Deep Learning-Based Image Captioning for Visual Impairment Using a VGG16 and LSTM Approach

Muhammad Talha Jahangir[*1], Naveed Ahmad[2], Samina Naz[3], Laiba[1], Sabahat Fatima[1], Sabahat Aslam[1]

[1]Department of Computer Science, MNS-University of Engineering and Technology, Multan, Pakistan.

[2]Department of Computer Science، NCBA & E, Multan Campus

[3]Faculty of Computing and Emerging Technologies, Emerson University Bosan Road Multan

[*]**Correspondence:** mtalhajahangir@mnsuet.edu.pk

Citation | Jahangir. M. T, Ahmad. N, Naz. S, Laiba, Fatima. S, Aslam. S, "Deep Learning-Based Image Captioning for Visual Impairment Using a VGG16 and LSTM Approach", IJIST, Vol. 6 Issue. 4 pp 1808-1825, Oct 2024

Received | Sep 29, 2024 **Revised |** Oct 18, 2024 **Accepted |** Oct 23, 2024 **Published |** Oct 25, 2024.

Visually impaired people face the challenge of gathering information about their surroundings. They are unable to make sense of visually presented information such as capturing images, reading sign boards, moving around especially when they are alone, or recognizing objects. This work proposes a novel approach for creating image captioning using two models, one is Convolutional Neural Networks Architectures (VGG16 and ResNet50), and the second is Long Short-Term Memory (LSTM). Using data augmentation and transfer learning on a custom dataset for this work, the system generated accurate image captions that includes a text-to-speech tool that will offer to read back responses to those who are blind or have low vision. The model showed excellent results in training with an accuracy of 90.16 %, and a validation loss of 17.66 %. In caption generation, it obtained the BLEU value ranging from 0.7788 to 0.1 indicating varied caption quality. In general, the average of the Accuracy and Loss results confirms the effectiveness of combining CNNs and LSTMs for improving image descriptions. This system generates such described robust environment for a visually impaired person that it can give the person more freedom to move around and interact with the environment.

Keywords: Image Captioning, Visually Impaired, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Text-to-Speech, Bilingual Evaluation Understudy (BLEU) Score.

Introduction:

To see the environment and gather visual information humans rely on their eyes. The visually impaired person uses assistive technology which is referred to as blind and visually impaired (BVI) people. For the 2.2 billion people who are suffering from vision impairment worldwide, it would be difficult daily if they couldn't see or identify objects in their environment [1][2]. This disorder can cause partial blindness or total blindness which has a consequent negative impact on an individual's capacity to read, recognize faces, and navigate their surroundings safely. These difficulties impair their freedom and their standard of living [3]. The increasing global incidence of visual impairments highlights the significant need for practical solutions that can enhance accessibility and offer substantial assistance [4]. According to the forecasts the number of people who are suffering from vision impairment has significantly growth, as shown in Figure 1, which displays a remarkable enhancement of blindness and severe vision impairment between 1990 and 2050 [5]. This pattern highlights how important it is to create solutions that are specially designed to help the people of the community who are visually impaired.

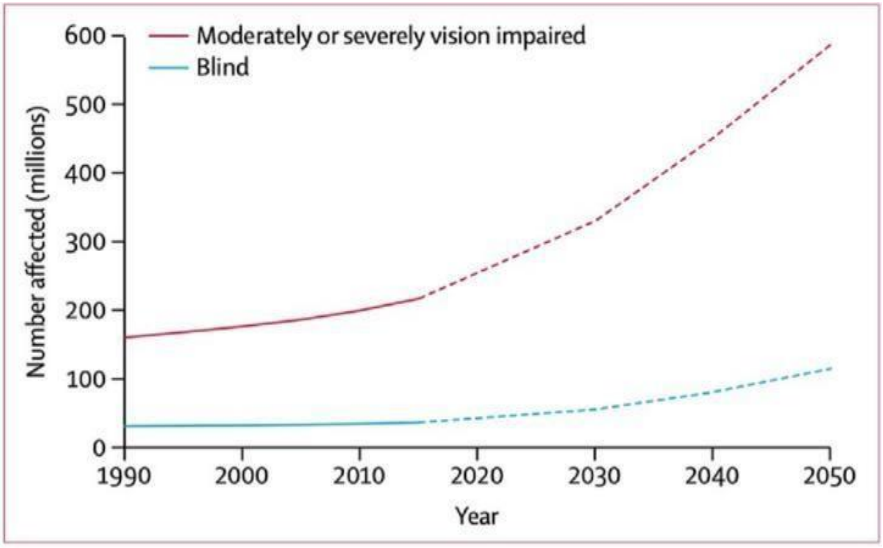


Figure 1. Statistical Diagram showing global trends and vision impairment [5]

Recent developments in Artificial Intelligence (AI) and deep learning have opened new avenues for addressing these problems [6]. Conventional picture description techniques, which mostly depend on human documentation or simple automated systems, frequently fail to provide visually impaired users with the comprehensive and accurate information that they require [7]. This restriction highlights the necessity for exponentially advanced systems that can provide comprehensive and contextually appropriate image descriptions [8]. To close these gaps, mostly LSTM networks and CNN are remarkable and effective for eliminating the colonial characteristics from the picture [9]. Even for producing textual descriptions that can be relevant and logical for the situation, LSTM is used [10]. By combining these techniques, modern picture captioning can be created to improve accessibility both visually and inwardly. The problems that are faced by visually impaired users can be addressed by recent developments in Artificial intelligence (AI) and deep learning [11]. Because of dependencies on human annotation and simple automated algorithms, conventional image description strategies fail to provide accurate and comprehensive information [12]. This focuses on the requirement of more and more complex systems that can provide comprehensive and contextually appropriate descriptions [13]. A solution is provided by deep learning technologies specifically CNN and LSTM networks. CNNs are very efficient at identifying complex characteristics in pictures [14].

Whereas LSTMs produce textual descriptions that are logical and suitable for the situation [15]. Advanced image captioning systems may be created, improving audiological

feedback and visual accessibility by combining these technologies. They can also be gradually integrated with text-to-speech functionality [16]. Hybrid systems are available that combine LSTMs for sequence generation and CNNs for feature extraction which is used to create more effective picture captioning models. These algorithms produce systematically integrated descriptions that are suited for visually impaired users. Due to this, visually impaired users understand complicated scenes and recognize so many items. The ability of these models to focus on influential sectors of an image has been further enhanced by the integration of attention mechanisms, producing captions that are more dependable [17].

Moreover, when we are training a small data then data augmentation methods including rotation, flipping, and scaling enhance the model generalization [18]. We use text-to-speech modules that can convert these captions into auditory input due to this visually impaired people's freedom and quality of life will greatly improve [19]. Considerable methodological progress was established in this study to improve the suggested picture captioning system. VGG and Reset architectures were trained from scratch using datasets, with the use of weight-zero VGG and Reset architectures the model was improved through data augmentation and transfer learning approaches to produce extremely accurate and reflective picture descriptions [20]. The system's peak performance characteristics were made possible in large part by these strategies which include a validation accuracy of 0.9015 and a validation loss of 0.1766. During training early stopping was also used to avoid over-fitting and guarantee that the model performs effectively when applied to the fresh data. The efficiency of the model was further confirmed by BLEU score evaluation which showed notable improvements in caption quality. The addition of text-to-speech technology is an important advancement which is used to generate captions to be spoken and gives visually impaired users audible feedback. This Image captioning system not only integrates with the mobile app, but we can develop the hardware for the visually impaired people.

The paper is organized into several key sections. Section 2 focuses on the Literature Review where we discuss the numerous studies relating to deep learning and computer vision for various tasks such as Navigation and Object detection. Section 3 outlines Model Development, emphasizing the design and implementation of Vgg16 and Resnet50 architectures for feature extraction. Section 4 includes the information about the dataset whereas section 5 describes the development of a Custom Data Set to handle multiple captions per image, including the creation of input sequences and management of batching and shuffling. In section 6 research methods like Data Augmentation, Feature Extraction, and Image captioning are explained. In section 7 evaluation matrix including Accuracy and BLEU score is assembled. In section 8 we discussed the results of our experiments and compared validation accuracy and loss, alongside analysis of training accuracy. Finally, Section 9 offers a comprehensive result interpretation analyzing the outcomes, comparing model performance, and discussing the implementation of the result.

Literature Review:

In recent years, significant advancements have been made in the field of assistive technologies for visually impaired individuals, particularly through the integration of Artificial Intelligence (AI). Studies have demonstrated the effectiveness of AI-based systems in enhancing mobility and accessibility for visually impaired people.

