International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

Image Caption Generator Using CNN-LSTM

Dr .Aziz . Makandar, Keerti.Suvarnakhandi

Professor Dept of Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura , India
MCA Student Dept of Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura , India

Date of Submission: 15-10-2022 Date of Acceptance: 31-10-2022

ABSTRACT— The goal of this project is to describe what's there in the given photograph or a picture. In this project we are creating the captions for an image given by the user, by using CNN(Convolution Neural Networks) and LSTM(Long Short -Term Memory) models. These are the Deep Learning Models. Here we are using the computer vision. In this project computer can identify the given image and generate an appropriate caption in the context.

Keywords— CNN, LSTM, Deep Learning Models, computer vision, Convolutional, Caption

I. INTRODUCTION

Image Caption Generator uses the concepts of natural language processing and computer vision to predict the given image and describe it in the English like language. This model is developed by two main models of deep learning , i.e. CNN and RNN-LSTM(Recurrent Neural Networks – Long Short -Term Memory).

In this project the captions are automatically generated from the given images, that's why this project has lot of advantages in the future. Some of the applications of this project are social media sites, autonomous cars means self driving cars, CCTV cameras, and in editing apps, so on.

In self driving cars ,where it could describe the scene around the car which is helpful for the blind one. [1] Because it can guide the people by converting the scene to caption and then caption to audio. Like this Image Caption Generator Project will be helpful.

And one more application is CCTV. Here are also this project will play an important role. i.e. in CCTV cameras the images are captured then our project will predict those images and if there could be any mischievous activity will be raised means alarms will started to ring.

Many editing apps and tools use this technique in many ways. Still many advantages are there for this technique.

II. LITERATURE SURVEY

Here, we are discussing about the three main types of existing image caption generator methods. They are retrieval based image captioning, template based image captioning, novel caption generator.[3]The template based image captioning method can generate the caption for the given image using fixed templates having blank slots. Later we were using computer vision techniques for generating the captions for the image.

Finally deep learning came into existence. This deep learning technique changed the Artificial Intelligence phase. This change helps to increase the speed, accuracy and performance of the image processing method.

III. METHEDOLOGY

In this project we are using two models of deep learning. They are, CNN and LSTM.

A. Convolutional Neural Networks :

The Convolutional Neural Networks are the deep learning neural networks , which is used for classification and identification of images. And those images will be represented in the form of 2D matrix. [2] And it analyze the image from left to right and top to bottom. CNN can import the important features from the images which can help to identify the pictures whether the given picture is a bird , superman or plane**.**

B. Long – Short Term Memory :

The Long – Short Term Memory are a type of Recurrent Neural Networks (RNN) ,which is known for the sequence prediction problems. It is helpful to identify the next word based on the previous text. The LSTM is used to overcome the problems of RNN because RNN has short term memory .

In Long – Short Term Memory , there is a forget gate, which helps to cut off the unrelated data.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

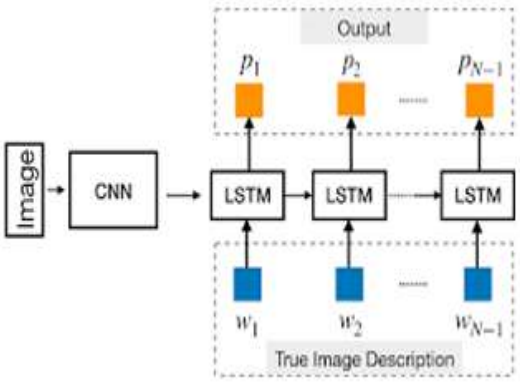

Fig1. Image Caption Generator Model

IV. PROBLEM DEFINITION

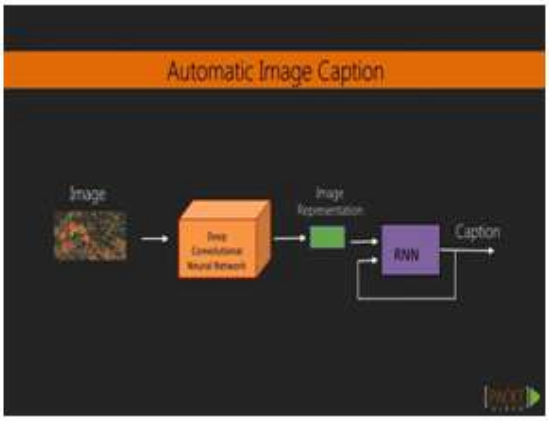

Fig2. Automatic Image Caption

The programme combines CNN and RNN, two key architectures that define properties, connections and objects in images and translate them into English.

CNN is an extractor that takes features out of an image that is provided.

The output of the CNN will be sent into the RNN – LSTM and then create a description and a caption.

