e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

VISUAL ATTENTION BASED IMAGE CAPTIONING

Kavyasri MN[*1], Sthuthi B Iyer[*2], Omana Prabhakar[*3], Rahul Jain HV[*4], Rohan K[*5]

[*1,2,3,4,5]Department of Computer Science and Engineering Malnad College of Engineering, Hassan, India.

ABSTRACT

Natural language descriptions of images are produced by a process called visual attention-based image captioning. Utilising deep learning models, this method creates captions for photos that accurately convey their information by automatically learning their visual and semantic representations. However, because of their inability to offer speech out, conventional image captioning systems for those with visual impairments are typically constrained. In this project, we propose a novel method that blends voice output generation with visual attention-based image captioning to produce written and spoken explanations of images. Our approach uses long short-term memory (LSTM) networks with attention mechanisms to create descriptive captions and convolutional neural networks (CNNs) to extract visual information from the input image. Furthermore, we incorporate a text-to-speech synthesis system to convert the generated captions into speech that addresses the limitations of traditional image captioning systems, making them more inclusive and accessible to a wider range of users.

Keywords: Deep Learning; Image Captioning; Convolutional Neural Networks; Recurrent Neural Networks; ResNet; Long Short Term Memory.

I. INTRODUCTION

Visual attention-based image captioning is a fascinating research area that combines computer vision and natural language processing to automatically generate descriptive captions for images. The goal of image captioning is to develop algorithms that can comprehend the visual content of an image and generate human-like textual descriptions that accurately convey the image's essence. This technology has numerous applications, including content retrieval, image understanding, and enhancing the accessibility of visual information.

To overcome these limitations, visual attention mechanisms have been introduced in image captioning. An attention mechanism inspired by the human visual system allows the model to focus on specific areas of the image when generating captions. This selective attention allows the model to align relevant image regions with corresponding words in the generated captions, resulting in more accurate and contextual descriptions. While visual attention has greatly improved the quality of captions, many existing systems still lack an important aspect: audio output. Most captioning models generate captions in character form, creating accessibility challenges for visually impaired users. To make captions richer and more accessible, it's important to integrate screen reader generation into the caption creation process. This creates a significant barrier for inclusivity and restricts the potential applications of image captioning in assistive technologies and multimedia accessibility.

II. LITERATURE SURVEY

The literature review takes into account various references to existing projects similar to this project. Liu, Shuang Bai, Liang Hu, Yanli Wang, Haoran et al. by the proposed method. [1], two deep learning models, namely convolutional neural network (CNN-RNN) based caption, convolutional neural network and convolutional neural network (CNNCNN) based caption. The CNN-RNN-based framework uses convolutional neural networks for encoding and recurrent neural networks for decoding. With the help of CNN, the images here are converted into vectors and these vectors are called image features. They are provided as input to recurrent neural networks. RNN uses NLTK libraries to get the actual subtitles for the project. In a CNN-CNN-based framework, only CNN is used for both image encoding and decoding. Here a dictionary is used which is mapped to image functions to get the correct word for a given image using the NLTK library. This is how a flawless title is made. From many models, giving convolutional techniques at once is definitely faster than constantly streaming and repeatedly repeating these techniques in a train. The CNN-CNN model has less training time compared to the CNN-RNN model. CNN-RNN model has more training time because it is sequential but has less loss compared to

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

CNN-CNN model4 In the method proposed by Ansari Han et al [2], they used a code decoding model for subtitling images. Here they mentioned two more models of captions which are: search-based caption and template caption. Search-based subtitling is a process where training images are placed in one space and the corresponding created captions are placed in another area, now in a new scale, correlations are calculated for the test image and the more valuable correlation text of the captions is taken. from the title dictionary for the given image as a caption. Prototype-based description is the technique they do in this paper. Here they used Inception V3 as an encoder and used an attention mechanism to generate subtitles and GRU as a decoder. In the method proposed by Subrata Das, Lalit Jain et al [3] This model is mainly based on how deep learning models are used to title military images. It mainly uses a CNN-RNN based framework. They used a primitive model to encode images and reduce the gradient descent problem, they used long-short-term memory (LSTM) networks.

III. METHODOLOGY

As we have observed, traditional CNN-RNN models suffer from vanishing gradient problems that hinder efficient learning and training of recurrent neural networks.

