Image Captions Generator using CNN and LSTM

School of Engineering & Technology Sharda University, Greater Noida

Supriya Kumari
School of Engineering and Technology,
Sharda University,
Greater Noida, India
2018009684.supriya@ug.sharda.ac.in

Mohit Rathore
School of Engineering and Technology,
Sharda University,
Greater Noida, India
2018015523.mohit@ug.sharda.ac.in

Julu Basnet
School of Engineering and Technology,
Sharda University,
Greater Noida, India
2018015987.julu@ug.sharda.ac.in

Dipanshu
School of Engineering and Technology,
Sharda University,
Greater Noida, India
2018014442.Dipanshu@ug.sharda.ac.in

***Abstract*- The realm of technology in the field of AI is progressing rapidly these days. Many research-based projects have been carried out and are still being carried out ,thanks to this advancement. Many studies have been done in the field of AI, and image caption creation is also a component of this research that is based on deep learning. There are a variety of activities that must be completed during the image captioning process, including identifying the items in the photographs, determining their semantic link, and translating the backdrop scene into the relevant phrases. The picture's information is generated automatically in artificial intelligence, which also includes computer vision and natural language processing. In order to assess the model's fluency and accuracy, the flickr8k dataset of 8000 photographs is used to describe the images. This shows that the model is appropriately captioning the photos.**

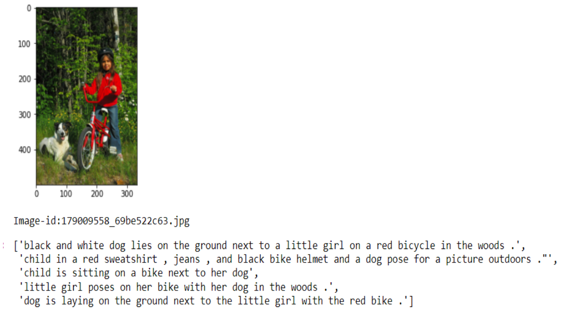
Keywords - Image Captioning, Natural Language Processing tool(nltk),CNN, LSTM ,Vgg16 model

I.INTRODUCTION

A natural language like English is used in Image Captioning, a method that generates appropriate captions for photographs. The CNN-LSTM model is used to generate image captions. This model generates natural language descriptions of photographs by analyzing their content. CNN and RNN are used to generate image captions (LSTM). As part of the project, the entire procedure is implemented on a jupyter notebook, although it may also be done in Google Colab.Only 8000 photos make up the Flickr-8k dataset, which is the total number of images on the site .When images are fed into a model, the model generates specific captions based on the images.In this manner, the model may be trained. The flickr-8k dataset also includes a caption file with about 40.000 captions. As seen in fig1, each image in this caption file has five possible captions.An image caption generator faces a significant challenge in overfitting training data due to large datasets like Microsoft Common Objects in Context (MOC) (MSCOCO). As tough as it may be, its positives aspects include helping those having visual impairments and

making it possible to post photographs with a variety of labels on the internet for free. On social networking sites like Instagram and Facebook, the user submits a picture and the captions for that image are automatically generated.As the identification of the objects in the images are identified easily by the humans,to do the same thing by the machine,deep learning approach is applied with the concept of CNN-LSTM model .Technology is a new ray of hope in the lives of hopeless visually impaired people, by using it we can upgrade the lives of visually challenged humans. AI is at the fore in this era of technology for visually challenged people. One of the sciences which can solve this problem is known as Image Caption Generation (ICG).

AI can create a machine that can exactly convey an image like a normal human has significant applications in the field of robotic vision, business, and many more. There have been various efforts taken to get a solution to this problem including template-based solutions which use image classification i.e. Labeling objects from a proper arrangement of classes that can be embedded into a dummy layout sentence.Computer vision examines an image thinking about it as a 2D array. Hence, Image captioning is a language translation problem described by Venugopalan. Using LSTM and GRU which contain internal processes and logic gates that keep information for a more extended time frame and pass only useful data, we can remove these issues. Abstracting pictures using natural languages is an elemental and challenging task.This has a huge potential result on upcoming aspects. For instance, it could assist visually impaired individuals with a better understanding of the substance of pictures on the web.



```
Image-id:179009558_69be522c63.jpg

['black and white dog lies on the ground next to a little girl on a red bicycle in the woods .',
 'child in a red sweatshirt , jeans , and black bike helmet and a dog pose for a picture outdoors ."',
 'child is sitting on a bike next to her dog',
 'little girl poses on her bike with her dog in the woods .',
 'dog is laying on the ground next to the little girl with the red bike .']
```

Fig1: Five different captions for each image.

