iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44289

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Audio Assistant Based Image Captioning System Using RLSTM and CNN

D. Akash Reddy [1], T. Venkat Raju [2], V. Shashank [3]
[1,2,3]*Dept of Computer Science Engineering, Sreenidhi Institute of Science and Technology*

Abstract-- As we know, visually impaired or partially sighted people face a lot of problems reading or identifying any local scenarios. To vanquish this situation, we developed an audio-based image captioner that will identify the objects in an image and form a meaningful sentence that gives the output in the aural form. Image processing is a widely used method for developing many new applications. It is also open source, so developers can use it easily. We used NLP (Natural Language Processing) to understand the description of an image and convert the text to speech. A combination of R-LSTM and CNN is used, which is nothing but a reference based long-short term memory which matches different text data and takes it as reference and gives the output. Some of the other applications of image captioning are social media platforms like Instagram, etc., virtual assistants, and video editing software.
Keywords-- Image processing, NLP, R-LSTM, CNN

I. INTRODUCTION

Given the fact that there are huge numbers of visually impaired people in the various parts of our world also called visually challenged or blind, which has no instant treatment, according to the latest data from the World Health Organization, more than 2.2 billion people have some kind of vision impairment. In this modern era where technology is evolving very fast, it's really important to make blind people understand the modern world and help them cope with their day-to-day lives. So, to perform these tasks, deep learning techniques are used, which are interlinked with machine learning and artificial intelligence. The combination of RLSTM and CNN is used to get the captions for the images. CNN is used to extract the image features; that is, it will identify the objects in the images and extract their features from the photographs, and RLSTM, which is based on RNN, acts as a decoder that is used to sequence the data in an ordered form and generate relevant captions for them. The challenges of deep learning are that each and every piece of data should be analyzed deeply as the interaction between human beings is done in natural language. It is really tough to make a human being understand by creating a system. The main aim of this paper is to find, recognize, and generate image captions using deep learning algorithms. Some more applications of this system are that it can be used in security cameras while there is suspicious activity going on. It can describe the scene. It can be used in newspaper articles to generate relevant captions. Some more applications are that self-driving cars can generate some kind of audio while parking and generating relevant captions, recommendations in editing and can be used on many social media platforms while posting content.

II. LITERATURE REVIEW

The System is integrated with two main layouts, CNN and RNN, describing attributes, relationships, and objects in the picture and putting them into sentences. CNN is a Convolutional Neural Network that decodes features from a given picture. CNN has three layers: a convolutional layer, also known as a convo layer, a polling layer, and a connected layer. The convolutional layer is also known as the image feature decoder, where it performs various operations and calculates the dot values. There is an architecture in the controller which converts all the -minus values to null values. The pooling layer is the layer where the size of the image will be reduced once the convo layer is executed. The fully connected layer is a layer that will be connected internally from one neuron to the other neuron, which involves biases and weights. Identification of the unusual features they have used on CNN, where only the starting word of the sentence gets matched, the sentence is carried forward the entire process with the previously gained word. It will be changed during this process. Information is kept in long short term memory so the sentence might not be correct every time to overcome this situation. LSTM is the approach to developing automated image captions for the user-given image. VCG16 is used for the encoding process and the RNN, which is a recurrent neural network, is used for getting proper description and RESNET is used for comparing the accuracy. It will predict each and every object in the image by using the pre trained model and it will also create some new words on its own based on the previously trained data this is time consuming and results in a relevant description creation. Multi-modal and deep visual semantics will be used to align sentence snippets which will be described through multi-modal analysis this will help to generate snippets and audio assistant caption generator model has been implemented by CNN and LSTM. This model is trained by using the Flickr data set, which will be helpful to generate captions accurately. By using the existing System it is error prone where the predicted output gets mismatched by its own captions

because it cannot identify the emotions in the picture and thus leads to errors.