To mitigate this gradient descent problem, this article proposes this model to improve caption generation efficiency and caption accuracy. Below is the architecture of our proposed model. [Fig. 1] In this paper, We are going to explain Resnet-LSTM model for image caption processing. Here Resnet architecture is used for encoding and LSTM for decoding. When the images are sent to Resnet (Residual Neural Network), image features are extracted then with the help of vocabulary that is built using training captions data, We will now train the model with these two parameters as input. Given below is the flow diagram of our proposed model in this paper.

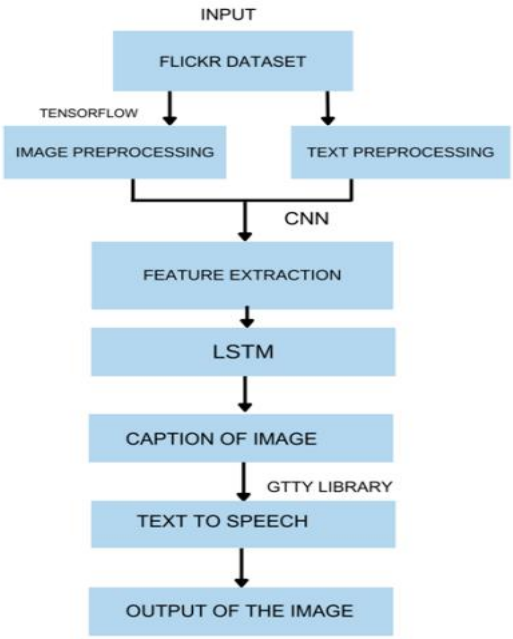


Fig 1: Block Diagram

The block diagram shown in is the high-level design architecture of the Visual-Attention based Image Captioning.

3.1 DATA SET COLLECTION

There are many data sets which can be used for training the deep learning model for generating captions in this paper we have used images from FLICKR 8K. We are using FLICKR 8K data set for training the model. FLICKR 8K data set works efficiently for training the Image Caption Generating Deep Learning Model. The FLICKR 8K data set consists of 8000 images in which 6000 images can be used for training the deep learning model and 1000 images for development and 1000 images for testing the model. Flickr Text data set consists of five captions for each given image which describes about the actions performed in the given images.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

3.2. IMAGE PREPROCESSING

After loading the data sets we need to preprocess the images in order to give this images as input to the ResNet. As we cannot pass different sized images through the Convolution layer like ResNet we need to resize every image so that they are in same size i.e; 224X224X3 .We are also converting the images to RGB by using inbuilt functions of cv2 library.

3.3. TEXT PREPROCESSING

After loading the captions for the images using FLICKR text data set we need to preprocess those captions so that there is no ambiguity or difficulty while generating vocabulary from the captions and also while training the deep learning model. We need to check whether the captions contain any numbers if found they must be removed and after that we need to remove white spaces and also missing captions in the given data set. We need to change all the upper case letters in the captions to the lower case in order to eliminate ambiguity during vocabulary building and training of the model. As this model will generate captions one word at a time and previously generated words are used as inputs along with the image features as input are attached at the starting and end of each of the caption to signal the neural network about the starting of the caption and ending of the captions during the training and testing of the model.

It is a low budget microcontroller board, which is programmable. While associating microcontroller board with computer we use USB cables, it has 14 digit input pins and 14digit output pins.

3.4. VOCABULARY BUILDING

We cannot pass the string captions directly as input to the neural network because neural network cannot process string as input so the captions which are in the form of strings to numbers for that process we need to build a vocabulary of numbers. This process is called encoding of captions. Firstly, After preprocessing of the captions given in the training data set we need to create new space where all words in every caption are taken. Now we have to give numbers to the words sequentially in the dictionary order. Now this space is called vocabulary library. With the help of this vocab library we will number each captions by numbering their words accordingly with vocab library. For a given caption each word is numbered by referring their values in already defined vocab library. For example: Let us consider a Vocab Library that we have built by numbering every unique word of the given training captions .Vocab Dictionary={a:1,aa:11,aam:2,……..,cat:450,…… ……..is:890…… ……..on:1120,…… ……..table:3770,…… ……..,the:5000,…… ……..} Now consider the caption={the cat is on the table}.Now this caption can be encoded into numbers using the dictionary and we can encode this caption as caption={5000 450 890 1120 5000 3770}. Now this encoded caption is passed into neural network(LSTM) for training the model to generate captions.

