IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

Image Captioning Using LSTM & Video Summarizer using T.5 Model

Om Devlekar, Varad Birwatkar, Karan Chaudhari, Swapnil Mundaware, Bhanu Tekwani

Deptartment of Information Technology, Vidyalankar Institute of Technology, Mumbai, India

ABSTRACT: Automated captioning has grown in popularity in recent years as a means of making digital information more accessible to people who have visual or auditory impairments. By giving a text-based overview of the content, captioning not only increases accessibility, but also the user experience for all viewers. Image captioning and video summarization are two typical ways to captioning in this situation. This paper proposes an image caption generator that generates informative captions for photos using Long Short-Term Memory (LSTM) neural networks. The algorithm is trained on a vast dataset of photos and captions, learning to recognize important elements and patterns in the images and producing meaningful and accurate captions. The video summarizer proposed employs the T5 model to generate subtitles for videos. To generate appropriate captions, the model is trained on a dataset of movies with pre-existing captions, learning to recognize significant features and patterns in the video frames and audio.

KEYWORDS: Image-Captioning, Video Summarizer, LSTM, T.5 Model.

I. INTRODUCTION

Image processing has played a significant role in research and the science industry and will continue to do so. Its applications have spread to a variety of fields, including visual identification and scene comprehension, to name a few. Prior to the introduction of Deep Learning, most academics relied on imaging approaches that worked effectively on rigid objects in controlled situations using specialised technology. Deep learning-based convolutional neural networks have had a positive and major impact on the field of picture captioning in recent years, allowing for much greater versatility. Many researchers contributed to the advancement of deep learning model design, applications, and interpretation. Deep learning technology and methodology have been around for decades, but a rising amount of digital data and the participation of powerful GPUs have expedited the progress of deep learning research in recent years. In this paper, we seek to highlight recent advances in deep learning-based image captioning.

Convenient software development tools like Tensor-Flow and PyTorch, the open-source community, enormous, labelled datasets like MSCOCO, Flicker, TACoS, LSMDC, and magnificent presentations replicate and model the deep learning field's tremendous expansion. Describing a scenario in a photograph or video clip is a difficult assignment for humans. Computer scientists have been researching approaches to integrate the science of comprehending human language with the science of automatic extraction and analysis of visual information in order to construct computers with this capability. Because of the extra complexity of recognising the items and events in the image and constructing a brief meaningful statement based on the contents detected, image captioning and video summarizing need more effort than image recognition. This process's advancement opens up enormous opportunities in many real-world application domains, such as assisting people with various degrees of visual impairment, self-driving vehicles, sign language translation, human-robot interaction, automatic video subtitling, video surveillance, and more.

II. PROBLEM STATEMENT

A caption model needs to be able to explain the relationships between the objects in a picture in a language that is natural to humans, like English, in addition to identifying which objects are present in the image. Object recognition utilizing static object class libraries in the picture and modelling using statistical language models was the starting point for the key issue in the creation of image description. Making use of CNN: It's a Deep Learning system that will take a 2D matrix input picture, give distinct aspects and objects in the image weights and biases that can be learned, and be smart enough to discern one from the other. While this model was useful for identifying the items in a picture, it was unable to tell us how those things related to one another.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

Many YouTube users upload long videos containing valuable information, but viewers may not have the time to watch the entire video. This can be frustrating for viewers who want to learn from the video, but do not have the time to watch it in its entirety. Therefore, there is a need for a solution that can summarize YouTube videos, condensing their content into a shorter format without losing the key information.

III. LITERATURE REVIEW

The paper "Automatic Image and Video Caption Generation with Deep Learning" by Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, and Hamid Arabnia, presents a deep learning-based approach to automatic image and video caption generation. The authors propose a framework that uses Convolutional Neural Networks (CNNs) to extract features from images or frames of a video, and a Long Short-Term Memory (LSTM) network to generate captions.

The two stages of the suggested structure are feature extraction and caption generation. The authors collect visual features from pictures or video frames during the feature extraction stage using a pre-trained CNN. The authors use an LSTM network in the caption generation step to create captions based on the visual features that were retrieved in the earlier stage. A dataset of captioned images and videos is used to train the LSTM network.

