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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

GENERATION OF CAPTION FROM IMAGE AND TEXT-TO-SPEECH CONVERTOR

Imdad Ali Dar [*1], Mayur Anil Bopche[*2], Shubhamsen Mandar Halde[*3], Prof. Rajesh A. Patil[*4]

[*123]Student, Department of Electronics and Telecommunications Engineering, Veermata Jijabai Technological Institute/Mumbai University, India.

[*4]Assistant Professor, Department of Electrical Engineering, Veermata Jijabai Technological Institute /Mumbai University, India.

ABSTRACT

The generated captions must now precisely reflect the image's graphical information and be highly syntactically understandable. Image captioning's purpose is to generate the best feasible description for an image automatically. If the scene or object is accurately recognised, as well as the relationship to the object and its properties are understood, image production can provide a meaningful description of the scene or object. The model was trained with Flickr-8k. We construct captions with a python GUI using the suggested merging method, which aggregates incomplete caption vectors with picture data first. RNN's hidden state vector is reduced by up to four times as a result of the caption , this generated caption can also be converted to text-to-speech using our technology, which is valuable for users who are blind or physically impaired. After captioned, the BLEU score is utilised for evaluation.

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

I. INTRODUCTION

The idea of transmitting information with a computer that mimicked the human mind had gotten a lot of interest and investigation. Machines, unlike humans, have a hard difficulty describing the same pictures. Captioning photos is a procedure that involves providing precise information and descriptions about the photographs. Objects, the backdrop environment, and the interactions between characters and scenes are only a few of the aspects that make up an image.

Language, like images, describes and offers information about the scenes depicted. One can obtain insight into the scenes and their significance around the world by creating captions from the photographs. With the help of the picture description system, visually impaired persons may soon be able to "see" the environment more clearly. It has gotten a lot of attention in recent years and has become one of the most important subjects in computer vision [6].

A mechanism for obtaining captions for photos is included in the technique we offer. The flicker 8K dataset is used with a mechanism that produces perfect captions from the image by combining the vector information of images with partial vectors of words and converting the outputted captions into speech. This allows people with vision impairment to receive information from anywhere in the world.

II. RELATED WORK

This study discusses three primary picture captioning methods: CNN-RNN, CNN-CNN, and reinforcement learning. They presented representative works and assessment criteria in addition to highlighting the benefits and limitations of each approach. [1]

Doshi et al. released the findings in 2018. They created a technique that allows blind persons to read through their paper. When the system extracts text from photos using the Google cloud vision API, which can detect text under a variety of situations, it returns a JSON format as a response. The text is received and then transformed into speech using the g TTS engine. After being saved as an mp3 file, the output text is turned into audio output in the form of synthetic speech.then played on an mp3 player. [2]

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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Vinyl et al. suggested the encoder-decoder framework for picture captioning. In truth, they are comparable in that they both extract and encode picture data using convolutional neural networks (CNNs) with different topologies. Using this strategy, a directional basis has been built for the analysis of captions for photographs. The encoded information extracted from the image by the convolutional neural network is decoded using an LSTM. Finally, based on the inputs, the decoder creates captions. [3]

In 2019, Baradar et al. created a Flask app that used machine learning. This model collects characteristics from the picture using VGG16 and Convolutional Neural Networks, which are subsequently computed by the Recurrent Neural Network. This model extracts the image's characteristics, which the Recurrent Neural Network subsequently calculates. The caption is then created using these words. The model is trained using data from Flickr-8k. This system allows users to look for photos. [4]

