

# Semantics aware abstractive multimodal summarization with multimodal output

Shaik Rafi

Asst. Professor

Computer Science and Engineering

Narasaraopeta Engineering College

(Autonomous)

Narasaraopet , India

[shaikrafinrt@gmail.com](mailto:shaikrafinrt@gmail.com)

Shaik Hidaitulla

Student

Computer Science and Engineering

Narasaraopeta Engineering College

(Autonomous)

Narasaraopet , India

[hidaitullashaik2002@gmail.com](mailto:hidaitullashaik2002@gmail.com)

Chakolti Chiranjeevi

Student

Computer Science and Engineering

Narasaraopeta Engineering College

(Autonomous)

Narasaraopet , India

[chiranjeevichakolthi@gmail.com](mailto:chiranjeevichakolthi@gmail.com)

Gandru Phanindra

Student

Computer Science and Engineering

Narasaraopeta Engineering College

(Autonomous)

Narasaraopet , India

[phanichowdary70298@gmail.com](mailto:phanichowdary70298@gmail.com)

**Abstract—** The abundance of textual content available on the internet, such as tales, news articles, and blogs, poses difficulties for users in terms of the time and effort needed to understand it. One way to improve user comprehension and cut down on reading time is by automatic abstractive text summarization. By utilizing multiple techniques suggested by the research community, this strategy condenses input text while maintaining meaning and contextual relevance. Semantic coherence and contextual understanding are two problems that still arise during the summarizing process. Multimodal abstractive text summarization, which integrates textual and visual information, has been presented as a solution to these problems. In this method, contextual connection features from text-related images are retrieved. We get rouge scores as R-1: 53.43, R-2: 39.01, R-L: 46.54.

**Keywords—** Deep learning, LSTM, Text Generation, Cosine Similarity. Text Summarization.

## I. INTRODUCTION

Two main strategies are the focus of automatic summarization. In the first method, known as extractive summarization, a text is distilled into a summary by locating and selecting important lines or phrases from the source document [1-2]. This method reorganizes already-written content to provide a cogent summary while preserving the original sentence structure and terminology. Abstractive summarization endeavors to craft fresh, compact, and logically structured summaries. This process entails comprehending the source material and articulating its core significance in novel ways, even if the precise wording or sentences are absent from the original text. Consequently, abstractive summarization has the capacity to

generate summaries that are more akin to human language, exhibiting fluency and coherence. Nevertheless, prior work in abstractive text summarizing has mostly concentrated on improving the resulting summaries' grammatical structure. As a result, these methods have frequently failed to provide users with an accurate summary of the content. The research community underscores the significance of coherence and semantics in sentence construction to produce summaries akin to human-generated ones. This emphasis extends to tackling challenges such as sentence structure, semantic comprehension, and syntactic understanding. Consequently, there's a rising demand for multimodal information integration, utilizing both text and visuals for efficient communication in summarizing content.

Figure 1 represents the sample output of Multimodal Abstractive Text Summarization, in which the encoder model generates the summary sequence by taking reference summary as input. To match an image for the generated summary, an algorithm called Cosine Similarity is used and based on highest score final image is considered as best matched image out of all the images that matches.

0.17285002893

0.26111648393

0.261116483933



Fig. 1. Example of Multimodal Abstractive Text Summarization

The goal of the rapidly developing subject of multimodal summarizing in research and technology is to use many data modalities to create succinct and enlightening summaries and retrieve relative image to it. These modalities include text, photos, music, videos, and any other kind of media that contains relevant data.

Multimodal summarization is the process of combining data from several modalities, like text and images, to produce a concise and well-organized summary. It comprises dissecting and extracting pertinent information from several sources, then combining them to create a thorough synopsis. This method provides a powerful way to record and visualize data from multiple modalities, which makes it easier to create summaries that are more insightful and impactful in a variety of fields. research, so named because it combines the realms of language and vision, is represented by the convergence of text and picture processing [9–10]. Multimodal fusion approaches create a cohesive representation that captures the interaction between text and image by merging textual and visual data. These fusion techniques improve the summarizing model's capacity to extract relevant data from both modalities, leading to summaries that are more thorough. Improvements in producing effective and insightful summaries have been shown by experimental studies exploring the integration of text and visual modalities. Semantic understanding and inadequate summary production are two problems that have not yet been addressed in multimodal summarization research. Thus, the goal of this research is to create multimodal abstractive text.

