International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 4, Jul.-Aug., 2023 pp: 191-196 www.ijemh.com

Image Caption Generator

Sanjana Reddy Gangam
Department of Electronicsand Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India.

Nikhitha Enugula
Department of Electronicsand Computer Engineering,Sreenidhi Institute of Science and Technology, Hyderabad, India.

P. Sravani
Department of Electronicsand Computer Engineering,Sreenidhi Institute of Science and Technology, Hyderabad, India.

Dr. M. Shailaja
Department of Electronicsand Computer Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India.

Date of Submission: 05-08-2023 Date of Acceptance: 19-08-2023

Abstract:
Image caption generator is a system of feting the surroundings of an photograph and annotating it with relevant captions the use of deep learning, and pc vision (CV).It consists of the labeling of a photograph with English key phrases with the assist of datasets furnished all through version training. Imagenet dataset is used to educate the CNN version referred to as Xception .In summary, the techniques defined are brainstorming and feature their very own characteristics, however all have the not unusual place drawback that they do now no longer make Intuitive characteristic observations on items or movements with inside the image, nor do they provide an stop-to-stop mature widespread version to clear up this problem. Previously suggested models were implemented and got very low accuracy. Photo caption The verbal description of such image is the focus of the popular artificial intelligence research field known as Generator. A well-formed sentence requires both semantic and syntactic knowledge of the language, therefore RNN and LSTM can be used to generate sentences more accurately. These extracted capabilities can be fed to the LSTM version which in flip generates the photograph caption.

Key Words: LSTM, RNN, CNN,Vgg16.

I.INTRODUCTION

In the field of AI, image caption generators offer photo expertise and a language explanation for each photo. Syntactic and semantic comprehension are both necessary for creating well-rounded sentences. It is exceedingly challenging to describe the content of an image with precisely constructed sentences; it could also have a significant impact on aiding people who are visually impaired in understanding images.

Since AI is currently at the core of the advance economy, it also serves as the foundation for this essay. Due to its remarkable accuracy results when compared to ML techniques that are already in use, the artificial intelligence Sub field of deep learning has recently attracted a lot of attention. The ability to create a meaningful statement from an image is a challenging challenge but one that can have significant benefits, such as improving the understanding of visuals for those who are blind.

Compared to photo classification, which has received the majority of attention in the CV field, the problem of picture captioning is far more difficult. The relationship between the objects in a photo should be captured in the description. Since the aforementioned semantic knowledge must be communicated in a language like English in addition to the visual interpretation of the image, a linguistic version is necessary. The two models were always sewn together in theprevious efforts.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 4, Jul.-Aug., 2023 pp: 191-196 www.ijemh.com

II. RELATED WORK

We present a synthesize output maker that locates and encodes items, characteristics, and courtship in a photo using a form of herbal language. The straightforward CNN design with 4 classes is depicted in fig1.

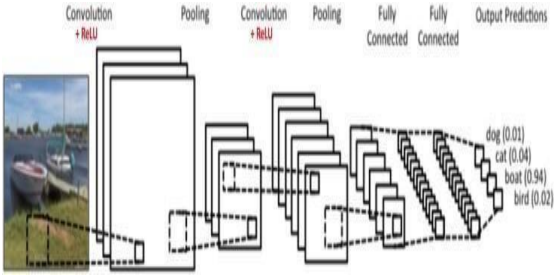

Fig 1: A straightforward CNN design

Therefore, we may combine such designs to create our model for the picture caption generator. It's also referred to as a "CNN- RNN" model. The image's capabilities are extracted using CNN .The CNN data will be used by LSTM to help create the image's contour. CNN with convolution Neural networks, a subset of deep neural networks, are capable of processing data that has been supplied as a 2D matrix. photographs may easily be displayed as a 2D matrix, hence CNN may be quite helpful when working with photographs.. CNN is primarily used to classify photos and determine whether they show a bird, an aero plane, Superman, etc. CNN with convolution Neural networks, a subset of deep neural networks, are capable of processing data that analyses photographs top to bottom and from left to right to extract the most crucial information before combining the features to categorise the pictures. Images that have been resized can be handled by it.

