IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

Image Caption Generator using Deep Learning

Farida Attar[1], Farzana Khan[2], Affan Ansari[3], Mujawar Saklen[4], Abubakr Shaikh[5], Danish Khan[6]
Assistant Professor, Department of Information Technology[1,2]
Students, Department of Information Technology[3,4,5,6]
M.H. Saboo Siddik College of Engineering, Byculla, Mumbai, India
farida.attar@mhssce.ac.in, farzana.khan@mhssce.ac.in, mohammedaffan.612008.it@mhssce.ac.in,
saklen.612037.it@mhssce.ac.in, abubakr.612047.it@mhssce.ac.in, danish.6119012.it@mhssce.ac.in

Abstract: *Image Caption Generation has always been a study of great interest to the researchers in the Artificial Intelligence department. Being able to program a machine to accurately describe an image or an environment like an average human has major applications in the field of robotic vision, business and many more. Automatic caption generation with attention mechanisms aims at generating more descriptive captions containing coarse to fine semantic contents in the image. This has been a challenging task in the field of artificial intelligence. In this paper, we present different image caption generating models based on deep neural networks, focusing on the various CNN techniques and analyzing their influence on the sentence generation. We have also generated captions for sample images and compared the different feature extraction and encoder models to analyse which model gives better accuracy and generates the desired results*

Keywords: CNN, RNN, LSTM , VGG, GRU, Encoder - Decoder, Image Captioning

I. INTRODUCTION

Generating accurate captions for an image has remained as one of the major challenges in Artificial Intelligence with plenty of applications ranging from robotic vision to helping the visually impaired.Long term applications also involve providing accurate captions for videos in scenarios such as security systems. “Image caption generator”: the name itself suggests that we aim to build an optimal system which can generate semantically and grammatically accurate captions for an image. Researchers have been involved in finding an efficient way to make better predictions, therefore we have discussed a few methods to achieve good results. Images are extensively used for conveying enormous amounts of information over the internet and social media and hence there is an increasing demand for image data analytics for designing efficient information processing systems. This leads to the development of systems with capability to automatically analyze the scenario contained in the image and to express it in meaningful natural language sentences. The BLIP (Blind Low-resolution Image Recognition) model is a computer vision model developed by Salesforce Research for recognizing objects and generating captions from low-resolution images. It's specifically designed to perform well on images with low resolution or poor quality, making it suitable for scenarios where high-resolution images are not available or feasible to use.

A good captioning system will be capable of highlighting the contextual information in the image similar to the human cognitive system. In recent years, several techniques for automatic caption generation in images have been proposed that can effectively solve many computer vision challenges. The primary purpose of the application is to demonstrate the capabilities of the BLIP model in generating captions for low-resolution images. It provides a user-friendly interface for users to interact with the model without requiring any deep knowledge of machine learning or computer vision techniques

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

II. LITERATURE SURVEY

Image Caption Generator using Deep Learning, NATIONAL INSTITUTE OF TECHNOLOGY SURATHKAL.
This model was trained and tested successfully to create accurate captions for the loaded photos. This is mostly a CNN and RNN model, in which the CNN will behave as an encoder and RNN will act as a decoder. This project is the application of deep learning. Using CNN and LSTM model we will first extract the features using CNN and generate the captions to the input image.The model makes use of scanned multiple frames of the image. Based on objects identified, an appropriate title is provided for the image.

TextMage: The Automated Bangla Caption Generator Based On Deep Learning,International Conference on Decision Aid Sciences and Application (DASA), 2022
In this paper we have presented an automated image captioning system, TextMage, that can perceive an image with a south Asian bias and describe it in Bangla. The model constructed for TextMage was heavily inspired from the first joint model ”Show and Tell: A Neural Image Caption Generator” from an architecture perspective.
Using the dataset that has been used in this paper and published, future works can include more newer methods for benchmark results.

Generating Image Captions using Deep Learning and Natural Language Processing, 9th International Conference on Reliability, Infocom Technologies and Optimization Amity University, Noida, India. Sep 3-4, 2021
Generation of image captions is found to be an essential tool as it can be used for dissimilar meadows for their different purposes. By generating captions for multiple images of the same file one can organize or arrange those files easily and quickly. The people who are blind or the ones who have low vision can understand the images by their caption or description provided by the image captioning process.