Pydala et al. [21] invented a smart stick for visually impaired persons. The system is trained to detect nearby objects, and for this reason, ultrasonic sensors are located. Additionally, it strengthens the visually impaired to navigate their surroundings safely and independently. The accuracy obtained by the model is almost 76.5% enhancing the trustworthiness of the system in real-world scenarios.

Upadhyay et al. [22] introduced a system named "Eye Road–An App that Helps Visually Impaired Peoples," an app designed to assist visually impaired people with advanced

technologies. The Dataset consists of 20,000 images divided into 40 classifications. Yolov5 and OCR are two examples of models that are used for object detection and text reading, respectively. OCR achieves a recall rate of 70% and reaches its peak accuracy of 80 to 90. This study emphasizes the useful advantages and efficiency of the "eye road app", highlighting the substantial potential of combining computer vision and machine learning technologies to improve mobility and independence for visually impaired users.

Kuriakose et al. [23] on developing a smartphone- advanced scene segmentation techniques, providing detailed feedback to the user and ensuring safer Navigation, uses 330,300 images for training, Methods used are CNN, GANs, and transfer learning giving precise accuracy rate of 88.6%.

Kumar et al. [24] reviewed various AI solutions designed to assist the visually impaired, emphasizing the role of deep learning in improving the accuracy and usability of these systems. Their work discussed models like SSD- MobileNetV2 and YOLO, which typically handle around 20 classes. This review provides a comprehensive overview of the current state of AI in assistive technologies, focusing on wearable devices and their integration with deep learning algorithms. To address the difficulties visually impaired people, encounter in indoor situations, Ajina et al. [25] developed an AI-assisted navigation system. The device uses "Deep-NAVI," a deep learning-based navigation assistant, to identify obstacles and lead users through unfamiliar areas. Three to four disease categories are targeted together with diabetic retinopathy, 80 classes of COCO datasets are trained, and Transfer learning techniques produce accurate predictions and faster model convergence the system is easy to use and handle.

Parenreng et al [26] invented an AI based on object detection using CNN. He utilizes 1000 images, methods like edge AI and Resnet for training and the model obtains an accuracy rate of 85%. It leverages to enhance the life of the visually impaired easily to navigate and make them move easily. Tamilarasan et al. [27] invented the "Blind Vision" system, which leverages AI to assist visually impaired individuals in navigating their surroundings. The system, designed to recognize 45 object categories, emphasizes the use of advanced AI algorithms to detect obstacles and provide real-time feedback to users.

Xie et al. [28] focus on liberating object detection with flexible expressions. He used a dataset of 10,578 images and obtained an accuracy rate (map) of 21.6% by utilizing techniques like REC and OVD. Faurina et al. [1] introduce the application of image captioning technologies to aid blind and visually impaired individuals in outdoor navigation. He used CNN and resnet512 architecture and obtained a precision rate of 91% on BLEU and 94.03% on ROUGH-L. It shows how sophisticated image captioning approaches can improve navigational help and improve the quality of life for visually impaired people, assisting their surroundings thoroughly and accurately. Table 1 shows the summary of the previous paper.

Table 1. Summary of previous papers

	Availability	Total Amount	Methods	Classes	Results
[21]	Yes	124,000	Ultrasonic Sensors, GPS	32	Precision (76.5%)
[22]	Yes	20,000	YOLOv5, OCR	40	Accuracy 85-90% Recall: 70%
[23]	Yes	330,300	CNNs, GANs, Transfer Learning	80	Accuracy 88.6% F1Score:12%
[24]	Yes	4500	YOLO, SSD, Mobile Net	20	Accuracy: 83% to 95.19%
[25]	Yes	1021	Resnet, deep learning.	3-4	Accuracy: 90%
[26]	Yes	1000	Edge AI, Res Net	20-50	Accuracy 85 %

[27]	Yes	45	Yolo v5, CNN	80	Accuracy, 83% to 95.19%
[28]	Yes	10,578	REC, OVD	422	Accuracy: map: 21.6%
[1]	Yes	2788	CNN, Resnet50	7	Accuracy:91%
Our Paper	Yes	126004	Vgg16,LSTM	24	Accuracy: 91.06%

Vgg16 Architecture:

The architecture of Vgg16 well-known for its accessibility and intensity that uses a series of convolutional layers with small 3x3 kernels which are followed by max-pooling layers. In our project, we retained a per-trained VGG model (e.g., VGG16 or VGG19) for its reliability in feature abstraction. The customization involved adjusting hyperparameters and discriminately reorientating certain layers to adjust the model to our specific dataset. This structure optimized the high-dimensional feature representations learned from large-scale datasets and customized the performance of our task. The architecture of Vgg16 is shown in Figure 2 [29].

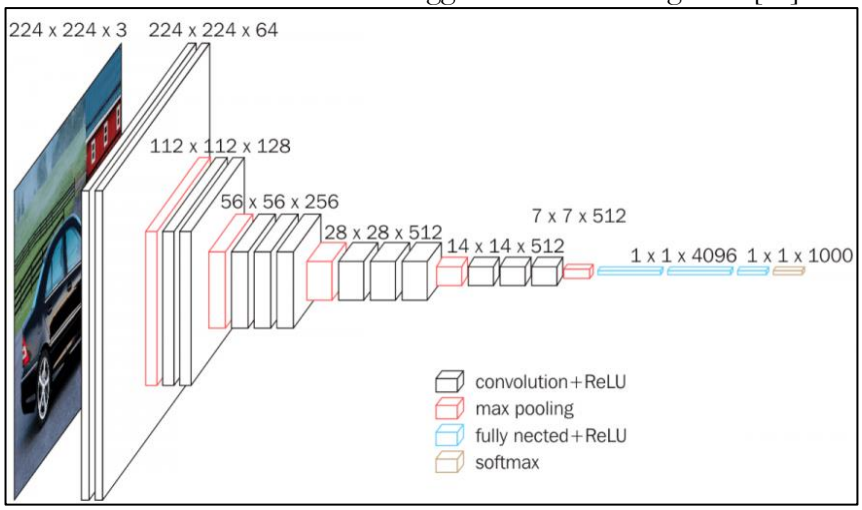


Figure 2. The Architecture of VGG16 Source [29]

Resnet Architecture: The Reset or Residual Networks deal with the obliterating optimization issues through residual connections that allow gradients to flow more efficiently during training shown in Figure 3 [30]. In this project,we retained a pre-trained Reset model (e.g., ResNet50) wherein the customization involve adjusting hyper-parameters and discriminately reorientating certain layers to adjust the modal to specific dataset characteristics.

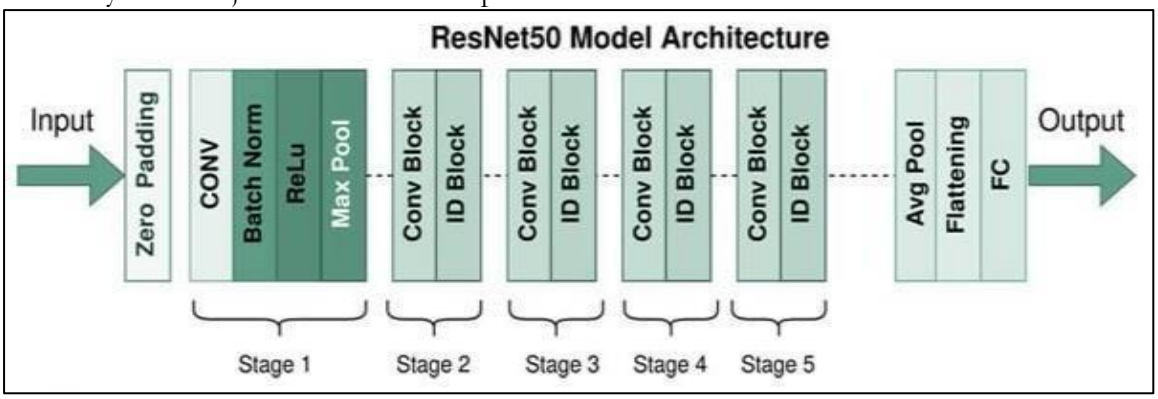


Figure 3. The Architecture of the Resnet50 Model [30]

Data Set:

The dataset [31] consists of images and corresponding captions stored in a CSV file. Each image is linked to one or more captions. Before training, images are preprocessed by resizing them to a fixed size and normalizing pixel values, while captions are cleaned, tokenized,

and converted into numerical sequences. The dataset is split into training and testing sets, which are given in Table 2. Thecaptions are tokenized with "start seq" and "end seq" tokens forboth sets and the size of the vocabulary is constantly maintained across the dataset.

