Real time Image Captioning Enhancing Accessibility for Visually Impaired

Mr.P.Ravi[1], Ms.Ch.Vasundhara[2], Mr.P.Subhash[3], Mr.K.Madhusudhan[4], Mr.N.Jashuva[5]
[1]*Assistant Professor, Department of Computer Science and Data Science, Raghu Engineering College, Visakhapatnam*
[2,3,4,5] *Student, Department of Computer Science and Data Science, Raghu Engineering College, Visakhapatnam*

Abstract: **The world becomes restricted for visual impaired people because they need additional help with accessing visual materials and making sense of visual elements. The proposed solution consists of real-time image conversion through visual data into descriptive audio output. Deep learning algorithms enable the system to extract image features from VGG16 and ResNet and DenseNet which function as pre-trained Convolutional Neural Networks (CNNs). After extracting information from images the Gated Recurrent Unit (GRU) network processes this data to produce relevant verbal descriptions. The TTS engine converts the created captions into spoken words so users can hear descriptive audio descriptions of their visual content. The integrated system combines computer vision techniques with natural language processing methods to generate correct descriptions which enables visually impaired people to better understand their environments. Users can interpret images without help through this independent system which requires minimal effort. Through its ability to offer inclusive interfaces the technology improves quality of life opportunities for people with visual impairments. The system provides users with an effective real-time audio description method which delivers easily understandable visual information.**

Keywords: Image Processing, Convolutional Neural Networks (CNN), Gated Recurrent Unit (GRU), VGG16, ResNet, DenseNet, Text-to-Speech (TTS), Assistive Technology, Natural Language Processing (NLP)

1. INTRODUCTION

Visual interpretation forms a fundamental basis through which human beings function in their daily activities. Visual impairment produces major obstacles which block people from recognizing surroundings and reading signs also restricts movements across environments. Real-time descriptions of visual scenes become impossible with assistive tools like screen readers and Braille although they offer text alternatives. Image captioning stands as an advanced computer vision method.

The image description technology based on computer vision effectively transforms visual content into spoken words which enables easy access to visual information for visually impaired people. Modern artificial intelligence and deep learning advancements improve how precisely and effectively systems generate image captions.

The models unite vision software with language processing technology to study image contents for the production of organized text descriptions with human-like characteristics. Technological limits for real-time image captioning persist due to natural lighting variations and object blocking and complex visual scenes. A complete system which includes feature detector and sequence predictor and speech synthesizer generates high-quality image descriptions. A deep learning framework containing VGG16, ResNet and DenseNet models functions as a feature extraction system which allows precise analysis of pictures.

The spatial object details along with context-based characteristics which are essential for precise caption generation are successfully extracted by these pre-trained convolutional neural networks. A sequence of features generated by the Gated Recurrent Unit-based Recurrent Neural Network creates descriptive captions from the extracted details. Clips of captured images are converted to audible speech output through Text-to-Speech technology for visual impairment auditory feedback. The main challenge arises from safeguarding the high quality of extracted features across multiple image types. The performance output of models gets impacted by various objects and background elements hence the integration of VGG16 together with ResNet and DenseNet networks strengthens results.

These three CNN architectures named VGG16, ResNet and DenseNet collaborate to improve robustness performance. Each architecture serves a distinct purpose because VGG16extracts features efficiently yet simply while ResNetexcels in deep feature processing alongside DenseNet that promotes effective layer information flow for preserving detailed features. The generation of meaningful text captions that maintain context proves to be another difficult task for the system. The traditional Recurrent Neural Network models encounter fading gradient problems which restricts their ability to identify extended relationships between data points but Gated Recurrent Units handle these issues through their gated mechanisms that enhance information transmission resulting in better natural and precise captions. The system faces one last challenge regarding its ability to provide real-time performance whereas user demand immediate feedback to make it usable in reality. The system maintains user satisfaction with natural and unbroken speech synthesis by using optimized TTS algorithms.

The deep learning system provides vital progress to assistive technology through its automatic real-time captioning mechanism for visually impaired users. This system employs AI automation to produce instantaneous accurate speech captions which function through CNN features and GRU generation along with TTS speech output synthesis methods. The innovative system provides users with an easy method to comprehend visual information via speech outputs. The use of VGG16, ResNet and DenseNet systems creates a robust feature extraction process and GRU-based text generation leads to better contextual captions.