## II. LITERATURE SERVEY

Several methods have been put up to overcome the drawbacks of text-only summarization by using multimodal methodologies. A technique based on word distribution was presented by Mikolov et al. [12] to produce succinct abstractive phrases. By creating a is their main goal. Furthermore, Mukherjee et al. Present a multitask learning strategy that emphasizes both on- and off-topic data in order to extract topic-related similarity associations from text and image modalities and produce indispensable for various image studies and provide more Zhu et al. [8] propose a multimodal objective function. This feature integrates multiple modalities—text and visual—to evaluate how well the produced summaries match the intended summarizing goals. In order to capture the connections between inputs and related summaries, this objective function considers both textual and visual intended summarizing goals. In order to capture the connections between inputs and related vocabulary from the source phrases, this technique makes it easier to create clear, intelligible sentences that are readable by humans. Similar to this, Mosa et al. [13] created shorter, , more concise sentences by extracting sentence weights from long source materials using a swarm optimization technique. Additionally, in long texts, tree-based methods have been applied to capture similarity between words or sentences with similar semantics These methods generate abstractive summaries by utilizing natural language generating algorithms Gupta et al. [14]. Furthermore,

employing RNN-based LSTM models, Rafi et al. [15] introduced the Teacher Force Technique to improve the effectiveness of networks in producing syntactically and semantically consistent abstractive summaries. However, while being trained for summarization tasks, conventional RNNs frequently operate slowly and less well than ideal. In order to solve this problem, the suggested method incorporates an Attention Decoder mechanism to enhance summary performance.

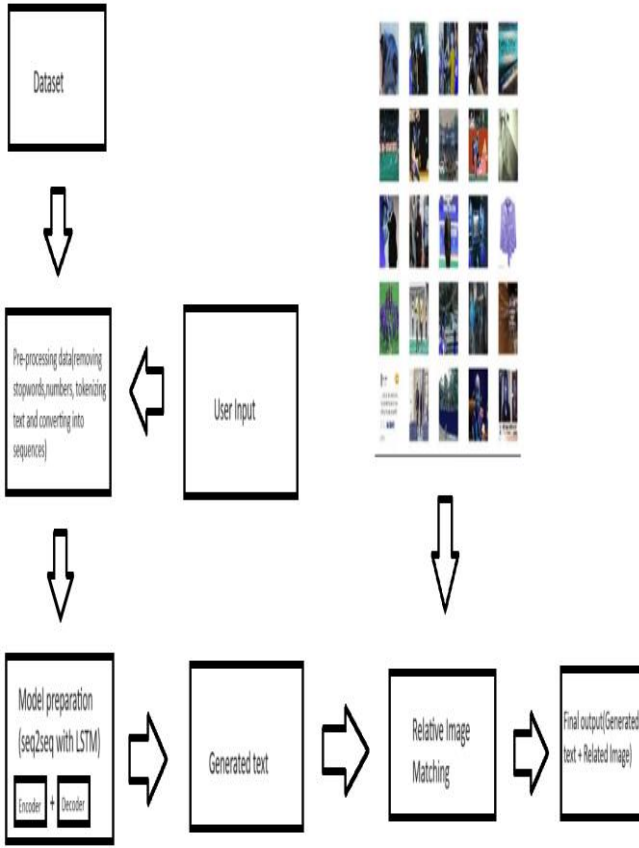
Deep learning relies heavily on the attention mechanism that Bahadanau et al. described. It helps identify important sentences for summary from original sources while reducing problems like recurrence in abstractive summarization. Sandeep et al. [17] proposed a topic-based image captioning system in their research article with the goal of obtaining semantic information from common items in photographs. The decoder produces semantically consistent captions by fusing these meanings with the text. Cross-attention is a multimodal summarization technique that Lu et al. [18] devised to integrate text and picture data. This approach, which uses CNN for image feature extraction and a transformer model for text encoding, captures the interaction between the two modalities to create a coherent multimodal summary. Haoran et al. [19] suggested use both intra- and inter- modality attention filters to extract important elements from text and image patches in a similar manner. Semantic problems are effectively addressed by generating a multimodal text summary through the employment of a hierarchical attention mechanism to merge these features.