Table 2. Dataset structure for training and testing

Category	Training Data	Testing Data
Images	80% of dataset	20% of dataset
Captions	Multiple per image	Multiple per image
Tokenization	Captions tokenized with 'start seq' and 'end seq' tokens	Captions tokenized with 'start seq' and 'end seq' tokens
Vocab Size	150	150
Max Length	17	17
Custom Generator	Handles multiple captions, input-output sequences, image features pairing, batching, and shuffling	Handles multiple captions, input-output sequences, image features pairing, batching, and shuffling

The maximum length of the caption is 15 words. A custom generator is used to manage multiple captions per image, handle input-output sequences, pair the features of the image with captions, and manage batching and shuffling. This structured strategy audit smooth and expedient model training and testing. The dataset utilized in this research initially consists of 24 distinct classes with a total of 1,600 instances. The distribution across these classes is shown in Figure 4 [31] is as follows: ATM-related (75), bench(120), bus(95), pharmacy (60), church (50), construction(70), door (85), Euro Banknote (40), food street(130), green signal (90), lift (55), luas (65), music related (80), playing kids (110), Pound Banknote (45), push button (50), red signal (100), stairs down (60), stairs up (65), trash related (90), wait (75), wallet (85), washroom (70), and wet floor (55).

In the context of the data-set scale, we started with the 8,000 original images. We applied augmentation techniques to enhance the dataset and improve model performance. The result of is a total of 126,004 augmented images. This augmentation process—through methods such as rotation, scaling, and flipping increases the dataset size and diversity. It plays a vital part in making our model more robust.

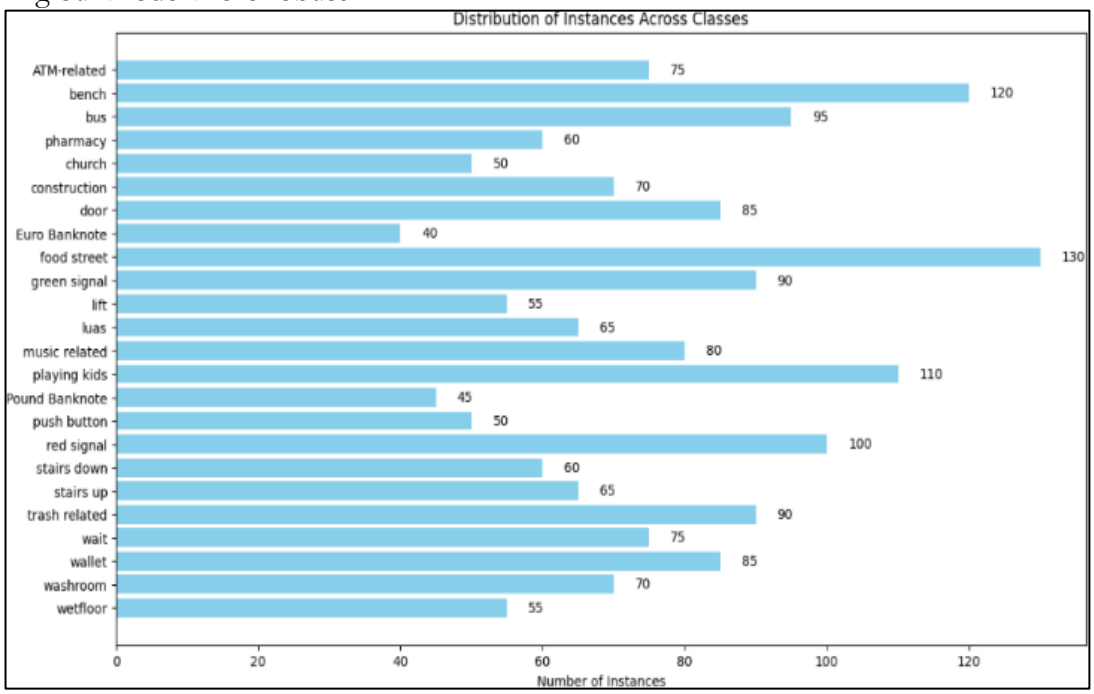


Figure 4. Distribution of Original and augmented images [31]

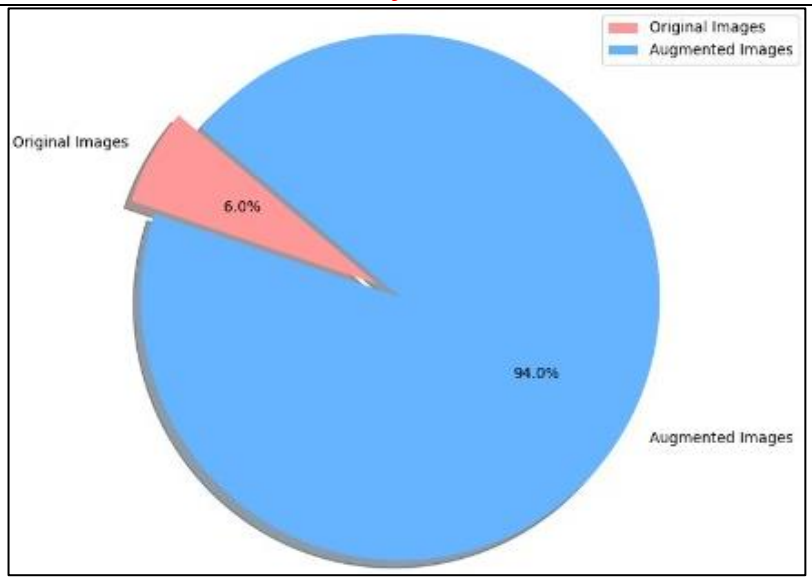


Figure 5. Distribution of Images Across Classes in Dataset

Figure 5 shows that the size of the dataset increases after applying Augmentation Techniques to each image.

Objective:

The research has the following objectives:

- Developed Image Captioning System using Vgg16 and LSTM.
- Data augmentation techniques are used to increase the quality of the caption.
- Converted the caption-generated text into the speech.
- Compare the performance of Vgg16 and Resnet50, to generate image captioning.

Novelty:

The following attributes show the novelty of our work.

- The dataset used in our work is based on 24 different categories.
- Different CNN architectures are used to increase the quality of captions.
- Integration of text with speech for visually impaired people.

Methodology:

In the proposed methodology (Figure 6), we used CNN for extracting features from the image while LSTM is used to generate captions from an image. VGG16 captures complex spatial hierarchies with its deep convolutional layers, while to improve feature learning by solving disappearing gradient issues, ResNet50 uses residual. Ensuring the extraction of these features, data is fed into a sequence model commonly LSTM network, which operates the visual information to produce precise and contextually relevant captions as shown in Figure 9. Finetuning of hyperparameters is utilized to obtain high performance, like learning rate and batch size. For evaluating the quality of generated captions BLEU score is used, showing insights into how well the caption conveys the meaning of images. This strategy efficiently integrates advanced captioning with Cutting-edge feature methods. By fusing a powerful captioning model along with sophisticated extraction techniques exceptional descriptive outcomes are obtained.

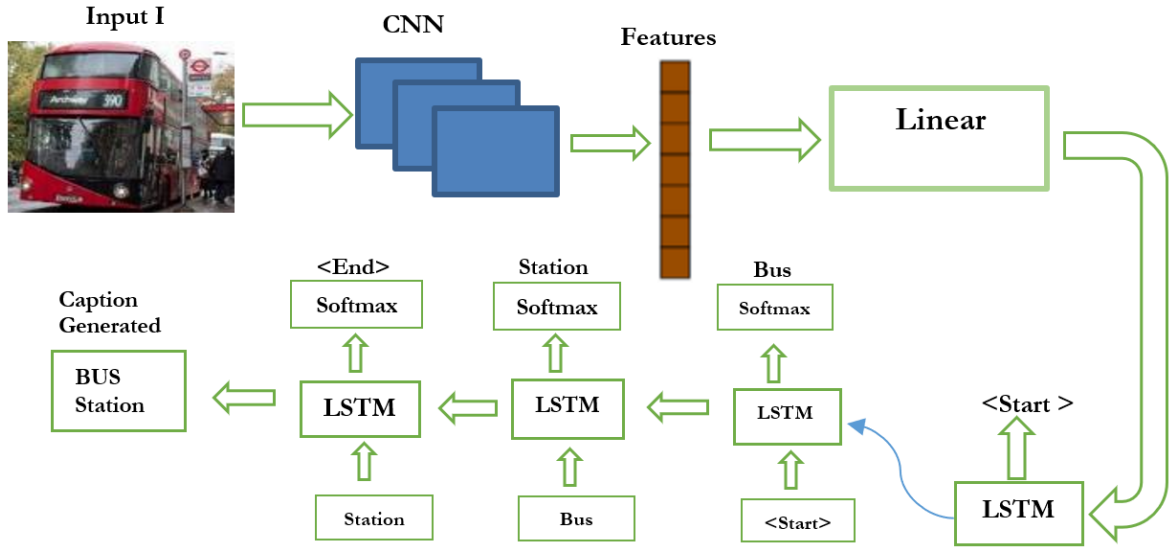


Figure 6. Image caption Generator

An image caption generator understands the content of an image and generates a relevant caption, with the help of computer vision and deep learning techniques. ImageNet dataset is used for the pre-trained models (Vgg16 and ResNet50). This enables them to acquire strong visual representations that contain attributes describing objects, texture, shape, and location of objects. Preliminary, an image goes further through the convolutional and pooling layers of these networks to obtain a descriptive feature. To generate captions and extraction of features then proceeded through the LSTM model. While the pyttsx3 library is used for converting generated captions into speech.