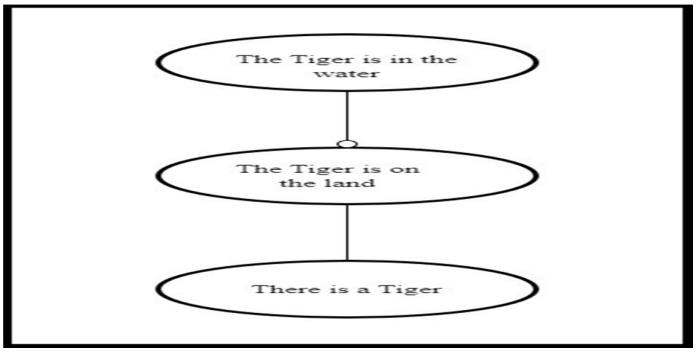


Fig. 1.1 Example of a predicted caption

III. SYSTEM ARCHITECTURE AND IMPLEMENTATION

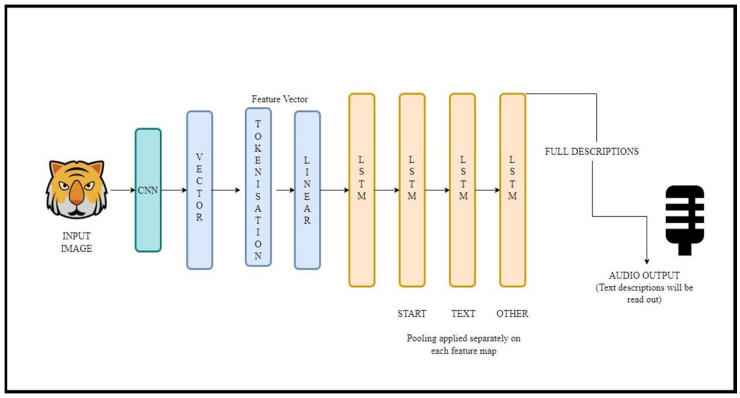


Fig. 3.1 Underlying Architecture of the proposed system

A. Four stages of the System

1) Feature Extraction: Extraction of features from the given image by tokenizing them and also generating vector features that are called embedding. The CNN model extracts the features from the actual picture, then the size is reduced and compressed so that it is compatible with LSTM, which is RNN.

```
features = {}
directory = os.path.join("/content", 'Images')

for img_name in tqdm(os.listdir(directory)):
    # load the image from file
    img_path = directory + '/' + img_name
    image = load_img(img_path, target_size=(224, 224))
    # convert image pixels to numpy array
    image = img_to_array(image)
    # reshape data for model
    image = image.reshape((1, image.shape[0], image.shape[1], image.shape[2]))
    # preprocess image for vgg
    image = preprocess_input(image)
    # extract features
    feature = model.predict(image, verbose=0)
    # get image ID
    image_id = img_name.split('.')[0]
    # store feature
    features[image_id] = feature
```

```
100% 8091/8091 [08:37<00:00, 16.63it/s]
```

Fig.3.2 Feature Extraction Process

2) Tokenization: The next stage after feature extraction is tokenization, which will identify each and every object in an image andassign it a keyword. This helps to decode the feature vector that is pushed to CNN. A sequence of words is created by predictionand then the captions are generated.

```
tokenizer = Tokenizer()
tokenizer.fit_on_texts(all_captions)
vocab_size = len(tokenizer.word_index) + 1
```

Fig. 3.3 Tokenization Process

3) Prediction: After the tokenization, the prediction will be done where the output gets generated before the prediction. The featurevectors will be decoded and generate a text caption.

```
from gtts import gTTS

# This module is imported so that we can
# play the converted audio
import os
#below code
# The text that you want to convert to audio
t=generate_caption("1000268201_693b08cb0e.jpg")

mytext = t[8:-7]
print(mytext)

# Language in which you want to convert
language = 'en'

# Passing the text and language to the engine,
# here we have marked slow=False. Which tells
# the module that the converted audio should
# have a high speed
myobj = gTTS(text=mytext, lang=language, slow=False)

# Saving the converted audio in a mp3 file named
# welcome
myobj.save("welcome.mp3")

# Playing the converted file
```