LSTM, is a type of recurrent neural network(RNN) that is excellent for collection forecast issues .We can predict what the next phrase will be based on the material that came before it. It has proven to be more potent than traditional RNN by getting through their short-term memory limitations. The LSTM can process applicable data throughout input processing and, using an overlook gate, it can eliminate non-applicable data.

III. METHODOLOGY

Certain procedures and techniques are employed when developing any system. The technology employed in this study is image processing with CNN for picture prediction and LSTM for caption generation.

A. Convolutional Neural Network (CNN)

CNN with convolution Neural networks, a subset of deep neural networks, are capable of processing data that Artificial neural networks that specialise in picture classification and recognition are known as Convolutional neural networks (ConvNets or CNNs). They have been heavily utilized for tasks including picture captioning, self-driving cars, and object detection. Yann Lecun made the initial discovery of ConvNets.

Figure1 depicts a fundamental ConvNets.
There are four major operations that may be used to illustrate the ConvNets architecture:
1. Non- Linearity(RELU)
2 . Convolution
3. Sub Sampling or pooling
4. Classification

Understanding how these operations function is essential to having a complete understanding of Convolutional Neural Networks because they are the fundamental building blocks of every Convolutional Neural Network. Below, we will go into more detail about each of these operations.
Each photo can be conceptualized as a matrix of pixel values. Red, green, and blue are the three channels that make up an image with a modern virtual digital dig cam. Think of these channels as three 2D matrices stacked on top of one another, one for each color, with pixel values ranging from 0 to 255 in each matrix.

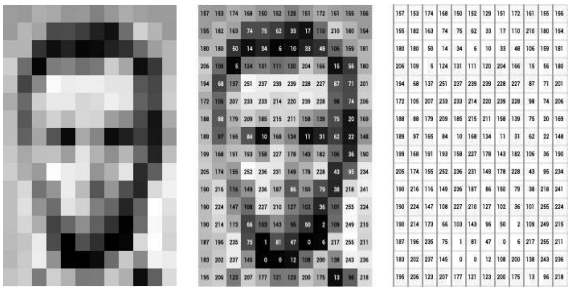

Fig 2 : A grayscale image as a numerical matrix.

B. VGG16 Architecture:

VGG is an acronym for the Oxford researchers that created this architecture, the Visual Geometry Group. The VGG structure is composed of blocks, and each block consists of 2D Convolution and Max Pooling layers. VGGNet comes in two variations, VGG16 and VGG19, each of which has sixteen and nineteen layers, respectively. Only VGG16, as indicated in fig. 3, was employed, though not exactly to the same dimensions. To extract characteristics from photos, we employ a layer with a

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 4, Jul.-Aug., 2023 pp: 191-196 www.ijemh.com

dimension of 7x7x512. Instead of placing them all in one file, we dump the features to a numpy file with the similar name as the photographs so that we can access them later.

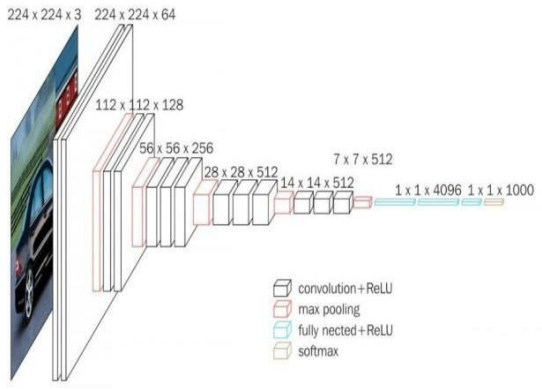

Fig 3:VGG16model

IV. IMPLEMENTATION

The steps involved in implementation are:

A. Get Dataset

The Flickr8k dataset is what we choose to use. There are 5 captions for each of the 8092 photos. There are various ways to caption an image, thus each image gets 5 captions. The training, testing, and evaluation subsets of this dataset each contain 6000, 1000, and 1000 images, respectively.