Preprocessing:

The processing phase focused on creating an image captioning system optimized for the visually impaired. The approach was divided into several key phases, including data preprocessing, feature extraction using pre-trained models, and the design and training of a custom model architecture. During the preprocessing phase, the visual.token.txt file has the names of the images and their captions, where the dataset is initially loaded. The 'image' and 'caption' columns of this file are transformed into a Data Frame using the Panda's package. Next, all of the text in the captions is converted to lowercase, and "start seq" and "end seq" are added to the beginning and end of each caption, respectively. This guarantees that all captions follow a uniform format as shown in Figure 7. Features retrieved from the dataset images using VGG16 and Reset50 models, with the best-performing model picked based on accuracy. The vocabulary size and maximum caption length are calculated from the processed captions. The dataset is then divided into training and testing sets, with 80% of the data utilized for training and 20% for testing in the original dataset and altered proportions in the augmented dataset. An LSTM model is deployed with early stopping to maximize training, ceasing when the validation loss does not improve for 100 epochs.

```
...  DataFrame shape: (8000, 1)
     DataFrame head:
                                                    0
     0  greensig1.jpg#0\tGreen signal please walk .
     1  greensig1.jpg#1\tGreen signal please walk .
     2  greensig1.jpg#2\tGreen signal please walk .
     3  greensig1.jpg#3\tGreen signal please walk .
     4  greensig1.jpg#4\tGreen signal please walk .
            Images                                      Captions
     0  greensig1.jpg  startseq green signal please walk  endseq
     1  greensig1.jpg  startseq green signal please walk  endseq
     2  greensig1.jpg  startseq green signal please walk  endseq
     3  greensig1.jpg  startseq green signal please walk  endseq
     4  greensig1.jpg  startseq green signal please walk  endseq
```

Figure 7. Data Frame with 'Image' and 'Caption' columns

In the preprocessing stage, the images were re-sized with the resolution of 224*224. Pixel values normalized to a range of [0, 1] to standardize the data which ensures consistent input scaling across all images. A series of data augmentation techniques were employed to enhance data-set diversity and improve model robustness.

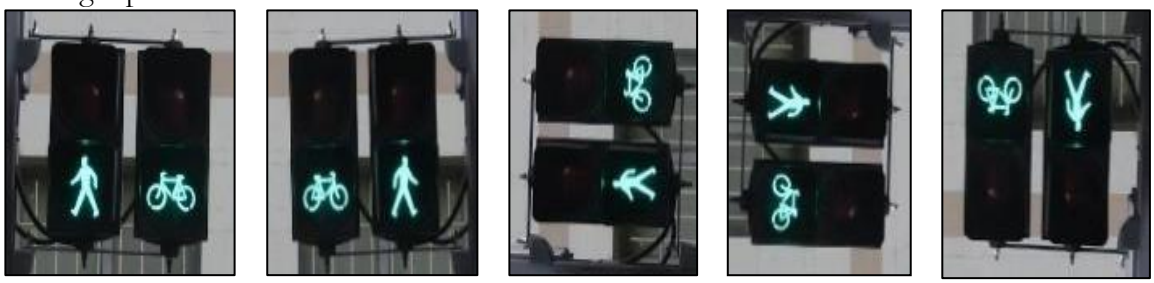
Data Augmentation:

Data augmentation strategies are applied to increase the data-set and improve model performance. Using the 'Image Data Generator' class in Kera's, several techniques are used to boost data-set distinctiveness. Optimization parameters include rotation range which randomly rotates images by up to 30 degrees; width_shift_range and height_shift_range which allow horizontal and vertical shifts up to 20% of the image dimensions; shear_range which applies shear transformations up to 20 degrees and zoom_range which enables zooming in or out by up to 20%. Further, empty pixels are filled using fill_mode along with nearest pixel values whereas, images are turned horizontally using horizontal_flip. Figure 8 [31] shows the different versions of an image after applying data augmentation. Data augmentation helps this model to generalize better across multiple settings, enhancing its performance and ability to manage a range of visual conditions.

Table 3. Parameters of Data Augmentation

Augmentation Technique	Description	Value/Range
Rotation Range	Rotate Angle	0 to 30 degrees
Width_shift_range	Shift horizontally.	Up to 20% of width
Height_shift_range	Shift vertically	Up to 20% of height
Shear_range	Shear angle.	Up to 20 degrees
Zoom_range	Zoom in/out	20% zoom
Horizontal_flip	Flip horizontally	True or False

Table 3 highlights the typical parameters which are applied during the data augmentation to enhance the diversity of the training set. The vocabulary size of the dataset is 150 and the maximum length of the captions is 21 words. For instance, a sample caption from the dataset might be: "start seq" green signal please walk "end seq". To enhance the robustness of the model and prevent overfitting, early stopping was implemented with patience of 100 epochs. This approach monitors the validation loss during training and halts the process over 12 epochs. By stopping early, the model avoids overfitting and maintains its generalization capabilities, ensuring it performs well on unseen data.



(a) (b) (c) (d) (e)

Figure 8. Example of Image Augmentation Techniques (a). Original Image; (**b**) 90°left rotation; (**c**) 90° right rotation; (**d**) 180° Horizontal rotation; (**e**) 180° Vertical rotation [31]

Table 4. Image count in different datasets

Category	Image Count
Normal Images	6004
Augmented Images	120,000
Concatenated Dataset	126,004

Table 4 summarizes the number of images in three stages of dataset preparation: the initial dataset of normal images, the dataset after augmentation, and the combined dataset with

both normal and augmented images.

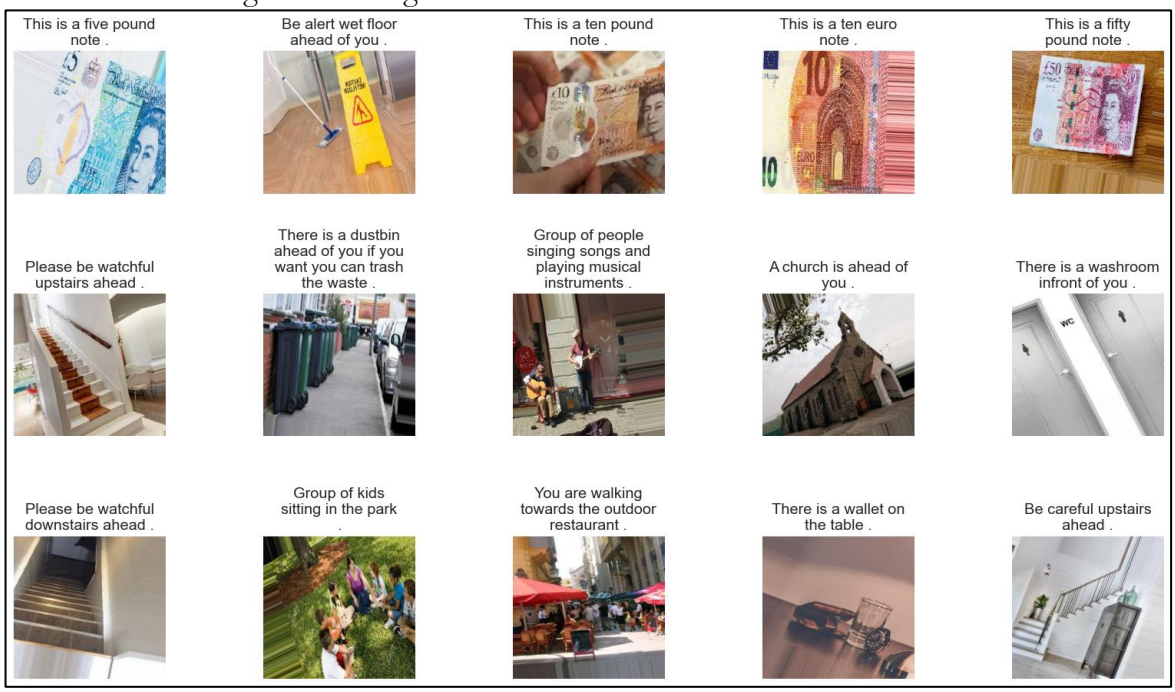


Figure 9. Images with Captions [31]

Figure 9 [31] shows augmented images along with their Captions generated to make it understandable, reliable, and easy for visually impaired persons to navigate their surroundings.