Users gain easy access to visual information after the TTS engine converts texts into natural-sounding speech. This research builds AI assistive technology that enhances independent functionality and digital accessibility for visually impaired people through addressing key accuracy and processing time and efficiency requirements.

2. LITERATURE REVIEW

[1] Priya et al. created a CNN-LSTM-based voice-generating image caption generator to caption speech automatically and convert text to speech for the visually impaired. The system utilized pre-trained models for prediction of sequences and extraction of features, using the Flicker dataset as the training corpus. The study highlighted the model's ability to produce accurate descriptions, making it more accessible for the visually impaired.

[2] The study by Goel et al. examined deep learning models for image captioning through an evaluation between three distinct CNN architecture types namely Xception and VGG-16 and ResNet50 for extracting features. The authors integrated LSTM with these descriptive captions by utilizing the Flickr_8k dataset for their evaluation process. ResNet50 yielded 0.84 BLEU score as the study established it as the top model because it exceeded other models in accuracy and avoided vanishing gradient problems.

[3]The research by Patnaik et al. employed EfficientNet as the primary CNN model to develop automated image captioning systems. Their work demonstrated how EfficientNet provides the benefits of complex computation reduction alongside high accuracy performance. They executed training of an encoder-decoder system with attention control through MSCOCO dataset to produce descriptive image captions. BLEU scores show that EfficientNet-based models generate superior captions than VGG16 and Inception by producing highly relevant results with proper grammar.

[4]Hemavathy et al. proposed a voice-enabled object recognition system using AI for visually impaired users. Optical character recognition, object recognition, and human recognition using a convolutional neural network (CNN) and K-nearest neighbor (KNN) were used for classification. The app used images from a smartphone camera and provided voice output in real-time using text-to-speech (TTS). The paper showed the feasibility of deep learning in improving accessibility so that visually impaired users can move around on their own.

[5]The researchers, Jahan et al. created an image-to-speech framework through the combination of CNN technology along with LSTM processing and Google Text-to-Speech (GTTS) functionality for assisting visually challenged users. The technical process started with CNN features extraction followed by LSTM-generated descriptions which GTTS converted into speech output. System evaluations demonstrated that the technology produced effective audio descriptions which proved its usefulness as an assistive technological solution.

[6]Vyawahare et al. created an image-to-sound translation system using convolutional neural networks (CNN) and optical character recognition (OCR) for assisting the blind. Their system utilized deep learning for object and text recognition and audio synthesis for real-time oral description. Their system was made to enhance accessibility by converting visual information into structured sound cues so that visually impaired users could navigate their surroundings on their own.

[7]The researchers, Hussan et al. investigated how deep learning can detect objects and recognize them in real-time for people who are visually impaired. Their system integrated deep neural networks alongside Google Text-to-Speech Google API for object identification followed by vocal description capabilities. The detection system applied a YOLO-based model to identify outdoor objects with 91 distinct categories in its recognition process. The research proved that assistive technology solutions function to boost the independent capabilities of visually challenged people in their daily activities.

[8]Ranganath et al. designed a voice-based object detection system for assisting visually impaired users. They used YOLOv3 for detecting objects and Web Speech API for voice description in real time. The system aimed at improving perception and navigation by means of accurate object identification and voice notification. The study established the viability of deep learning approaches for accessibility enhancement in visually impaired users.

[9]Nasir et al. developed an Android app using deep learning to enable visually impaired individuals to identify objects. Their approach used a single-shot detector (SSD) and a convolutional neural network (CNN) with a pre-trained MobileNet V2. The system enabled users to capture images, which were analyzed and narrated using audio output. The study demonstrated that the model was highly accurate and worked offline, and therefore it could be a reliable tool for visually impaired users.

3. MATERIALS AND METHODS.

3.1.Dataset

Flickr8k dataset consists of 8,091 images with five human captions per image and is a widely used testbed for image captioning. The images are a diverse set of objects, scenes, and activities, thereby enabling models to learn rich visual contexts. The dataset was divided into 6,000 images for training with the remaining for validation and testing. The captions consist of having multiple descriptions for an image, thereby enhancing captioning model generalization. The images were resized to (244,244). In the case of captions, a tokenizer was built and spaces and non-essential punctuation were removed. A word-to-index and vice versa mapping was then built.