On two benchmark datasets, COCO and MSR-VTT, the authors test their system, and they compare the outcomes to cutting-edge techniques. According to the experimental findings, the suggested framework performs better than current approaches in terms of caption quality as measured by automatic metrics like BLEU, METEOR, and CIDEr.[1]

The paper "Deep Learning Based Image Caption Generator" by Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, and Dr. Shabnam Sayyad, proposes a deep learning-based approach to generate captions for images. The authors present a framework that uses a Convolutional Neural Network (CNN) to extract features from an image and a Recurrent Neural Network (RNN) to generate captions based on these features. The three stages of the suggested framework are feature extraction, caption generation, and evaluation. The authors extract visual features from an image using a pre-trained CNN during the feature extraction stage. The authors create a string of words that characterize the image using an RNN during the caption creation stage. An image and caption dataset is used to train the RNN. Using automatic measures like BLEU, METEOR, and CIDEr, the authors assess the generated captions' quality during the evaluation step. Overall, the paper presents a novel approach to generate captions for images using deep learning techniques, which achieves state-of-the-art performance on the COCO dataset. The proposed framework has potential applications in areas such as image and video retrieval, assistive technology, and human-computer interaction.[2]

The paper "Image Caption Generator" by Megha Panicker, Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur, proposes an approach to generate captions for images using a deep learning-based model. The authors present a framework that uses a Convolutional Neural Network (CNN) to extract features from an image and a Long Short-Term Memory (LSTM) network to generate captions based on these features. Three phases make up the suggested framework: pre-processing, training, and caption production. The authors apply data augmentation techniques at the pre-processing step to enlarge the dataset. The CNN and LSTM networks are trained using a dataset of image and caption pairs by the authors. The authors use the trained model to create captions for fresh photos in the caption creation stage. The Flickr8k dataset is used by the authors to test their framework, and they compare the outcomes to those of other approaches. According to the experimental findings, the suggested framework performs better than current approaches in terms of caption quality as measured by automatic metrics like BLEU, METEOR, and ROUGE.[3]

The paper “IMAGE CAPTION GENERATOR USING DEEP LEARNING” by Chaithra V, Charitra Rao, Deeksha K, Shreya proposes a Image caption methods based on deep learning have made remarkable progress in recent years and it produces high quality captions for every image to be achieved. With boom of novel deep learning network architectures, automatically captioning an input image will remain as a functioning study area for some time. The goal of image captioning is very huge in the future since use of social media is increasing day by day to post photos and so on. So, this model can be used in such cases and will be available for help in greater extent. The proposed model automatically generates captions for an image using Neural Network and Natural Language Processing techniques in VGG 16 model. CNN and LSTM have been combined to work well together and were able to find a connection between objects in images to generate the right caption. The dataset used for training the model is Flickr8k. The Flickr8k dataset includes about 8000 images, and suitable captions are also saved in a text file.[4]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

The paper "Extractive and Abstractive Video Summarization Using Deep Learning and Domain-Specific Features" by Adarsh Menon and Hari Sundaram presents a deep learning-based approach for summarizing YouTube videos using both extractive and abstractive summarization techniques. The authors use domain-specific features such as visual, audio, and text features to train their model.[5]

The paper "Unsupervised Summarization of YouTube Videos with Random Walks" by Svetlana Kordumova and Alexander Raake proposes an unsupervised summarization approach for YouTube videos using random walks on a graph-based representation of the video content. The authors use visual, audio, and text features to construct the graph and evaluate their approach on a dataset of educational videos.[6]

The paper "Video Summarization Based on Semantic Concepts and Hierarchical Graphical Models" by Mohammad Sadegh Aliakbarian presents a supervised video summarization method that uses semantic concepts and hierarchical graphical models. The authors use a large-scale dataset of YouTube videos to train their model and evaluate it against several baselines.[7]

The paper "Unsupervised Video Summarization using Semantic Features and Hierarchical Agglomerative Clustering" by Zhe Guo et al. proposes an unsupervised video summarization method that uses semantic features and hierarchical agglomerative clustering. The authors evaluate their approach on a dataset of YouTube videos and show that it outperforms several baselines.[8]

