e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

CAPTION GENERATION FROM IMAGE AND TTS CONVERTION

Amey Parkhe[*1], Akshat Rathod[*2], Omkar Salunkhe[*3], Shweta Sharma[*4]

[*1,2,3]Student, Department Of Computer Engineering, Atharva College Of Engineering/Mumbai University, India.

[*4]Assistant Professor, Department Of Computer Engineering, Atharva College Of Engineering/Mumbai University, India.

ABSTRACT

It is now imperative that the generated captions accurately reflect the graphical information of the image, and they be highly syntactically readable. The goal of image captioning is to automatically create the best description possible for an image. Image generation can give a meaningful description of the scene or object if the scene or object is correctly recognized, as well as if the relationship to the object and its attributes is understood. Flickr-8k is used to train the model. Using a python GUI, we generate captions using the proposed merging method, which first takes the partial caption vectors and aggregates them with the image data. The resulting caption reduces RNN's hidden state vector up to four times, reducing RNN's hidden state vector by four times is beneficial to long-term memory. Our system also enables text-to-speech conversion of this generated caption, which is also useful for blind or physically disabled users. BLEU score is used after captioning for evaluation.

Keywords: Image Captioning, Long Short-Term Memory, Feature Extraction, Text To Speech, Fliker-8k.

I. INTRODUCTION

The concept of using a machine similar to the human mind to transmit information had received a great deal of attention and research. In contrast to humans, machines have a difficult time describing the same images. Creating captions for images is a process that provides detailed information and description about the images. In an image, there are several different elements, such as objects, the background environment, and the relationships between individuals and scenes.

Similar to images, language describes and provides information about the scenes in the images. By producing captions from the images, one can gain an insight into the scenes and their significance around the world. It may soon be possible for visually impaired people to "see" the world more clearly with the help of the image description system. In recent years, it has gained increasing attention and become one of the most important topics in computer vision [6].

The method we propose includes a system to obtain captions for images. By combining the vector information of images with partial vectors of words, and converting the outputted captions into speech, the flicker 8K dataset is used with a mechanism that produces perfect captions from the image, allowing people with vision impairment to receive information anywhere in the world.

II. RELATED WORK

There are three main image captioning methods discussed in this paper: CNN-RNN, CNN-CNN, and reinforcement learning. In addition to demonstrating the benefits and challenges of each system, they provided representative works and evaluation metrics. [1]

The study was published in 2018 by Doshi et al. They developed a system that facilitates reading for blind people through their paper. A JSON structure is retrieved as the response from the system when it extracts the text from images using the Google cloud vision API that can recognize text under multiple conditions. Using g TTS engine, the text is obtained and then converted into speech. Finally, the output text is converted into audio output in the form of synthetic speech after being stored as an mp3 file and then played on an mp3 player. [2]

The encoder-decoder structure was proposed by Vinyl et al. for captioning of images. In fact, they are similar in that they both rely on convolutional neural networks (CNNs) with differing structures to extract and encode image information. A directional foundation has been laid for the study of captions for images using this method. An LSTM is used to decode the encoded information that the convolutional neural network extracts from the image. Last but not least, the decoder generates captions based on the inputs. [3]

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

A Flask app using machine learning was implemented by Baradar et al. in 2019. By using VGG16 and Convolutional Neural Networks, this model extracts the features from the image that are then calculated by the Recurrent Neural Network. These words are then used as the caption. Data from Flickr-8k is used to train the model. Users can search for images using this system. [4]

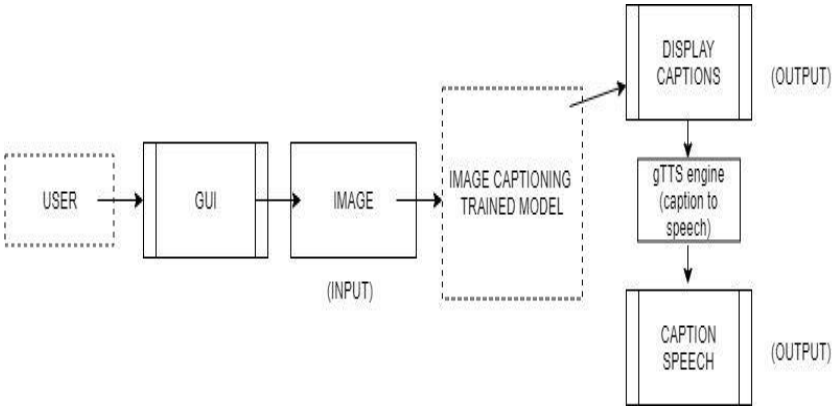
By aggregating web documents that provide image locations, Aker and Gaizauskas propose an approach to automatically caption geotagged images. The patterns are dependent on various scenarios, and they used higher ROUGE scores than n-grams, resulting in dependency relations.[5]