Image Caption And Speech Generation, Second International Conference on Augmented Intelligence and Sustainable Systems IEEE Xplore Part Number : CFP23CB2-ART ; ISBN : 979-8-3503-2579-9, 2023
The proposed deep learning approach is used to generate captions for the images and GTTS API for converting captions into speech. In the proposed method, the sequential API of Keras is used with TensorFlow as a backend for implementing the proposed deep learning architecture. The proposed model has achieved a BLEU score of 52.7%, a very high-quality, adequate translation.

III. PROPOSED SYSTEM

A. Problem statement

Every day the world is searching new techniques in the field of computer science to upgrade human limitations into machines to get more and more accurate and meaningful data. The way of machine learning and artificial intelligence has no negative slope it has only the slope having positive direction. To develop a system for users, which can automatically generate the description of the images using deep learning. The problem introduces a captioning task, which requires a computer vision system to both localize and describe silent regions in images in natural languages.

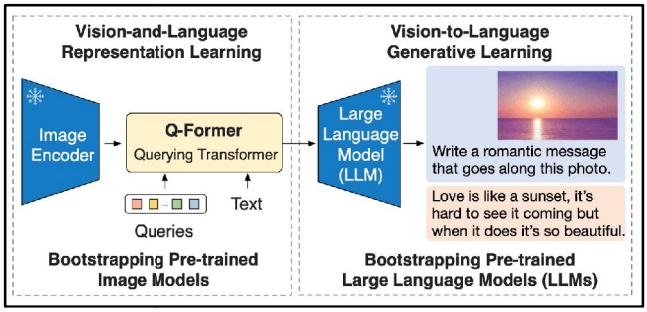
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

B. Methodology

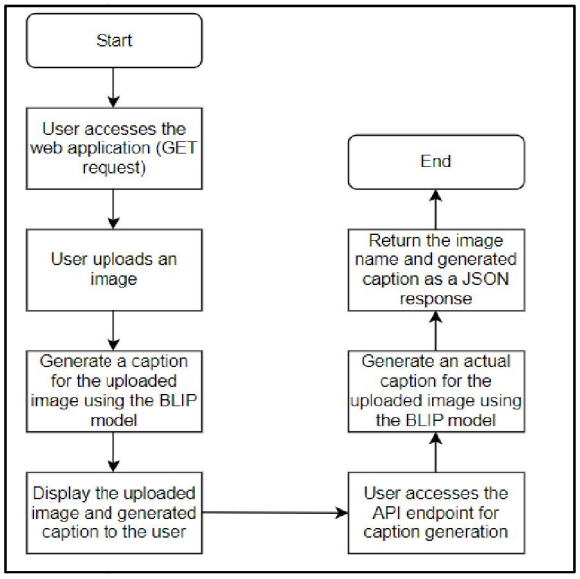
The model used in the provided code is the BLIP (Blind Low-resolution Image Recognition) model, specifically the implementation provided by the Hugging Face Transformers library. This model is used for generating captions from images, particularly designed to handle low-resolution images effectively. The BLIP model is pretrained on a large dataset of low-resolution images and their corresponding captions, allowing it to understand and describe the contents of such images accurately. Setting up the Flask Application: The code starts by importing necessary libraries and initializing a Flask application.

C. BLIP Model

Loading the BLIP Model
Image Caption Generation Functions:
generate_caption(image)
convert_image_to_base64(image)
Routes
GET Request
POST Request
API Route (/api/generate_caption)
Error Handling
Running the Flask Application



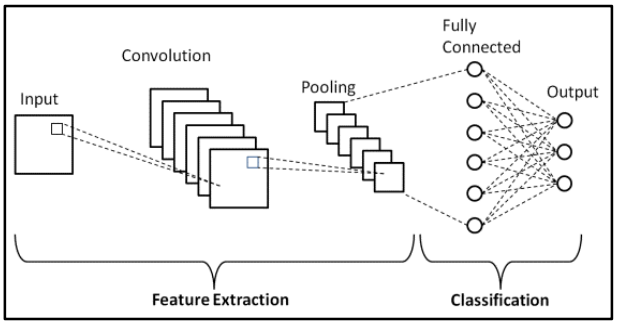
Flowchart



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

CNN Architecture

CNNs have revolutionized the field of computer vision and have been instrumental in achieving state-of-the-art performance on various image-related tasks, including object detection, image segmentation, and image captioning. They have also found applications in other domains such as natural language processing and speech recognition through techniques like transfer learning and feature extraction.n host 0.0.0.0 and port 5000.