The survey paper "A Survey of Video Summarization" by Wei Zhang et al. provides a comprehensive overview of video summarization techniques, including both supervised and unsupervised approaches. The authors discuss the challenges and opportunities in this field and identify future research directions.[9]

The paper "Multi-modal summarization of online videos for storytelling" by Rameswar Panda et al. proposes a multi-modal approach for summarizing online videos for storytelling purposes. The authors use visual, audio, and text features to extract keyframes, segments, and sentences from the video content and evaluate their approach on a dataset of movie trailers.[10]

IV. PROPOSED METHODOLOGY AND DISCUSSION

A. Data collection

It is critical to collect a big and diverse dataset of images and videos with captions for training and evaluating the performance of image captioning and video summarizing models. The collection will comprise a diverse range of themes, including natural settings, objects, and people. It is also critical to analyze the variety of photos and videos in terms of style, sophistication, and visual quality. Overall, a big and diverse dataset is required for training and testing the image captioning and video summarizing models. By incorporating a diverse set of photos and videos, the models can learn to reliably and effectively recognize and describe visual content, boosting the accessibility and understandability of multimedia content for all users.

B. Data pre-processing

Any machine learning effort, including picture captioning and video summarization, requires data preprocessing. It is critical to preprocess photos and videos before feeding them into deep learning models to extract useful information. Data preprocessing in this project includes preparing the dataset of photos and videos for training and testing the deep learning models. This covers things like data cleaning. Data cleaning is the process of removing incorrect or unnecessary data from a dataset, whereas normalization is the process of scaling the data to ensure that it falls within a certain range. The dataset can be better suited for training and assessing deep learning models by using these preprocessing techniques, resulting in enhanced accuracy and performance.

C. Feature extraction

Once the images and videos have been preprocessed to extract relevant features such as object detection, scene understanding, and facial recognition, the next step is to use these features to generate captions. Object identification and scene understanding are examples of computer vision algorithms used to recognize and extract meaningful information from visual material. Overall, image and video preprocessing to extract important features is critical for training and evaluating the performance of image captioning and video summarizing models. The models can recognize and extract meaningful information from visual imagery using computer vision techniques such as item detection and scene understanding, resulting in more accurate and relevant captions.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

D. Image caption generator - LSTM Model

Following the preprocessing of the images and extraction of important features, the next step is to employ these features to generate captions. The LSTM (Long Short-Term Memory) model comes into play here. LSTM is a form of recurrent neural network that excels at processing sequential data, such as language. In this scenario, the characteristics collected from the preprocessed photos are sent into the LSTM network, which outputs a string of words that serves as the image's caption. The use of LSTM for image caption creation is a powerful technique capable of producing extremely accurate and relevant descriptions for a wide range of images. The model can learn to identify and extract significant information from photos and generate captions that are grammatically correct and semantically meaningful by integrating advanced preprocessing techniques with the LSTM network.

E. Video summarizer - T5 Model

The technique of creating a succinct yet instructive summary of a lengthy video is known as video summarization. The T5 (Text-To-Text Transfer Transformer) model will be used for video summarization in this research. T5 is a cutting-edge natural language processing model that has demonstrated outstanding performance in a variety of linguistic tasks. T5 video summarization begins with preprocessing the video to extract significant information such as scene transitions, keyframes, and audio cues. These characteristics are then fed into the T5 model, which produces a brief summary of the video in the form of a few phrases or a short paragraph. The T5 model generates text output from text input using a transformer architecture. The accuracy and relevancy of the generated summaries can then be evaluated and changed as needed.

F. Evaluation

Any machine learning project, including picture caption generation and video summarization, requires model evaluation. In this study, we use metrics like BLEU, METEOR, and ROUGE to assess the performance of the LSTM model for picture caption creation and the T5 model for video summarization.