In order to extract important characteristics at different granularities from both text and image modalities, Jieli et al. [20] uses a multimodal hierarchical attention mechanism in their investigation. Making non-redundant multimodal multimodal summaries that are pertinent to the issue. To direct the multimodal summarization training process, textures, forms, and object parts, which makes them summaries, this objective function considers both textual and visual modalities. Zhang et al. [22] describe a transformer model-based location- aware multimodal summarization technique that modalities Through the stacking of numerous fusion blocks, these locations are integrated with text, and the resulting encoded data is then subjected to bilinear pooling to generate a multimodal summary. In a similar vein, Rafi et al. [23] presents the idea of core word fusion attention, which determines crucial words by combining word importance and semantic analysis methods. An attention mechanism is then used to gather and integrate visual information from images. In the end, the approach creates abstractive summaries that capture pertinent information by concentrating on important core words and employing visual data. As can be seen from the above literature review, multimodal summarizing has attracted more attention in research than text- only summary. Nevertheless, none of the publications particularly address and integrate both difficulties at the same time.

Consequently, our goal in multimodal abstractive text summarization is to address and resolve these related problems.

### III.METHODOLOGY

This study's main goal is to use textual and visual data to generate multimodal abstractive text summaries. The sequential-to-sequential model is used to accomplish this, along with the integration of an attention mechanism to extract global features from images for contextual understanding and GloVe embeddings to capture text global features for semantic representation. The Multimodality Image Text (MIT) layer integrates these two information modalities, combining contextual information from images with semantic information from text to generate multimodal abstractive summaries. Figure 2 represents the workflow of the abstractive multimodal summarization, in which the text is preprocessed and added to the glove embeddings to train the model that helps in generating the sequences and the decoder model is responsible for the decoding of the sequences to human readable format.



#### A . Global Features of Text

One of the main purposes of GloVe (Global Vectors for Word Representation) embedding is to represent words as dense vector representations while still capturing semantic links within sentences. These graphics are produced using statistical co-occurrence data that has been collected from large text datasets. They provide a dispersed representation of words that captures their contextual relationships and subtleties of meaning. Contextual signals from nearby words or full phrases are taken into account throughout this

process, which helps the model efficiently understand both local contextual nuances and global semantic linkages. GloVe representations, in contrast to certain embeddings, encode words according to their general co-occurrence patterns throughout the training corpus. Consequently, they offer a static representation of word semantics that is independent of the particular context in which the words are used.

In spite of this, GloVe embeddings efficiently capture word associations and similarity, enabling natural language processing models to understand and make use of semantic connections.

#### B. Preprocessing Textual Data

In the exploration of the expansive MSMO dataset, a thorough examination was conducted across its 44,000 text files. These files contained various textual elements, including distinct sections like body, head, and summary. Focusing on the body and summary sections for analysis, the text from each file was systematically transferred into a structured CSV file. This method ensured that the content was accurately categorized under the corresponding headings. This systematic approach facilitated the preparation of the data for subsequent analysis, streamlining the process and enhancing clarity in the extracted information.

```
[2] df = pd.read_csv("/content/drive/MyDrive/rafibodysummaryfinal (1).csv", encoding="iso-8859-1").reset_index(drop=True)
```

|       | @body  | @summary  |
|-------|--|---|
| 0     | Five years ago Giuliana Rancic had a double ma...  | In 2012 she had a double mastectomy as she bat... |
| 1     | Genetically-modified cows are being bred witho...  | Scientists have created the cows to protect fa... |
| 2     | She carved out a role as a pathbreaking profes...  | Ivanka Trump was a constant presence at her fa... |
| 3     | A couple who made \$ 1million from cocaine blew... | Carl and Donna Honey-Jones used drug dealing e... |
| 4     | Though he did n't come dressed as a scary curs...  | Joel Lynch , an Iowa State University senior ...  |
| ...   | ...  | ...   |
| 44002 | It was once the largest and wealthiest city in...  | Pictures taken in 1800s in Constantinople have... |
| 44003 | West Ham vice-chairman Karren Brady insists sh...  | West Ham vice-chairman Karren Brady shares the... |
| 44004 | The echoes of demonstisation rang through an N...  | Mujahid is accused of arranging meetings and L... |
| 44005 | The mother of a serial killer who stabbed thre...  | Joanna Dennehy stabbed three men to death befo... |
| 44006 |  | @summary  |

44007 rows x 2 columns

After obtaining the raw textual data, the next crucial step was preprocessing, essential for ensuring data integrity and relevance. This involved various transformations aimed at refining the text for analysis. Firstly, extraneous elements such as stop words and numerical figures were removed to reduce noise. To maintain uniformity, all text was converted to lowercase, minimizing discrepancies due to letter casing variations. Furthermore, superfluous special characters and redundant whitespace were meticulously excised to enhance text clarity and coherence.

