Deep Learning Based Captioning of Toys in a Smart Monitoring System

Kritika Jindal*, Govind Agarwal*, Abishi Chowdhury*, Vishal Krishna Singh†,
Rahmat Ullah†, Mujeeb Ur Rehman‡, Amrit Pal*
*School of Computer Science and Engineering, Vellore Institute of Technology Chennai, India
†School of Computer Science and Electronics Engineering, University of Essex, Colchester Campus, Colchester, UK
‡ School of Computer Science and Informatics, Institute of Artificial Intelligence, De Montfort University, Leicester, UK

***Abstract*—The domain of image captioning has attracted increased interest in recent times due to advancements in computer vision technology and the incorporation of deep learning models, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These developments empower the creation of more precise and contextually comprehensive descriptions of images. This research aims to adapt deep learning to address the challenge of image captioning particularly for toys. A new dataset is curated in the research by sourcing copyright free images from websites featuring diverse categories of toys. Through augmentation techniques, the images are enhanced to promote dataset generalization and robustness, culminating in a comprehensive collection of images spanning distinct classes, each meticulously annotated with manually crafted captions. Feature extraction was performed using pre-trained VGG16, DenseNet201, ResNet50, and ResNet101. These models were fine-tuned to achieve optimal performance on the collected dataset. The language model utilized was LSTM. For extending the image captioning methodology to video captioning, YOLO was implemented to detect objects within video frames. Additionally, to assist visually impaired children and create a more inclusive environment, the captions were translated to audio using Google Text-to-Speech. The approach was evaluated with BLEU score and ResNet101+LSTM yielded the highest BLEU-1 score of 0.975825 outperforming the other proposed approaches.**

***Index Terms*—Deep learning, Image captioning, Video captioning, CNN, LSTM**

I. Introduction

Image captioning is a constantly evolving field with ongoing study aimed toward improving quality, diversity, and interpretability [1]. The goal of image captioning is to develop algorithms and models that automatically generate meaningful and cohesive captions without human intervention. The final aim is to provide captions that clearly convey the information, background, and key elements shown in the image [1]. To provide a meaningful natural language description, a deeper comprehension of the image is required beyond classification and object recognition. Image captioning combines Computer Vision and Natural Language Processing, two key areas of Artificial Intelligence [2]. Researchers are interested in this area due to its practical applications and integration of two major AI fields: natural language processing and computer vision [3]. Image captioning can address real-world problems involving the blind, autonomous automobiles, academic bots, and military applications [4]. Its applications extend beyond mere image recognition, opening up possibilities for enhancing accessibility, education, and overall user experience. One particularly impactful application lies in captioning images of toys for visually impaired children. Young children are extremely sensitive to the stimuli in their environment during the early stages of their lives [5]. Regardless of the developmental issues they face, visually impaired children are entitled to the same basic experiences of childhood that sighted children have such as enjoying play and learning [6]. Toys are enjoyable and stimulating objects that amplify a child's imaginative thinking, but they serve a purpose beyond mere entertainment. Although many parents consider toys to be a source of distraction, they actually provide an enjoyable and interactive means of combining essential stimuli that enhance the development of physical, social, artistic, emotional, and cognitive skills [5]. The barriers visually impaired children face are multifaceted, encompassing not only academic content but also the recreational aspects crucial for their holistic development.

It is relatively difficult for machines to generate human-like descriptions of images and videos, yet humans can effortlessly identify their surroundings and explain any image or video scenario in their native language [7]. Even while machines are somewhat capable of identifying different human activities from video frames and photos, it is still difficult to automatically describe visual situations for intricate and prolonged human activities [7]. Traditional approaches fail to detect and identify minute elements in images and videos [7]. Recognizing this challenge, the focus of the proposed research project is to leverage the capabilities of deep learning-based image captioning technology. Encoder-decoder systems are widely employed in deep learning and are highly effective for image captioning. The usage of encoder-decoder structures for image captioning has advanced significantly, with LSTM emerging as a crucial decoder for creating word sequences [8]. This methodology comprises of two primary elements: an encoder and a decoder. The encoder is commonly a Convolutional Neural Network (CNN). Convolutional neural networks (CNNs) are used for visual feature extraction, which entails taking high-level visual representations out of images [8]. Deep learning with convolutional neural networks (CNN)