CNNs , process data with the input having a two – dimensional matrix like shape.

The input layer, convo layer, pooling layer, fully-connected layers, softmax and output layers are only a few of the numerous layers in the CNN model.

CNN's input layer is a picture. Image data is shown as a 3D matrix.

Convo Layer is sometimes referred to as a feature extractor, and it uses convolutional calculates the dot products after the operation.[6] ReLU is a layer below the Convolutional layer that zeroes out all negative values.

The pooling layer one where the image's volume is diminished once the runs the convolution layer.

A connection layer is fully connected layers ,that links a layer of neurons to a layer of neurons in another layer, involving neurons, weights, and biases.

Use of the So ftmax layer allows for the multi-classification of objects utilising formulas.

The encoded result is sent to the LSTM model in the output layer, which is the final layer of the CNN model.

Recurrent neural networks (RNN) use the output from earlier steps as input feed to running procedures. [5] Long Short Term Memory (LSTM) is a prolonged using a modified RNN to forecast the sequence depending on the where it remembered all the previous steps and the anticipated at each stage, a sequence.

It takes the necessary data from the processing of inputs, forget gate, and elimination of the Non-essential data.

V. PROPOSED WORK

Work is proposed in three phases.

A. Extraction :

Images are being extracted for their various features. Vector features, also referred to as embeddings, are produced. The CNN model takes out the original photos' characteristics before being reduced to smaller and feature vectors compatible with RNNs. It also goes by the name Encoder.

B. Tokenization :

Tokenization: The application's next stage is RNN, which decodes the feature vectors from CNN were fed to it. The order of the words is as follows: is assumed and regardless of how the captions are produced.

C. Prediction :

The final stage after tokenization is prediction. the vectors in this are decoded, and the get prediction command generates the final output ()function.

VI. FLOW OF THE PROJECT

- Importing the libraries.
- Setting up the GPU memory for training purposes.
- Importing the dataset of images and their corresponding captions.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

- Plotting a few dataset photos and their captions.
- Removing captions to allow for more analysis cleaning the captions so they can be used later the top 50 terms from the cleaned dataset are plotted.
- Extracting features by loading the VGG16 model and weights.
- Features extraction.
- Plotting related pictures from the collection.
- Tokenizing the captions for additional use.
- Putting the texts and pictures through the necessary processing by the model.
- Construction of the LSTM model.
- LSTM model training.
- Making a loss value plot producing captions using the BLEU score to assess performance.

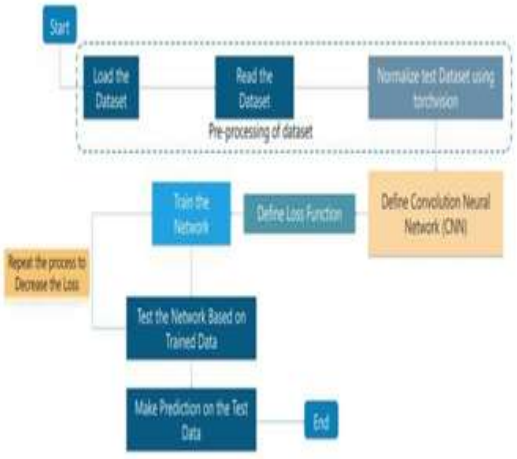

Fig3. Data Flow of Image Caption Generator Model

VII.DATA SETS

We used the Flickr 8k dataset for the image caption generator application.

This dataset includes a variety of images with numerous different types various settings and circumstances.

There are 8000 images in the Flickr 8k dataset, and each one has five captions.

We divided the 8000 image dataset into three groups: 6000, 1000, and 1000.

Each image has separate training, validation, and testing sets, accordingly dimensions.

A. Flickr 8k Dataset :

An openly accessible benchmark for image-to-set instructions is the Flickr 8k dataset.

This collection contains 8000 images, each with five captions. These pictures collected from several Flickr communities [4]. Each caption features a thorough a description of the things and things that happened in the picture.

The main file of our dataset, Flickr8k.token, which contains the names of the images and their corresponding captions, separated by newlines ("n"), can be found in the Flickr 8k text folder.