*3.2.*Preproccesing

We resized the images to (244,244). The captions, we tokenized and removed spaces and unwanted punctuation. We then created a word-to-index and vice versa mapping.

3.3.Convolutional Neural Network

Convolutional Neural Networks (CNNs) are also a type of deep learning software that is inherently well-suited to the processing of data that has both a spatial and a temporal relationship. CNNs, like other types of neural network, also add an additional level of complexity with the inclusion of convolutional layers. This makes them particularly well-suited to process this type of data naturally. There are three main components to a CNN:

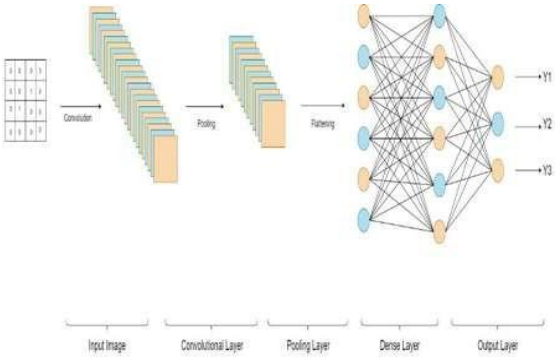


Fig. 1. CNN Architecture

1)Convolutional layer: The convolutional layer is the most important part of a Convolutional Neural Network, where linear and nonlinear processing is done in a bid to extract features from an array of numbers referred to as a tensor. The convolutional layer employs a small numerical array, referred to as a mask or kernel, in an effort to scan the input.

2)Pooling layer: The pooling layer passes a kernel or filters across the input image. A pooling layer provides a default operation that reduces the image dimension through max or average pooling.

3)Fully connected layer: The fully connected layer of the CNN classifies images according to the features extracted from the previous layers. In fully

connected, all input neurons are connected to all active neurons.

3.4. Types of CNN Used:

- *VGG16 Net:* It also has 138 million parameters. Replaces large kernel filters in Convolutional layers to 11 and 5 in the First and Second layers. Input to this model is fixed to 224*224 RGB image. It has 16 layers, as the name suggests, that have some weights. It is very good for Deep Learning for setting benchmarks on any particular task.
- *Dense Net*: It connects each layer to all subsequent layers, allowing feature reuse and gradient passing. DenseNet eliminates the vanishing gradient problem, with enhanced training efficiency. Dense Blocks and Transition Layers enable feature propagation, avoiding redundant parameters. DenseNet is more accurate than ResNet and VGG16 but with fewer parameters. Its efficient design makes it a good fit for deep learning tasks, like image captioning.
- *ResNet:* ResNet with two or more-digit number is the ResNet model with a specific number of Neural Network Layers. ResNet is able to train over thousands of layers and work very efficiently. ResNet employs its backbone in Batch Normalization. It enhances its performance with a bottleneck residual block structure. It safeguards the network from vanishing gradient issues with the Identity condition.

3.5.Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural network whose current input is taken from the output of the previous step as the input of the current step. RNNs are capable of creating models for time-series and sequential data issues such as stock price prediction and text creation. RNNs' Hidden Layer or State is an important parameter that stores information regarding the input sequence.

Gated Recurrent Unit: The Gated Recurrent Unit (GRU) is a variant of Recur- rent Neural Network (RNN) that serves as a simpler alternative to Long Short-Term Memory (LSTM). Invented in 2014, GRU leverages gating mechanisms to regulate information transfer between cells, making it faster and more memory-efficient than LSTM. However, in datasets with longer sequences, LSTM is more accurate than GRU. GRU is commonly utilized in natural language processing, speech recognition, and other sequential data applications.

3.6.Model Creation

We made a train and test split with the train test split library on image path and captions. The data was divided in 80:20 per cent, respectively. We constructed the encoder, which is CNN, that is in charge of extracting and vectorising features. Then we utilize an interface to bridge the encoder and decoder. The input to the GRU is a vector concatenated with an embedded vector. As an

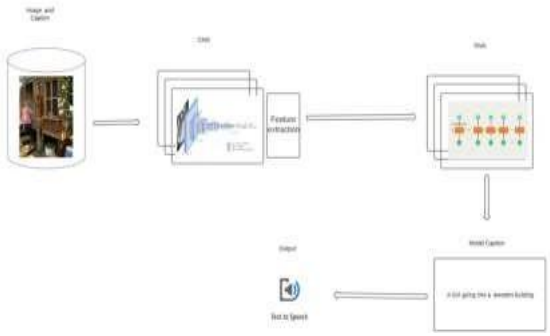


Fig. 2. System Overview

output, it produces the caption.The same procedure was performed for all the Cnn models with GRU.
The size of the epoch was 25. And this was evaluated with the Bleu metric. The best Bleu score of 0.8 was provided by ResNet50.