This depicts the picture of a little girl on a red bicycle in the forest.Each picture consists of five different suitable captions of it.An image caption generator faces a significant challenge in overfitting training data due to large datasets like Microsoft Common Objects in Context (MSCOCO). As tough as it may be, its positives include helping those with visual impairments and making it possible to post photographs with a variety of labels on the internet for free. On social networking sites like Instagram and Facebook, the user submits a picture and the captions for that image are automatically generated.Also,it could give more correct and conservative data of images in situations like picture sharing in social networks or video surveillance systems.

II. RELATED WORKS

A Generative adversarial Network model presented by Md. Zakir Hossian et.al. has been shown to boost the production of caption generators. Using both real and false data, the model is trained and tested. To create fictitious visuals, this model uses a text-to-image generator. Attention-based image captioning was trained on both actual and fake images, and generated captions for each. Proposed work yielded two advantages for its author. As an example of image captioning for fake images, it illustrates that it is capable of generating the appropriate output. Another reason for this is that extra photographs are used in training, so the best captions provided for the original image are enhanced. They are designed to use attention-based GAN to produce initial synthetic images of text material. They used a combination of actual and generated photos to train and assess the model's ability to correctly caption images. Captioning for both actual and fictional photographs is a common focus of their research. Based on BELU evaluation metrics, here is the result of the caption generated using several models.[1]

According to Soheyla Amirian et al., "Automatic video and image caption generation" has been proposed using deep learning. Images can be captioned using CNN, LSTM (RNN), GRU, and other natural language processing (NLP) techniques (gated recurrent unit).The encoder and decoder architecture for image captioning can also be used for video captioning. When it comes to describing video content, there are two main processes. For the first phase, it is necessary to comprehend the object, which is accomplished by utilising the DL model to identify the activity, performer, and object in the video clip. An image-based video clip is provided in the form of a series of frames. Feature vectors are constructed from the video clips, and these feature vectors are then used in a second stage to provide information about the video clips. Using the object determined in the first phase, they build an aggregate of DL architecture for encoding and decoding state in the second step, which determines what is taken from a grammatically accurate sentence. BELU, METEOR, CIDEr, and other metrics are used to evaluate the generated caption and the model.[2]

It was suggested by Tanti et.al that a neural network-based model could be used to generate image captions, however instead of accessing existing captions, they created new ones using a recurrent neural network and long short-term memory. Pre-trained picture characteristics are typically used in these models (CNN). As a result, a grouping of these terms results in captions that are meaningful in relation to the image.[3]

In addition to the "hard stochastic attention" and "soft deterministic attention" models, Kelvin Xu et. al. suggested a model that automatically learns to explain the content of images. The researchers employed an attention model to study a variety of features of an image. Additionally, the model proposed in this study is able to self-learn and alter its perspectives on crucial elements in the image while delivering a descriptive sentence in which every word is related and makes complete sense. BELU and METEOR datasets are used to evaluate the model's accuracy. For example, they saw the modularity and attentiveness as valuable in various contexts for the encoder and decoder approaches.[4]

Image Caption Generator was developed by N.Kumar et al. using a deep learning approach [5]. Using a combination of image processing and computer vision, it aims to generate captions for an image. This approach recognises the relationships between people, objects and animals in an image and captures semantic meaning. It is used to identify, recognise, and offer captions for regional objects. The proposed approach is based on deep learning in order to improve the current system. The Flickr-8k dataset was used to test this strategy. The present image caption generators were unable to provide captions that were both succinct and informative.[5]

A software attention model based on ResNet50 and LSTM is used in this system. They used ResNet50,CNN(encoder) and LSTM(recognition) (decoder). The image was encoded using CNN, and decoded using LSTM (generation of descriptive sentences). In order to boost performance, they combined the soft attention model with the LSTM. For the overall process/model, stochastic gradient descent is used for convenience of use in training. The result is an increase in the number of relevant captions generated by machine.[6]

Chen et al. presented the GRU technique, which implements MATLAB using C++ in Caffe. In order to generate captions, various pre-trained models such as AlexNet, VggNet, and GoogleNet are applied to the image datasets. The evaluation is carried out using the COCO dataset. During the process of updating the LCRN method's pipeline, something went wrong. Based on the software architecture's mathematical computations, more performances are done. The MATLAB implementation is complete, however the time required to complete the GRU implementation is prohibitive. The BLEU score is an evaluation statistic for determining the degree of similarity between captions written by humans and those created by models. [7]