Feature Extraction:

Feature extraction was implemented for employing pre-trained models including VGG16 and Res Net which are known for their robust picture recognition skills. VGG16's deep convolutional layers and max-pooling operations capture comprehensive spatial hierarchies while Res Net's residual blocks manage the deconstruction gradient problem, increasing feature learning through shortcut connections. Images were re-sized and normalized to meet the specifications of the model. The recovered feature vectors are shown in Figure 10. with dimensions (1, 7,7, 512), were then employed as inputs for the captioning model, helping the development of precise and contextually relevant image descriptions. This approach strengthens pre-trained networks, detaching the requirement for substantial training from scratch while ensuring precise feature extraction.

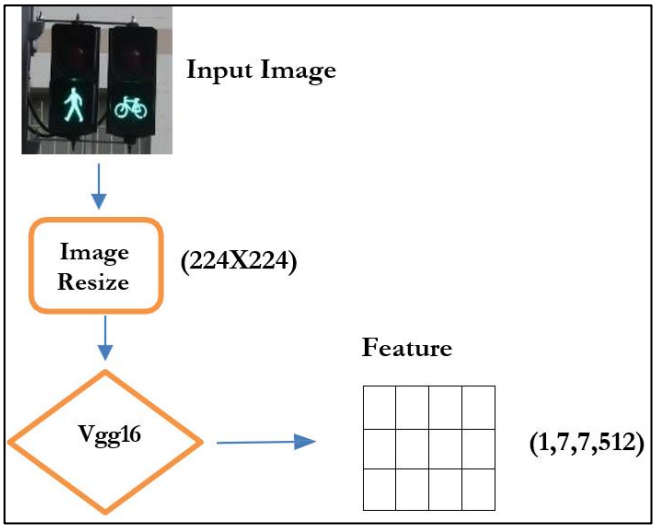


Figure 10. Extracting features from the input image

Evaluation Metrics:

To obtain the accuracy and evaluation of the model, we use some tools to measure their presence or to make the model work more perfectly and effectively. Tools used in this experiment are,

- Accuracy
- BLUE

Accuracy:

One of the most important evaluation metrics for evaluating a classification model's performance is accuracy. Its definition is the ratio of the total number of instances to the number of accurately predicted instances. In terms of math, it is stated as:

$$\text{Accuracy} = \frac{\#\text{ of correct prediction}}{\text{total \# of points}} \text{ X } 100$$

BLEU Score:

A commonly used metric for assessing the quality of generated text in natural language processing tasks, such as image captioning, is the BLEU (Bilingual Evaluation Understudy) score. By evaluating the n-gram precision a measure of how well man words or sequences in the generated captions match those in the reference captions BLEU compares the machine-generated captions to the human-created reference captions.

Formula:

$$\text{BLEU} = BP\ X\ exp\frac{1}{N}\sum_{n=0}^{N}\log in$$

Whereas,

$$p_n = \frac{\text{Number of Ingram tokens in the system and reference translations}}{\text{Number of Ingram tokens in system translations}}$$

Experiment and Results: A high-performance system was used for the experiments according to the following specifications:

- **CPU:** AMD Ryzen 9 5900X
- **GPU:** NVIDIA GeForce RTX 4080 SUPER 16G VENTUS 3XOC
- **Memory:** 32 GB RAM

Res Net without Data Augmentation:

By using Resnet50 and LSTM, the validation loss dropped from 0.4928 to 0.21977 by Epoch 27, while the training accuracy grew steadily from a starting point of training accuracy of 0.6801 and a validation accuracy of 0.6712 at Epoch 1. The validation loss was further improved to 0.19127 with the use of a learning rate adjustment at Epoch 33 as shown in Figure 10. The accuracy of the model increased further, but advancements slowed, and the model was early stopped when it attained its ideal performance without overfitting.

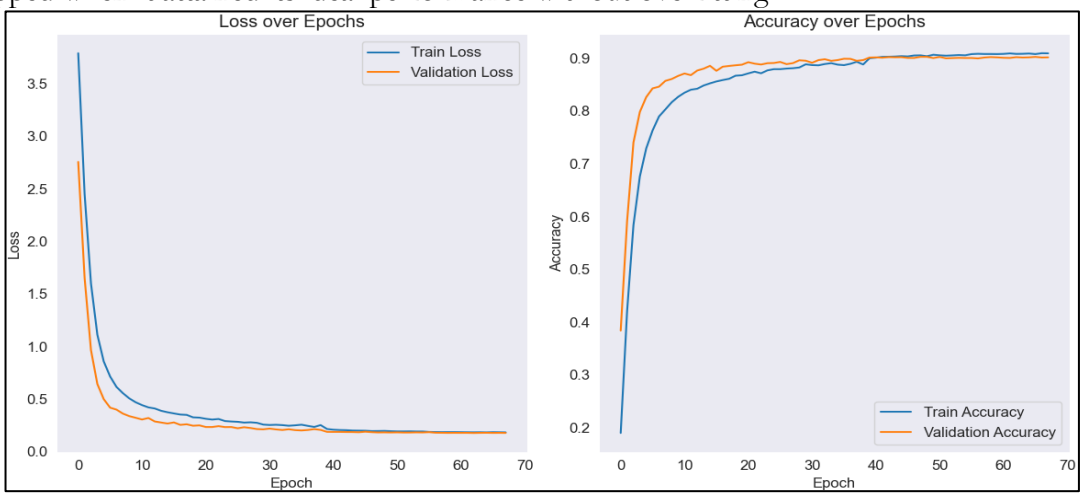


Figure 11. Performance using ResNet50 and LSTM for Image Captioning

Res Net50 with Data Augmentation:

All images, original and augmented are preprocessed and scaled to meet the specifications of ResNet50's input. All image's features are taken out and put in a dictionary. Then, to make training data loading more efficient, the `Custom Data Generator` class is used. Batching, tokenizing captions, padding sequences, and data shuffling are all handled by this class. Two inputs are used to build the model: one for caption sequences and one for image features. While the caption sequences are processed through an LSTM layer after an embedding layer, the image features are flattened and run through a dense layer. Callbacks for model checkpointing, early pausing, and learning rate modifications are part of the training process. Based on validation loss, the best model is preserved during the training phase, which is tracked. The performance graph uses resnet50 with data augmentation and LSTM form image captioning is shown in Figure 12.

Vgg16 with Data Augmentation:

In this study, we extracted features from a dataset image using the VGG16 model. To guarantee consistency and readability, the captions underwent preprocessing. The VGG16 model, which was pre-trained on the ImageNet dataset and did not include its top classification layers was used to extract features from images. To meet the input criteria of VGG16, the images underwent preprocessing and were scaled to a fixed dimension. Rotations, flips, and scaling were incorporated as augmentation approaches to add variety and improve the model's generalization across various image situations. To control the data loading procedure during model training, a unique data generator was created. This generator handled padding, conversion of caption text into integer sequence, and one-hot encoding. For training and testing purposes subset of the data set is reserved. For sequence processing model contains LSTM layers, for text an embedding layer, and image feature extraction for merging textual or image information. In order to prevent overfitting and maximize training, we trained the model using categorical cross-entropy loss, The model's performance is measured using a range variety of callbacks, aiming to boost its accuracy and robustness while generating captions for images. For Image Captioning, the performance graph uses Vgg16 with data augmentation shown in Figure 13.

