*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 12 Issue 4 ǁ April 2024 ǁ PP. 62-67*

# Image Caption Generator

Dr. Anubhuti Mohindra, Prityusha Priyadarshi
*Dept. Of Computer Science & Engineering*
*Jaypee Institute of Information Technology, India*

***Abstract -*** *People who are visually or aurally impaired frequently experience a variety of hurdles in their daily life as a result of inaccessible infrastructure and social issues. They could have trouble following conversations, comprehending visual information, or reading signs. Simple things like going shopping, travelling, or even just navigating around their own area can become quite tough due to these challenges. Individuals who are blind or deaf can benefit from an android software that creates image captions by giving them a tool to better comprehend and manage their surroundings. Using an intuitive tool for creating captions and descriptions for photos, the android app we suggest in this study intends to improve the quality of life for people who are blind or deaf. The app takes pictures with a Smartphone camera and then use sophisticated algorithms to decipher the visual information of the picture and produce a natural language description. Additionally, a narrator can choose to read the caption out loud. The programme generates the caption by first extracting the image's visual characteristics using a deep learning architecture called VGG16, and then feeding those characteristics into an LSTM model. The created caption is then shown to the user after the image has been processed on a distant server via the cloud. It is a useful tool for people who are visually or hearing impaired to navigate daily life because it is portable, simple to use, and produces results quickly.*

***Keywords- visually impaired, image captions, natural language description, deep learning architecture, VGG16, LSTM model, cloud processing***

---

Date of Submission: 01-04-2024 Date of acceptance: 08-04-2024

---

## I. Introduction

One of the toughest AI problems has been teaching a computer to recognize objects and describe them in language. Recently, models that generate captions for photos have been developed thanks to improvements in Deep Learning techniques, data accessibility, and computer capacity. Both image processing and natural language processing ideas are used in this. Although challenging, the capacity to automatically explain image information using good English sentences might be immensely useful, for example, for those who are blind. Deep Learning techniques have produced encouraging results for this task, frequently by predicting a caption given a snapshot using single end-to-end model. Since this kind of learning can extract information from unstructured data, it can be especially helpful for applications in the real world.

The creation of android apps that can automatically create captions or descriptions for photographs is one example of such an application. Such a software aims to make it possible for users to comprehend an image's content without having to actively examine it. This can be especially helpful in cases where an image is too complicated to grasp at a glance or for people who are visually impaired. Aside from developing captions for social media photos, image caption creation can also be used for a number of other purposes, including indexing images for search engines, providing alternative text for images on the web, and more.

It is far more difficult to automatically create captions for photographs than it is to categorise or identify the things in an image. In order to describe an image, one must not only identify the things in it but also describe their characteristics, actions, and interactions. Previous visual recognition research has concentrated on classifying or categorising images, which is insufficient for creating captions. Instead, the semantic information needed for visual comprehension must be expressed using a language model. When given an image as input, our suggested model generates a string of words that describe the image. We create sentences using an RNN as a decoder and a pre-trained CNN as an image encoder. In addition to making a contribution to the field of image identification and analysis, we anticipate that our research will offer useful insights for the creation of image caption-generating Android apps.

## II. Literature Survey

In [1] Haoran Wang et. Al provides an invaluable resource for grasping the intricate details of image captioning techniques. The paper commences with a clear introduction to talks about the significance of image caption generation, illustrating its applications in enhancing human-computer interaction and accessibility for

visually impaired individuals. The emergence of attention mechanisms and transformer architectures is explained well, showing how they have revolutionized this field by capturing intricate image-text relationships and enabling more contextually relevant captions. Given the potential implications of AI-generated captions, addressing concerns related to bias, misinformation, and ownership could enhance the paper's relevance and broaden its impact. In [7] J. Lu, et. al present an innovative approach to image captioning that introduces the concept of adaptive attention through a visual sentinel mechanism. The authors recognize the challenge of focusing attention on relevant regions of an image while generating captions. To address this, they propose an adaptive attention mechanism using a visual sentinel. By integrating the visual sentinel, the paper's approach aims to tackle the limitation of conventional attention mechanisms that attend to all image regions indiscriminately. The model learns to identify salient regions that contribute most to the caption's context, resulting in more coherent and semantically relevant descriptions. While the paper is very insightful, more details on how the sentinel is trained and how it modulates attention dynamically in response to different image features would provide better understanding of the proposed approach.