With the data thus processed and sanitized, the subsequent challenge lay in partitioning it into distinct subsets suitable for training, testing, and validation purposes. Adhering to best practices in data science, I meticulously segregated the dataset into these subsets, ensuring a balanced distribution of samples to facilitate robust model evaluation. This involved employing established techniques for randomization and

stratification to mitigate the risk of bias and ensure the generalizability of the subsequent analyses.




The text data was carefully processed, divided, and converted into a structured format using tokenization. GloVe embeddings were then incorporated to enrich the data's semantic context and improve its analytical potential. This fusion of techniques created a multi-dimensional representation of the text, laying the groundwork for further analysis and modeling tasks.

### C. Extraction of Relative Images

In the realm of multimodal data analysis, particularly concerning datasets comprising images paired with textual descriptions, aligning predicted summaries with actual captions is crucial for comprehensive understanding and analysis. To facilitate this alignment process, cosine similarity emerges as a valuable tool in natural language processing (NLP).

Researchers utilize cosine similarity to quantitatively measure the similarity between a predicted summary and image captions. This approach facilitates the extraction of relevant textual information aligned with the summary, enhancing data interpretation coherence. By harmonizing textual and visual elements, cosine similarity aids in efficiently identifying and extracting pertinent sentences from captions. This method improves dataset interpretability and utility for downstream tasks such as image extraction. Image captioning, content retrieval, and semantic understanding. Its integration into research methodologies underscores its significance in facilitating comprehensive analysis and interpretation of multimodal datasets. Table 1 represents the sample outputs of the Multimodal abstractive summarization.

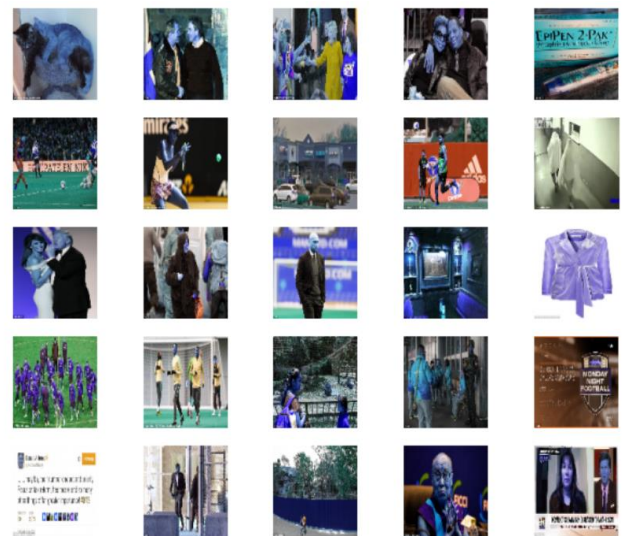
Table 1. Some of the outputs of generated multimodal summarisation

| Generated Multi-modal Summary  | Reference Summary   | Matched Images  |
|--|---|---|
| Police police alerted<br>usus us involved crash<br>tongariro forest park | Police alerted possible<br>gunman us grant high<br>school oklahoma          |  |
| Police police uk us<br>involved crash<br>tongariro forest park<br>new    | Dakar rally winding<br>way across miles three<br>countries south<br>america |  |
| Riley alessandra found<br>perfect fashion<br>formula                     | Tassels pompoms<br>fringing scarves nt keep<br>necks warm                   |  |

### D. Generating Multimodal Summaries

By starting its input state with the multimodal context vector, the LSTM decoder creates summaries using the sequential-to-sequential model. It learns to generate coherent and grammatically correct summaries by training on large-scale text datasets, where it picks up on language patterns, sentence structures, and semantic linkages. This method guarantees that the decoder fully understands the input data, making it easier to include pertinent information from many sources while preserving semantic consistency throughout phrases. As a result, by creating abstract representations, it is essential in providing multimodal abstractive text summaries of the highest Caliber. Figure 3 represents the images that are available in the images dataset which consists of around 8625 images which are related to different topics such as: politics, sports, protests.....etc.

