IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CAPTIONING OF IMAGES FOR VISUALLY IMPAIRED PERSON

Unnati Khanapurkar, B. Uthkarsh Jaiswal, K. Devashish Singh, Shaikh Asif Ahmed
Assistant Professor, Student, , Student, , Student
Department of CSE
Methodist College of Engineering and Technology, Hyderabad, India

Abstract: Image captioning has always been a great source of help for visually impaired by generating captions for the given image. But limiting it to the captions won't be that helpful for the visually challenged. In this project we tried to give voice to our generated captions by using the concept for TTS that is text-to-speech which is more impactful and practical. To accomplish caption generation and to implement Deep learning architecture we have used Tensorflow and Keras. it is challenging to generate captions that have right linguistic properties because it requires sophisticated level of image understanding. In this project, we used VGG16 deep learning architecture for the purpose of feature extraction for the images. The generated caption's quality and accuracy are evaluated using BLEU score.

Index Terms: TTS (test-to-speech), Deep Learning,Tensorflow, Keras, VGG-16, LSTM, BLEU Score

I. INTRODUCTION

Caption generation for the given image has always been a fascinating model in the era of computer vision and machine learning. Understanding images, extracting features and translating visual scenes in the images into plain text and converting the plain text into speech are all elements of this project. The goal of this project is to generate appropriate captions for a given image and converting them into speech which enhances the experience of visually impaired. To generate accurate captions for the given input image, our project uses convolution neural networks (CNNs) and recurrent neural networks (RNNs) or their branched models. CNNs like VGGs provide architecture for extracting features from the images and RNNs like LSTM model is used for the purpose of caption generation. For the purpose of extracting features, we use CNN model like VGG16 and removed the last two layers as those layers are used for the purpose of classification. RNN model like Long-short-term memories (LSTM) are especially helpful in sequential prediction. Images and captions from large, labelled data sets, such those in Microsoft COCO and Flickr, offer details about the events and objects. The generated captions' appropriateness and relevance to the data set captions can be evaluated using metrics such as BLUE score. Lastly, we can use TTS libraries like gTTS, pyttsx3 can be used for the purpose of text-to-speech conversion. Over the past few decades, a number of pertinent research articles have sought to conduct this task, but they have encountered a number of difficulties, including linguistic problems, cognitive absurdity, and irrelevant content. In order to get over those problems, we developed this method, which uses computer vision and natural language processing techniques to extract pertinent content and properly structure sentences, making the model usable by visually impaired individuals.

II. LITERATURE REVIEW

[1]An advanced approach to producing textual descriptions of images is examined in the review paper "Image Captioning Model Using Attention and Object Features to Mimic Human Image Understanding". The model is improved via attention methods and uses CNNs for feature extraction, YOLOv4 for object identification, and GRUs for sequence generation. Important datasets are Flickr30k and MSCOCO. GRUs have inadequate long-term dependency handling, which makes them difficult to use for long, cohesive descriptions even though they achieve great accuracy for short captions. The quality of captions has been greatly enhanced by incorporating sophisticated object identification and optimizing attention techniques.

[2]Using the MS COCO dataset, the publication "Deep Learning-Based Image Captioning for Visually Impaired People" uses RNNs for sequence creation and EfficientNet-B3 for feature extraction. The model helps those with vision impairments by turning generated captions into voice. Although it works well for text-to-speech conversion, its vocabulary is limited, which makes the captions less emotive. The significance of attention processes for enhanced caption accuracy has been underscored by recent developments. Vocabulary problems must be resolved in order to produce more diverse and contextually relevant captions, which will increase the usefulness of picture captioning systems for visually impaired users.

[3]Utilizing LSTMs for sequence generation and VGG16 for feature extraction, the study "Auto Image Caption Generator for Visually Impaired People" is assessed on the MS COCO (2014) dataset. While the model can produce precise captions to assist people with visual impairments, its usefulness is limited because it is unable to convert text to audio. Caption quality has

improved with recent breakthroughs in image captioning, including attention techniques. It is imperative to tackle the text-to-speech conversion limitation in order to develop a complete solution that can provide correct captions as well as audible descriptions, hence increasing its usability for visually impaired users.

[4]The paper titled "Image Caption Generator Using VGG and LSTM For Visually Impaired" evaluates the use of LSTMs for sequence creation and VGG16 for feature extraction using Flickr 8K and Flickr 30k datasets. For visually impaired users, the model's practical utility is limited because it cannot convert text to speech, even though it can generate captions that are mainly accurate. Caption quality has improved due to recent improvements, such as attention processes. It is imperative to tackle the text-to-speech conversion limitation in order to develop a complete solution that is more beneficial for visually impaired people by producing descriptive audio in addition to precise captions.

[5]ECANN, which combines CNNs with LSTMs for caption creation, is introduced in the publication "An Accurate Generation of Image Captions for Blind People Using Extended Convolutional Atom Neural Network" and is tested on the Freiburg Groceries and Grocery Store Dataset. Although the model performs well on this particular specific dataset, it struggles on more generic datasets. The quality of captions has improved with recent innovations, such as attention techniques. For more widespread practical applications, the model's generalizability must be improved outside particular domains, highlighting the necessity for additional study in this field.

III. REFERENCES

[1]Ponnaganti Rama Devi, Mannam Thrushanth Deepak, Morampudi Lohitha, M.Surya Chandra Raju , K. Venkata Ramana, “Image Caption Generator Using VGG and LSTM For Visually Impaired” International Journal of Advances in Engineering and Management (IJAEM) Volume 5, Issue 4 April 2023

[2]MuhammadAbdelhadie Al-Malla, Assef Jafar and Nada Ghneim, “Image captioning model using attention and object features to mimic human image understanding” Al-Malla et al. Journal of Big Data (2022).

[3]Thivaharan S, Vasanthakumar A, Vishal K, Vishnudarshan S, “Deep Learning Based Image Captioning In Regional Language Using CNN AND LSTM” International Research Journal of Engineering and Technology (IRJET Volume: 10 Issue: 05 | May 2023

[4]ChristopherElamri, Teun de Planque, “Automated Neural Image Caption Generator for Visually Impaired People”2016

[5]R. Kavitha , S. Shree Sandhya , Praveena Betes , P. Rajalakshmi , and E. Sarubala, “Deep learning-based image captioning for visually impaired people” E3S Web of Conferences 399, 04005 (2023)

[6]SimaoHerdade, Armin Kappeler, Kofi Boakye, Joao Soares, “Image Captioning: Transforming Objects into Words”, In proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019),2019.

[7]L. Zhou, C. Xu, P. Koch, and J. J. Corso, “Image caption generation with text conditional semantic attention,” arXiv preprint arXiv:1606.04621, 2016.