## III. Proposed Solution Approach

In this section, we outline the proposed approach using the LSTM model and the VGG16 deep learning architecture. It takes two steps to caption an image. Extraction of the visual features from photographs with more in-depth content is the first step in creating captions that are as human-like as possible. Sending these properties to the NLP model is the next step. As illustrated in the subsequent subsections, we propose merging the LSTM model with the VGG16 architecture to complete both of these procedures.

The app is created with Java and Android Studio, and Realtime Firebase Database is used to store the model's generated captions and the images that were clicked.

This makes it possible for the data to be easily retrieved for additional analysis and for the photographs and annotations to be processed in real-time.

The app will be developed using Java for Android app development. Java is a popular programming language for Android app development and provides the necessary tools and libraries to develop high-quality apps.

Overall, the proposed solution involves using VGG16 and LSTM for generating image captions, Java for Android app development, and Firebase Realtime Database for storing images and their captions. This solution will provide an efficient and user-friendly way for users to generate captions for their images.

### A. VGG 16 Architecture

The Visual Geometry Group (VGG) at University of Oxford developed the convolutional neural network (CNN) architecture known as VGG16. This deep learning model, which is frequently employed in image classification and object identification tasks, is taught to identify objects in images. The model includes 16 layers total, including 13 convolutional layers and 3 fully connected layers, as indicated by the "VGG16" designation. Small convolutional filters (3x3) and deep architectures with several layers are features of the VGG16 architecture. While the deep architecture enables the model to learn complicated features from the photos, the tiny filters are utilised to catch fine details in the images. A further method used by VGG16 is known as max pooling, which lowers the spatial resolution of the image by selecting the highest value from a group of neighbouring pixels. This enhances the model's functionality by lowering the amount of parameters in the model.

VGG16 can recognise many objects and their features because it has been trained on a big dataset of photos and the labels that go with them. After the model has been trained, it may be used to extract features from fresh photos, which can then be applied to activities like object identification and image categorization.

VGG16 serves as a feature extractor while creating image captions. The concept is that since the VGG16 model has already learnt to identify objects and their features in photos, it can be used to extract those features from fresh images and feed them to a different model (such an RNN-LSTM) that has been trained to produce captions. Due to the model's ability to consider the objects and features in the image, it can now produce captions that are more accurate and descriptive.

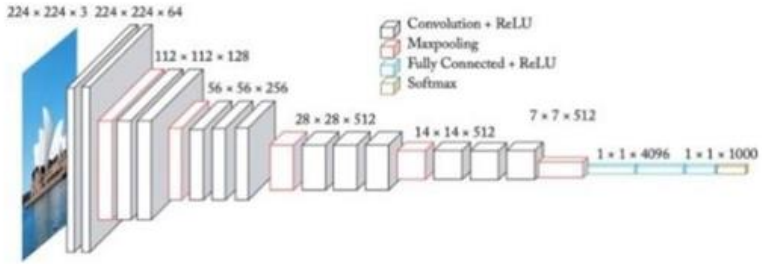



Fig 2.1 Architecture of VGG16

**B. LSTM (Long Short-term Memory)**

Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) type are frequently employed in sequential data modelling and natural language processing (NLP). The issue of vanishing gradients in conventional RNNs, which can make it challenging to learn long-term relationships in sequential data, is addressed with LSTMs.

Memory cells, gates, and input/output layers make up an LSTM network. An LSTM's memory cells hold data for a longer period of time, and the gates control when data is added to or removed from the memory cells. This enables LSTMs to keep track of data from earlier in the sequence and use it to guide predictions in the future.

An LSTM model is trained on a dataset of images and their accompanying captions in the context of creating image captions. The LSTM is trained to predict the next word in the caption given the picture attributes and the words that have come before using the image features retrieved from the VGG16 model as input. This enables the model to provide captions that reflect the contents of the image and are both coherent and semantically relevant.