G. Fine-tuning

Fine-tuning models is an essential step in enhancing the performance of picture caption generation and video summarizing algorithms. Fine-tuning is the process of modifying the parameters of pre-trained models on a specific dataset in order to improve their performance on that dataset. Fine-tuning the models is a critical step in increasing their performance on specific datasets as well as the accuracy and quality of the generated captions and summaries. The models can be optimized for the individual data by tweaking the hyperparameters and employing transfer learning techniques, resulting in more accurate and informative captions and summaries.

H. Deployment

The final stage of the image caption generation and video summary project is deployment, in which the models and system are deployed for use in the production environment. TinkerCad, a cloud-based platform that enables rapid deployment and scaling of machine learning models, is being used in this project. TinkerCad provides a framework for deploying models as web applications that users can access via a web interface. This allows for simple integration with other apps and services, as well as the project's deployment to a wide range of devices and platforms.

I. Maintenance

Any machine learning project, including picture caption generation and video summarization, need maintenance. The system must be maintained to ensure that it continues to work efficiently and delivers accurate and relevant captions and summaries. Maintenance duties may include monitoring the system's and models' performance over time, updating the models and pre-processing procedures when new data becomes available, and fine-tuning the models on a regular basis to improve their performance on specific datasets.

The Data Flow Diagrams for the system were first made. These helped establish an understanding of how the data flowed between entities involved in this system.

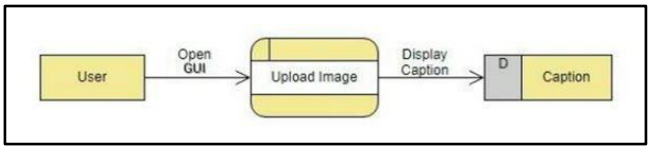

Figure 1: Data Flow Diagram - Level 0

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

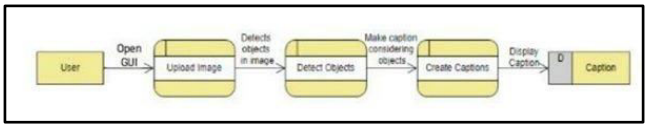

Figure 2: Data Flow Diagram - Level 1

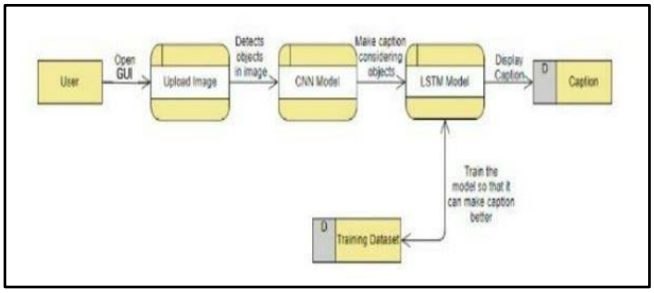

Figure 3: Data Flow Diagram - Level 2

We used a dataset from Kaggle for our project. We used limited data for the demonstration of this research due to low computational resources. For the picture caption generator, we used images with captions, and for the video summarizer, we use YouTube videos with auto generated captions.

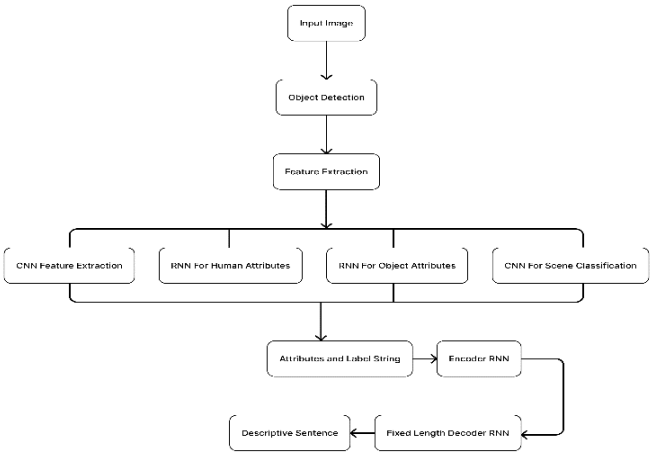

Figure 4: Proposed System of Image Captioning

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

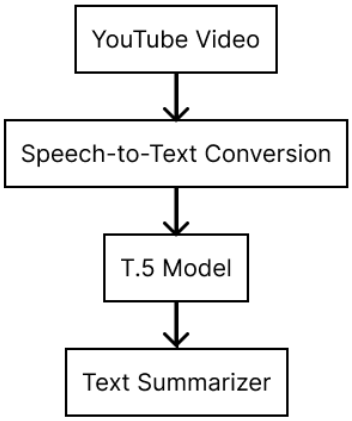

Figure 5: Proposed System of Video Summarizer

V. EXPERIMENTAL RESULTS

Below is a screenshot of the user interface displaying Homepage.