Wang and his team in this overview have summarized all aspects of image caption generation and have discussed briefly the model framework proposed in recent years to solve the tasks as well as how they have implemented the attention mechanism to improve the system. This study summarized the many datasets as well as the most common evaluation criteria for images caption generation. [6]

A system was proposed by Yang et al. that produces a natural image automatically, which aids in understanding them. A multi-modal neural network model was proposed in which the objects are detected, as well as localized as humans do, and the image is described accordingly. [7]

A model for creating captions for images with CNN and LSTM was developed in this paper using the principles of the Long Short-Term Memory and Convolutional Neural Network models. LSTM extracts the word vectors from the image vectors and CNN extracts the image vectors. [8]

III. IMPLEMENTATION AND METHODOLOGY



Using Python and machine learning techniques, we have developed a web app using GUI. In this model, RNN and LSTM are used along with the emerging model to produce image captions. It outputs the most preferable captions from the Flickr-8k dataset, and it also uses the TTS engine to convert the generated captions into speech for visually impaired users.

3.1 DATA SET

In addition to the 8000 images provided by Flickr 8K, there are also the 18,000 images contained by MS Coco, the 30,000 images in Flicker-30K, etc. For our preliminary use, we have taken the Flickr 8K set. 8000 images are included in this dataset, each with 5 captions, which are divided into 6000 images for training sets, 1000 for dev sets, and then another thousand for the test sets.

3.2 DATA PREPROCESSING

Every image must be converted into a corresponding vector that can be fed to the system as input. Using the Inception V3 model (CNN), we can then convert the images into fixed-size vectors, which can then be loaded into our system using transfer learning techniques. We aim to give you a fixed-length fixed-length vector containing information about every image instead of classifying it. A 1K different sets of images were classified using this model which was trained on the ImageNet dataset. Feature engineering is the process of automatically identifying features from a dataset. To obtain the bottleneck features for every image, we seek to remove the last SoftMax layer from the model.

It is necessary to encode the whole caption into a vector and to represent it as an integer index so that we can predict it. For captions, we encode the words into vectors by using RNNs as our target variables.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

3.3 Data preparation and optimization

Images are converted into 2048-point vectors and captions are indexed.

Figure 1: Training image of first caption

Figure 2: Training image of second caption

Figure 1 is the vector feature for the first image and Figure 2 is the vector feature for the second image:

As an example, let's take the two tokens for the captions "startseq" and "endseq" These tokens appear in both captions "startseq" is a picture of two white puppies sitting on a chair "r" is a picture of an end." Second caption: A young girl with a dog "endseq" plays with startseq."

In the vocabulary, there are: white, puppy, endseq, chair, on, sat, with, startseq, playing, dog Anndex each word now as follows:

white-1, puppy-2, endseq-3,with-7, startseq-8, playing-9, dog-10 , chair-4, on-5, sitting-6.

Let's convert it into a supervised learning problem.

A set of data points has been set up as D = [Ai, Bi], in which Ai is the feature vector of the variable 'i' and Bi is the variable target.

In order to predict the likelihood that each word in a prefix will be the next word in the prefix, the RNN passes its final state to a feedforward layer. After providing the image vector and the first two words as input, we attempt to predict the third word; then, we provide the image vector and the third word as input, and so on.

Therefore, we can summarize the data matrix for an image and its corresponding image. The RNN then processes the sequence. A batch process makes them equal lengths. The dataset was optimized using SGD with a generator function to look at the loss on data batches, which ensures that the whole dataset won't need to be stored in memory.

To train the model, we used a pre-trained GLOVE model that created an embedding matrix for every index mapped to 200 long vectors.

LSTMs, like RNNs, are specialized recurrent neural networks that process partial captions.

A backpropagation algorithm will be used to update the model and it will caption the image using the vector and partial caption vectors, and select the best caption by picking the words with the highest probability.