Images can be used to generate new words using multilayer approach proposed by Lakshminarasimhan and others.Using the flickr 8k dataset, the evaluation is carried out.The model has been evaluated using the Bleu score.The extraction method is carried out using a multimodal recurrent neural network.The CNN is modeled after the VGGNet algorithm.To begin, the image that is fed into the model is transformed into a word vector, which is then fed into the LSTM, which results in the right creation of sentences.Efforts are currently being made to improve upon this model's findings for the next challenge.They gave data (pictures) during training, and an appropriate caption was generated based on this.The Vgg model is trained, and then some mathematical operations are used to lower the loss function during training inorder to identify all of the items in the data.There is one drawback to their model: it takes longer to train on GTX 1050 and GTX 760 graphics cards with 4 GB of RAM.[8]

Encoder-decoder(CNN-LSTM) has been used by Mathur P.et.al in their model utilising GoogleNet pretrained model.They used CNN as an encoder and LSTM as a decoder in their technique. The entire process was implemented with the help of Google's numerical computation library. Pre-processing methods based on graphs were implemented by the researchers since it increased the speed of image pre-processing by six-fold. Preprocessing files linked to checkpoints were merged to generate three buffer files that let them separate three separate modules during training. They used the Flickr 8k dataset to develop their model. They switched to the MSCOCO dataset for training after experiencing some difficulties with the training dataset. BLEU, ROUGE, METEOR, and CIDEr scores are used in the evaluation process. As opposed to obtaining cutting-edge performance, they placed a high priority on speeding up caption development. Finally, they developed a smartphone application that may be used in real-world situations. This model's most difficult challenge is that all training and performances are conducted on one model at a time.[9]

Domain-specific-image caption generator[10] was proposed by Seung.et.al.Protege was used to build the domain-based ontology. Training and evaluation are done using the MSCOCO dataset; BLUE and METEOR scores are used to generate the captions. To create the embedding space from the dataset's images and text, they utilised a canonical correlation analysis. The dataset's semantic and visual knowledge and information can be gathered in two ways: top down and bottom up. R-CNN was used to extract the characteristics of the objects. Vgg-19, the most recent version, is used for feature and picture extraction, whereas "Glove" is used for text extraction. They constructed an attribute predictor based on correlation analysis. The MSCOCO dataset and the Mask R-CNN model were used to detect the image's objects. "This method is not an end-to-end method for semantic ontology," one of the drawbacks noted here. As a result, for the coming days, they're planning to make an end-to-end semantic ontology method.[10]

A model based on a deep learning method was proposed by Komal Kumar.et.al. Computer vision and image processing are both included in this concept. [11] Both a convolutional neural network (CNN) and recurrent neural network (RNN) are used in this model (RNN). An RNN and a CNN are used to extract features from images, based on their properties. The model is trained on the Flickr8k dataset. In this method, it is possible to show how the objects are related to each other and how they are located in relation to the background, but the use of RNN causes a vanishing gradient problem. A hybrid model for picture captioning can be built in order to improve the accuracy of the captions.[11]

A "Multilayer dense Attention Model for Image Caption Generation" has been proposed by Chien-Ming Chen et.al.Recurrent convolutional neural networks (R-CNNs) are used to encode descriptive text, while long short-term memory (LSTMs) are used to encode the multilayer dense attention model. It is used in reinforcement learning to select the model parameters. There are two sorts of attention processes in the encoder-decoder structure: top down and bottom up. On Imagenet, they applied the pre-trained ResNet-101 model in order to pre-train the bottom-up attention method to extract image feature models and suggested that it uses the top-down attention mechanism to calculate the relevance of each visual characteristic and generate the associated text in each time series, The attention LSTM and the language LSTM are combined in a multilayer dense attention model that has been found to be the most successful in a number of evaluation cases. The text generation model is composed of three LSTM layers. Reward learning uses strategy gradient optimization to pick model parameters. There are two sorts of attention mechanisms: top-down attention and bottom-up attention in the model architecture.For extraction the features of the objects presented in the provided images i.e datset. For this purpose, they used the pre-trained ResNet-101 model on Imagenet and indicated that it uses a top-down attention mechanism to calculate the importance of each visual characteristic and generate the related text for every time sequence. With the use of the LSTM (Linguistic Spatial Temporal Memory), this study delivers the most effective multilayer dense attention model. The text generation model consists of three layers of LSTM neural networks.With the approach of both the proces(top-down and bottom-up approach),the model of the whole process gets completed.[12]