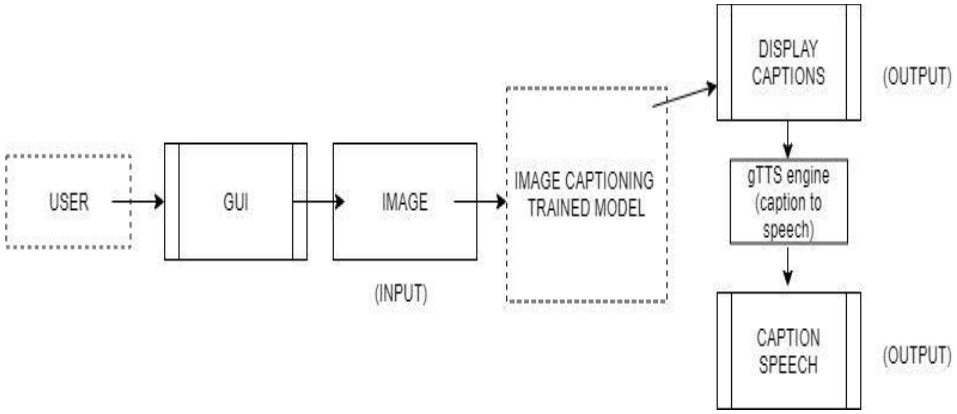
Aker and Gaizauskas suggest a method for automatically captioning geotagged photos by collecting online articles that contain image locations. The patterns are dependant on a variety of contexts, and they employed ROUGE scores that were greater than n-grams, resulting in dependence relationships. [5]

Wang and his colleagues have reviewed all elements of picture caption creation in this overview, as well as briefly discussing the model framework proposed in recent years to tackle the tasks and how they have integrated the attention mechanism to improve the system. This study summarised the numerous datasets as well as the most frequent picture caption generating assessment criteria. [6]

Yang et al. suggested a technique that automatically generates a natural visual that assists in comprehending them. A multi-modal neural network model was presented in which items are identified and located in the same way as humans do, and the image is described as a result. [7]

This research uses the ideas of the Long Short-Term Memory and Convolutional Neural Network models to construct a model for writing captions for photos using CNN and LSTM. The image vectors are extracted using LSTM, and the image vectors are extracted by CNN. [8]

III. IMPLEMENTATION AND METHODOLOGY



We created a web app with a GUI utilising Python and machine learning algorithms. To generate picture captions, RNN and LSTM are used with the emerging model in this model. It uses the TTS engine to translate the generated captions into voice for visually challenged users, and it generates the most preferred captions from the Flickr-8k dataset.

3.1 DATA SET

There are additional 18,000 photographs in MS Coco, 30,000 images in Flicker-30K, and so on, in addition to the 8000 images given by Flickr 8K. We've used the Flickr 8K collection as a starting point. This dataset contains 8000 photos, each with five captions, separated into 6000 images for training sets, 1000 for development sets, and another thousand for test sets.

3.2 DATA PREPROCESSING

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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Every image must be turned into a vector that can be used as input by the system. We may then transform the pictures into fixed-size vectors using the Inception V3 model (CNN), which can subsequently be put into our system utilising transfer learning techniques. Instead of categorising a picture, we want to provide you a fixed-length fixed-length vector containing information about it. This model, which was trained on the ImageNet dataset, was used to classify 1K distinct sets of photos. The technique of autonomously identifying characteristics from a dataset is known as feature engineering. We want to remove the final SoftMax layer from the model to get the bottleneck features for each image. To anticipate the caption, we must encode the entire caption into a vector and represent it as an integer index. We use RNNs as our target variables to encode the words into vectors for captions.

3.3 Data preparation and optimization

The descriptions are indexed and the images are transformed to 2048-point vectors.

Figure 1: Training image of first caption

Figure 2: Training image of second caption

The vector feature for the first image is shown in Figure 1, and the vector feature for the second image is shown in Figure 2:

Take the two tokens for the captions "startseq" and "endseq" as an example. Both captions include these tokens. ""r" is a picture of an end." "startseq" is an image of two white pups sitting on a chair.

A little girl with a dog named "endseq" plays with startseq in the second caption."

There are words like white, puppy, endseq, chair, on, sat, with, startseq, playing, and dog in the lexicon.