By greedily selecting the words with maximum probability, the user interface ultimately chooses the preferred captions for given images based on the image vectors and the partial caption vectors. Additionally, our system also converts generated captions into voice speech through the use of a TTS engine (text-to-speech conversion), which will also be helpful to the blind. Ultimately, the system would provide the captions for the image as a final output, as well as a simultaneous translation of the caption output into speech.

IV. TECHNICAL REQUIREMENTS

The Flickr_8K and Flickr_30K datasets will be used for generating image captions. We used this dataset along with the MSCOCO, flicker_8k, and flicker_30k datasets since other large datasets were available, but it would take a lot of time for systems only supporting CPUs to train the network. Flickr8k/Flickr30k. There are 8,000 photos in Flickr8k dataset, 6000 trained images already in the dataset, 1000 image verifications, and 1000

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

image testings. Flickr8k dataset contains images from Yahoo's photo albums site Flickr. The Flickr30k dataset consists of images from Flickr. Fliker_30k contains 31,783 images. 28000 images which is already trained.1000 images testing. validation set has 1000 images. The corresponding dataset contains 5 sentences for each image.

MSCOCO is the Microsoft COCO Captions dataset, which is created by the Microsoft Team, and it is intended for scenes. Images are captured from complex scenes, and images are used to perform different tasks, such as image recognition, image segmentation, and image description. Each dataset uses Amazon's Mechanical Turk service to generate sentences for each image. It generates at least five sentences for each image and there are more than 1.5 million sentences in total. It contains 20M images with 500 million annotation files that provide information on the images and their descriptions. The overall training data set contains 82,783 images, the test set contains 40,775 images, and the validation set contains 40,504 images. These numbers can be found in the following table.

Table 1: Summary of the number of images in each dataset.

Dataset name	Train	Size Valid	Test
MS COCO	82783	40504	40775
Fliker8k	6000	1000	1000
Fliker30k	28000	1000	1000

The BLEU engine is not designed for solving the image caption problem, it is designed for evaluating the error rate for machine translation. This is the most commonly used evaluation indicator. BLEU is used to analyze a translation statement's correlation to the reference translation statement. They are compared using the correlation between the two translation statements. In the case of machine translation, the performance depends on the accuracy of the machine translation statement vs. the human professional translation statement. The closer the machine translation statement is to the human professional translation statement, the better the performance. Machine translation is applied to this task when the multiple images are translated into the multiple source languages. By analyzing longer matching information, BLEU considers n-grams rather than words as the granularity of the search. BLEU scores improve with the number of n-grams. It is the BLEU's ability to recognize an n-gram rather than the word that is of benefit.

V. MODEL ARCHITECTURE

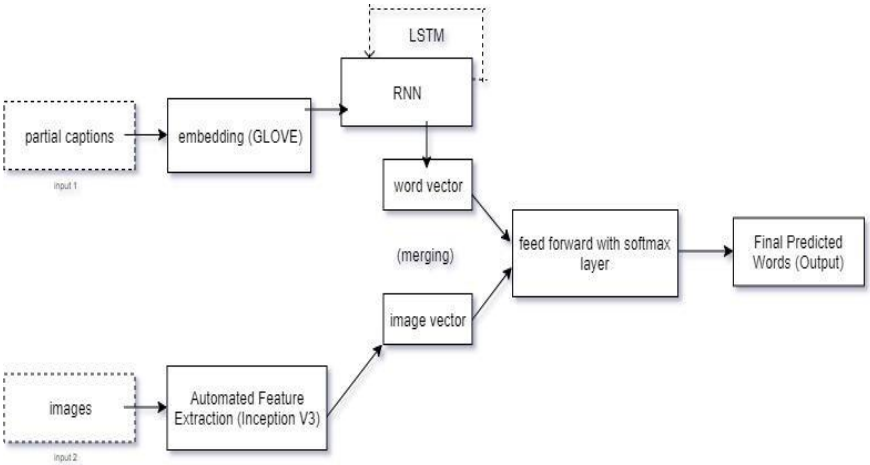


Figure 3: Proposed model architecture.

We propose a merging model to reduce the hidden state vector of RNN up to four times as there are two input parameters, which are image vector and partial captions.

The image is not part of RNN's subnetworks, since RNN only processes linguistic information in one prefix as a whole. Using the GLOVE embedding function, the image vector and the word vector are combined, and both are available for the SoftMax output layer after encoding. A recurrent neural network called an LSTM layer is used by the RNN to process partial captions. This approach does not expose the RNN to the image vector instead, it introduces the image vector or vectors into the final layer model, which is encoded using InceptionV3, all these vectors are then merged simultaneously with SoftMax and provided with the best captions.