3.7. Model Architecture

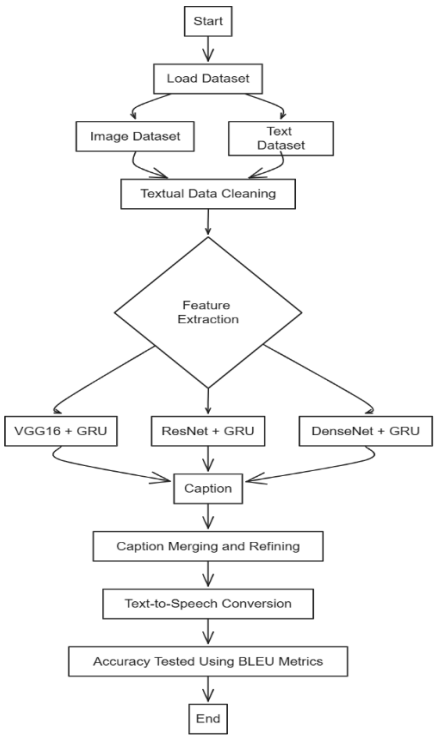


Fig.3.Architecture model of image captioning system

4. RESULTS AND DISCUSSION

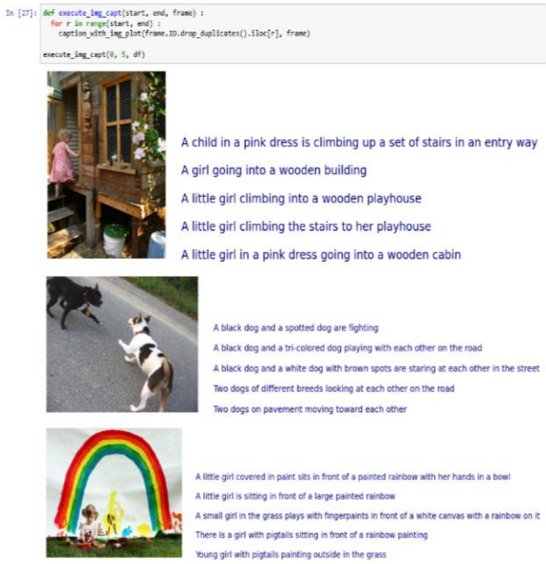
The model was trained with the dataset as the input. CNN and GRU were utilized in the project to generate captions for images: CNN for image feature extraction and GRU for caption generation. The model was able to summarize the content of an image upon training. The outcomes are as follows.

A. Xception-GRU

As indicated, the image's caption generator labels the image with a caption, for instance, "A man holds a dollar bill in front of his face while posing in fronof a street band." In some situations, though, the model generates incorrect captions, as indicated, whereby the caption is not in the image content. The accuracy of Xception-GRU is gauged by its feature extraction process, which fails to extract finer details at times. The model attains 51% accuracy for 10 epochs, and improving the number of epochs,and as the number of epochs is increased,the accuracy is 79%.

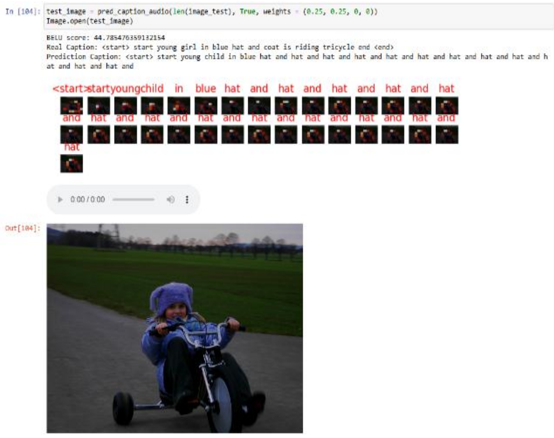
B.VGG16-GRU

Two dogs are playing or fighting, as we can see. The model can determine the most significant objects within the image and caption them appropriately with descriptions such as "A black dog and a tri-colored dog playing with each other on the road." It demonstrates the model's ability to determine color and object interaction, which is beneficial for the blind. The model, however, comes up with the incorrect caption, such as "A black dog is running on the grass," which is incomplete, explaining the action within the image. The last model with 50 epochs has 75% accuracy, which is better compared to earlier versions.