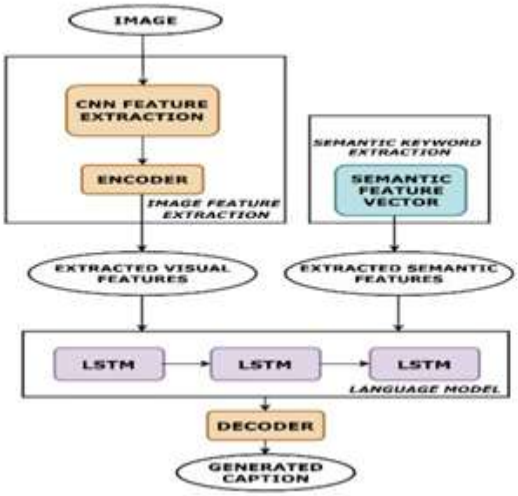

Fig4. System Architecture of the Image Caption Generator Model

VIII. SEMANTIC KEYWORDS EXTRCTION

The method of automatically determining the words that best represent the subject of a document, in this example, a grammatically accurate image, is known as Semantic keyword extraction.

The generated caption ought to include numerous more precise information on the image's characteristics, like hue.

By utilizing Semantic Keyword Extraction will improve caption quality and the generated captions will be more correctly written. [7] Upon receiving the perceptron network receives these as input in the form of vocabulary or keywords.

This network has multiple layers, each of which is identified by a unique performance. Layers like the dropout layer, softmax layer, and dense layer are in charge of identifying properties, while the model will benefit from their assistance in producing relevant captions.

The involved actions are:

Step 1: First, punctuation will be removed from the dataset to clean it.

Step 2: Data loading and preparation in step two.

Step 3: Transforming text into vectors

Step 4: Developing an own vocabulary

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

Step 5: Selecting the top 400 words from the dictionary you produced in step 4 we'll go with semantics.
Step 6: A multi-layered vector developed in the previous step is applied network.

IX. LANGUAGE MODEL

CNN-LSTM is the language model we employed since the experimental the data shows that LSTM out performs RNN for text and image production processing because the vanishing gradient issue is not present.

The LSTM resembles a more advanced RNN, which is created to address the vanishing RNN gradient problem.

They have feedback connections that enable us to process lengthy data streams, such as speech or video. They have a many uses in speech recognition, picture recognition, and other fields video processing and the recognition of handwriting. This model receives as input the image features that were retrieved from the photos in earlier steps.

With the aid of the tokenizer class, the sequence processor will produce captions. LSTM layers feature a memory component that can identify which word after the first word, the words that have meaning may follow. The image vector will be used as input by the model to produce a caption. In this instance, the InceptionV3 model of the CNN is used to obtain the vector.

X. LIBRARIES USED

Tensorflow: This free package uses Python and other frameworks to facilitate deep learning.
Keras: A deep learning library for Python that is open-source.
Pillow: A Python Imaging Library (PIL) called Pillow adds functionality for opening, picture editing and image storing.
Numpy: The Numpy library is used to work with arrays.
Matplotlib: Python library for making animated and static visualizations framework.

XI. DATA AND EVALUATION METRICS

Several annotated picture datasets are available for the task of captioning photos. The two most popular ones are Flickr and Pascal VOC dataset, MSCOCO Dataset and 8K. The dataset for captioning Flickr 8K images is employed in the suggested model. 8,092 images make up the dataset known as Flickr 8K. photos from the website Flickr.com.

This dataset includes a variety of daily activities and the subtitles that go with them. each item in the first after the image is identified, a description based on the things in a picture.[4] From this corpus of 8,000 photos, we divided them into three separate sets. 6000 pictures make up the training data (DTrain), while the each of the development and test datasets has 1000 photos.

The potential of the image-caption pairs to connect previously undiscovered photos and captions with one another must be assessed in order to evaluate the image-caption pairs. The model that creates natural language sentences can be evaluated by accomplished using the BLEU (Bilingual Evaluation Understudy) Score. It explains how human-generated sentences differ from natural sentences. It is frequently used to assess how well machine translation performs.

XII. PHASE OF TRAINING

A pair of input photos and their results are provided during the training phase to the relevant captions that fit the paradigm for image captioning.

VGG is a model that trained to recognize any conceivable thing in a picture. The LSTM component of the model, however, is trained to predict each word in the phrase after it has seen the image and all of the words before it. Each caption has an addition of two extra identifiers to represent the beginning and finish of the sequence.

XIII. IMPLEMENTATION

Create a caption generator first before producing captions. Beam search is used by the generator to produce better sentences generated. The generator passes the prior state of each iteration of the LSTM (where the image embedding is the initial state) and earlier sequence to produce the following softmax vector.

The show and tell model will then be loaded and used along with the caption generator above to provide potential sentences. together with their log probability, be printed. A graph can be performed on any compatible device once it has been defined.

These are the photo features: pre-calculated and saved using the pretrained model.[2] These qualities are then included to our model as an interpretation of a particular image, the dataset to lessen the duplication of processing each image through the every time we wish to test a different language model setup on the network.

The image features are also preloaded in real time. Because the VGG net was present and used for

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

object identification, Keras 2.0 was used to create the deep learning model.

