REAL TIME IMAGE CAPTION VOICE GENERATOR

M. P. R, M. Anu and D. S, "Building A Voice Based Image Caption Generator with Deep Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India Weblink

Term Paper Submission for the Course

IT414- Data Warehousing and Data Mining

Even 2022-23

By **Group 23**

AJAY SINGH(201IT205)

AJAY KUMAR(201IT106)

DEEPAK MINA(201IT117)

SATYAM VATS(201IT156)

JATIN KHOLIYA(201IT226)

under the guidance of

Dr. Shrutilipi Bhattacharjee

DEPARTMENT OF INFORMATION TECHNOLOGY NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA SURATHKAL, MANGALORE -575025

APRIL,2023

ABSTRACT

Google, the medical sector, and other industries use image processing, one of the most cutting-edge technologies available. The free and open-source tool that this technology offers, which every developer can buy, has recently attracted a lot of programmers and developers. Image processing helps in extracting a lot of information from a single image because it is currently utilized as the main method for collecting information from images, processing such images for different purposes, and executing different operations on them. Natural language processing (NLP) is a technique that is used to translate the description of a picture when constructing voice-based image captions. Finding the optimum caption for an image is the main objective of the suggested research, so combining CNN and LSTM is the best approach for this project. After being obtained, the description will be translated into the text, and the text will then be given voice. Image descriptions are the best alternative for blind people who are unable to understand sights. A voice-based image caption generator can produce the descriptions as speech output if the user's vision cannot be corrected. Future research will increasingly focus on image processing, mostly for life-saving purposes.

Keywords: NLP (natural language processing), CNN (Convolutional neural network), LSTM (Long short-term memory), RNN(recurrent neural network), ResNET(residual neural network), VCG(Visual Geometry Group)16.

CONTENTS

LIST OF FIGURES

.

INTRODUCTION

Using artificial intelligence methods to produce natural language descriptions of images is a cutting-edge technology that can be used to build voice-based image caption generators. For people with visual impairments who depend on audio descriptions to understand visual content, this technology can be especially helpful. Convolutional neural networks (CNN) are used to extract features from the input image during the creation of a voice-based image caption generator. A recurrent neural network (RNN), such as a long short-term memory (LSTM), is then used to generate the caption. While the RNN creates a natural language description of the image based on the learned features, the CNN learns to recognise the objects and features in the image. Using deep learning for this task has a number of benefits, one of which is that the model can learn the underlying patterns and correlations between the photos and their descriptions rather than depending on predetermined rules. This method produces more accurate and conversational descriptions of the images. A sizable dataset of images and captions is needed in order to create a voice-based image caption generator. With the use of these words and images, the model is trained to provide precise and evocative captions for brand-new images that it has never seen before. The model is then adjusted on a validation set, and its performance is assessed using a different test set. Building a voice-based picture caption generator presents a number of difficulties, not the least of which is how to provide natural language descriptions that truly capture the context of the image. In addition, it may be challenging to obtain high accuracy due to linguistic ambiguity and the subjective nature of describing visuals. Despite these difficulties, there are a lot of potential advantages to this technology. For instance, those who are visually handicapped can more easily access visual content with the use of this technology, and researchers can use it to learn more about how people view and describe images. Additionally, a vast number of possible uses for the technology exist in industries like healthcare, education, and entertainment. Deep learning image caption generators have made major advancements in recent years, and the technology is expanding quickly. Researchers are continually experimenting with novel methods to increase the precision and effectiveness of these models, such as adding attentional mechanisms and mixing various input modalities. Voice-based picture caption generators are anticipated to advance further as deep learning technology develops, making them more capable and practical for a range of uses.

LITERATURE SURVEY

According to the findings of this study, photo captioning models use a [1] encoder-decoder architecture, with the encoder getting input from abstract image feature vectors. Utilizing [1] feature vectors derived from an object detector's region proposals is one of the finest methods. This paper presents the Object Relation Transformer, which extends this approach by explicitly introducing information about the spatial relationships between input-detected things through geometric attention.

To increase the effectiveness of producing image captions, a method to train picture representation was proposed in this study. To extract visual representation, they apply a deep [2]fisher kernel and [2]transfer learning CNN to words. In this presentation, they create phrases using [2]LSTM to show an improvement in performance.

We found that whereas RNN or LSTM are used to produce language, CNN is used to understand image contents and recognise objects in images. The most often used data sets are [3]flicker 8k and flicker 30k, [3]MS COCO, which is used in every research. The matrix used most commonly across all research is the BLEU assessment matrix. Additionally, it was found that LSTM and CNN performed better than RNN and CNN. We found that the two most promising approaches for implementing this model are an attention mechanism and an encoder-decoder, and that combining both of them can considerably enhance results.

This research offers a [6] hybrid object identification technique and combines it with an image annotation algorithm in order to investigate the disciplines of pattern recognition and machine learning. Since the majority of automated image annotation research has relied on a single feature detection technique, it has the same drawbacks as the feature detection technique.

This study focuses on the [10] Image2Text system, a real-time captioning system that can offer human-level natural language explanations for an input image. They provide a machine translation-like sequence-to-sequence recurrent neural networks (RNN) model for producing

image captions. Contrary to most earlier research, which uses CNN (convolutional neural network) properties to represent the full image.

MOTIVATION: -

Visual content can be made more accessible for those who are blind or visually impaired by using image captions. These people will be able to comprehend and interact with the information being delivered more effectively if text descriptions of the material in photographs are provided.

Captions for photographs can help with searchability and indexing of the content. Search engines can more accurately categorize and rank visual information by using text descriptions, which makes it simpler for users to find pertinent photos.

Additionally, by giving context and meaning to visual content, image captioning can improve communication. Images are frequently unclear or susceptible to interpretation, but captions can help to make sure that the intended message is received and clarified.

PROBLEM STATEMENT

With the use of LSTM and CNN algorithms, we are developing a system that will first automatically provide a precise and descriptive description for an image in text format, which we will then translate into voice. People who are blind or visually handicapped can benefit greatly from this kind of system since it allows them to comprehend an image's significance without ever seeing it.

EXISTING SOLUTION

First, we will use the user's input to identify the object, extract certain key features from an image, and store feature vector values before feeding that information to CNN. Following this step, the long short-term memory layer will be reached in order to forecast sequence sentences. Here, the softmax function is utilized (this is the softmax layer) to precisely anticipate the output and to address the overfitting problem. This layer's output ranges from [0, 1], which indicates probabilities. The probabilities being calculated are invalid if the output values are outside the intended range, therefore the system will not correctly predict the appropriate caption for the image it is analyzing.

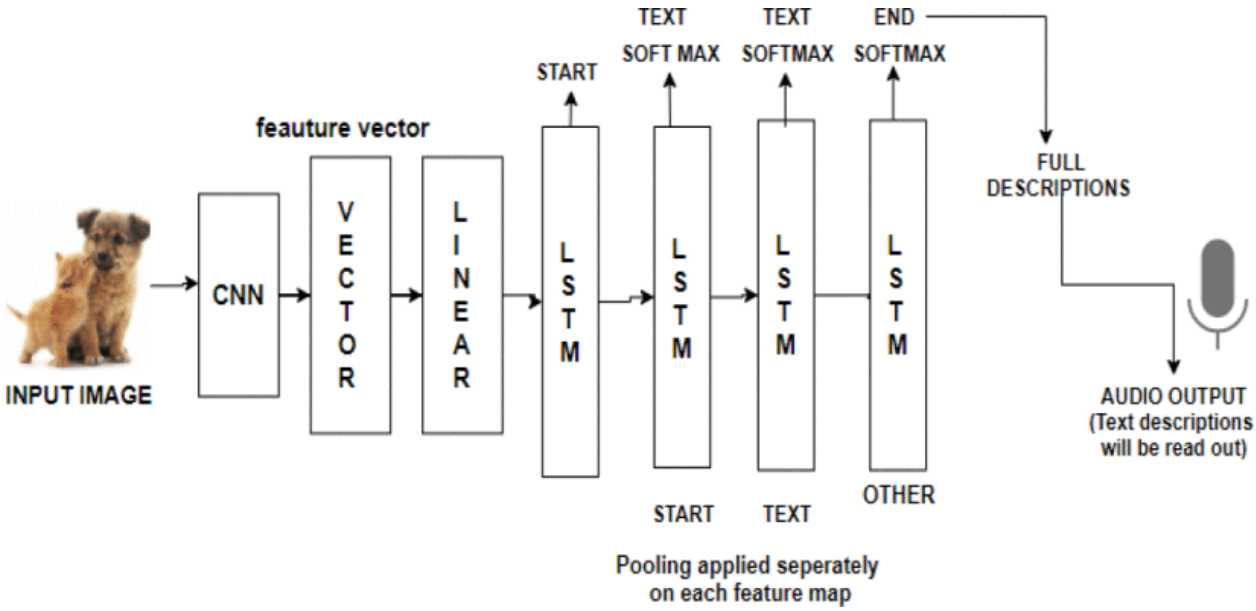

Fig1: Existing Model

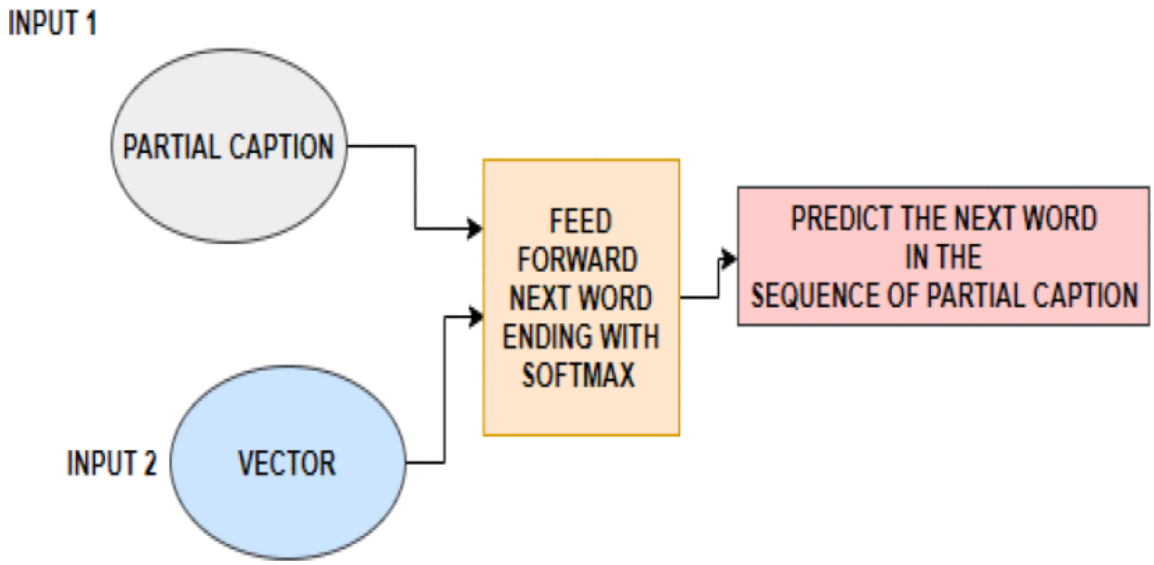

Fig2: RNN FLOW

CONVOLUTIONAL NEURAL NETWORK: -

Considering the algorithms Convolutional Neural Network is referred to as CNN. It is a specific kind of deep learning method that is mainly used for computer vision and image recognition applications. CNNs are built to automatically recognise features and patterns in images and are inspired by the way the human brain analyzes visual information.

The convolutional layer, which applies a series of teachable filters to the input image and extracts pertinent information, is a critical component of a CNN. The Rectified Linear Unit (ReLU), a non-linear activation function, is applied to the output of the convolutional layer to increase the network's ability to discriminate between different objects.

Following the convolutional layers, the network typically consists of fully connected layers and pooling layers, which minimize the spatial dimensions of the feature maps.

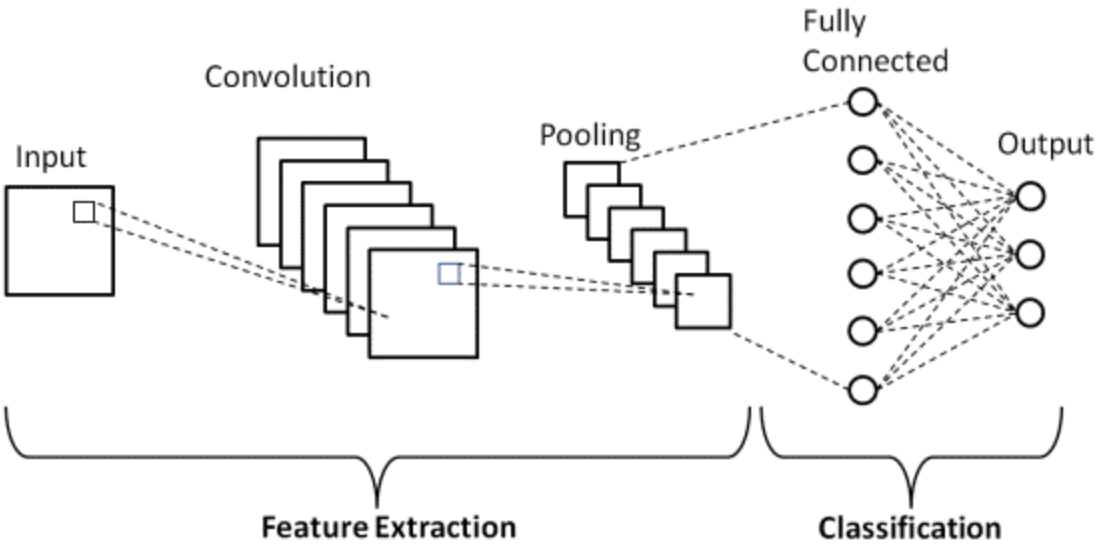

Fig 3: CNN Architecture

LONG SHORT-TERM MEMORY: -

the long short-term memory, or LSTM. It is a particular kind of recurrent neural network (RNN) architecture made to deal with long-term dependencies and maintain data over a lengthy period of time.

Memory cells, a critical component of LSTM networks, enable the network to selectively store or forget information depending on the input and the current state. Gates—learned parameters that control the information flow via the network—control the memory cells.

The input gate manages the flow of information into the memory cell; the forget gate chooses which information should be removed from the memory cell; and the output gate manages the flow of information out of the memory cell.

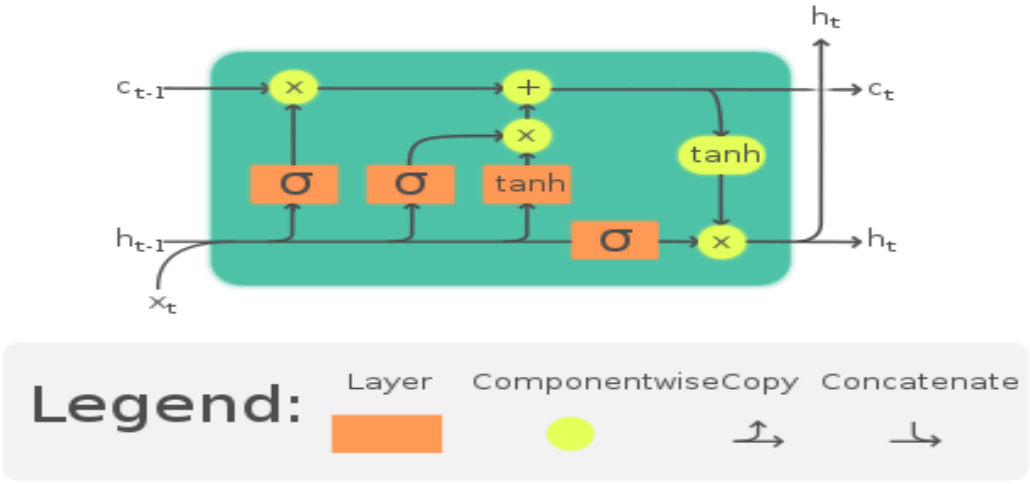

Fig 4: LSTM Structure

NEW SUGGESTIONS

- This paper generated a voice for the caption in one language only so we can add options for multiple languages. It will be helpful for people who speak different languages.

- This model can be made more advanced when it can generate captions in real-time. For example, if some user is describing some image so the model should be advanced enough to generate a caption at that time only without much delay.

IMPLEMENTATION

The image captioning system that we have built makes use of a webcam which captures the frame and extracts the features , generates the caption and then generates the voice for the caption. The project mainly consists of three parts

Dataset:

We will be using the Flickr 8K Flickr dataset, which consists of 8000 unique images and each image will be mapped to five different sentences which will describe the image.

Feature Extraction:

The image from the webcam will be taken for the feature extraction from the image. For the extraction of the features from the image we are using a pre-trained VGG16 model. This model is trained on the ImageNet dataset which contains over 1 million images across different categories. This model is capable of detecting objects, doing image segmentation and image recognition. The total 16 layers are there in the VGG16 model but have used 15 layers. We did not include the last dense layer for our project. So we used this model on the images and it will generate a feature vector of size 4096 for each image.

Tokenizing Vocabulary:

We have used Flickr 8K datasets that contain 8091 images and 5 captions for each image. Using all captions present in the datasets we have generated the vocabulary of size 8485. Using this vocabulary, a word index dictionary is created where each word has been given an index on the basis of their frequency. For example, words having maximum frequency will be given the least index and similarly it will follow for the other words.

Model building:

The model has three components: the first one is the image feature layer, the second one is the sequence feature layer and the last one is the decoder.

In the image feature component the first layer is a dropout regularization layer which takes a feature vector of size 4096 as input and dropout rate is set as 0.4 means 40 percent of the input

will be randomly set to 0 to avoid overfitting. Then output of this layer will be fed as input to the dense layer where we will be using 256 neurons and relu as activation function.

In the sequence feature layer, the first layer is an embedding layer which takes an input sequence of integers where each integer represents a word. This layer takes vocab size, embedding dimension for each word that we took as 256 and mask-zero parameter which is set as true which means any input value equal to 0 will be considered as padding. This layer is followed by one dropout layer and one LSTM layer with 256 output units.

In the last component we add the output of the image feature layer and sequence feature layer. Then it is followed by one fully connected layer with 256 output units and relu activation function. After this there is an output layer that has an output unit equal to vocab size and activation function used is softmax which converts the output to a probability distribution over vocab size.

Voice generation:

The caption that we will get for an image will be converted to voice using google tts(text to speech) library.

RESULTS

We have calculated two BLEU scores using two different weights for our model. In the first bleu score we use a unigram weight of 1.0 and all n-gram weights set to zero. This means that only the precision of individual words in the predicted caption is taken into account. Blue score-1 that we got was 0.154. In the second BLEU score, we use a bigram weight of 0.5 and a unigram weight of 0.5, with all other n-gram weights set to 0. This score considers both the precision of individual words as well as the precision of word pairs in the predicted caption. Blue score-2 was 0.087.

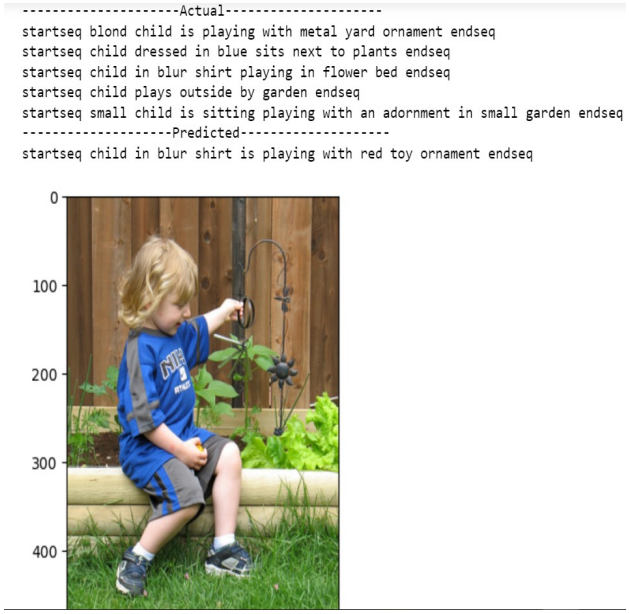

Fig 5-Example1

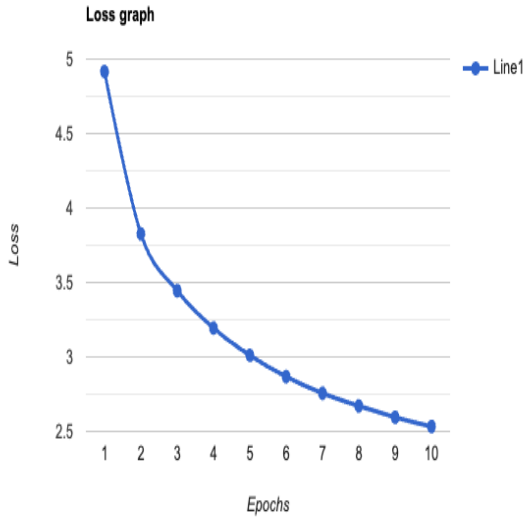

Fig 6- Loss Graph

```
In [27]: capt
Out[27]: 'startseq man wearing glasses and sunglasses is smiling endseq'
```

CONCLUSION

We have developed an image caption voice generator model using Flickr 8K dataset with VGG16 and RNN LSTM for image caption generator and Google text-to-speech technology to read out the description. Proposed project is advantageous for people with limited vision to understand images. In future when image augmentation techniques will be advanced then it can achieve exceptions. Image processing will be the subject of more and more research in the future, largely to save lives.

REFERENCES

[1] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares Yahoo Research San Francisco, CA, 94103

Link:-https://proceedings.neurips.cc/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-Paper.pdf

[2] Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim and In So Kweon, "Sentence Learning Deep convolutional neural Network for Image Caption Generation", *13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)-2016*.

Link:- https://ieeexplore.ieee.org/abstract/document/7625747

[3]Murk Chohan1 , Adil Khan2 , Muhammad Saleem Mahar3 Saif Hassan4 , Abdul Ghafoor5 , Mehmood Khan6

Link:-https://thesai.org/Downloads/Volume11No5/Paper_37-Image_Captioning_using_Deep_Learning.pdf

[4] Y. Ushiku, T. Harada and Y. Kuniyoshi, "Automatic sentence generation from images", *(ACM)Multimedia*, 2011.

Link:- https://dl.acm.org/doi/abs/10.1145/2072298.2072058

[5] S. Horiuchi, H. Moriguchi, X. Shengbo, and S. Honiden, "Automatic image description by using word-level features", *International Conference on Internet Multimedia Computing and Service(ICIMCS)-(2013)*.

Link:- https://dl.acm.org/doi/abs/10.1145/2499788.2499823

[6] K. Shivdikar, A. Kak and K. Marwah, "Automatic image annotation using a hybrid engine", *IEEE India Conference*, 2015.
Link:- https://ieeexplore.ieee.org/abstract/document/7443338

[7] R. Shetty, H.R. Tavakoli, and J. Laaksonen, "Exploiting scene context for image captioning", *Vision and Language Integration Meet Multimedia Fusion*, 2016.
Link:-https://dl.acm.org/doi/abs/10.1145/2983563.2983571

[8] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image captioning with both object and scene information", *ACM Multimedia*, 2016.
Link:-https://dl.acm.org/doi/abs/10.1145/2964284.2984069

[9] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs", *ACM Multimedia*, 2016.
Link:-https://dl.acm.org/doi/abs/10.1145/2964284.2964299

[10] C. Liu, C. Wang, F. Sun and Y. Rui, "Image2Text: a multimodal caption generator", *ACM Multimedia*, 2016.
Link:-https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/Image2Text.pdf