Fill in the blanks with the following words: white-1, puppy-2, endseq-3, with-7, startseq-8, playing-9, dog-10, chair-4, on-5, sitting-6.

Let's turn it into a problem of supervised learning.

D = [Ai, Bi], where Ai is the feature vector of the variable I and Bi is the variable target, has been used to create a set of data points.

The RNN transmits its final state to a feedforward layer in order to estimate the likelihood that each word in a prefix will be the next word in the prefix. We attempt to predict the third word after supplying the image vector and the first two words as input; then we offer the image vector and the third word as input, and so on.

As a result, the data matrix for a picture and its related image may be summarised. The sequence is subsequently processed by the RNN. They're all the same length thanks to a batch procedure. The dataset was optimised utilising SGD and a generator function to look at data batch loss, ensuring that the entire dataset would not need to be retained in memory.

We utilised a pre-trained GLOVE model to train the model, which generated an embedding matrix for each index translated to 200 long vectors.

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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

LSTMs are specialised recurrent neural networks that handle partial captions, similar to RNNs.

The model will be updated using a back propagation technique, which will caption the picture using the vector and partial caption vectors, and choose the best caption by selecting the words with the highest probability.

The user interface determines the preferred captions for provided pictures based on the image vectors and partial caption vectors by greedily picking the words with the highest probability. Additionally, our system uses a TTS engine (text-to-speech conversion) to turn created captions into spoken speech, which will be useful to the blind. Finally, the system would give captions for the image as well as a simultaneous translation of the caption output into voice as a final output.

IV. TECHNICAL REQUIREMENTS

Image captions will be generated using the Flickr 8K and Flickr 30K datasets. We combined this dataset with the MSCOCO, flicker 8k, and flicker 30k datasets since additional big datasets were not available, and training the network on computers with just CPUs would take a long time. Flickr8k/Flickr30k. In the Flickr8k dataset, there are 8,000 photographs, 6000 trained images, 1000 image verifications, and 1000 image testings. Flickr8k is a collection of photographs from Flickr, Yahoo's photo-sharing service. The Flickr30k dataset is made up of Flickr photos. There are 31,783 photos in Fliker 30k. There are 28000 images that have already been trained. A total of 1000 photos were tested. There are 1000 photos in the validation set. Each picture has five phrases in the matching dataset.

MSCOCO stands for Microsoft COCO Captions, a dataset generated by the Microsoft Team and designed for scenes. Images are taken from complicated settings and utilised for a variety of activities including image identification, image segmentation, and image description. For each image, Amazon's Mechanical Turk service is used to produce text. Each image generates at least five sentences, for a total of more than 1.5 million sentences. It has 20 million photographs and 500 million annotation files, which include information on the images and their descriptions. There are 82,783 photos in the entire training data set, 40,775 images in the test set, and 40,504 images in the validation set. The following table contains these figures.

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 isn't meant to solve the picture caption problem; rather, it's meant to assess machine translation error rates. This is the most widely utilised criterion for assessment. The BLEU algorithm is used to examine the relationship between a translation statement and the reference translation statement. The correlation between the two translation statements is used to compare them. In the case of machine translation, the correctness of the machine translation statement vs. the human professional translation statement determines the performance. The greater the performance, the closer the machine translation statement is to the human professional translation statement. When numerous photos are translated into different source languages, machine translation is used to complete the operation. BLEU considers n-grams rather than words as the granularity of the search by assessing lengthier matching information. The quantity of n-grams increases the BLEU score. The capacity of the BLEU to distinguish an n-gram rather than a word is advantageous.