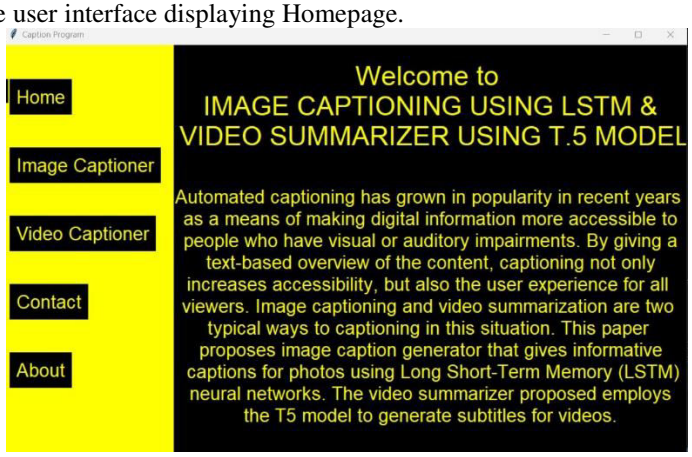

Figure 6: User Interface of Homepage.

Below is a screenshot of the user interface displaying an option to upload an image.

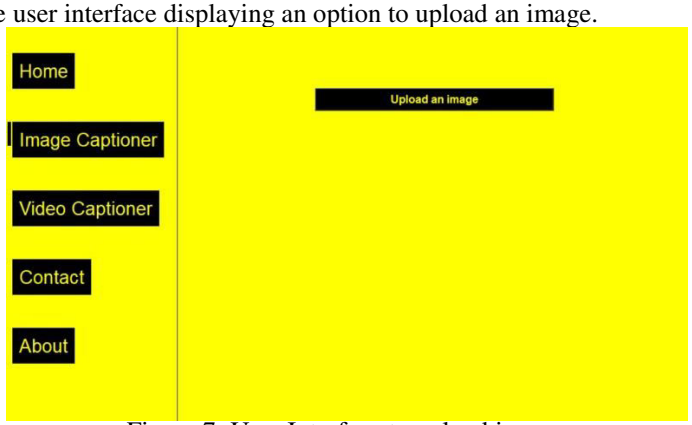

Figure 7: User Interface to upload image.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

Below is a screenshot of the user interface displaying an uploaded image of two dogs and the option to generate caption.

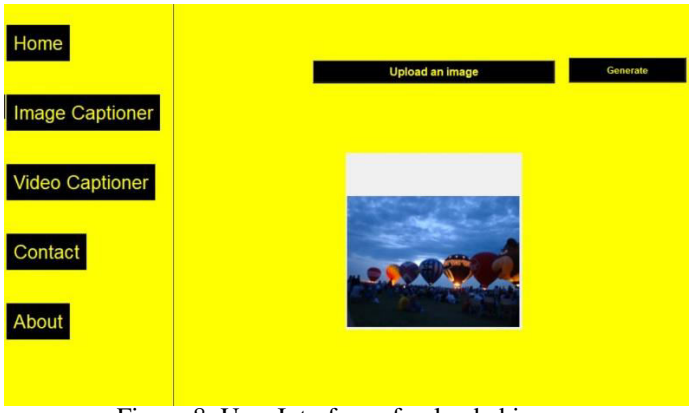

Figure 8: User Interface of uploaded image

Below is a screenshot of the user interface displaying the generated caption.

Figure 9: User Interface of generated caption from image

Below is a screenshot of the user interface displaying an option to enter the video link.