III.PROBLEM STATEMENT

If we use static object class libraries in the inputs for picture captioning, the problem begins with the detection of the object. CNN is used today since it is based on a deep learning algorithm. Images in 2D form are sent to the CNN in this procedure. The image's aspects are assigned distinct biases and weights, which aids in the distinguishing of individual objects or the overall image. The CNN-LSTM architecture, which incorporates both computer vision and machine translation, was used in this study. Sentences with several meanings can be generated this way. The employment of a recurrent neural network is recommended. The vanishing gradient problem is addressed by LSTM, which is an RNN variant. LSTM helps to keep memories long-lasting and solves the problem of sequence prediction. This problem will be addressed by utilizing CNN as an encoder and LSTM as a decoder. The encoder first encodes the image, and the encoder then combines it; similarly, the encoder encodes the text, and the decoder then feeds it.By the help of encoder-decoder we'll overcome the problem .The combination of both the encoded image and encoded text caption by the encoder feeds to the LSTM(decoder).We'll build a architecture by merging both the CNN and LSTM which then trains the neural network.In this model,before having the predictions,the image depiction can be collaborated with the concluding RNN state.

Though it is the challenging task,it's benefits ranges from aiding the person having sight problem to enabling different labels of the images to the internet with cost saving .Nowadays on social medias like instagram and facebook,it is applicable like when user uploads the images for posting the the suitable captions for the uploaded images get created.

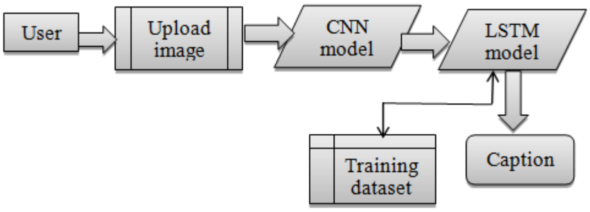


Fig2:Visual representation of this approach

IV.PROPOSED SYSTEM

We propose a CNN-LSTM model to generate the captions..The underlying principle is that photos are presented. as input first They are then recognised and a corresponding caption is provided depending on the image.Caption generation begins with tokenization, a technique that breaks down an object's glyphs into smaller components like words, symbols, and other elements.Tokens are created from the strings and then saved to a file.First and foremost, the data set is utilised for training, testing, and validation in data processing.Because certain data may be duplicates, this technique is used to remove the duplications and acquire only the original data (pure data).In the case of object detection, whatever the picture's objects are will be identified.The LSTM model is utilised in this process. Things in the photo are identified when images are uploaded. LSTM is fed the extracted feature and then uses it to generate the words based on that feature. There is a third phase of this assignment, which is the generation of many phrases, in which the objects are first recognised for the generation of words. There is an output in the form of a phrase as a result of the addition of each word to the previously formed words.We provide a model that uses CNN-LSTM neural networks to automatically identify and describe the images in a given scene. For object detection, a variety of pre-trained models are used, and the CNN-LSTM model is used to generate captions for the images. Pre-trained models are used for object detection in Transfer Learning. The CNN is used to identify specific objects and concepts in images, while the RNN is used to feed the sequential data to itself. The encoder and decoder are typically referred to as CNN and RNN, respectively.

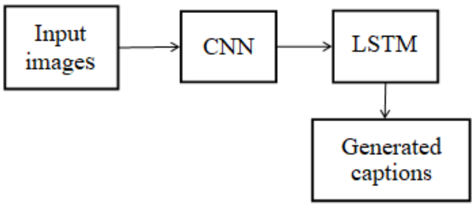


Fig 3:working process of the model

V.CNN-LSTM ARCHITECTURE

Our model consists of both the CNN and LSTM methods.

A.Convolutional Neural Network

It is the deep learning algorithm in which the user provides the image as input. A variety of items can be seen in the provided image. Images may be differentiated from each other because CNN gives each object various weights and biases.For the most part, CNN is required in order to properly classify the photos.The CNN is made up of a number of neural networks, each of which is comprised of a specific number of layers.Convolutional, pooling and fully linked layers are all included in this set.There are many ways in which these levels are interconnected.To begin, information is provided by the user (images).In the CNN design, the convolutional layer is a key component, and the filters applied to it help to create different activation features in the images it processes.A parameter that we'll be using is w (width), h (height), and d (dimension) (depth).Assuming this is a three-dimensional system, we could have kernels or filters of any size.If we have a 3*3 filter, we can use it in several ways, such as 1*1, 5*5, etc.