Fig. 3. Experimental Setup



### E. Dataset

The MSMO dataset [8] is used to efficiently address multimodal abstractive text summarization difficulties because of its rich content, which includes text articles, photographs, and accompanying captions. With a large training set of 293,965 articles, 10,355 validation pairs, and 10,261 test pairs, this dataset has an average of 6 captions and 723 linked articles. Figure 4 represents the graphical representation of the words present in the textual data that is taken from an article from MSMO dataset.

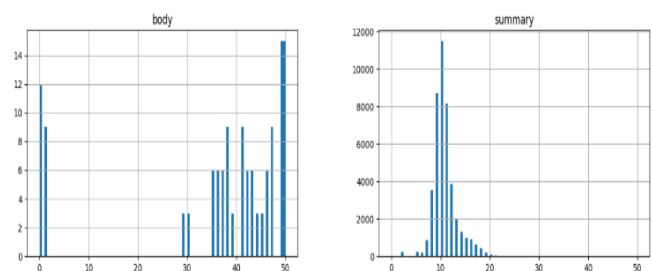


Fig. 4. Counting Words in Textual Data



## F. Accuracy

The accuracy of this project is calculated in term of rouge score, which means ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of automatic summarization and machine translation outputs by comparing them to reference summaries or translations. The ROUGE scores measure the overlap between the generated text and the reference summaries in terms of n-gram overlap, word overlap, and other metrics. Below fig-5 represents resultant scores.

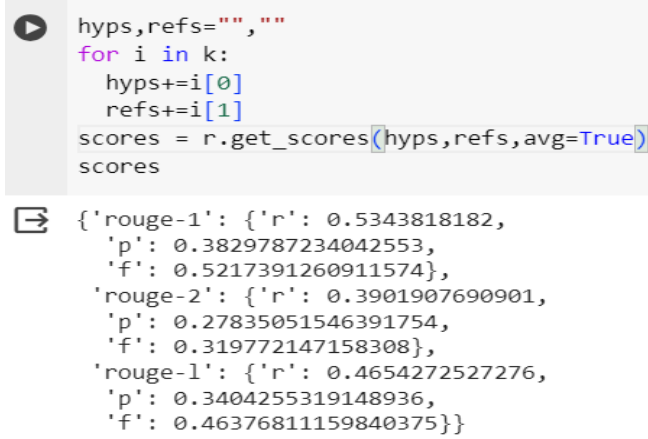


Fig. 5. Rouge Scores

The accuracy and loss are visualized as below while evaluating the seq2seq with LSTM model.

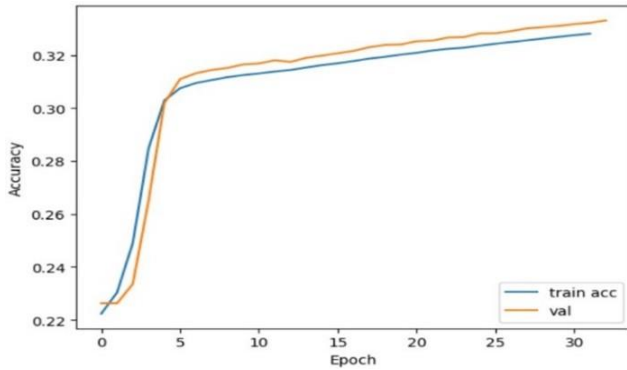


Fig. 6. Accuracy results

Figure-6 represents the accuracy standards that are visualized in a graphical format

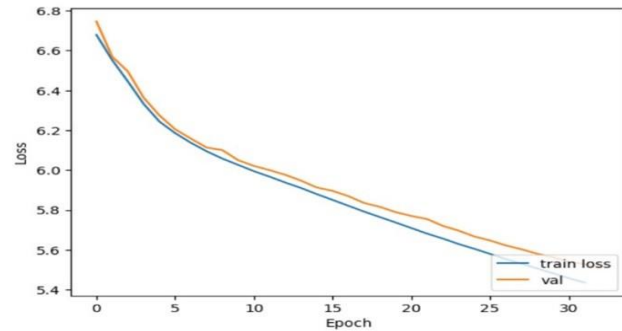


Fig. 7. Loss result

Figure-7 represents the loss standards that are visualized in a graphical format

## IV. IMPLEMENTATION

GloVe embeddings are used to extract global text features during the implementation phase, which makes it easier to identify semantics by converting words into numerical vectors. These vectors capture the semantic links between words and are usually 300 dimensions in size.