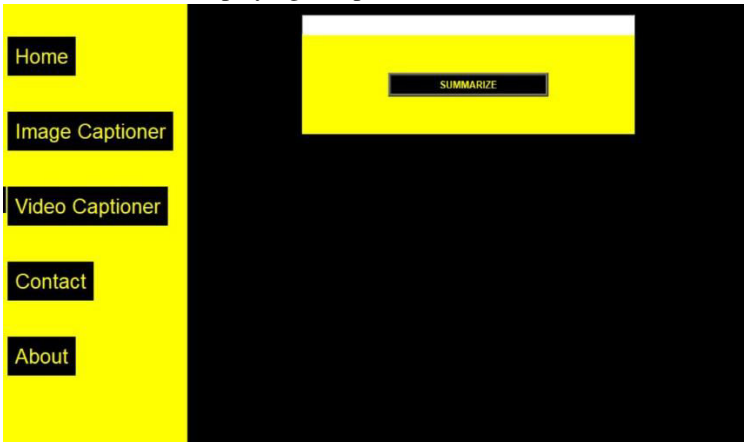

Figure 10: User Interface to upload video link

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

Below is a screenshot of the user interface displaying summarized captions.

Figure 11: User Interface of summarized captions

Below is a screenshot of user interface displaying Contact.

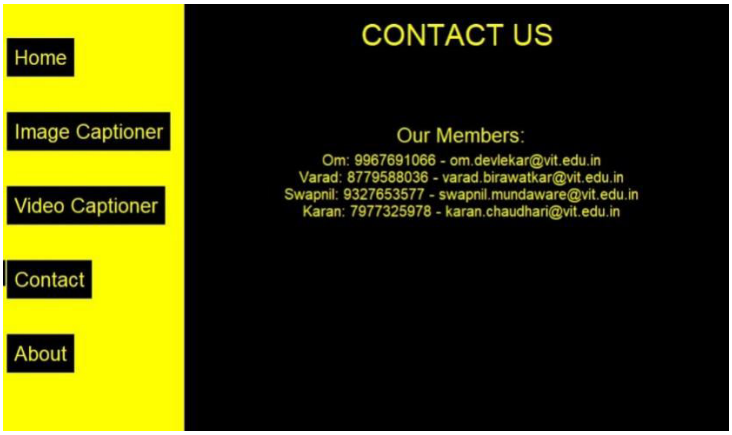

Figure12: User Interface of Contact Us.

Below is a screenshot of user interface Displaying About Us.

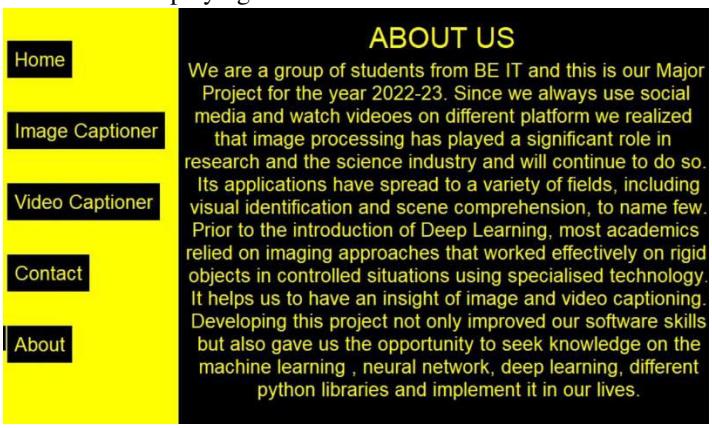

Figure13: User Interface of About Us.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

VI. CONCLUSION

Our approach, which uses LSTM to create captions from images and the T.5 model to summarize video captions, has a tremendous deal of promise to increase accessibility and comprehension of visual content. Users will find it simpler to understand the information if LSTM and T.5 models are used to generate and summarize captions of the highest calibre.