Assuming that it has the same depth as the input, then there will be a fixed number of kernels, which is basically how many activation filters we have Inputs are transformed into outputs by applying a set of kernels, and the outputs have the same spatial dimension as the inputs and a depth equal to the number of kernels.Stride is a factor that affects the

spatial dimension.Users can choose how Stride processes the pixels. If stride is set to 2, it means that every other pixel in the image will be processed, resulting in output dimensions W and H that are two multiples of two. This is known as down sampling. As an example, one filter may be used to extract the edges of an image and another to extract the colors.

Non-Linear Activation Layer:

This layer receives input data as the first step. Non-linearity in the dataset removes overfitting probabilities and makes the dataset more adaptable to a real-world instance where the relu is involved, therefore everything that is less than zero will be made zero.

Pooling:

Because pooling receives input from the preceding layer (as in this case, W by H by n), its output will vary depending on the window size. Let's assume that the user's convolutional layer window size is 2*2 and we reduce it to 1 pixel. Pooling can be done in a variety of ways. Max.pooling is the most widely used approach. Consider input data with a 2*2 window size.Max pooling on this results in a non-linearity and down-sampling of 9 from the given number series, after which the maximum value is picked out.In this case, down sampling is critical.Pooling can minimize the size and computation difficulty, as well as the memory and computation complexity, of CNN, which is well-known for its hefty computing and memory demands.

Fully connected layer:

This layer's primary purpose is to identify the final output categories, which explains its placement at the output stage. When the output from the previous layer is used as input for the next layer, it creates the output. This is a thin layer of nodes that connects the inputs to the coefficients of the model. Algorithms are used to select the top three coefficients from the output data once they have been added in.

B. Long Short Term Memory:

To begin, it pulls in data from CNN and uses that to build a textual description of the image that was provided. To address the STM issue, a customized version of RNN was developed.

The neural transmission network had a few flaws, which led to the invention of the RNN.

- Fails to handle Sequential Data
- Process only current input
- Can't remember previous input

The solution to these problems is LSTM. LSTMs are designed in such a way to overcome the problem of gradient vanishing.

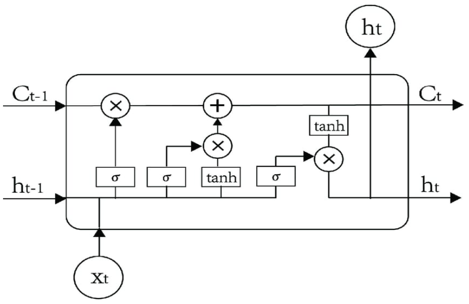


Fig 4: LSTM model

These three gates accept the current input 'x' and associate it with a vector before applying a sigmoid. It is possible to apply (C) as a new successor value to the cell state. It is now C(t) = tanh(C(t)) is obtained by applying this state to the output gate. O(t).
The bi, Wi, bf and Wf are the LSTM parameters.

C. Encoder Decoder overview

Encoder :
The CNN model is also referred to as an encoder since it performs CNN operations on an image's encoded vector. A CNN may construct the input image by including it in a fixed length under the encoder-decoder structure for the captioned image.

Decoder: To generate the output, the decoder uses LSTM(RNN) as an input and a feed-forward neural network (activation function) as an output. It decodes the encoded data (information) one word at a time to generate the image. The image is fed into the feed-forward neural network (activation function).

VI.EXPERIMENTS AND IMPLEMENTATION

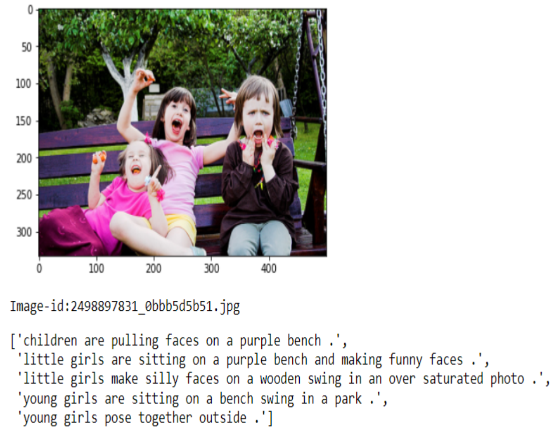
1. DATASET

Here,in this project,we have used flickr-8k dataset.There are many other large dataset like flickr-30k dataset ,MSCOCO datasets etc but we used only flickr-8k dataset because it is realistic and small dataset which helps to train the model easily as it takes less time as comparison to other datasets.If we talk about MSCOCO dataset,it’s fact that we can build a good model using this dataset but the problem is it takes a lot of time ,almost one week to train the model.

Name of the datasets	**Size**		
	Train	*Valid*	*Test*
Flickr 8k	*6000*	*1000*	*1000*
MSCOCO	*8273*	*40504*	*40775*

As part of the design process for this tool, we looked at a selection of images from the un splash and interest websites, and then added relevant keywords and captions to our dataset. We worked with tens of thousands of photographs to train the AI and identify and provide captions for a wide variety of different images, including scenery, animals, locations, and more.

Flickr 8k dataset comprises of 8000 images,each paired with 5 captions for each and every images .Out of 8k images,6000 images are used for training the model,1000 datasets are used for validation and 1000 datasets are used for testing the model



```
Image-id:2498897831_0bbb5d5b51.jpg

Out[7]: ['children are pulling faces on a purple bench .',
         'little girls are sitting on a purple bench and making funny faces .',
         'little girls make silly faces on a wooden swing in an over saturated photo .',
         'young girls are sitting on a bench swing in a park .',
         'young girls pose together outside .']
```

Fig 4. Image Caption

2)Preprocessing for image descriptions:

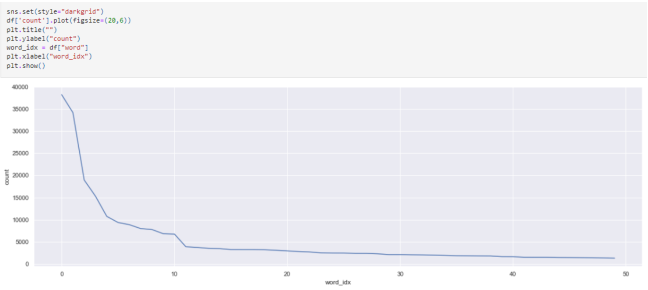
Each image contains five descriptions (captions). The important function here is Clean description, which takes all of the descriptions and cleans them up:

- Removing punctuations
- Removing Words that contain numbers

- Converting all description in lowercase

- Removing special tokens(such as ‘%’, ‘$’, ‘#’, etc.)

We tokenized our data and employed a fixed vocabulary size of 8,464 words.The top 50 most often used words are as follows:

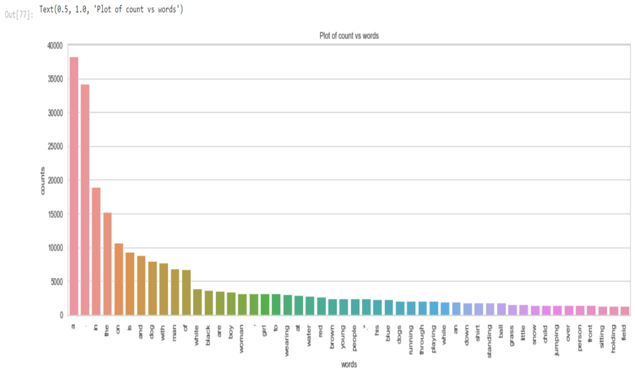


```
sns.set(style="darkgrid")
df['count'].plot(figsize=(20,6))
plt.title("")
plt.ylabel("count")
word_idx = df["word"]
plt.xlabel("word_idx")
plt.show()
```

Fig 5. word_idx and count plot

Distribution of word and count:

The top 50 words are distributed as follows:



```
Out[77]: Text(0.5, 1.0, 'Plot of count vs words')
```

Fig 6. Distribution of word and count

3)Preprocessing for images:

It's obvious that we can't feed an image directly into our model, therefore the first step is to do some preprocessing: Reduce each image to (299 * 299) for the Xception model and (224 * 224) for the VGG16 model. 2- Flatten it, and 3- Image pixel sizing (normalization)

For the implementation,we are using a jupyter notebook and

python, the deep learning model is built using the Keras 2.0 framework. The Tensorflow library is used as a backend for the Keras framework when constructing and training deep neural networks. TensorFlow, a deep learning library, was created by Google. The neural network was honed with the help of Google Colab.

We also used the following modules:

1)Keras preprocessing image module

2)keras VGG16 module

3) keras pad_sequences(..) module

4)keras layers module

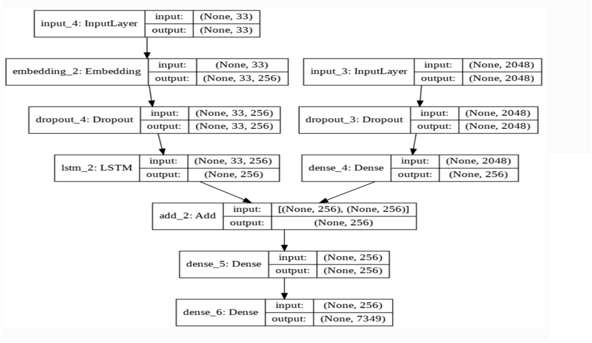


Fig7:Implementation of the model(model plot)

VII. EVALUATION:

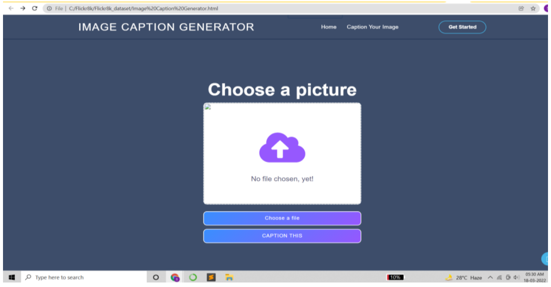
Initially, there was a disagreement over which convolutional feature extractor should be used. As a result of their training on the Imagenet dataset, we chose VGG16 and Xception for identical decoding strategies. Instead of classifying an image, we want to create a fixed-length informative vector for each one. Automatic feature engineering is the name given to this method. Each image is then reduced to a 2048-length vector by removing the final softmax layer (bottleneck features). The second problem is to compare a single model to an ensemble. There have been studies showing that ensembles boost performance; however, our results only cover the performance of a single model.Only the BLUE Score's VGG16 features were used in our analysis

Metric	VGG16
BLUE-1	0.460179
BLUE-2	0.320112
BLUE-3	0.100245
BLUE-4	0.040572

Table: BLUE score

VIII.RESULTS:

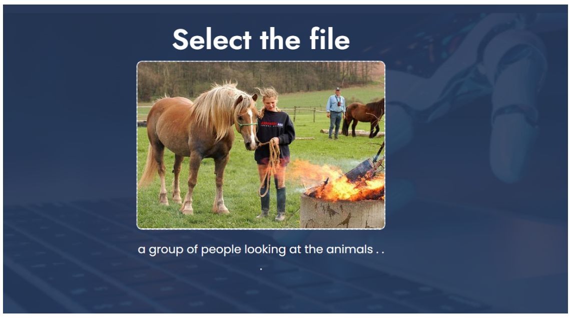
For our purposes, a BLUE score is employed. For assessing the quality of machine-translated text, this algorithm has been utilized. Our generated captions can be checked for quality using BLEU. BLEU can be used in any language. [0,1] is the range. The better the caption quality, the higher the score.



Choose a picture

No file chosen, yet!

Fig6: Upload Image



Select the file

a group of people looking at the animals . .

Fig 7. Generated Caption

IX.CONCLUSION AND FUTURE IMPROVEMENT

We utilized a deep learning strategy to caption pictures in this research. An effective BLEU score of 46.0 percent was achieved with the Vgg 16 model using the sequential API of Keras and the Tensorflow backend.Our results can be improved by applying adjustments to the VGG16 Model, such as:Using a largest set of data, an attention module is added to the model's architecture, for example.increasing the number of hyper parameters that can be tuned (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.).The cross validation set can be used to discover more about overfitting.Here is use of Beam Search instead of Greedy Search when inferring.

In the upcoming days,the importance of the image caption generator is going to be boundless in many different fields like in industries,in the vast social media. As there is the vast need of image caption generators nowadays, there must be some improvements in this field for better scope in future.Some of the models used nowadays are not giving accurate results as mostly we are also using small datasets. Instead of using small datasets,we can use MSCOCO(i.e.a large scale object detection,segmentation and captioning datasets)which helps in getting a more accurate model.COCO captions consists of 1.5 million captions describing over 1,30,000 images. As we are not getting good accuracy,a complete research must be done for some methodologies to get better results which can be better in future scope.

X.IMPROVEMENT AS PER REVIEWER COMMENTS

After discussing this project idea with seniors and experts, some changes in work-flow and algorithms are done. After suggestions we added some more captions and also we tested images of different scenarios for better generation of captions of the tool. assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. Thanks for sharing their pearls of wisdom with us during the course of this research, although any errors are our own and should not tarnish the reputations of these esteemed persons.

XI.REFERENCES

[1]Md.zakir hossain, Ferdoussohel,Mohd fairuz shiratuddin,Hamid laga,Mohammed bennamoun- “Text to Image Synthesis for Improved Image Captioning” IEEE Access(2021),Digital Object Identifier 10.1109/ACCESS.2021.3075579.