Corresponding author: Amrit Pal (email: amrit.pal@vit.ac.in).

has demonstrated remarkable success in computer vision and machine learning in recent years [9]. Convolutional Neural Networks (CNNs) are very suitable for problems linked to images, as they have the ability to collect both hierarchical and spatial information present in the image. Conversely, the decoder is typically implemented as a Recurrent Neural Network (RNN) or a Transformer. The retrieved visual attributes are sent to the language modeling component, which may use recurrent neural networks (RNNs) such as Transformer models or Long Short-Term Memory (LSTM) networks [8]. The decoder utilizes the context vector produced by the encoder to build a series of words that compose the image description. The research's approach hinges on the integration of deep learning techniques, specifically tailored for image and video captioning, into the realm of assistive technology. The generated captions are converted to audio in order to facilitate a better environment. This research aims to accomplish the following:

- Compile and annotate a dataset consisting of diverse toy categories in order to ensure that all age groups of children are represented.
- Explore and address the challenges associated with extending image captioning technology to dynamic content like videos of toys.
- Generating the audio module describing the toy item in the video.

The paper is organized in the following format. First, we will address the literature survey, which entails a comprehensive analysis of existing research, theories, and procedures that are pertinent to image captioning. The suggested architecture, as seen in Figure 1, is then thoroughly discussed , after which the results are assessed and, ultimately, the primary conclusions of this work are deduced.

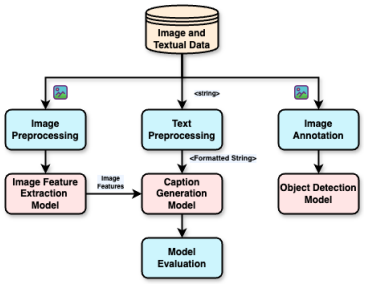


Fig. 1: Illustration of the different modules of the proposed architecture.

II. Related Work

Existing research in image captioning demonstrates diverse methodologies to address challenges related to accuracy, efficiency, and application-specific requirements. Many studies employ Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) such as LSTMs or GRUs for caption generation. For instance, models like those proposed by [10], [11], and [12] leverage CNN-LSTM architectures with pre-trained feature extractors (e.g., VGG, ResNet, or ResNeXt) to achieve strong BLEU scores on datasets such as MSCOCO, Flickr8k, and Flickr30k. Innovations such as attention mechanisms [13] and semantic reconstructions [12] further enhance caption relevance by improving contextual understanding. Specialized approaches for non-English languages like Bangla [14] and Hindi [15] highlight the adaptability of encoder-decoder frameworks with tailored datasets and evaluation metrics.

Hybrid methods and dataset-specific adaptations continue to push the boundaries of image captioning. Strategies like combining generation and retrieval-based approaches [16], integrating transformer-based encoders [17], and leveraging bidirectional models [18] address unique needs across applications. Comparisons between different encoder-decoder setups [19], [20] underline the importance of model selection based on dataset characteristics and use cases. Evaluations using multiple metrics, including BLEU, ROUGE, METEOR, and CIDEr, reveal the potential of attention-enhanced and feature-rich architectures to outperform standard models. Collectively,

TABLE I: Notation description.

Notation	Description
$D_{original}$	Original dimension of image.
D_{max}	Maximum possible dimension of image wile resizing.
$Scale_{img}$	Dimension scale of original image w.r.t W_{max}.
$Aspect - ratio_{img}$	Minimum scale up or down of the original image.
D_{new}	Dimension for resizing image.
D_{target}	Dimension to be achieved using padding.
Pad_{img}	Padding to be added image.
$caption_{old}$	Caption after text preprocessing.
$caption_{final}$	Caption after adding start and end tokens.
C_l	Classification category of annotation box.
X_c	X coordinate of center of annotation box.
Y_c	Y coordinate of center of annotation box.
w_b	Width of annotation box.
H_b	Height of annotation box.
fog_t	Forget gate of LSTM at time t.
W_{fog}	Weight for Forget gate of LSTM.
X_t	LSTM input at time t.
h_{t-1}	Hidden state of LSTM at time t-1.
b_{fog}	Bias for Forget gate of LSTM.
i_t	Input gate of LSTM at time t.
W_i	Weight for input gate of LSTM.
b_i	Bias for input gate of LSTM.
$inter_t$	Intermediate cell state of LSTM at time t.
W_c	Weight for cell state of LSTM.
b_c	Bias for cell state of LSTM.
C_t	Updated cell state of LSTM at time t.
C_{t-1}	Cell state of LSTM at time t-1.
out_t	Output gate of LSTM at time t.
W_{out}	Weight for output gate of LSTM.
b_{out}	Bias for output gate of LSTM.
h_t	Hidden state of LSTM at time t.
Num_{cap}	Number captions generated.
n_{frame}	Number of frames in one second.
t_{len}	Time duration of video in seconds.
$Video_{cap}$	Final caption of the video.
max_{count}	Frequency of unique captions generated per frame.
BP	Brevity penalty.
$exp(w_n$	Weight assigned to n-gram precision.
$\sum_{clipped_count}$	Cumulative clipped count of n-gram.
$\sum_{count}$	Cumulative count of n-gram.