We employed the MSMO dataset as the core text source, meticulously curating and preprocessing the data to ensure its suitability for our research. Utilizing a sequence-to-sequence (seq2seq) architecture with LSTM units, we crafted an encoder-decoder model. The encoder generates summary sequences based on user input, while the decoder transforms them into human-readable formats, facilitating the creation of concise and coherent summaries crucial for effective communication.

To optimize our models, we harnessed the computational power of Google Colab, specifically leveraging a Tensor Processing Unit (TPU) machine for scalability and efficiency. Further enhancing performance, we fine-tuned parameters such as the Adam optimizer with a learning rate of 0.001 and a dropout rate of 0.3. Additionally, we constrained the length of generated summaries to ten words, balancing brevity with informativeness to cater to users' preferences for succinct yet informative content.

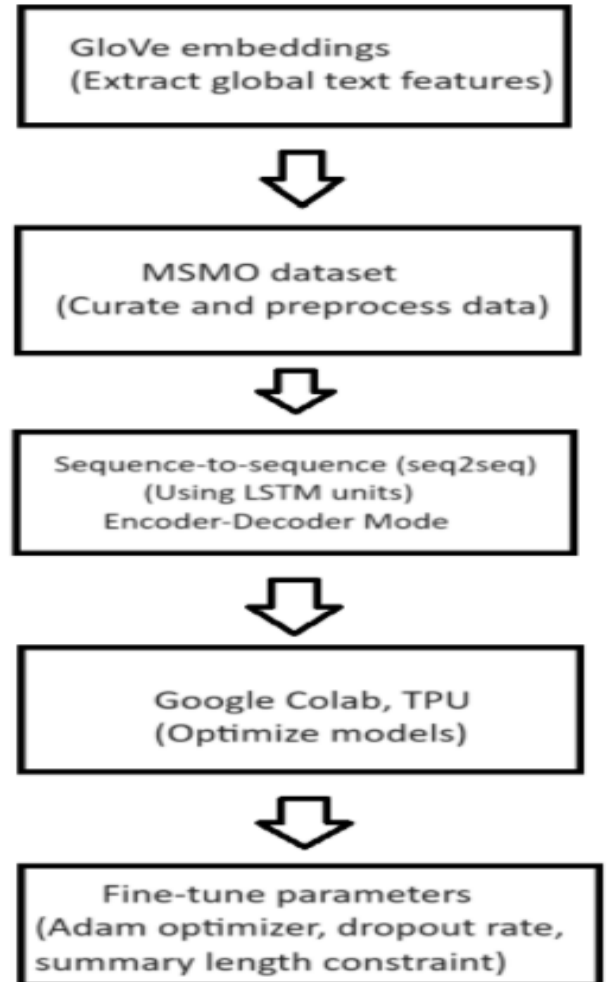


Fig. 8. Implementation workflow

Figure-8 represents the step-by-step work-flow of the implementation process. It explains the architecture of the steps involved and what are the steps are actually involved in the process.

## V. RESULT

### A. Table of Resultant scores

In order to create multimodal abstractive summaries, the sequential-to-sequential model is utilized to generate semantic relationship summaries. These summaries are then assessed using ROUGE [26] ratings. The ROUGE metrics provide a numerical evaluation of the degree of alignment between the reference summary and the generated summary. Greater quality summaries that include more information from the reference are indicated by higher ROUGE scores. These metrics are widely used in summarizing competitions as a standard benchmark and to evaluate automatic summary systems. The multimodal summaries that are produced are evaluated in comparison to the most recent findings. ROUGE scores, which measure the overlap between system-generated summaries and human reference summaries, are often used metrics to assess the efficacy of automatic summarizing systems. These scores include Rouge1, rouge 2, rouge n. Table 2 represents the rouge score levels of the output models that are developed previously and the present model, showing the difference in the values.