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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

V. MODEL ARCHITECTURE

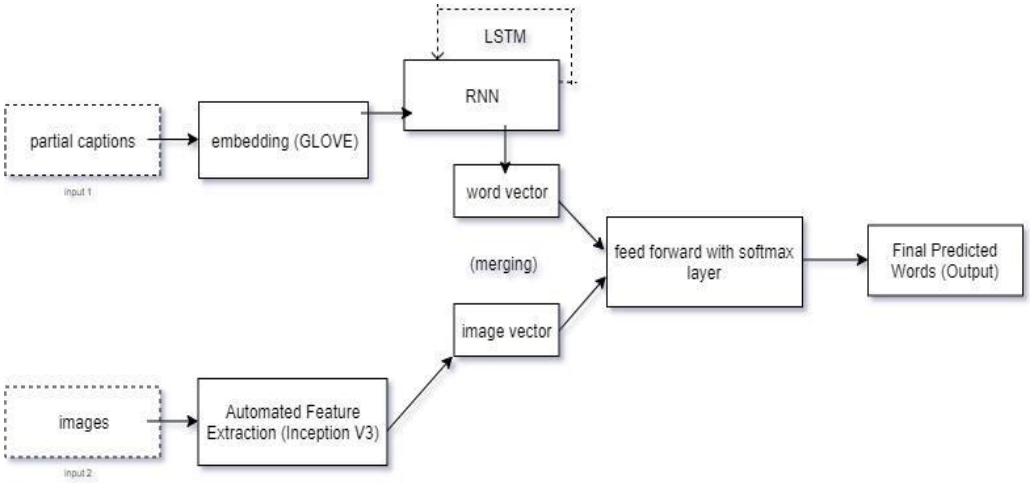


Figure 3: Proposed model architecture.

Because there are two input parameters, picture vector and partial captions, we suggest a merging approach that can lower the RNN's hidden state vector by up to four times.

Because RNN only analyses linguistic information in one prefix as a whole, the picture is not part of RNN's subnetworks. The picture vector and the word vector are joined using the GLOVE embedding algorithm, and both are accessible for the SoftMax output layer following encoding. The RNN processes partial captions using an LSTM layer, which is a recurrent neural network.

Instead of exposing the RNN to the picture vector, this technique, It adds the picture vector or vectors to the final layer model, which is encoded using InceptionV3. All of these vectors are then combined with SoftMax at the same time and given the best captions.

The recommended captions will be generated by greedily picking the words with the highest probability and greedily generating the preferred captions from the picture vectors and partial caption vectors. In order to turn the generated caption into spoken speech, the system additionally includes a TTS engine (text to speech conversion). Those who are blind or crippled will benefit from this..

VI. RESULTS AND EVALUATION

The conditioned dataset is used in this assessment, and the photos are submitted to the system to generate the most desired captions. Some of the preliminary testing and assessment findings, as well as accompanying photos, are shown below.

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

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:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Greedy: motorcyclist is riding an orange motorcycle

The captions are also used to produce a spoken output.

VII. CONCLUSION

We conclude our early work on picture caption creation using the flicker 8k dataset in this part, and we illustrate our suggested technique for automating the process of extracting the most desired captions from the photos given. Our method not only utilises machine learning to generate captions for the pictures, but it also transforms the captions to voice, allowing persons with visual impairments to access human-centered captions using both our system and partial captioning vectors.

VIII. FUTURE SCOPE

The scope of this area is immense since this technology may be used to automate machines and provide results that are comparable to those produced by the human mind. The objective is to improve the system's accuracy in order to make it more human-like in the future. The accuracy of the system will also be improved by using datasets with a large quantity of relevant data that will become accessible in the future. Eventually, the system will be able to learn and acquire domain-specific outputs, which will improve its accuracy and allow it to provide field-specific results. This technology can help in a variety of fields, including medicine, where it can help doctors analyse x-rays or MRI images, traffic and surveillance, where it can help the visually impaired understand their environment and surroundings using images, and human-computer interaction, computer vision, and so on. It may be feasible for writers to improve the system in the future so that it may be a useful tool for extracting specific information from photos via voice output, which will serve as a thorough guide for users and the general public.

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

[10] Automatic Caption Generation for News Images
Publisher: IEEE
Authors: Yansong Feng; Mirella Lapata

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

[12] 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

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

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