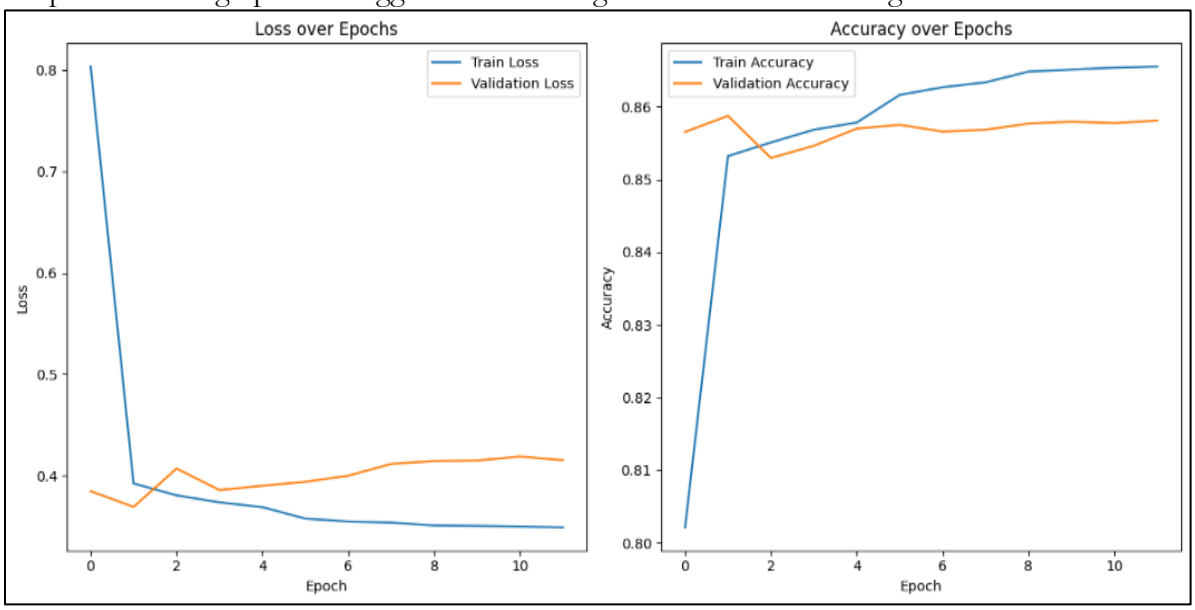


Figure 12. Performance using ResNet50 with Data Augmentation and LSTM for Image Captioning

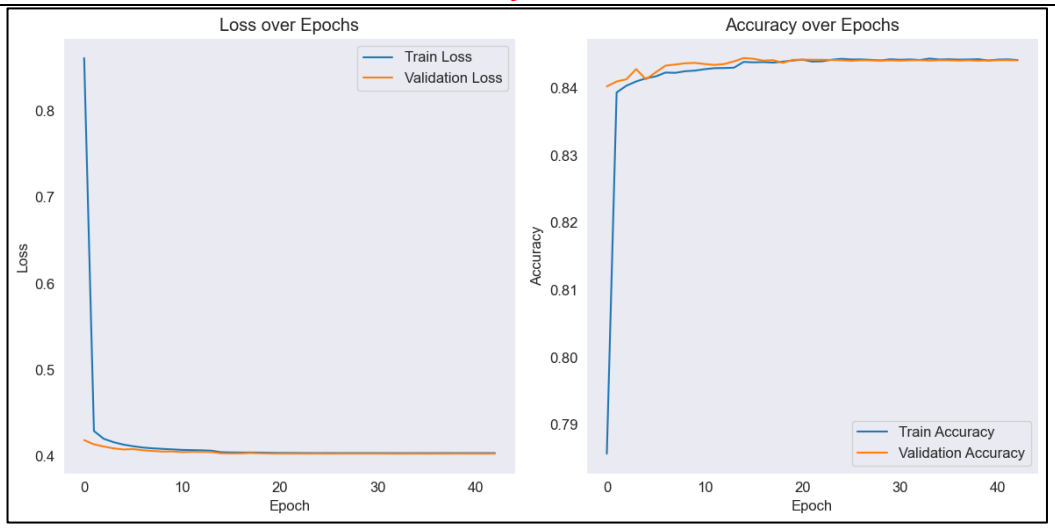


Figure 13. Performance using Vgg16 with Data Augmentation and LSTM for Image Captioning

Training of VGG16 without Data Augmentation:

For captioning images, we use a VGG-based architecture. Initially, the dataset is split into training and testing data, and then using image attributes and tokenized captions custom model is assembled. We use a data generator to handle batches of image-caption pairs for training the model. Key callbacks are used to optimize the training process, including early stopping, reducing the learning rate, and model checkpointing. The model restores weights from epoch 58, when the best validation performance was noted, triggering early halting at epoch 63. With matching losses of 0.1776, the model obtained a training accuracy of 0.9106 and a validation accuracy of 0.9001. Following training, the top-performing model is saved to the designated directory, and matplotlib is used to show the training history, which includes trends in accuracy and loss shown in Figure 14.

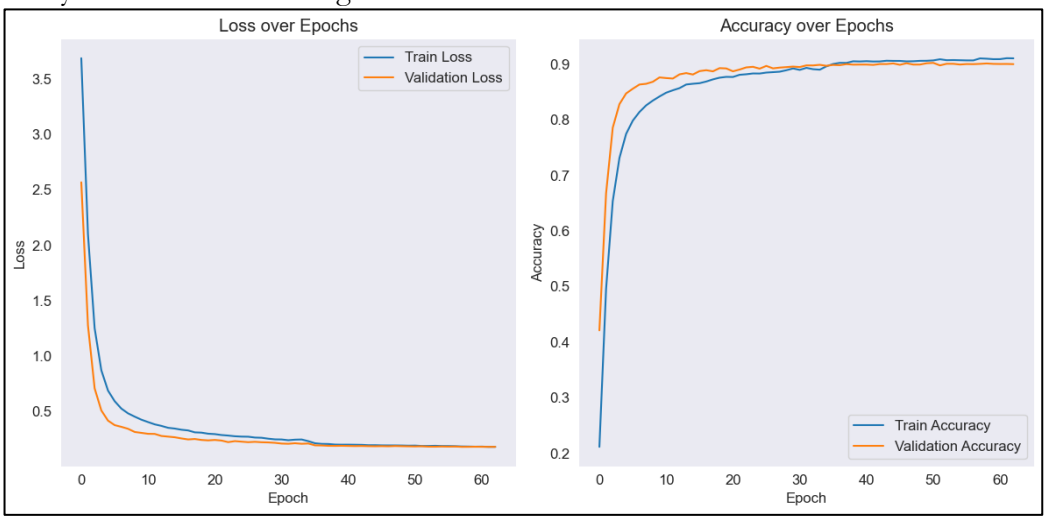


Figure 14. Performance using Vgg16 and LSTM for Image Captioning

BLEU Score of VGG16 Trained on Scratch:

The BLEU scores for the VGG model that was trained using ImageNet differ considerably. The similarity between the actual caption "Be watchful wet floor ahead of you," and the predicted caption "Please mind your steps wet floor ahead of you" has a modest BLEU score of 0.446. With a score of 0.779, the prediction "be watchful upstairs ahead" compared to "please be watchful upstairs ahead" showed better alignment. A BLEU score of 1.0 indicated

that perfect matches, like "this is a ten-pound note" and "this is a ten-pound note" reflected precise predictions.

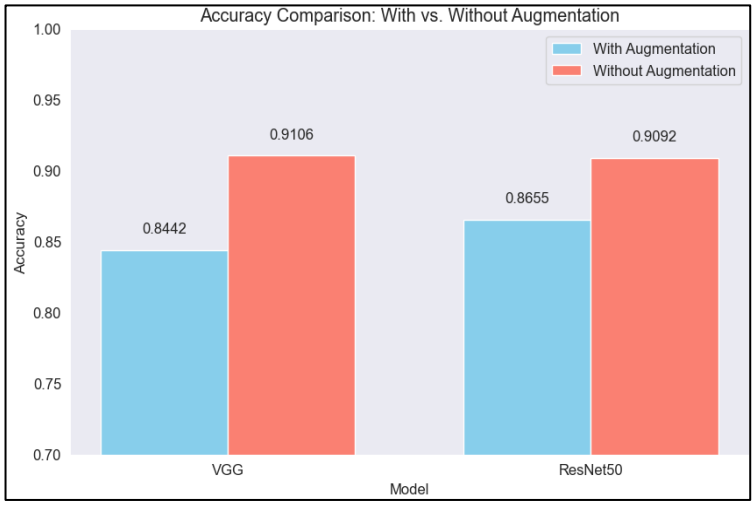


Figure 15. Accuracy Comparison with vs without Augmentation

Comparing the VGG16 and ResNet50 Models:

Comparing the VGG16 and ResNet50 models with and without augmentation three comparative bar charts are created. The diagrams show training and validation accuracy. In terms of accuracy and training loss the speed of ResNet50 is better than VGG16 after augmentation is applied, while validation accuracy remains the same for both models. Both models display somewhat improved accuracy and reduced loss in the absence of augmentation. VGG and ResNet50 model performance was evaluated both with and without data augmentation. Without augmentation, the VGG model outperformed its augmented version. According to this, the accuracy of the ResNet50 model was 86.55% with augmentation and lower at 90.92% without it as shown in Figure 15. Although both models gain from augmentation, these findings show that the VGG model has a more noticeable decline in performance.