Table 2. State of art results compared with our model

| S.No | Method   | R-1   | R-2   | R-L   |
|------|--|-------|-------|-------|
| 1    | Abstractive Text-Image summarization using Multi-Modal RNN [9] | 32.64 | 12.08 | 23.88 |
| 2    | MSMO [11]  | 41.11 | 18.31 | 37.74 |
| 3    | Abstractive Text Summarization Using Multimodal Information    | 52.33 | 34.18 | 45.33 |
| 4    | Proposed Model   | 53.43 | 39.01 | 46.54 |

### B. Resultant Output-1(Summary Generation)

```
# Single Prediction
# striker Christiano Ronaldo has become manchester united top scorer in this season
text = np.array(["striker Christiano Ronaldo has become manchester united top scorer in this season"])
#text_2_seq = np.array(x_tokenizer.texts_to_sequences([text]))
#print(type(text_2_seq),text_2_seq)
seq = x_tokenizer.texts_to_sequences(text)
pad_seq = pad_sequences(seq, maxlen=max_text_len, padding='post')

gen = decode_sequence_seq2seq_model_with_just_lstm(
    pad_seq.reshape(1, max_text_len), encoder_model,
    decoder_model)

print(gen)

1/1 [=====] - 0s 260ms/step
1/1 [=====] - 0s 99ms/step
1/1 [=====] - 0s 50ms/step
1/1 [=====] - 0s 56ms/step
1/1 [=====] - 0s 82ms/step
1/1 [=====] - 0s 46ms/step
1/1 [=====] - 0s 40ms/step
1/1 [=====] - 0s 37ms/step
manchester united news latest old trafford
```

Fig. 9. Generated Summary

The Figure-9 shows the generated summary for the given sentence. This summary is generated to simplify the user's input string whether it might be a paragraph or a sentence like in fig-9. This summary is generated by encoder model as mentioned in the implementation and methodology sections and the end generated sequence of encoder model is sent to decoder model to decode it and get a human-readable text.

### C. Resultant Output-2(Matching Relative Image)

Relevant Image-->  
Matched with cosine similarity of: 0.5163977794943222



Fig.10 . Relevant matched Image

The Figure-10 displays the relevant image that is matched to the summary generated from output-1. This image is taken from a large dataset which consists of 8625 images in it. The algorithm used for matching the resultant image is Cosine Similarity as mentioned in the methodology section previously.

In addition to summarizing user input and providing relevant images using cosine similarity matching, this project may also incorporate natural language processing techniques to enhance the accuracy and coherence of the summaries. By analyzing the semantic meaning and context of the input, the system can generate more informative summaries tailored to the user's needs.

Furthermore, it could implement an image retrieval system that retrieves images from a database based on the keywords extracted from the input text, ensuring the relevance and diversity of visual content. The project might also explore methods for evaluating the quality of the summaries and images, possibly through user feedback mechanisms or automated evaluation metrics.

Overall, the integration of these advanced techniques aims to offer users a comprehensive and engaging experience, facilitating better understanding and retention of information.

## VI. CONCLUSION

Research into multimodal abstractive text summarization, a cross-modal undertaking needing a deep understanding of input data, has been made possible by the investigation of abstractive summarization difficulties. In order to preserve sentence meaning, the suggested method extracts global text features using GloVe embeddings. Furthermore, contextual elements are extracted from text-related images to improve the comprehension of key visual features. After that, a sequence-to-sequence model is trained on these fused features to provide accurate multimodal abstractive text summaries. In an effort to improve user pleasure when seeing text-related images, the research also makes recommendations for future directions, such as expanding multimodal summarization to include multimodal output. The resultant score generated is pretty low which is almost only around 50% which is not suitable for real time implementation among people.