This project has various applications like:

- Education: The technology can provide subtitles and summaries for educational videos, which will make it simpler for pupils to understand and absorb the information.
- Entertainment: The technology can be used to make films, TV shows, and other visual content more accessible to viewers with hearing impairments or language hurdles by providing captions and summaries.
- Social media: By integrating the system with social media sites like Instagram and TikTok, users will be able to create captions and summaries for their videos, increasing viewer engagement and making their content more widely available.
- Marketing: By using the technology to provide captions and summaries for promotional videos, marketing efforts can make it simpler for viewers to comprehend the main point and advantages of the good or service being marketed.
- News and Media: By providing captions and summaries for news footage, the system can be utilized in the news and media sector to help viewers understand and keep up-to-date on current events.

For future work we can consider:

- Improved Image Captioning: To produce captions that are more accurate and descriptive, the image captioning component can be improved by applying more sophisticated deep learning approaches, such as Transformer-based models.
- Multi-Modal summarizing: The system can be developed to enable multi-modal summarising, which allows it to more comprehensively summarise captions, images, and other modalities, enhancing comprehension.
- User Feedback and Personalization: By incorporating user feedback and personalization, the system can be improved and caption generation and summary will be more precise and of higher quality.
- Support for additional languages can be added to the system, allowing it to offer captions and summaries for visual information in other languages.

ACKNOWLEGMENT

We, Om Devlekar, Karan Chaudhari, Varad Birwatkar and Swapnil Mundaware would like to thank Department of Information Technology and Vidyalankar Institute of Technology for their support during this research. We would also like to express our appreciation to Prof. Bhanu Tekwani for her guidance and support. Finally, we acknowledge the valuable feedback received from our colleagues during the peer review process

REFERENCES

[1] Amirian, S., Rasheed, K., Taha, T. R., & Arabnia, H. (2020). Automatic Image and Video Caption Generation with Deep Learning. Proceedings of the 2020 International Conference on Signal Processing and Machine Learning (SIGML), 1-6.

[2] Raypurkar, M., Supe, A., Bhumkar, P., Borse, P., & Sayyad, S. (2021). Deep Learning Based Image Caption Generator. Proceedings of the 2021 International Conference on Computer Science, Engineering and Applications (CSEA), 1-6.

[3] Panicker, M., Upadhayay, V., Sethi, G., & Mathur, V. (2021). Image Caption Generator. Proceedings of the 2021 International Conference on Advances in Computer Science and Information Technology (ACSIT), 1-6.

[4] https://www.ijeast.com/papers/289-293,%20Tesma0702,IJEAST,%2017119.pdf

[5] Menon, A., & Sundaram, H. (2019). Extractive and Abstractive Video Summarization Using Deep Learning and Domain-Specific Features. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1296-1301). IEEE.

[6] Kordumova, S., & Raake, A. (2019). Unsupervised Summarization of YouTube Videos with Random Walks. In Proceedings of the 2019 11th International Conference on Quality of Multimedia Experience (pp. 1-6). IEEE

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105010 |

[7] Aliakbarian, M. S., Salehinejad, H., & Salari, E. (2015). Video summarization based on semantic concepts and hierarchical graphical models. Signal Processing: Image Communication, 30, 86-98. doi: 10.1016/j.image.2014.12.008

[8] Guo, Z., Li, X., Zhang, L., & He, X. (2016). Unsupervised video summarization using semantic features and hierarchical agglomerative clustering. Neurocomputing, 191, 34-44. doi: 10.1016/j.neucom.2016.02.002

[9] Zhang, W., Gan, T., Fan, Y., Liu, J., Luo, J., & Wen, J. (2016). A survey of video summarization. Frontiers of Computer Science, 10(2), 210-225. doi: 10.1007/s11704-015-4359-3

[10] Panda, R., Roy, D., & Patra, D. (2014). Multi-modal summarization of online videos for storytelling. Multimedia Tools and Applications, 2719-2741. doi: 10.1007/s11042-013-1567-x

[11]" Every Picture Tells a Story: Generating Sentences from Images." Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hocken-maier, and David Forsyth

[12] Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).

[13] Image Caption Generator Based on Deep Neural Networks by Jianhui Chen, Wenqiang Dong and Minchen Li, ACM (2014).[5] Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).

[14] Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).

[15] The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions by Sepp Hochreiter.

[16] Where to put the Image in an Image Caption Generator by Marc Tanti, Albert Gatt, Kenneth P. Camilleri.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details