The VGG model retained a validation accuracy of 90.01% without augmentation and 84.41% with augmentation when analyzing validation accuracy shown in Figure 16. Without augmentation, the validation accuracy of the ResNet50 model was 90.15%; with augmentation, this dropped to 85.81%. The two models' validation accuracy trends are consistent, indicating that the performance loss brought on by augmentation is comparatively homogeneous.

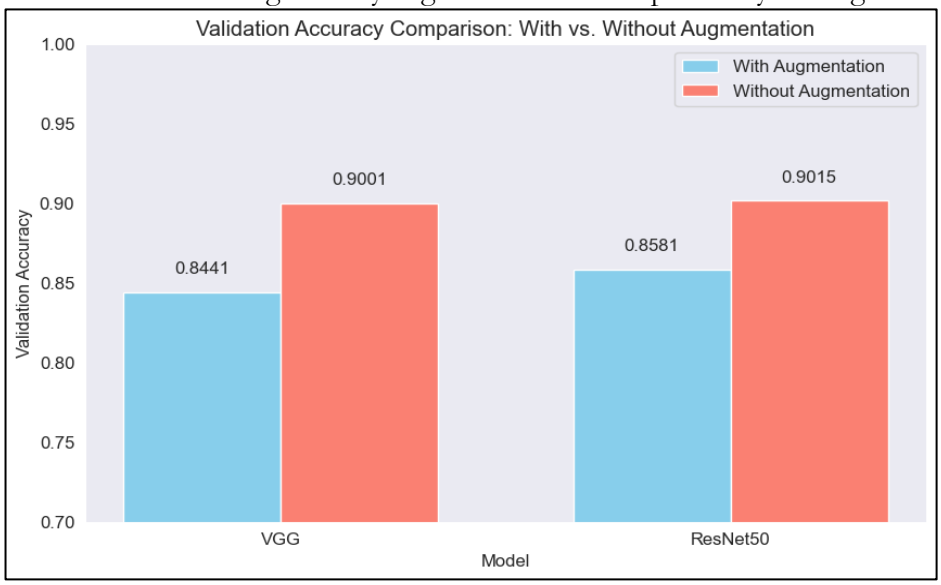


Figure 16. Validation Accuracy Comparison with vs without Augmentation

When it came to training loss, the VGG model showed a loss of 0.1776 without augmentation and a loss of 0.4037 with it. With and without augmentation, the training loss of the ResNet50 model was 0.1818 and 0.3489, respectively. The VGG model reported a validation loss of 0.1776 without augmentation, which increased somewhat to 0.4030 with augmentation. With augmentation, the validation loss of the ResNet50 model increased to 0.4152 from 0.1776. These findings show that, especially for the VGG model, augmentation significantly increases loss as shown in Figure 17. This implies that, although augmentation enhances generalization, it may also add complexity, which raises loss.

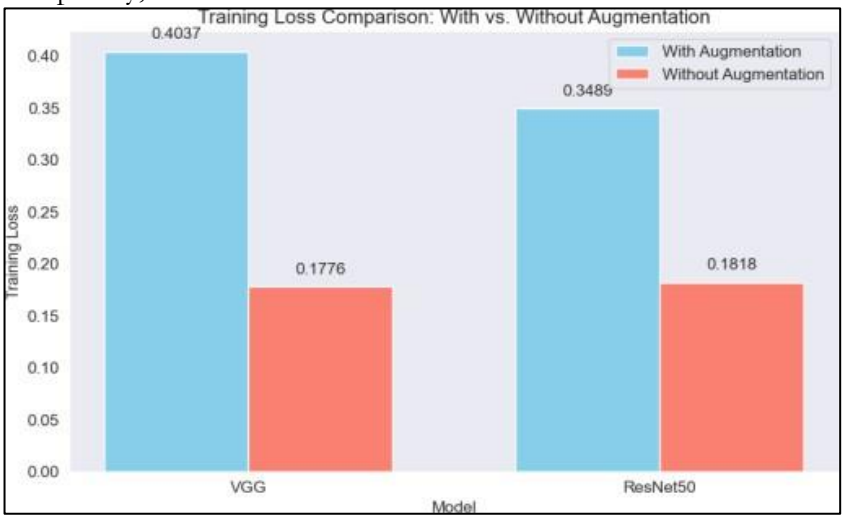


Figure 17. Training Loss Comparison with or without Augmentation

Key details about the performance of the VGG and ResNet50 models are revealed by contrasting them in both augmented and non-augmented scenarios. While training the VGG model without augmentation accuracy rate obtained was 0.9106, which was slightly greater than Resnet50's 0.9092. Resnet50 performed better than VGG in terms of validation accuracy obtaining measures as 0.9015 in comparison to 0.9001. This demonstrates that resnet50 is capable of generating unseen data better than VGG, although VGG16 is suitable for the training set. This indicates that the validation accuracy and training loss (0.1818 vs 0.1776) of resnet50 is higher. Given that both models' validation losses are similar (0.1776), ResNet50 may be less prone to overfitting than VGG. Both models saw a decrease in accuracy when data augmentation was used, although the VGG model was more severely affected. The training accuracy of VGG decreased to 0.8442, while the validation accuracy dropped to 0.8441, indicating a major difficulty in adjusting to the enhanced data. The rise in loss metrics—a training loss of 0.4037 and a validation loss of 0.4030, respectively—which show a significant increase over the non-augmented condition, further supports this. ResNet50, on the other hand, demonstrated greater resistance to augmentation, as seen by a decline in accuracy to 0.8581 for validation and 0.8655 for training. Even though ResNet50's validation loss went up to 0.4152—higher than its training loss of 0.3489—the model continued to perform better than VGG in enhanced scenarios. The result shows the change in the capacity of two models founded by the state of data. Even though the performance of VGG is remarkable in non-magnified backgrounds, faces problems with augmented data. In comparison, Resnet50 handles the complications more accurately than VGG, even though in non-augmented conditions its performance is less than VGG16. It shows that Resnet50 is better for dealing with model accuracy trained from data augmentation in the real world. Table 5 shows the model comparisons obtained after treating data with or without augmentation.

Table 5. Vgg16 without augmentation excels over Resnet 50

Model	Condition	Accuracy	Validation Accuracy	Training Loss	Validation Loss	Time Taken (seconds)
Vgg16	Non-Augmented	0.9106	0.9001	0.1776	0.1776	272 sec
Res Net 50	Non-Augmented	0.9092	0.9015	0.1818	0.1776	7267sec
Vgg16	Augmented	0.8442	0.8441	0.4037	0.4030	7267sec
Res Net 50	Augmented	0.8655	0.8581	0.3489	0.4152	3156 sec

Discussion:

Some of the conventionally available facilities such as hardware navigation sticks for the visually impaired are somewhat inadequate. They are equipped with ultrasonic sensors for obstacle detection; however, they only have audio prompts if the previously mentioned conditions are met. This kind of conditional coding allows the simple execution of procedures but fails to handle actual scenarios. New technologies that are presented in the field of deep learning enable the development of more complex assistive systems. Such systems can handle full environments, instead of merely responding to specific stimuli. This broader interpretation of navigation spaces is more advantageous to the visually impaired. In-depth, deep learning improves physical perception and understanding of the surroundings. They are empowering free mobility for the blind and the visually impaired since they allow more integration. Table 1 shows the summary of all the previous papers. Table 1. shows that deep learning techniques are used for image captioning. According to Table 1, our proposed work can classify more than 23 classes with enhanced accuracy than others. Our methodology not only performs object detection or classification tasks but it is also used to generate image captioning with speech to visual impairment people.

Conclusion and Future Work:

This research has demonstrated the promise of deep learning when it combines CNN networks with LSTM networks in improving image captioning for the visually impaired. In this part, the experimental result indicates that transfer learning and data augmentation combined with the dataset are successfully used for developing a more accurate and enhanced image description with a high accuracy of 0.9106 and a low validation loss of 0.1766. Also, there's a text-to-speech feature that makes generated captions much more accessible. Future works will involve using other CNN architectures like Inception and Efficient Net to increase the reliability and accuracy of the captioning model. Moreover, description quality can be improved, and distinctive visual features can be identified through the addition of attention mechanisms and transformers.

Recommendation and Limitations:

- Our dataset used only 24 sets of classes, while we can increase it more.
- In the dataset, we have few captions based on a single class, so we need to add more captions for each class from the dataset.
- We reached 91% maximum accuracy; we can increase the performance of our model by applying different deep learning state-of-the-art approaches.