## VII. REFERENCES

- [1] Mao, Xiangke, Hui Yang, Shaobin Huang, Ye Liu and Rongsheng Li. "Extractive summarization using supervised and unsupervised learning." *Expert Syst. Appl.* 133 (2019): 173-181.
- [2] Du, Yongping, Qingxia Li, Lulin Wang and Yanqing He. "Biomedical- domain pre-trained language model for extractive summarization ." *Knowl. Based Syst.* 199 (2020): 105964.
- [3] Cai X, Shi K, Jiang Y, Yang L, Liu S. HITS-based attentional neural model for abstractive summarization. *Knowledge-Based Systems.* 2021 Jun 21;222:106996.
- [4] Chopra, Sumit, Michael Auli and Alexander M. Rush. "Abstractive Sentence summarization with Attentive Recurrent Neural Networks." *Association for Computational Linguistics* (2016).
- [5] Nallapati, Ramesh, Bowen Zhou, C'icero Nogueira dos Santos, C, aglar Gu'lc,ehre and Bing Xiang. "Abstractive Text summarization using Seq- to-seq RNNs and Beyond." *Conference on Computational Natural Lan- guage Learning* (2016).
- [6] Xiao, Liqiang, Hao He and Yaohui Jin. "FusionSum: Abstractive sum- marization with sentence fusion and cooperative reinforcement learning." *Knowl. Based Syst.*(2022).
- [7] , Kexin, Logan Lebanoff and Fei Liu. "Abstract Meaning Representation for Multi-Document summarization" (2018).
- [8] Zhu, Junnan, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong and Chang Liang Li. "Multimodal summarization with Guidance of Multi- modal Reference." *AAAI Conference on Artificial Intelligence*(2020).
- [9] Chen, Jingqiang and Hai Zhuge. "Abstractive Text-Image summariza- tion Using Multimodal Attentional Hierarchical RNN." *Conference on Empirical Methods in Natural Language Processing* (2018).
- [10] Chen, Jingqiang, Hai Zhuge. "Extractive summarization of documents with images based on multimodal RNN." *Future Gener. Comput. Syst.*(2019).
- [11] Zhu, Junnan, Haoran Li, Tianshan Liu, Yu Zhou, Jiajun Zhang and Chengqing Zong. "MSMO: Multimodal summarization with Multimodal Output." *Conference on Empirical Methods in Natural Language Pro- cessing* (2018).
- [12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed representations of words and phrases and their com- positionality". In *26th International Conference on Neural Information Processing Systems(NIPS'13)*.
- [13] Mosa, Mohamed Atef, Arshad Syed Anwar and Alaa El-deen Hamouda. "A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms." *Knowl. Based Syst.* 163 (2019): 518-532.
- [14] Gupta, Som and S. K. Gupta. "Abstractive summarization : An overview of the state of the art." *Expert Syst. Appl.* 121 (2019): 49-65.
- [15] Rafi, Shaik and Ranjita Das. "RNN Encoder And Decoder With Teacher Forcing Attention Mechanism for Abstractive summarization ." 2021 IEEE 18th India Council International Conference (INDICON) (2021): 1-7.
- [16] Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *CoRR abs/1409.0473* (2014).
- [17] Dash, Sandeep Kumar, Shantanu Acharya, Partha Pakray, Ranjita Das and Alexander Gelbukh. "Topic-Based Image Caption Generation." *Arabian Journal for Science and Engineering* 45 (2020): 3025-3034.
- [18] Lu, Qiduo, Xia Ye and Chenhao Zhu. "MTCA: A Multimodal sum- marization Model Based on Two-Stream Cross Attention." 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI) (2022): 594-601.
- [19] Li, Haoran, Junnan Zhu, Tianshan Liu, Jiajun Zhang and Chengqing Zong. "Multimodal Sentence summarization with Modality Attention and Image Filtering." *International Joint Conference on Artificial Intel- ligence* (2018).
- [20] Qiu, Jielin, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao and Hailin Jin. "MHMS: Multi- modal Hierarchical Multimedia summarization ." *ArXiv abs/2204.03734* (2022).
- [21] Mukherjee, Sourajit, Anubhav Jangra, Sriparna Saha and Adam Jatowt. "Topic-aware Multimodal summarization ." *AAACL/IJCNLP* (2022).
- [22] Zhang, Zhengkun, Jun Wang, Zhe Sun and Zhenglu Yang. "LAMS: "A Location-aware Approach for Multimodal summarization (Student Abstract)." *AAAI Conference on Artificial Intelligence* (2021).
- [23] Rafi, S., Das, R. Topic-guided abstractive multimodal summarization with multimodal output. *Neural Comput & Applic* (2023).
- [24] Szegedy, Christian, Vincent Vanhouck e, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna. "Rethinking the Inception Architecture for Com- puter Vision." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 2818-2826.
- [25] Kuchaiev, Oleksii and Boris Ginsburg. "Factorization tricks for LSTM networks." *ArXiv abs/1703.10722* (2017).
- [26] Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." *Annual Meeting of the Association for Computational Linguistics* (2004).
- [27] Rafi, Shaik and Ranjita Das. "A Linear Sub-Structure with Co- Variance Shift for Image Captioning." 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI) (2021): 242-246.