C. DenseNet-GRU

In picture, a young girl is sitting in front of a painted rainbow, playing with finger paints. The caption generated, "A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it," accurately describes the image. This highlights ResNet-GRU's superior feature extraction capabilities. Similarly, shows another example where a child in a pink dress is climbing a set of stairs. The model correctly captions it as "A little girl climbing into a wooden playhouse," demonstrating high accuracy. With a BLEU score of 44.78 and a final accuracy of 84%, ResNet-GRU outperforms the other models in generating contextually relevant captions.



D. Model Performance Analysis

The loss plot shows the training loss and test loss across several epochs. Loss is high at the start but declines as the training progresses, showing improved learning. Comparison of models is shown in Table 1.

MODEL	ACCURACY SCORE		
	10 EPOCHS	*20 EPOCHS*	*50 EPOCHS*
Xception-LSTM	51%	66%	79%
VGG16-LSTM	54%	64%	75%
ResNet50-LSTM	60%	71%	84%

TABLE 1. Accuracy score of model v/s epoch.

The procedure was performed on various CNN models, and the performance was determined based on the Bleu score from the table and graph above, we can observe that the ResNet-GRU model performed the best.

5. CONCLUSION

This research seeks to develop an intelligent system with deep learning technology to assist people with visual impairment. The system consists of the following: CNN extracts useful information, and GRU detects visual input. The system utilizes a speech synthesis module to provide users with voice messages of information. According to the research, GRU models are best for image description and prediction. Despite the limitations, such as utilizing the Flicker8k datasets, this intelligent system simplifies the understanding of text and images for people with visual impairment. Future research can utilize larger datasets to learn and acquire image descriptions from real-time images and their descriptions, Thus improving the system.

6. REFERNCES

[1] M. P. R, M. Anu and D. S, “Building A Voice Based Image Caption Generator with Deep Learning,,” 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India,, pp. 943-948, 2021.

[2] N. Goel, A. Arora, P. Kashyap and S. Varshney, “An Analysis of Image Captioning Models using Deep Learning,,” International Conference on Disruptive Technologies (ICDT), Greater Noida, India,, pp. 131-136, 2023.

[3] S. V. Patnaik, R. Mukka, R. Devpreyo and A. Wadhawan, “Image Caption Generator using EfficientNet," 2022 10th International Conference on Reliability,,” Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India,, pp. 1-5, 2022.

[4] J. Hemavathy, A. S. Shree, S. Priyanka and K. Subhashree, “AI Based Voice Assisted Object Recognition for Visually Impaired Society,” International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, pp. 1-7, 2023,.

[5] A. Jahan, S. Shadan, Y. Fatima and N. Sultana, “Image Orator - Image to Speech Using CNN, LSTM and GTTS,” International Journal for Research in Applied Science and Engineering Technology, vol. 11, no. 6, pp. 4473-4481, 2023.

[6] P. D. G. Vyawahare, A. Gadge, S. Cholkhane, A. Mishra and S. Anturlikar, “Image to Audio Conversion for Blind People Using Neural Network,” Ijraset Journal For Research in Applied Science and Engineering Technology, vol. 11, no. 12, pp. 2-11, 2023.

[7] M. I. T. Hussan, D. Saidulu, P. T. Anitha, A. Manikandan and P. Naresh, “Object Detection and Recognition in Real Time Using Deep Learning for Visually Impaired People.,” International Journal of Electrical & Electronics Research (IJEER), vol. 10, no. 2, pp. 80-86., 2022.

[8] D. M. Y. Babu, A. Jatavath, G. Y. K. Reddy and P. A. Kumar, “Object Detection System with Voice Alert for Blind,” Ijraset Journal For Research in Applied Science and Engineering Technology, vol. 11, no. 5, pp. 2-11, 2023.

[9] H. M. Nasir, N. M. A. Brahin, M. M. M. Aminuddin, M. S. Mispan and M. F. Zulkifli, “Android based application for visually impaired using deep learning approach,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 4, pp. 879-888, 2021.