or analyzed during the current study.

## Declarations

**Conflict of interest** The authors have no conflict of interest.

## References

- Rafi S, Das R (2021) A linear sub-structure with co-variance shift for image captioning. In: 2021 8th International conference on soft computing and machine intelligence (ISCFMI), pp 242–246
- Yang S, Wang J, Deng B, Azghadi MR, Linares-Barranco B (2021) Neuromorphic context-dependent learning framework with fault-tolerant spike routing. *IEEE Trans Neural Netw Learn Syst* 33:7126–7140
- Yang S, Tan J, Chen B (2022) Robust spike-based continual meta-learning improved by restricted minimum error entropy criterion. *Entropy* 24:455
- Yang S, Linares-Barranco B, Chen B (2022) Heterogeneous ensemble-based spike-driven few-shot online learning. *Front Neurosci* 16:850932
- Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization
- See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. *arXiv, arXiv:1704.04368*
- Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend
- Nallapati R, Zhou B, dos Santos C N, Çaglar Gülçehre, Xiang B (2016) Abstractive text summarization using sequence-to-sequence RNNs and beyond
- Ganesan KA, Zhai C, Han J (2010) Opinosis: a graph based approach to abstractive summarization of highly redundant opinions
- Khan A, Salim N, Farman H (2016) Clustered genetic semantic graph approach for multi-document abstractive summarization. In: International conference on intelligent systems engineering, pp 63–70
- Wu P, Zhou Q, Lei Z, Qiu W, Li X (2018) Template oriented text summarization via knowledge graph. In: 2018 International conference on audio, language and image processing (ICALIP), pp 79–83
- Dash SK, Sureshchandra YV, Mishra Y, Pakray P, Das R, Gelbukh A (2020) Multimodal learning based spatial relation identification. *Computación y Sistemas* 24:1327–1335
- Zhu J, Li H, Liu T, Zhou Y, Zhang, J, Zong C (2018) MSMO: multimodal summarization with multimodal output
- Zhu J, Zhou Y, Zhang J, Li H, Zong C, Li C (2020) Multimodal summarization with guidance of multimodal reference
- Chen J, Zhuge H (2018) Abstractive text-image summarization using multi-modal attentional hierarchical RNN
- Li H, Zhu J, Liu T, Zhang J, Zong C (2018) Multi-modal sentence summarization with modality attention and image filtering
- Blei DM, Ng A, Jordan MI (2001) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41:391–407
- Yang S, Gao T, Wang J, Deng B, Azghadi MR, Lei T, Linares-Barranco B (2022) SAM: a unified self-adaptive multicompartmental spiking neuron model for learning with working memory. *Front Neurosci* 16:850945
- Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 26:3111–3119
- Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. *CoRR, abs/1409.0473*
- Çaglar G, Ahn S, Nallapati R, Zhou B, Bengio Y (2016) Pointing the unknown words. *ArXiv, arXiv:1603.08148*
- Barzilay R, McKeown KR (2005) Sentence fusion for multi-document news summarization. *Comput Linguist* 31:297–328
- Binwahlan MS, Salim N, Suanmali L (2010) Fuzzy swarm diversity hybrid model for text summarization. *Inf Process Manag* 46:571–588
- Rossello G, Basile P, Semeraro G, Ciano MD, Grasso G (2016) Improving neural abstractive text summarization with prior knowledge (position paper)
- Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, neural image caption generation with visual attention, attend and tell
- Liu C, Sun F, Wang C, Wang F, Yuille AL (2017) MAT: a multimodal attentive translator for image captioning
- Kiros R, Salakhutdinov R, Zemel RS (2014) Multimodal neural language models
- Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2017) Semantic compositional networks for visual captioning. *IEEE Conf Comput Vis Pattern Recogn* 2016:1141–1150
- Dash SK, Acharya S, Pakray P, Das R, Gelbukh A (2020) Topic-based image caption generation. *Arab J Sci Eng* 45:3025–3034
- Lu Q, Ye X, Zhu C (2022) MTCA: a multimodal summarization model based on two-stream cross attention. In: 2022 2nd International conference on computer science, electronic information engineering and intelligent control technology (CEI), pp 594–601
- He B, Wang J, Qiu J, Bui T, Shrivastava A, Wang Z (2023) Align and attend: multimodal summarization with dual contrastive losses. *ArXiv, arXiv:2303.07284*
- Qiu J, Zhu J, Xu M, Derroncourt F, Bui T, Wang Z, Li B, Zhao D, Jin H (2022) MHMS: multimodal hierarchical multimedia summarization. *ArXiv, arXiv:2204.03734*
- Mukherjee S, Jangra A, Saha S, Jatowt A (2022) Topic-aware multimodal summarization
- Libovický J, Helcl J (2017) Attention strategies for multi-source sequence-to-sequence learning. *ArXiv, arXiv:1704.06567*
- Calixto I, Liu Q (2017) Incorporating global visual features into attention-based neural machine translation
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition, pp 2818–2826
- Inception v3. <https://cloud.google.com/tpu/docs/inception-v3-advanced>
- Wang L, Yao J, Tao Y, Zhong L, Liu, W, Du Q (2018) A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *ArXiv, arXiv:1805.03616*
- Lahat D, Adali T, Jutten C (2015) Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc IEEE* 103:1449–1477
- Jangra A, Saha S, Jatowt A, Hasanuzzaman M (2021) Multimodal supplementary–complementary summarization using multi-objective optimization. In: Proceedings of the 44th

- international ACM SIGIR conference on research and development in information retrieval
42. Jangra A, Jatowt A, Saha S, Hasanuzzaman M (2021) A survey on multi-modal summarization. *ACM Comput Surv*
  43. Li K, Zhang Y, Li K, Li Y, Fu YR (2019) Visual semantic reasoning for image-text matching. In: *IEEE/CVF international conference on computer vision*, pp 4653–4661
  44. Cho K, van Merriënboer B, Çaglar Gülçehre, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation
  45. Kuchaiev O, Ginsburg B (2017) Factorization tricks for LSTM networks. *ArXiv*, [arXiv:1703.10722](https://arxiv.org/abs/1703.10722)
  46. Pathak A, Pakray P, Das R (2019) LSTM neural network based math information retrieval. In: *Second international conference on advanced computational and communication paradigms (ICACCP)*, pp 1–6
  47. Rafi S, Das R (2021) RNN encoder and decoder with teacher forcing attention mechanism for abstractive summarization. In: *2021 IEEE 18th India council international conference (INDICON)*, pp 1–7
  48. Nallapati R, Zhai F, Zhou B (2016) SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents
  49. Belwal RC, Rai S, Gupta A (2021) Text summarization using topic-based vector space model and semantic measure. *Inf Process Manag* 58:102536
  50. Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.