Because they are effective at handling sequential data, like the word sequences in captions, LSTMs are frequently used in the creation of image captions. Additionally, as image captions frequently express intricate interactions between image objects, LSTMs have the capacity to learn long-term dependencies in the data. A caption generator for images can be created that produces accurate and descriptive descriptions by combining VGG16 for feature extraction and an LSTM for caption production.

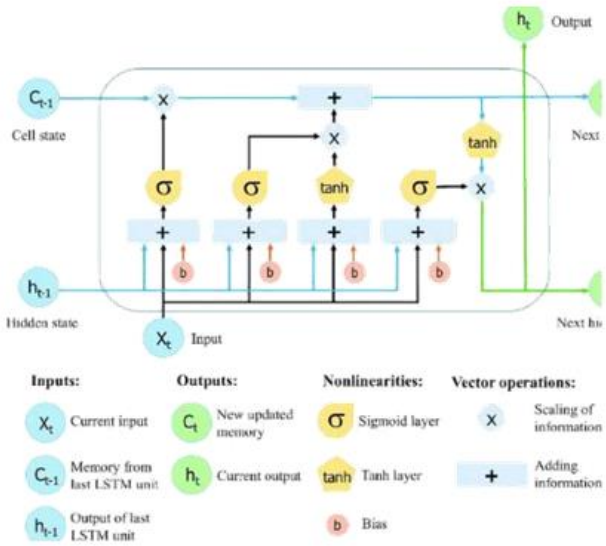



Fig 2.2 Architecture of LSTM

## IV. Results

### A. Dataset

To make the image caption generator, we'll use the Flickr 8k dataset. Instead of using the bigger Flickr 30K and MOCOCO datasets, which can take weeks to train the network, we will use the smaller Flickr 8k dataset. A large dataset allows us to build better models. The Flickr 8k dataset contains numerous images depicting various situations and settings. The dataset's 8000 images each have five linked descriptions. Using 6000, 1000, and 1000 pictures each, we created training, validation, and testing sets from the data. The sizes of the pictures range widely.

### B. Final data

The model has been trained for 20 epochs. As number of epochs used are more, it helps to lower the loss to 2.48. Some results generated are as shown in Fig. 3.1. By using the Flickr8k dataset for training model and running test on the 1000 test images available in dataset results in BLEU = 0.544012.

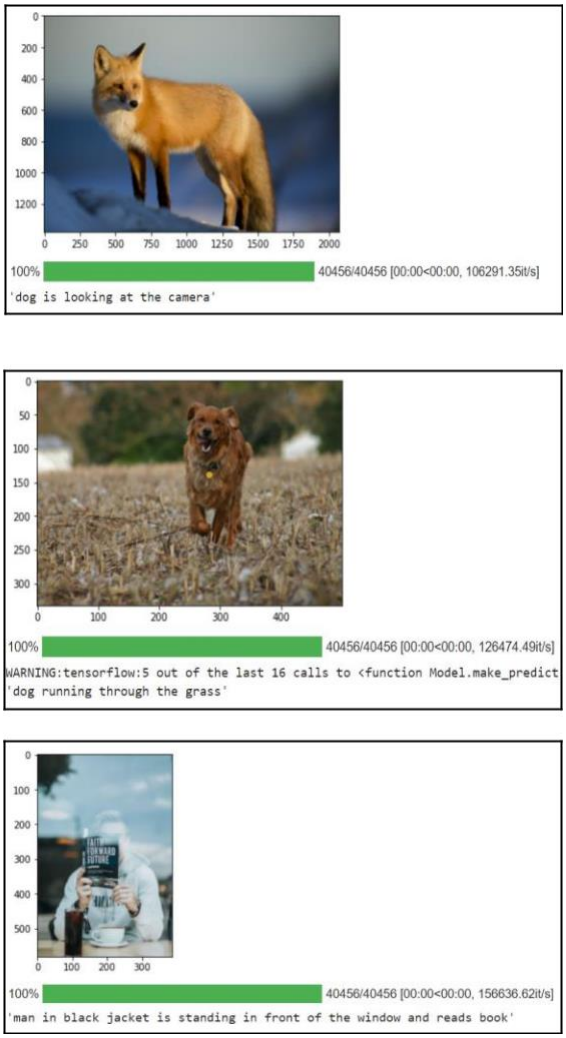
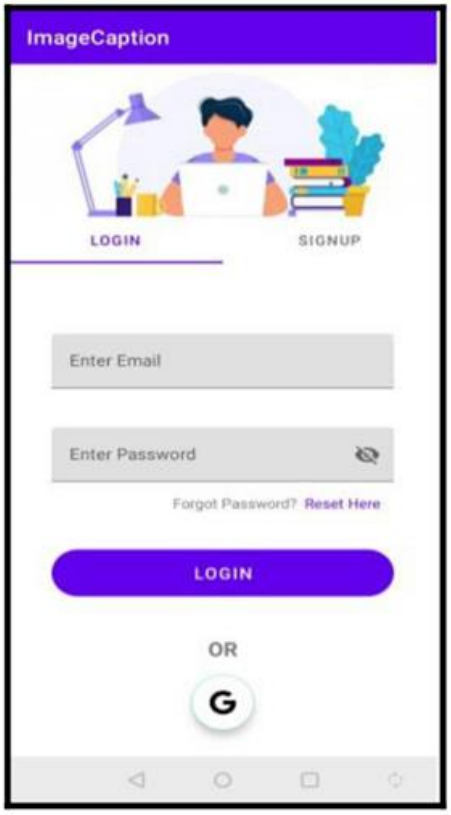