IV. RESULT

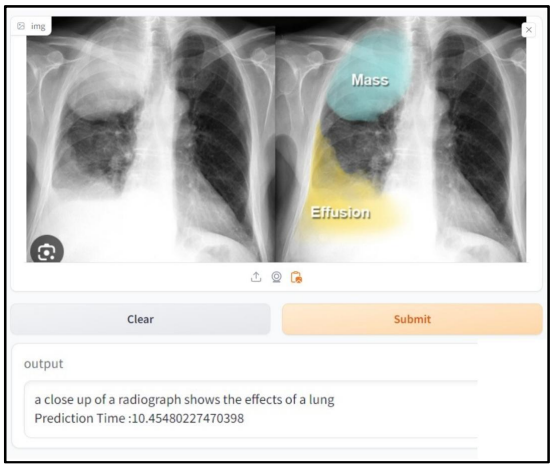
The input images along with respective generated captions are shown in the below figure. The result in terms of generated captions shows accuracy and reliability of the proposed model.

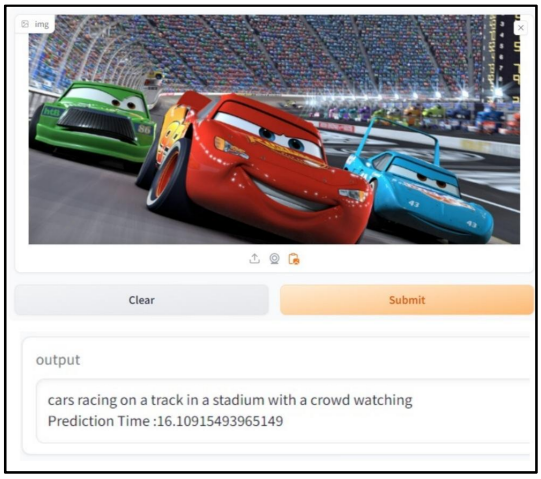
Once we upload the image of which caption is to be generated is uploaded and then our model will generate the captions automatically based on the image

But sometime, due to some inconvenience of modelling or poor image quality can decrease the accuracy of the system.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

The model is trained on predefined blip model and therefore the accuracy of model is much more than other caption generation model





V. CONCLUSION

We have presented a deep learning model that tends to automatically generate image captions with the goal of not only describing the surrounding environment but also helping visually impaired people better understand their environments.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

Our described model is based upon a CNN architecture. Our project will be used for successfully caption generation of an image and its implementation will be done in the next semester. So, finally this would be more helpful for the visually impaired people and in order to get more accuracy, we can use bigger datasets. Since the model utilizes its dataset to identify the objects, large sets of data are bound to improve the results, and more suitable captions can be generated.

VI. FUTURE WORK

Our model is not perfect and may generate incorrect captions sometimes. In the next phase, we will be developing models which will use Inceptionv3 instead of VGG as the feature extractor. Then we will be comparing the 4 models thus obtained i.e. VGG+GRU, VGG+LSTM, Inceptionv3+GRU, and Inceptionv3+LSTM . This will further help us analyze the influence of the CNN component over the entire network. The future work is to make the system more accurate in generating captions and error free and efficient.

AUTHORS' CONTRIBUTION

- Farida Attar: Conceptualization, Supervision.
- Farzana Khan: Supervision, Guidance.
- Affan Ansari:Methodology,Formal analysis, Resources.
- Saklen Mujawar: Formal analysis, Visualization, Validation.
- Abubakr Shaikh: Formal analysis, Investigation,Validation.
- Danish Khan: Investigation, Resources.

REFERENCES

[1] CS771 Project Image Captioning by Ankit Gupta , Kartik Hira, Bajaj Dilip.
[2] "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
[3] Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).
[4] 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 Vinyl, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).
[6] Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).
[7] The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions by Sepp Hochreiter.
[8] Where to put the Image in an Image Caption Generator by Marc Tanti, Albert Gatt, Kenneth P. Camilleri.
[9] Sequence to sequence -video to text by Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko.
[10] Learning phrase representations using RNN encoder-decoder for statistical machine translation by K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengi.
[11] TVPRNN for image caption generation .Liang Yang and Haifeng Hu.
[12] Image Captioning in the Wild: How People Caption Images on Flickr Philipp Blandfort, Tushar Karayil, Damian Borth, Andreas Dengel,German Institute for Artificial Intelligence, Kaiserslautern, Germany.
[13] Image Caption Generator Based On Deep Neural Networks Jianhui Chen ,Wenqiang Dong, Minchen Li ,CS Department. ACM 2014.
[14] BLEU: A method for automatic evaluation of machine translation. I