3.5. DEFINING AND FITTING THE MODEL

After collecting the data set and preprocessing the images and captions and building vocabulary. Now we have to define the model for generation of captions. Our proposed model is ResNet(Residual Neural Network)-LSTM(Long Short Term Memory) model. In this model Resnet is used as encoder which extract the image features from the images and converts them into single layered vector and pass them as input to LSTM's . Long Short Term Memory is used as decoder which takes image features as input and also vocabulary dictionary to generate each word of the caption sequentially.

3.5.1. RESNET

50 With the introduction of transfer learning (using knowledge gained in training network on one type of problem and applying the knowledge in another problem of same pattern) using deep neural networks like RESNET(Residual Neural Network) which is a pretrained model for many image recognition and classification became easy. We use this ResNet model in place of Deep Convolutional Neural Network because ResNet is a pretrained model on ImageNet data set to classify the images. So by using the concept of transfer learning we are reducing the computation cost and training time. If we have used CNN which is not pretrained then the computation cost would have increased and the model takes more time to learn. By using ResNet pretrained model we are also increasing the accuracy of the model. Resnet50 consists of 50 deep convolutional neural network layers. ResNet50 is the architecture of Convolutional Neural Network that we are using in Image Caption Generation Deep Learning Model. The last layer of Restnet50 is removed as it gives classification output

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

and we are accessing the output of the o layer before the last one in order to get the image features as output single layered vector because we don't need classification output in this paper. The ResNet is preferred compared to traditional deep convolutional neural networks because the ResNet contains residual blocks which have skip connections that ultimately reduce the vanishing gradient problem in CNN and ResNet also decreases the loss of input features compared to CNN. ResNet is having better performance and accuracy in classification of images and extracting image features compared to traditional CNN ,VGG. Below is the figure representing the working of ResNet block and its importance compared to traditional CNN.

3.5.2. RESNET

50 With the introduction of transfer learning (using knowledge gained in training network on one type of problem and applying the knowledge in another problem of same pattern) using deep neural networks like RESNET(Residual Neural Network) which is a pretrained model for many image recognition and classification became easy. We use this ResNet model in place of Deep Convolutional Neural Network because ResNet is a pretrained model on ImageNet data set to classify the images. So by using the concept of transfer learning we are reducing the computation cost and training time. If we have used CNN which is not pretrained then the computation cost would have increased and the model takes more time to learn. By using ResNet pretrained model we are also increasing the accuracy of the model. Resnet50 consists of 50 deep convolutional neural network layers. ResNet50 is the architecture of Convolutional Neural Network that we are using in Image Caption Generation Deep Learning Model. The last layer of Restnet50 is removed as it gives classification output and we are accessing the output of the o layer before the last one in order to get the image features as output single layered vector because we don't need classification output in this paper. The ResNet is preferred compared to traditional deep convolutional neural networks because the ResNet contains residual blocks which have skip connections that ultimately reduce the vanishing gradient problem in CNN and ResNet also decreases the loss of input features compared to CNN. ResNet is having better performance and accuracy in classification of images and extracting image features compared to traditional CNN ,VGG. Below is the figure representing the working of ResNet block and its importance compared to traditional CNN. The traditional CNN consists of Convolutional Layer, ReLU (Rectified Linear Unit)Layer and Pooling Layer. After passing the input through the traditional CNN the output is as follows: H(x)=f(wx+b) or H(x)=f(x) where H(x) is the output value and x is the input and w is the weights that are multiplied and b is the bias that is added and f() is the activation function. We see that input is not equal to output in case of traditional CNN .So if we apply this to extract image features or classify the images there will be error in the result and the accuracy is low. When it comes to ResNet model the skip connections are the core of this model. This skip connections are the short cut path that is followed by the gradient to reach the output layer. When this skip connections are applied the output is equal to input .i.e; H(x)=x+f(x) where f(x)=0 as represented by the above figure. So we can observe that when the images are passed through ResNet model the output will be equal to input without any bias or weights added .Thus when ResNet is used for image feature extraction there is no much loss of data or image features. Hence ResNet is better in extracting image features than traditional CNN model. The Residual Neural Network has various layers in it like Convolution layer, ReLU(Rectified Liner Unit),Batch Normalization, Pooling layer, and Flatten .The description and working of various layers of Residual Neural Network is given below: Convolution Layer, When the image is passed through the convolution layer ,then the image is converted to pixel values. Image filters(feature map) are applied on the image and convolution operation is performed. The output of this convolution layer is passed through the ReLU layer. When the convoluted image matrix is passed through the Rectified Liner Unit layer. It applies ReLU activation function and modifies the pixel values. ReLU Activation Function output={input when input>=0,0 when input