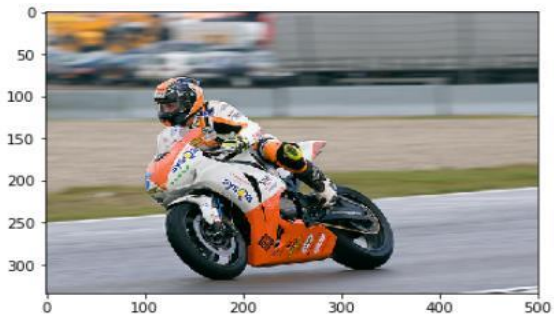
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

By greedily selecting the words with maximum probability, greedily constructing the preferred captions from the image vectors and partial caption vectors will ultimately generate the preferred captions. A TTS engine (text to speech conversion) is also included in the system in order to convert the generated caption into voice speech. This will be helpful to those who are blind or disabled.

VI. RESULTS AND EVALUATION

In this evaluation, the conditioned dataset is taken into consideration, and the images are given to the system to output the most preferred captions. Below are some of the results of preliminary testing and evaluation, along with their images.

Greedy: white crane with black begins to take flight from the water



Greedy: motorcyclist is riding an orange motorcycle

A speech output is also generated from the captions obtained.

VII. CONCONCLUSION

In this section, we finish our preliminary work on the generation of image captions based on the flicker 8k dataset, and we demonstrate our proposed solution for automating the process of getting the most preferred captions from the images provided. Our system not only uses machine learning to produce captions for the images but also converts the captions into speech, so people with visual impairments can get the captions that are human-centred using both our system and the partial captions vectors.

VIII. FUTURE SCOPE

As this technology can be used to automate machines and generate outcomes similar to what the human mind produces, the scope of this field is enormous. The goal is to make the system more human-like in the future by increasing its accuracy. Using datasets with a massive amount of relevant data that will become available in the future will also increase the accuracy of the system. Eventually, the system will be able to train and gather outputs that will not be generic in nature, but domain-specific, which will enhance its accuracy and will allow it to produce field-specific outcomes. Various fields can be helped by this technology such as the medical field, which can assist doctors in analyzing x-rays or MRI images, the field of traffic and surveillance, which can help the visually impaired understand their environment and surroundings using images, and the field of human-computer interaction, computer vision, and so on. It may be possible in the future for authors to make advancements in the system so that it can be a great tool for getting detailed information from the images provided with voice output that will serve as a comprehensive guide for the users and the population in general.

IX. REFERENCES

[1] Shuang Liu, Image Captioning Based on Deep Neural Networks, MATEC Web of Conferences 232, 01052 (2018) Available: https://doi.org/10.1051/matecconf/201823201052.

[2] Doshi, Text Reader for Visually Impaired Using Google Cloud Vision API, international journal of innovative research in technology (IJIRT). Vol. 4, 5/18.

[3] Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 39. 4/17.

[4] Shaunak Baradkar, Cap Search - "An Image Caption Generation based search", International Research Journal of Engineering and Technology (IRJET) Vol. 6, 4/19.

[5] Ahmet Aker, generating image descriptions using dependency relational patterns, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1250–1258, Uppsala, Sweden, 11- 16 July 2010. c 2010 Association for Computational Linguistics.

[6] An Overview of Image Caption Generation Methods, Hindawi Computational Intelligence and Neuroscience Volume 2020, Article ID 3062706, 13 pages https://doi.org/10.1155/2020/3062706

[7] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available:

https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df

[8] Image Caption Generator using Big Data and Machine Learning, International Research Journal of Engineering and Technology (IRJET), Vol.7, 4/20.

[9] Image Caption Generation Using Deep Learning Technique Publisher: IEEE Authors: Chetan Amritkar; Vaishali Jabade Automatic Caption Generation for News Images Publisher: IEEE Authors: Yansong Feng; Mirella Lapata

[10] A parallel-fusion RNN-LSTM architecture for image caption generation Publisher: IEEE Authors: Minsi Wang; Li Song; Xiaokang Yang; Chuanfei Luo

[11] Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap Publisher: IEEE Authors: Soheyla Amirian; Khaled Rasheed; Thiab R. Taha; Hamid R. Arabnia

[12] Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs Authors: Shizhe Chen, Qin Jin, Peng Wang, Qi Wu;

[13] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9962-9971