Fig 3.1 Evaluation Results

Fig 3.2 and 3.3 demonstrates a portable smartphone based platform for image captioning controlled by software, named as ImageCaption, developed in Android Studio. A simple and user-friendly interface is designed to provide a simple operation for visually and hearing impaired.



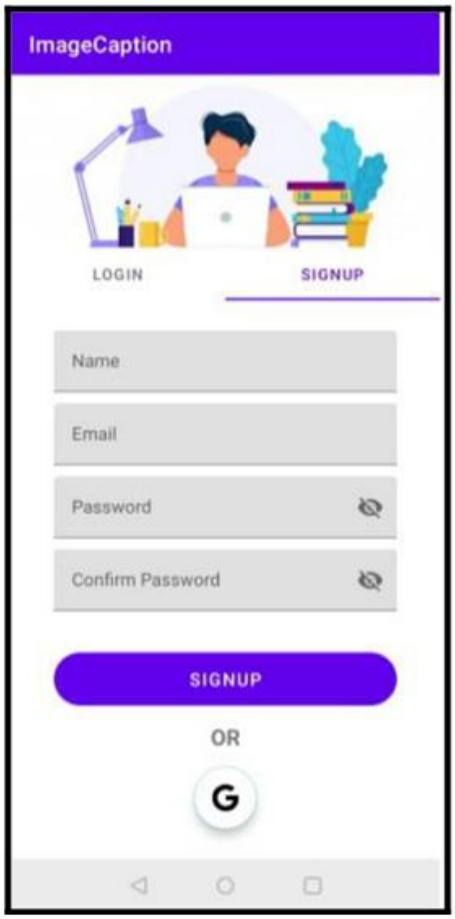
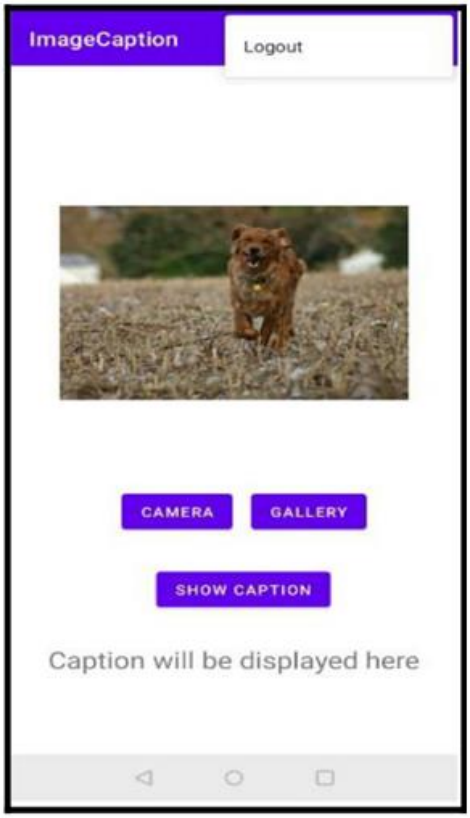