Fig. 3.4 Prediction of text output

4) Text to Speech: The final stage is all about getting the audio output from the predicted text. We used text to speech, where anaudio segment module is used to convert the text to speech.

```
from pydub import AudioSegment
#below code
song = AudioSegment.from_mp3("/content/welcome.mp3")
song.export("final.wav", format="wav")
```

Fig. 3.5 Text to Speech Conversion

IV. RESULTS

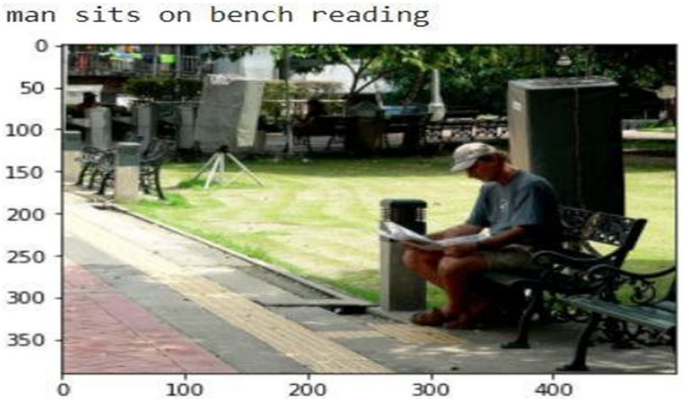


Fig. 4.1 Example of Predicted output

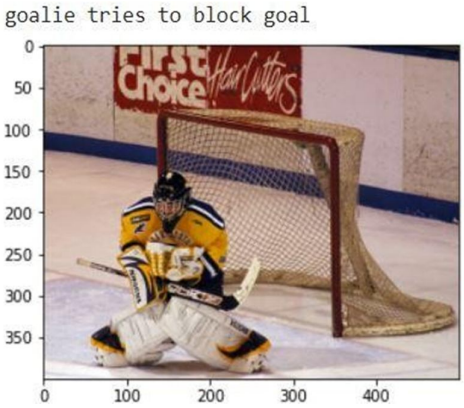


Fig. 4.2 Example of Predicted output

Fig. 4.3 Audio Output of the Predicted Image

V. CONCLUSION AND FUTURE WORKS

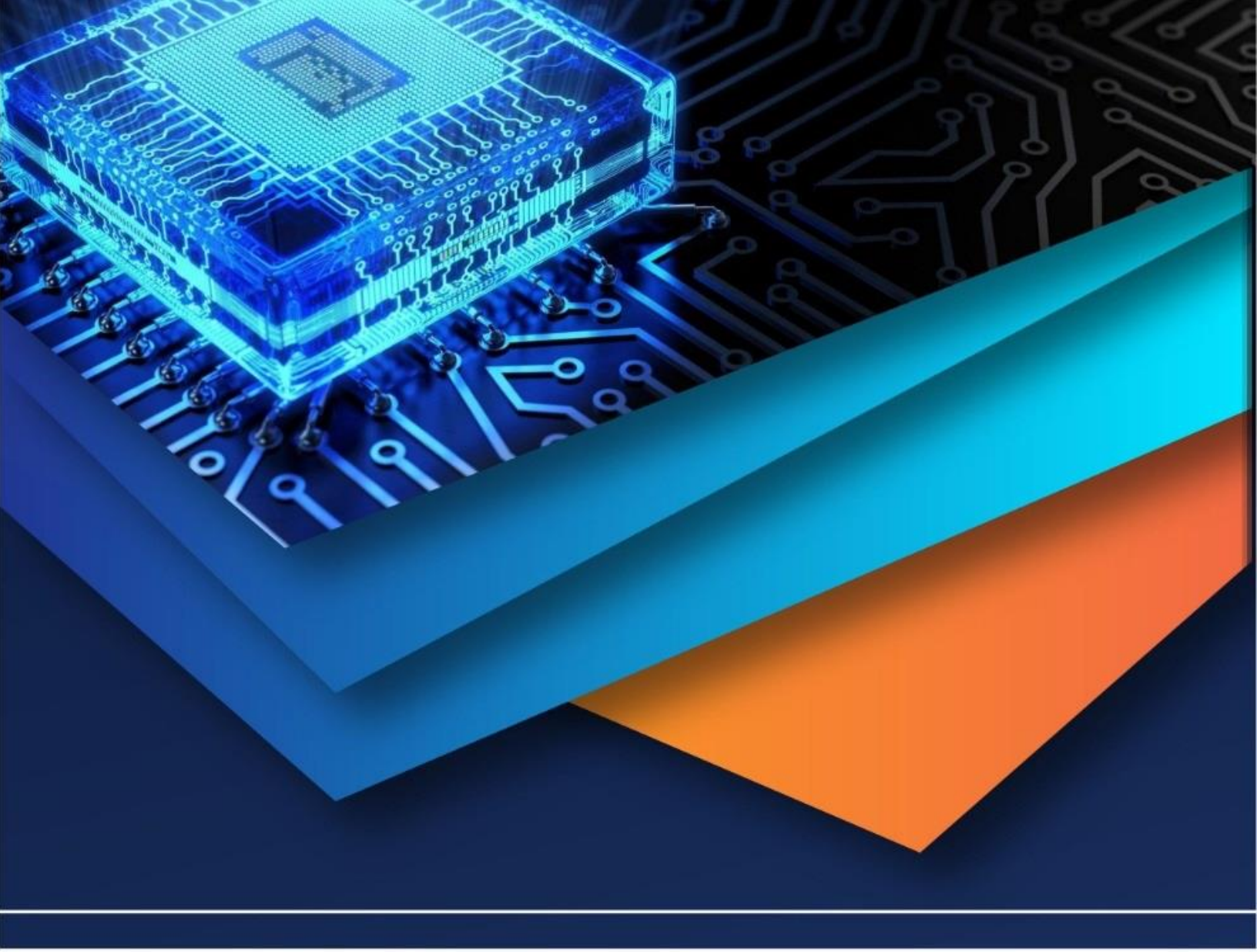
In this paper, we have successfully developed an audio-based assistant for image caption generation. This project will undoubtedlybenefit the blind because it will provide audio output. Some of the future scopes of this are that accuracy can be increased by training with large datasets and video caption generators can be developed. Modifications can be made in the identification of theimages to recognize them more clearly. Learning from the user feedback will also be added. This model represents the exact representation and works with most of the object pictures.

AKNOWLEDGEMENT

This research was made possible under the guidance, support, and motivation provided by our faculty, who have our esteem to pursue our interests in the field of image processing. We are thankful to Mrs. Doddi Srilatha, Assistant Professor, Dept of CSE, SNIST; and Dr. Sundaragiri Dheeraj, Assistant Professor, Dept of CSE, SNIST.