[2] Soheyla Amirian et.al.- “Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap” IEEE Access(December 4, 2020)Digital Object Identifier 10.1109/ACCESS.2020.3042484.

[3]K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov,R. Zemel, Y. Bengio. “Show, attend and tell: Neural image caption generation with visual attention” In International conference on machine learning, pp. 2048-2057, 2015.

[4]M. Tanti, A. Gatt and K.P. Camilleri. "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator" arXiv preprint arXiv:1708.02043, 2017.

[5] N. Kumar, D. Vigneswari, A. Mohan, K. Laxman, J. Yuvaraj. “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach”, In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 107-109, 2019.

[6]Shobia L,Pradheesa R, Prof.Kala.Image Captioning using deep learning methodsInternational Journal of Research in Social Science and Humanities.Volume 9.Issue 2021.

[7]JianHui ,Chen,Wen Qiang Dong,Minchen Li.Image Caption generator using deep learning approach,International conference on deep learning,2019.

[8]Lakshminarasimhan Srinivasan,Dinesh Shreekanthan,Amutha A.L.Image captioning based on deep learning approach. International Journal of Applied Engineering Research.ISSN 0973-4562,Volume 13.

[9]Mathur P.Gill A.,Yadav,A.,Mishra,A.and Bansode,N.K .Real time captioning based on videos using deep learning.International Journal of Engineering Research and Technology.

[10]Seung-ho han and Ho-Jin Choi.Domain Specific Image Generator System: A Deep Learning approach.5th international conference based on advanced computing and communication .

[11].N. Komal Kumar , D. Vigneswari , A. Mohan , K. Laxman , J.Yuvraj.Detection and Recognition of Objects in Image Caption Generator with semantic Ontology.2020,IEEE International Conference on Big Data and Smart Computing

[12] Chien-Ming Chen, Eric Ke Wang , Xun Zhang , Fan Wangi, Tsu-Yang Wu. “Multilayer Dense Attention Model for Image Caption” IEEE Access 2019,Digital Object Identifier 10.1109/ACCESS.2019.2917771.

[13]A. Mathews, L. Xie, and X. He, “SemStyle: learning to generate stylised image captions using unaligned text,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.

[14]“A gentle Introduction to deep learning Caption Generation Models”, by Jason Brownlee, November 22 2017, For deep learning Natural Language Processing.

[15]Vinyals, Oriol, et al. ”Show and tell: A neural image caption generator.” Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.

[16]Liya Ann Sunny , Sara Susan Joseph, Sonu Sara Geogy , K. S. Sreelakshmi , Abin T.Abraham.”Image Caption Generator”.International Journal of Recent Advances in Multidisciplinary Topics Volume 2, Issue 4, April 2021.

[17]Eric ke Wang,xun Zhang ,Fan Wang,Tsu-yang Wu ,and Chien-ming Chen -"Multilayer Dense Attention Model for image caption" (2019).

[18]Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 0, 0, Article 0 (October 2018), 36 pages. Computing methodologies→Machine learning; Neural networks.

[19] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, “A guide to convolutional neural networks for computer vision,” Synth. Lectures Comput. Vis., vol. 8, no. 1, pp. 1–207, Feb. 2018

[20]K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[21]Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306 (2014)

[22]Md.zakirhossian,Ferdoussohel,Mohdfairuzshiratuddin, Hamid laga,Mohammed bennamoun-Text to Image Synthesis for Improved Image Captioning (2021)

[23]Akash Verma1, Harshit Saxena1, Mugdha Jaiswal1, Dr. Poonam Tanwar2- IEEE-Intelligence Embedded Image Caption Generator using LSTM based RNN Model (2021)

[24]O.Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge", IEEE transactions on Pattern Analysis and Machine Intelligence, 2016.

[25]"Anilkumar Holambe, Dr Ravinder C Thool, Dr S.M Jagade - Printed and Character & Number Recognition of Devanagari Script using Gradient Features. International Journal of Computer Application Volume 2 No- June 2010."

[26]A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

AUTHORS

First Author-
Supriya Kumari4th Year B.tech IT Student, Sharda University.
Email -2018009684.supriya@ug.sharda.ac.in

Second Author-
Julu Basnet 4th Year B.tech IT Student, Sharda University.
Email-2018015987.julu@ug.sharda.ac.in

Third Author-
Mohit Rathore 4th Year B.tech IT Student, Sharda University.
Email-2018015523.mohit@ug.sharda.ac.in

Fourth Author-
Dipanshu 4th Year B.tech IT Student, Sharda University.
Email -2018014442.dipanshu@ug.sharda.ac.in