B. Prepare Photo Data

VGG16 and ResNet50, two pre-trained CNN models, are used to extract features from images. Since classification of photos is not what we are interested in, we delete the final layers from these models. The way the photos are represented is something that interests us. We compute all of the characteristics and save them to a file rather than extracting them as needed. Features vgg16.pkl and features Rssnet.pkl are two separate files where we saved the extracted features for each model. These models need photos that are exactly 224 pixels in size. Images required to be scaled before being turned into arrays and then reshaped. 2048 × 2048 pixel vectors make up the retrieved features. The employed CNN taught model Rsnet50 and LSTM and vgg16 is shown in Figure 4.

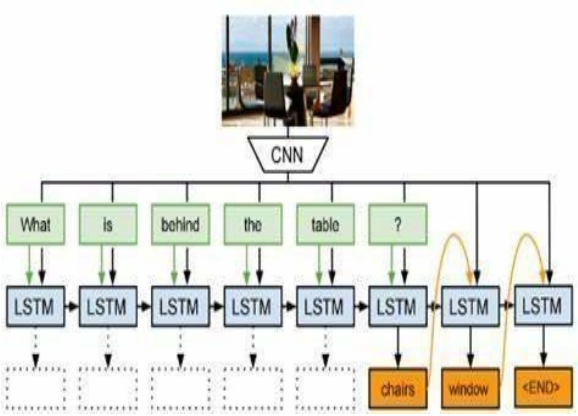

Fig 4: LSTM+ResNet50architecture

C. Prepare Text Data

First, we load all of the image descriptions. We produce a vocabulary that links descriptions to image names. We have to tidy up the image descriptions before preparing the text data. All words were changed to lower case for this, and all punctuation, words longer than one character, and words including numerals were also eliminated. Next, we compile the distinctive words from the descriptions into a lexicon (VGG16). The generated vocabulary is 8,763 words in length. In order to use them later, we save the descriptions to the file descriptions.txt. The unique terms from the blurbs are then used to generate a vocabulary (ResNet50). The generated vocabulary is 1848 words in length. The descriptions are then saved to the descriptions.txt file so that we can use them later.

D. Load Data

We create a function to load the data for the train, validation, and test subsets. In the files Flickr8k.train, Images.txt, Flickr8k.dev, Images.txt, and Flickr8k.test Images.txt, these subsets are predefined. Without paying close attention, we enter the image keys, descriptions, and features into the notebook. Because we'll take out them afterwards, we don't load features during the attention condition. "The model will produce a caption word by word while accounting for the words that came before."Therefore, the generation must begin and stop with the initial and final words. Because of this, we added the words "start" and "end" to the descriptions as the first and last words.

E. Encode Text Data

To convert vocabulary terms from the vocabulary to numbers, we utilize a tokenizer. This tokenizer is saved in the file tokenizer.Pkl it for later use. We also choose the vocabulary size and the maximum

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 4, Jul.-Aug., 2023 pp: 191-196 www.ijemh.com

description length so that we can utilise them later.

F. Define Model

We made the decision to test various Recurrent Neural Network topologies and contrast the outcomes. Figures 18 show that although the models are different, both feature a recursive Long Short-Term Memory network. ReviewFigure 5.

G. Fit Model LSTM+VGG16

Here, we fit the model to the available data. We keep track of validation and training loss. Because fitting can take a very long time, we save the model with the lowest validation loss so we can utilize it later. To save time and prevent over fitting, we halt training if the validation loss rises in two straight epochs. The model is given the output word, a list of words, and image attributes. We generate every potential input sequence and output word for each image and description by totaling single word at a time until we reach the end.

There are two inputs for the RNN model. There are Embedding, Dropout, and LSTM layers in the text sub model. There are two layers—Dropout and Dense—in the picture sub model. In order to train the model utilising the RAM available in Google Collaboration, the final two dense layers have a size with soft max, and this is followed by the addition of these two sub models. The data generator uses picture ids to select the photos that go with the captions after receiving the training data that has been jumbled, added layers for the fig. 5 trained RNN model.

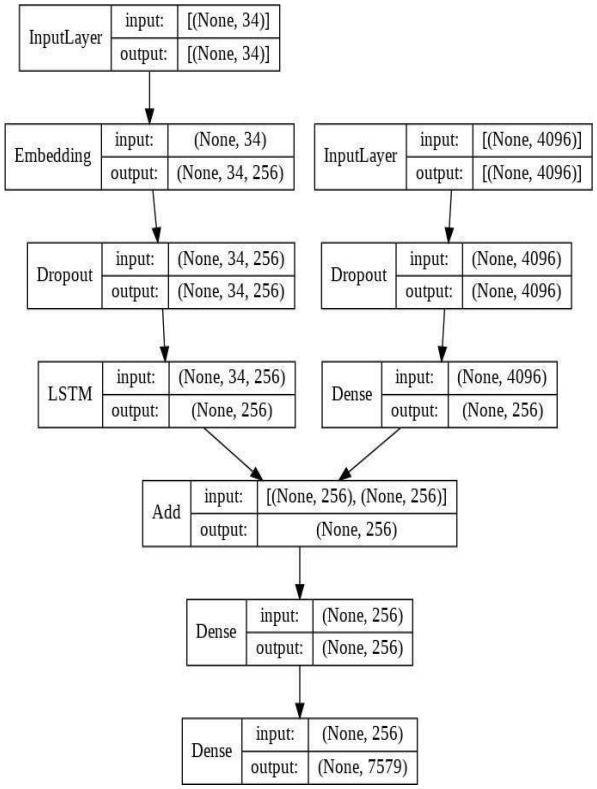

Fig 5. Vgg16 Trained Model design

H. Evaluate Model

Beam search and sampling were the two methods we utilised to create descriptions of the photographs. The best word is selected for sampling at each time step till the finish. At each time step, Beam Search takes into account the top k sentences. For every one of them, it foretells the subsequent words. The likelihood of receiving the description with the highest probability often increases as k grows.

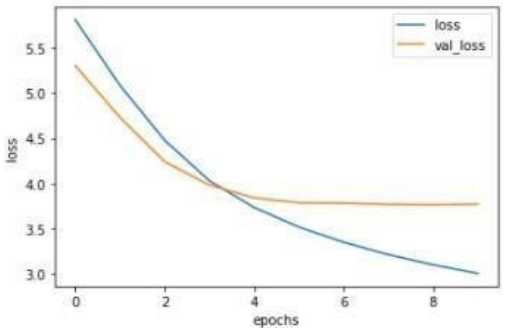

Fig 6: RNN1 and VGG16 loss and val loss levels at variousepochs

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 4, Issue 4, Jul.-Aug., 2023 pp: 191-196 www.ijemh.com

I. Generate Captions

You can observe how well our models perform when tested against real data by looking at the captions we make for a few of the test set's images. For each picture, we give the five original captions, the outcomes of sampling, and the results of beam searches when k=3 and k=5. As a reference for all the created captions, we also display BLEU-1 scores. For each image, we present the 5 original captions, the results of sampling, and the findings of beam searches when k=3 and k=5.As a reference for all the created captions, we also display BLEU-1 scores. These are some of the outcomes we obtained using VGG16 and the second RNN model with certain test photos. The trained model is captioned with one of our best captioning images shown in fig 7.

Fig 7. Most excellent captioning

Captions
1: Dog in brown and white standing outside in the snow
2: Dog is observing a little close to the ocean.
3: A Fluffy Dog shakes its wet coat to try and dry itself..
4: Brown and White Dog shaking itself dry.
5: The huge White and Brown Dog shrugs off the water .
Dog is swimming in the water while sampling (BLEU-1: 58.4101)
In Beam Search k=3, a white and brown dog is seenswimming in the lake (BLEU-1: 72.7273).
In Beam Search k=5 (BLEU-1:70.0000), a brown andwhite dog is swimming in shallow water.