The Keras framework is installed with the Tensorflow library as a backend for deep neural network construction and training. TensorFlow is an intricate Google has established a learning library.

It offers a diverse array of platform for algorithm execution, capable of running on little power both large-scale distributed systems with devices like mobiles many GPUs. TensorFlow uses graphs to define the structure of our network definition.

XIV. OUR MODEL

Our model is divided into three primary sections:

A. Image Feature Extraction:

Due to the model's success in object detection, the features of the photos from the Flickr 8K dataset are extracted using the Xception model. Given that it is more accurate than VGG16. Later, we'll see that. Since this model configuration learns relatively quickly, the Xception is a convolutional neural network made up of 36 nodes. These go through a Dense layer's processing to create a 2048-vector element representation of the image, which is then forwarded to the LSTM layer.

B. Sequence processor:

A sequence processor serves as a word embedding layer and is responsible for handling text input. The embedded layer includes a mask to disregard padding data and algorithms to extract the necessary text features. The last step in the picture captioning process is connecting the network to an LSTM.

C. Decoder:

The model's final stage combines the input from the image extractor and sequence processor phases using a separate operation before feeding it to a 256-neuron layer and producing a final output.

Dense layer built from the text data processed in the sequence processor phase that generates a softmax prediction of the following word in the caption over the whole vocabulary. The network's architecture can be used to analyse how text and image flow

XV. RESULTS

Fig5. Input

Fig4. Output

XVI. APPLICATIONS

- The use of virtual assistants
- Editing recommendations
- Image Coding
- Self-driving vehicles
- Social networking
- Using an application for natural language processing

XVII. CONCLUSION

To effectively generate the correct captions for the imported photos, the model has been trained and tested.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 10 Oct. 2022, pp: 1001-1006 www.ijaem.net ISSN: 2395-5252

The suggested model is based on multi-label categorization and generates captions using a CNN-RNN technique, where CNN serves as an encoder and RNN as a decoder.

In this post, we examined deep learning approaches to image captioning. It highlighted the benefits and drawbacks of, showed a general block diagram of the key groupings of, and showed how to categorise image annotation systems. Each metric and dataset's advantages and disadvantages have been listed separately.

While comprehensive image labelling methods that can produce high-quality labels for virtually every image have not yet been developed, deep learning-based image labelling systems have made substantial advancements in recent years.

Automated captioning will continue to be a hot area of research for some time, especially with the advent of new deep learning network architectures. It takes use of the captions, which are stored in a text file, and the roughly 8000 photos from the Flickr 8k collection.

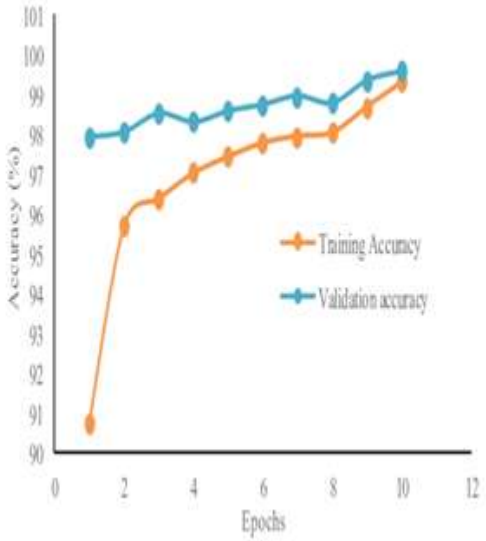

Fig5. Accuracy Level of the Result

XVIII. FUTURE AMIS

VGG16 is the CNN architecture in use. Future improvements could enable multiple target description of the captions. A range of languages should be used in the created caption.

Using larger datasets and several CNN architectures, such as LeNet, AlexNet, GoogLeNet, ResNet, and others, the model is trained and tested. For the model to be as accurate as possible, the values produced by BeLU, or BeLU scores, must be high.

REFERENCES

[1] Automatic Image Captioning Using Convolution Neural Networks and LSTM, R. Subash, November 2019.

[2] Domain-Specific Image Caption Generator with Semantic Ontology, Seung-Ho Han and Ho-Jin Choi (2020).

[3] Camera Caption: A Real-Time Image Caption Generator by Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode.

[4] Image captioning: Transforming Objects into Words, Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares (june 2019).

[5] Deep learning-based Image Caption Generator by Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, and Dr. Shabnam Sayyad (March 2021).

[6] Show and Tell: A Neural Image Caption Generator, Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015)

[7] Tao Mei, Yehao Li, Zhaofan Qiu, Ting Yao, and Yingwei Pan. enhancing the captioning of images with attributes. Pages 4904–4912 of the 2017 IEEE International Conference on Computer Vision (ICCV).