REFERENCES

[1] R. Subash November 2019: Automatic Image Captioning Using Convolution Neural Networks and LSTM.
[2] Karpathy, A., & Fei-Fei, L. (2017). Deep Visual Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis andMachine Intelligence, 39(4)664–676.
[3] Ushiku, Y., Harada, T., Kuniyoshi, Y.: "Automatic sentence generation from images". In: (ACM)Multimedia (2011)
[4] Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, Dr. Shabnam Sayyad (March 2021): Deep learning-based Image Caption Generator
[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2015): Show and Tell: A Neural Image Caption Generator
[6] McNee, S.M., Riedl, J. and Konstan, J.A. (2006) 'Being accurate is not enough: how accuracy metrics have hurt recommender systems', ACM CHI '06Extended Abstracts, ACM, pp.1103–1108.
[7] Linden, G., Smith, B. and York, J. (2003) 'Amazon.com recommendations: item to item collaborative filtering', IEEE Internet Computing, Vol. 7, No. 1,pp.76–80.
[8] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[9] Primkulov S., Urolov J., Singh M. (2021) Voice Assistant for Covid-19. In: Singh M., Kang DK., Lee JH., Tiwary U.S., Singh D., Chung WY. (eds) Intelligent Human Computer Interaction. IHCI 2020. Lecture Notes in Computer Science, vol 12615. Springer, Cham. https://doi.org/10.1007/978-3-030-68449-5_30



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY