2022 International Conference for Advancement in Technology (ICONAT)
Goa, India. Jan 21-22, 2022

Image Caption Generation using Deep Neural Networks

Sudhakar J
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India
sudhakarj17114@it.ssn.edu.in

Viswesh Iyer V
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India
visweshiyerv17126@it.ssn.edu.in

Sree Sharmila T
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India
sreesharmilat@ssn.edu.in

***Abstract*— In recent years, computer vision has made significant progress, primarily in the field of image classification and object detection and recognition. Describing the image content automatically using natural languages is challenging and has a tremendous potential impact. Here, the idea is to extract features from an image, generate captions, and convert the generated captions to speech. This work systematically analyses deep neural networks based image caption generation. With an image as an input, the model can output an English sentence that describes the content in the image by CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), and sentence generation. The generated caption is converted to audio using Google's Text to Speech (gTTS). These models are built on the Flickr 8k dataset consisting of 8000+ images. Usually, human beings tend to describe a scene using natural languages which are compact and concise. However, machine vision systems describe the scene/image by taking an image that is a two-dimensional array.**

Keywords—Image Captioning, Deep Neural networks, CNN, RNN, Text-to-Speech.

I. INTRODUCTION

Humans are capable of processing a large amount of information in an instant. This information are most probably pictures, videos, and anything in written format. Every image has a large amount of information through which humans decipher it and process it, and their natural language is used to describe an image. Any individual can generate multiple captions for the same image. If the same can be achieved through machines, it paves the way for simplifying multiple coherent tasks. However, generating captions for images is a very tedious and demanding task for the machinery of today's world. Generating a caption using a machine includes a basic understanding of natural language processing and differentiating different objects, and correlating them. Earlier approaches were based on defined syntax, but this restricts the type of sentences created. Exploiting from the advancements in the field of image classification and object detection, it becomes feasible to automatically generate captions ranging from one or more sentences to understand the content of an image, which is image captioning [1]. In present circumstances, many well-designed deep networks are used in very massive databases. Many architectures such as GoogLeNet, which s a 22-layer deep CNN, ResNet, and many types of VGG have been introduced. The most commonly used datasets for image caption training are Flickr datasets, as shown in Fig. 1, which includes thousands of processed images.

In this paper, various existing image captioning models have been studied and how they generate a caption for the images. We have also documented the results of our implementation of the models we used (VGG16 and ResNet50) with comparison.

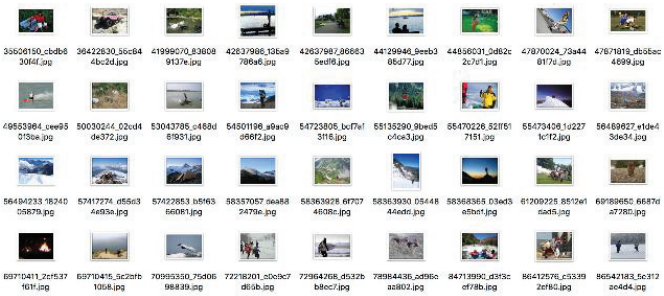


Fig. 1. Flickr 8k Image dataset

II. RELATED WORK

Human beings are competent because of their reasoning and intelligence by combining relationships in images and objects. Creating an Image captioning system that mimics human language is a very challenging task. A single image can be described by more than one sentence, which can be used as a caption, leading to text summarization in NLP (Natural Language Processing).

There are many ways to generate a caption for an image. The most common methods are a generative-based method and retrieval method. One of the best models of retrieval method was proposed and implemented by Girish Kulkarni, Vicente Ordonez, and Tamara L Berg, and it is called the Im2Txt model [4]. Their system consists of two parts – Image matching and Caption generation.

An input image is provided to the model, and consequently, matching images will be retrieved from the database, which contains the images and their appropriate captions. Once the images are found, it is compared with high-level objects from the original input images and matching images. The main disadvantage of such a retrieval-based method is that it can only generate captions already available in the dataset, and it can't generate genuine novel captions.

The limitations of retrieval-based method [7] are solved by generative-based models. It is used to create novel captions for the images. They are either pipeline-based models or end-to-end models. The Pipeline-based model uses two separate and distinct learning processes where it first identifies objects in an image and then provides the result for modeling task. In end to end based model, both language modeling and image recognition models are performed together. Both parts of the model learn simultaneously in an end-to-end system. They are usually created using a combination of CNN and RNN.

2022 International Conference for Advancement in Technology (ICONAT) | 978-1-6654-2577-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICONAT53423.2022.9726074

The show and Tell model proposed by Vinayals et al., [3] is a generative end-to-end model. It is of the forerunner models that is used as a reference in image captioning as it uses recent advancements in captioning images and recognizing images. It uses a combination of LSTM cells and Inception version 3 (v3) model.

All the above works pave the way for enhancing the models to develop image captioning systems. Using CNN and RNN is the most feasible and effective way to caption an image through a dataset.

Our contribution to the existing models is by training through Flickr8k datasets and obtaining weights of the trained dataset through which image captioning can be done. Conversion of the generated caption of the image to speech for various useful amenities for visually impaired and image recognition in self-driving cars.

III. IMAGE CATION GENERATION SYSTEM

Humans have advanced levels of reasoning and are experienced in generating captions by incorporating objects and their relationship in an image. However, creating a captioning system that precisely mimics humans is a challenging task.

A. System Architecture

The Fig. 2 shows the architecture for image captioning is based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [6].

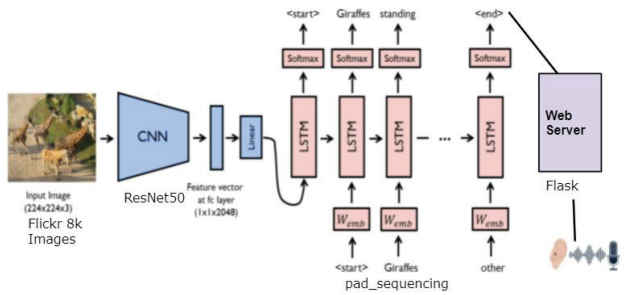


Fig. 2. System architecture of image captioning model

Convolutional Neural Network**,** usually called CNN or ConvNet, is a class of deep neural networks commonly applied to analyze images. In the model used, ResNet50[8] is used as a CNN model since it prevents degradation and vanishing gradient problems [5] in the neural nets during intensive training and helps in maintaining good accuracy. It is 50 layers deep, and it can be optimized for increased depth. Recurrent neural networks (RNN) is a class of deep neural networks that are helpful in modeling sequence data. They use patterns to predict the subsequent possible outcome. In the model used, Long Short Term Memory (LSTM) model is used as the RNN model, as shown in Fig. 3.

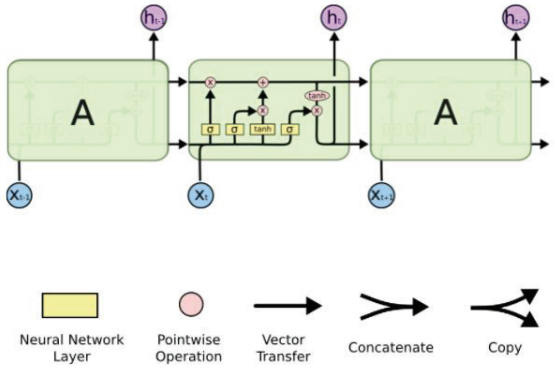


Fig. 3. Architecture diagram of LSTM (RNN)

B. Datasets

To predict any outcome of a system, training datasets are a crucial factor. For caption generation, there are many image datasets available. The most common datasets are the Flickr dataset, Pascal dataset, and MSCOCO Dataset [2]. In this work, the Flickr8k dataset is used. This dataset contains a collection of different activities that are carried out throughout the day with their related captions. First, every object in the image is labeled and followed by the description based on the objects mapped to an image. Flickr8k dataset contains around 8091 images gathered from six different Flickr groups.

C. Implementation and Training Procedure

The features in an image are extracted by training the images from the datasets using convolutional neural networks. Images are taken from the Flickr8k dataset and are fed into the ResNet50 model, where image classification and vectors of the images are mapped. There are usually two kinds of residual connections, and each has its calculation. The identity shortcuts *(x)* can be directly used when the input and output are of the same dimensions [9], as shown in Equation (1).

$$y = F(x, \{W^i\}) + x \quad (1)\ [9]$$

The shortcut still performs identity mapping when the dimensions vary, with extra zero entries padded with the increased dimension. The projection shortcut is used to match the dimension (done by *1*1* Conv) using the following Equation (2).

$$y = F(x, \{W^i\}) + Wx \quad (2)\ [9]$$

Subsequently, the trained images are fed into RNN for captioning of the images.

D. Data pre-Processing of Captions

In machine learning, data pre-processing is the easiest and cleanest method to clean the data to get error-free and unified data. During data training, captions are the target variables or outputs that the model is training to predict. Using the trained weights of the dataset, it becomes easier to test for various samples of data.

E. Text to Speech Conversion

Once the model generates the captions, Text-to-Speech creates very humanlike raw audio data. It has a broad category of custom voices to choose from. It is simply incorporated into the system using gTTS API that converts the caption to speech.

IV. RESULTS AND DISCUSSION

Multiple models were tried to train the dataset for better results. The study experiments were carried on Flickr8k Dataset.

A. Training Procedure using VGG16

The Flickr8k dataset contains 8091 images. Initially, the CNN model used is the VGG16 network framework, as

shown in Fig. 4 with image size *224 X 224*. Using VGG16 as a model [10], an estimated 29 percent was the training accuracy for the Flickr8k dataset. Image is passed through different layers of convolutional neural network with the kernel size of *3*3*. Convolutional layers are followed by three fully connected layers (the first two have 4096 channels, and the third has 1000 channels).

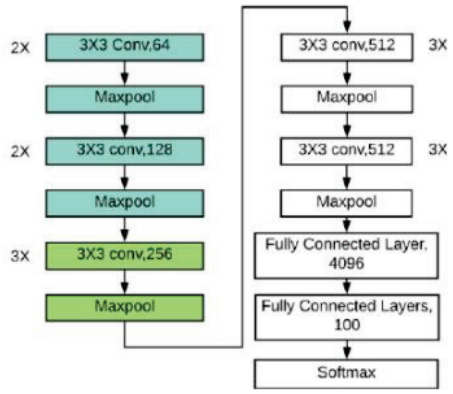
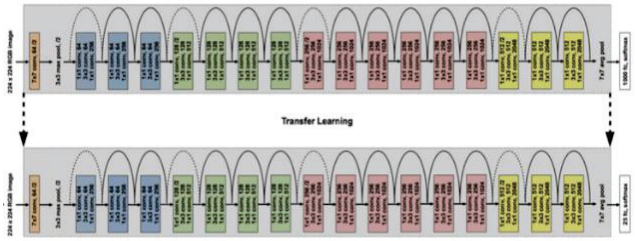


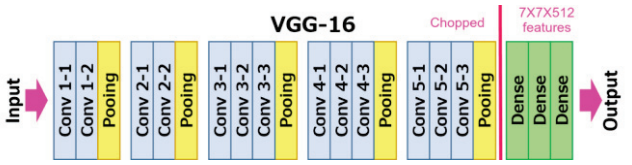
Fig. 4. VGG16 Architecture layers

B. *Training procedure using ResNet50*

ResNet50, also called Residual Networks, was used as a CNN to train the dataset [11]. When ResNet50 is used as a model (Fig. 5), an approximate 45% was obtained for training the model for 20 epochs, and 73% accuracy was obtained for training the model for 50 epochs.



(a)



(b)

Fig. 5. Feature extraction in (a) ResNet50 network (b) VGG-16 network

TABLE I. RESULTS & ACCURACY

Architecture	Online Data	Dataset Name	Training Accuracy
VGG(Existing Model)	8091	Flickr 8k	0.29 (50 Epoch)
ResNet50	8091	Flickr 8k	0.45(20 Epoch)
ResNet50	2624	Flickr 8k (Animals & Scenery)	0.73 (50 Epoch)

On generating the captions of the images, the accuracy is tabulated in Table 1. Table 1 concludes that the ResNet50 achieves an average accuracy of 79%, which is more accurate and better than VGG16 (29%).

V. CONCLUSION AND FUTURE WORK

Image captioning is a very challenging and demanding problem in various scenarios in real-time. This paper focuses on captioning an image using a Flickr8k dataset using ResNet50 as a convolutional neural network and LSTM as a recurrent neural network. Experimental analysis, testing, and training of datasets were done for both VGG16 and ResNet50 models. The results show that ResNet50 models perform better than VGG16 with an accuracy of 73% with ResNet50 and 29% with VGG16. The end caption is further converted from text to speech using gTTS.

Future works will focus on training for a larger number of images and datasets to improve the model's overall accuracy.

REFERENCES

[1] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012)

[2] Sandeep Kumar Dash, Shantanu Acharya, Partha Pakray, Ranjita Das1, Alexander Gelbukh, "Topic Based Image Caption Generation," Arabian Journal for Science and Engineering (2019).

[3] Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).

[4] Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).

[5] The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions by Sepp Hochreiter.

[6] Vaidehi Muley, Varsha Kesavan, Megha Kolhekar, "Deep Learning based Automatic Image Caption Generation," Institute of Electrical and Electronics Engineers (2020).

[7] Vijayaraju, Nivetha, "Image Retrieval Using Image Captioning," San Jose State University (2019).

[8] Zhengkui Wang, Xiao Yue, Yan Chu, Lei Yu, Mikhailov Sergei, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention" (2020).

[9] Xiangyu Zhang, Kaiming He, Shaoqing Ren, Jian Sun "Deep Residual Learning for Image Recognition," Microsoft Research, (2015).

[10] Liang Bai, Shuang Liu, Yanli Hua, Haoran Wang "Image Captioning Based on Deep Neural Networks" (2018).

[11] San Pa Pa Aung, Win Pa Pa, Tin Lay New, " Automatic Image Captioning using CNN and LSTM-Based Language Model," (2020).