Fig 3.2 Login and signup pages



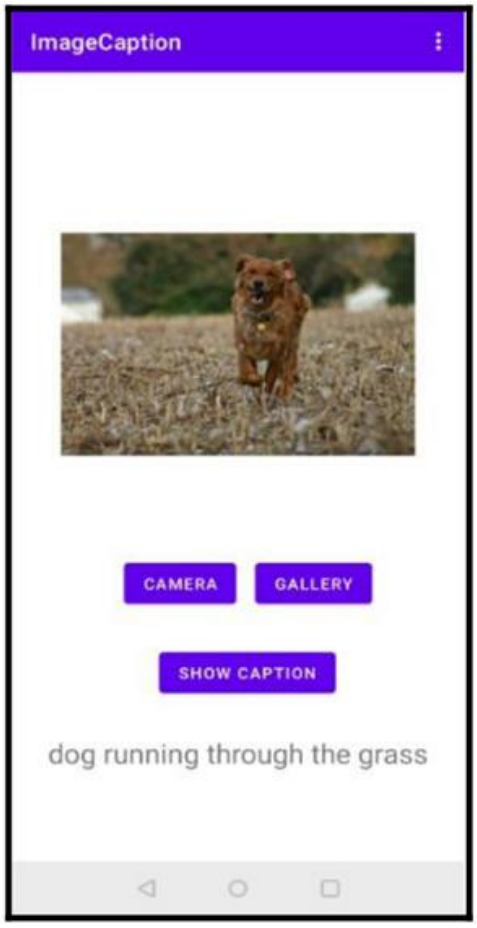




Figure 3.3 Homepage of ImageCaption App

The proposed system for image caption generation uses a combination of a pre-trained VGG16 model and an LSTM model. The VGG16 model is used to extract the visual attributes of an image, which are then fed into the LSTM model to generate a caption. Once the model was trained, it was integrated into the android app. Sample caption generated by the app is shown in Figure 2. These captions, such as "a dog running through grass" are close to the visual content of the images and have a natural-looking text.

In conclusion, the proposed system demonstrates potential for use in image captioning for visually and hearing impaired individuals.

## V. Conclusion

In this research paper, we proposed a model for an image caption generator android app that utilizes the CNN VGG16 for feature extraction and the RNN LSTM for model creation. The app is built using Android Studio and Java, and utilizes Realtime Firebase Database for storing image clicked and caption generated by the model in the backend. Our proposed approach was tested on the Flickr 8k dataset and then combined with our specially designed Android application "ImageCaption." The user either chooses an image from the gallery or takes a brand-new picture with their smartphone's camera. The chosen image will be uploaded to Firebase before being sent to the remote server that implements our suggested image captioning strategy. To show the caption, the created captions are sent back to the app using Firebase.

The results show that the model is able to accurately and concisely describe the content of the images and the app performed well in terms of user experience. This research provides valuable insights for the development of image caption generator android apps and contributes to the field of image recognition and analysis.

The app will be further enhanced to incorporate a number of features, including the ability to listen to the captions, the ability to translate the English captions into Hindi or other languages, and support for the iOS operating system.

## References


[1]. HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020).

[2]. B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, (international Journal of Advanced Science and Technology- 2020 ).

[3]. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, “Show and Tell:ANeuralImageCaption Generator”,(CVPR 1, 2- 2015).

[4]. Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, “Pointing novel objects in image captioning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 497–12 506.

[5]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, et al., "Show, attend and tell: Neural image caption generation with visual attention", Proceedings of the International Conference on Machine Learning (ICML), 2015.

[6]. J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, “Smart guid- ing glasses for visually impaired people in an indoor envi- ronment,” IEEE Transactions on Consumer Electronics, vol. 63, no. 3, pp. 258–266, 2017.

[7]. J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[8]. ”Every Picture Tells a Story: Generating Sentences from Images.” Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen

[9]. Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).

[10]. Andrej Karpathy and Li Fei-Fei, "Deep Visual-SemanticAlignments for Image Description Generation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, Apri1 2017.

[11]. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, et al., "Every picture tells a story:Generating sentences from images", European conference on computer vision, pp. 15-29, 2010.