We observed during training that VGG16 produced thebest outcomes for us.
However, there is still much room for development, as seen, for instance, in the captioning of Figure 8. Additionally, we can see that more work needs to be done because, although having a description that is largely accurate, it does not closely resemble any of theoriginal ones.

V. RESULTS

The CNN method and long short term memory are used to build the model for automatically writing captions for photos..8000 pictures and associated captions were captured for this model. The model, which we trained and evaluated, now generates captions for the pictures. 70% scale accuracy was attained by the model. In order to provide the user with an interactive UI, the user interface was constructed using the Flask API. The output can be seen through the user interface, which has buttons to upload images and drop-down menus to select models. The output is determined by taking into account the uploaded image and the model that was selected, meaning that if we provide the image, the caption will be generated by the trained model that was selected. Figure 8 illustrates how the image caption generator project functions.

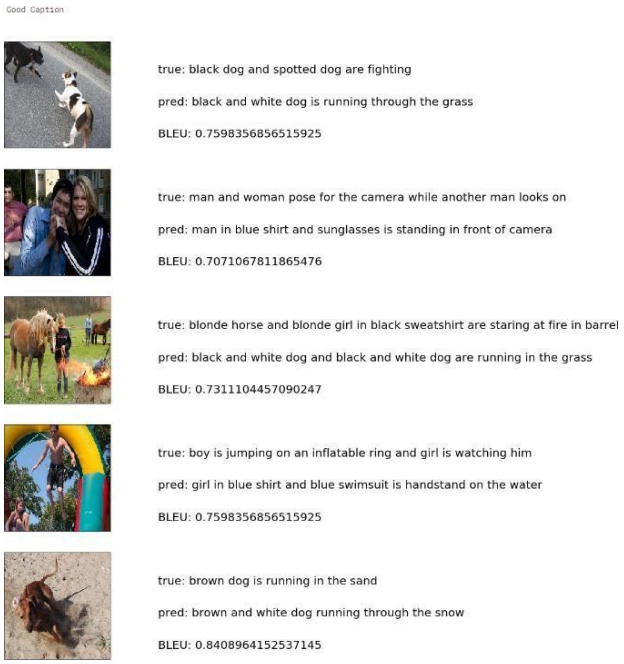

Fig 8: Outputs

VI. CONCLUSION

After training our model using VGG16 and LSTM validation loss is less for this model when compared to all other models. The BLEU score for the predicted captions are at an average of 0.72which is better among any other models. We can improve our model's accuracy by increasing number of epochs and using better pre trained (saved model (.h5 file)).

VII. FUTURE ENHANCEMENT

We will advance our work by making our model better so that it can generate captions even for the live video . Only the image gets captions in our current setup. Since captioning live video frames is entirely GPU-based, it is not practical to use standard CPUs for this. With application cases that are extensively applicable in practically every domain, video captioning is a hot topic in research that will revolutionize people's lifestyles. The most complicated jobs, such video surveillance and other security tasks, are automated. By applying the model to a larger dataset, the model's vocabulary will considerably expand, improving the model's accuracy. The accuracy of the categorization work can also be increased by using somewhat more modern architecture, such as Google Net, which lowers the mistake rate in language production.

REFERENCES

[1]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney , Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv: 1707.07998(2017).

[2]. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision andPatternRecognition.5561–5570.

[3]. LisaAnne Hendricks, Subhashini Venugopalan,Marc usRohrbach, Raymond Mooney, KateSaenko, Trevor Darrell, Junhua

[4]. Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[5]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on LearningRepresentations (ICLR).

[6]. Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing.

[7]. A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional imagesentencemapping.NIPS,2014.

[8]. R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In arXiv:1411.2539,2014. R.Kiros and R.Z.R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop,2013.

[9]. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR,2011.

[10]. P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL,2012.

[11]. P. Kuznetsova, V. Ordonez, T. Berg, and Y.Choi. Treetalk: Composition and compression of trees for image descriptions. ACL, 2(10), 2014.