3.6 TEXT TO SPEECH

gTTs is a flexible tool that is used in converting the text into spoken words. The generated text can be then converted into mp3 file. The gTTs tool supports many languages. We need to pass the text to the gTTs object that is an interface to Google Translate's Text-to-Speech API.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

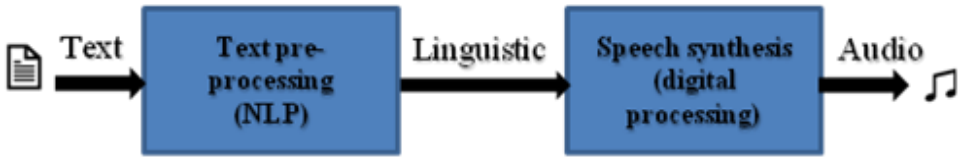
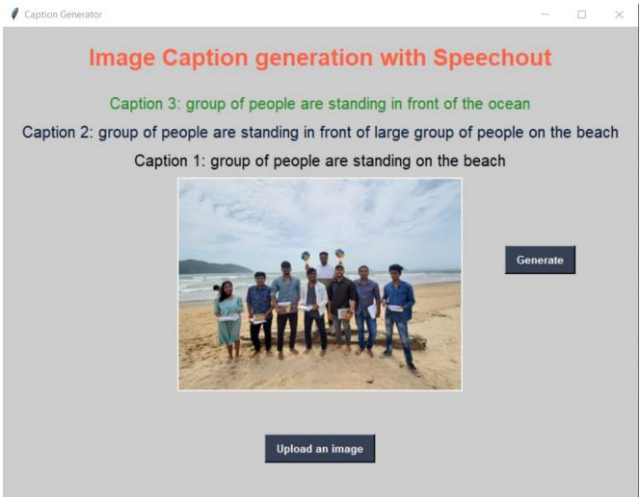


Fig 2.1 Text to speech conversion block diagram

IV. RESULT ANALYSIS

After defining the model, the model is trained for few number of epochs. As per the analysis, initial epochs training resulted in low accuracy and the generated captions were not related to the test images. Increase in epochs training increased the accuracy rate. In the below given example, the input image is fed into the Inception v3 model for feature extraction. The visual features are then combined with textual context using an RNN with an attention mechanism. The model attends to different regions of the image while generating each word of the caption, resulting in a descriptive and contextually relevant caption. The actual output may vary depending on the specific implementation, training data, and the complexity of the image being captioned.



V. CONCLUSION

In conclusion, visual attention-based image captioning using the Inception v3 model is a powerful technique that combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to generate descriptive captions for images. The Inception v3 model is used to extract visual features from the input image, and these features are combined with textual context using an RNN with an attention mechanism. By incorporating visual attention, the model can dynamically focus on different regions of the image while generating each word of the caption. This attention mechanism allows the model to capture relevant details and generate more accurate and contextually appropriate captions. Visual attention-based image captioning using Inception v3 has shown promising results in generating captions that are descriptive and meaningful, capturing both the visual content of the image and the textual context. It has numerous applications in areas such as image understanding, assistive technology, and content generation in various domains.

VI. REFERENCES

[1] Liang Hu Yanli Wang Haoran Liu, Shuang Bai. Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052., 2018.

[2] N. Tagougui A. Hani and M. Kherallah. Image Caption Generation Using A Deep Architecture. International Arab Conference on Information Technology (ACIT), 2019.

[3] L. Jain S. Das and A. Das. Deep Learning for Military Image Captioning. 21st International Conference on Information Fusion (FUSION),, 20.