References:

[1] R. Faurina, A. Jelita, A. Vatresia, and I. Agustian, "Image captioning to aid blind and visually impaired outdoor navigation," IAES Int. J. Artif. Intell., vol. 12, no. 3, pp. 1104–1117, Sep. 2023, doi: 10.11591/ijai.v12.i3.pp1104-1117.

[2] K. Pesudovs et al., "Global estimates on the number of people blind or visually impaired by cataract: a meta-analysis from 2000 to 2020," Eye 2024 3811, vol. 38, no. 11, pp. 2156–2172, Mar. 2024, doi: 10.1038/s41433-024-02961-1.

[3] S. K. West, G. S. Rubin, A. T. Broman, B. Muñoz, K. Bandeen-Roche, and K. Turano,

"How Does Visual Impairment Affect Performance on Tasks of Everyday Life?: The SEE Project," Arch. Ophthalmol., vol. 120, no. 6, pp. 774–780, Jun. 2002, doi: 10.1001/ARCHOPHT.120.6.774.

[4] N. Awoke et al., "Visual impairment in Ethiopia: Systematic review and meta-analysis," https://doi.org/10.1177/02646196221145358, vol. 42, no. 2, pp. 486–504, Dec. 2022, doi: 10.1177/02646196221145358.

[5] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," Cogn. Robot., vol. 3, pp. 54–70, Jan. 2023, doi: 10.1016/J.COGR.2023.04.001.

[6] R. Ratheesh, S. R. Sri Rakshaga, A. Asan Fathima, S. Dhanusha, and A. K. Harini, "AI-Based Smart Visual Assistance System for Navigation, Guidance, and Monitoring of Visually Impaired People," Proc. 9th Int. Conf. Sci. Technol. Eng. Math. Role Emerg. Technol. Digit. Transform. ICONSTEM 2024, 2024, doi: 10.1109/ICONSTEM60960.2024.10568710.

[7] R. Gonzalez, J. Collins, C. Bennett, and S. Azenkot, "Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People," Conf. Hum. Factors Comput. Syst. - Proc., May 2024, doi: 10.1145/3613904.3642211/SUPPL_FILE/PN8235-SUPPLEMENTAL-MATERIAL-1.XLSX.

[8] R. M. Silva et al., "Vulnerable Road User Detection and Safety Enhancement: A Comprehensive Survey," May 2024, Accessed: Oct. 24, 2024. [Online]. Available: https://arxiv.org/abs/2405.19202v3

[9] R. N. Giri, R. R. Janghel, S. K. Pandey, H. Govil, and A. Sinha, "Enhanced Hyperspectral Image Classification Through Pretrained CNN Model for Robust Spatial Feature Extraction," J. Opt., vol. 53, no. 3, pp. 2287–2300, Jul. 2024, doi: 10.1007/S12596-023-01473-7/METRICS.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June-2015, pp. 3156–3164, Oct. 2015, doi: 10.1109/CVPR.2015.7298935.

[11] D. K. Kumar, A., Nagar, "AI-Based Language Translation and Interpretation Services: Improving Accessibility for Visually Impaired Students," As Ed. Transform. Learn. Power Educ., 2024.

[12] N. Thakur, E. Bhattacharjee, R. Jain, B. Acharya, and Y. C. Hu, "Deep learning-based parking occupancy detection framework using ResNet and VGG-16," Multimed. Tools Appl., vol. 83, no. 1, pp. 1941–1964, Jan. 2024, doi: 10.1007/S11042-023-15654-W/METRICS.

[13] J. H. Huang, H. Zhu, Y. Shen, S. Rudinac, A. M. Pacces, and E. Kanoulas, "A Novel Evaluation Framework for Image2Text Generation," CEUR Workshop Proc., vol. 3752, pp. 51–65, Aug. 2024, Accessed: Oct. 24, 2024. [Online]. Available: https://arxiv.org/abs/2408.01723v1

[14] V. Gorokhovatskyi, I. Tvoroshenko, and O. Yakovleva, "Transforming image descriptions as a set of descriptors to construct classification features," Indones. J. Electr. Eng. Comput. Sci., vol. 33, no. 1, pp. 113–125, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp113-125.

[15] J. R. B. Da Silva, J. V. B. Soares, L. R. Gomes, and J. R. Sicchar, "An approach to the use of stereo vision system and AI for the accessibility of the visually impaired," 2024 Int. Conf. Control. Autom. Diagnosis, ICCAD 2024, 2024, doi: 10.1109/ICCAD60883.2024.10553945.

[16] T. Nandhini, P. Kalyanasundaram, R. M. Vasanth, S. H. Raj, and K. S. Keerthi, "Deep Learning Enabled Novel Blind Assistance System for Enhanced Accessibility," Proc. - 2024 4th Int. Conf. Pervasive Comput. Soc. Networking, ICPCSN 2024, pp. 153–158, 2024, doi: 10.1109/ICPCSN62568.2024.00034.

[17] A. Bhattacharyya, M. Palmer, and C. Heckman, "ReCAP: Semantic Role Enhanced Caption Generation." pp. 13633–13649, 2024. Accessed: Oct. 24, 2024. [Online]. Available:

https://aclanthology.org/2024.lrec-main.1191
[18] W. Zeng and W. Zeng, "Image data augmentation techniques based on deep learning: A survey," Math. Biosci. Eng. 2024 66190, vol. 21, no. 6, pp. 6190–6224, 2024, doi: 10.3934/MBE.2024272.
[19] Z. Wen and L. Guo, "Efficient Higher-order Convolution for Small Kernels in Deep Learning," Apr. 2024, Accessed: Oct. 24, 2024. [Online]. Available: https://arxiv.org/abs/2404.16380v1
[20] O. Kolovou, "MACHINE TRANSLATION FROM ANCIENT GREEK TO ENGLISH: EXPERIMENTS WITH OPENNMT," Jun. 2024, Accessed: Oct. 24, 2024. [Online]. Available: https://gupea.ub.gu.se/handle/2077/81765
[21] B. Pydala, M. K. Reddy, T. Swetha, V. Ramavath, P. Siddartha, and V. S. Kumar, "A Smart Stick for Visually Impaired Individuals through AIoT Integration with Power Enhancement," pp. 409–419, Jul. 2024, doi: 10.2991/978-94-6463-471-6_40.
[22] N. M. Upadhyay, A. P. Singh, and A. Perti, "eyeRoad – An App that Helps Visually Impaired Peoples," SSRN Electron. J., May 2024, doi: 10.2139/SSRN.4825671.
[23] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments," Expert Syst. Appl., vol. 212, p. 118720, Feb. 2023, doi: 10.1016/J.ESWA.2022.118720.
[24] S. Kumar et al., "Artificial Intelligence Solutions for the Visually Impaired: A Review," https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-6519-6.ch013, pp. 198–207, Jan. 1AD, doi: 10.4018/978-1-6684-6519-6.CH013.
[25] A. Ajina, R. Lochan, M. Saha, R. B. K. Showghi, and S. Harini, "Vision beyond Sight: An AI-Assisted Navigation System in Indoor Environments for the Visually Impaired," Int. Conf. Emerg. Technol. Comput. Sci. Interdiscip. Appl. ICETCS 2024, 2024, doi: 10.1109/ICETCS61022.2024.10543550.
[26] J. M. Parenreng, A. B. Kaswar, and I. F. Syahputra, "Visual Impaired Assistance for Object and Distance Detection Using Convolutional Neural Networks," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 8, no. 1, pp. 26–32, Jan. 2024, doi: 10.29207/RESTI.V8I1.5491.
[27] "BLIND VISION-USING AI." Accessed: Oct. 24, 2024. [Online]. Available: https://www.researchgate.net/publication/378304881_BLIND_VISION-USING_AI
[28] C. Xie, Z. Zhang, Y. Wu, F. Zhu, R. Zhao, and S. Liang, "Described Object Detection: Liberating Object Detection with Flexible Expressions," Adv. Neural Inf. Process. Syst., vol. 36, Jul. 2023, Accessed: Oct. 24, 2024. [Online]. Available: https://arxiv.org/abs/2307.12813v2
[29] "VGG16 - Convolutional Network for Classification and Detection." Accessed: Oct. 24, 2024. [Online]. Available: https://neurohive.io/en/popular-networks/vgg16/
[30] "The Annotated ResNet-50. Explaining how ResNet-50 works and why… | by Suvaditya Mukherjee | Towards Data Science." Accessed: Oct. 24, 2024. [Online]. Available: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758
[31] "Image Captioning for Visually Impaired people." Accessed: Oct. 24, 2024. [Online]. Available: https://www.kaggle.com/datasets/aishrules25/automatic-image-captioning-for-visually-impaired/data