these studies pave the way for more robust, efficient, and adaptable image captioning systems that meet the demands of various languages, domains, and user contexts.

III. Proposed Architecture

A. Data Collection

The data gathering phase is critical for developing the image and video captioning system to address the challenges that visually impaired children encounter while accessing educational and recreational items, notably toys. A custom dataset was constructed for this study by collecting images from several copyright-free sources to establish the groundwork for creating a comprehensive and effective deep learning model. Copyrights free images were obtained from sources such as Pexels [21], PxHere [22], PixaHive [23], and Unsplash [24]. The images feature a variety of colors and toys, including soft toys, toy cars, toy trains, and sports toys like badminton rackets, cricket bats, and basketballs. A total of 581 images were collected, representing a collection of 20 distinct categories. Each image in the dataset was assigned a unique ID, and captions were manually generated after augmenting the images. Figure 2 depicts some sample image from several classes.

Fig. 2: Example images from the collected dataset demonstrating different classes.

B. Image Processing

The image preprocessing step is an important phase in developing the image and video captioning system for enhancing the accuracy and efficiency of the system. Primarily the images were standardized to a particular format namely JPG to introduce uniformity in across the dataset. The images were then cropped based on the toy object present and then they were all resized to 420x420 pixels based on aspect ratios to preserve the robust features and facilitate computational efficiency. Aspect ratio represents the proportional relationship between the height and width of an image. The images were resized based on aspect ratio in order to prevent the image from getting stretched or compressed out of proportions. For this the first the scale of the image and the aspect ration based on which resizing will be done was calculated by using equation 1 and 2. The new dimensions of the image were calculated using equation 3. Following aspect ratio-based resizing, the images were padded with white borders to preserve their original dimensions, the dimensions of the white border to be added for each image was calculated using equation 4. This process resulted with images of dimension 512x512 for the dataset as a whole. This methodology ensures that the deep learning model captures the important visual aspects of the images. The dataset comprises a total of 581 images spanning 20 classes. These images were augmented to introduce generalization and variations in the toy images. The operations such as flipping, rotation and color modification were used for data augmentation. A total of 3,581 images were generated from the original 581 images collection to form the final dataset. All the images were given unique IDs in the form of an integer and manual captions were written for each image at the end of image processing phase. This enhanced the images hence contributing to the development of the deep learning model.

$$Scale_{img} = \frac{D_{max}}{D_{original}} \tag{1}$$

$$Aspect - ratio_{img} = min(Scale_{img}) \tag{2}$$

$$D_{new} = Scale_{img} * Aspect - ratio_{img} \tag{3}$$

$$Pad_{img} = D_{target} – D_{new} \tag{4}$$

C. Text Preprocessing

Text preprocessing is essential for refining and organizing the textual data present in the form of captions for each image before inputting them into the system. Initially all the captions were converted to lowercase to maintain consistency. Following this the special characters and punctuation were eliminated to simplify the textual data and decrease the complexity. Redundant words and stop words were eliminated to further streamline the textual data and focus on primary language characteristics. Furthermore, white space normalization was performed by replacing many consecutive white space characters with a single space, to facilitate uniformity. Then the text was subjected to tokenization that is breaking it down into tokens or words to enable the formation of organized sequences which provides deep learning model with structured inputs that can be efficiently interpreted. To add additional context and increase model interpretation, each caption is preceded by a special 'startseq' token and followed by a 'endseq' token, indicating the beginning and end of the caption as shown in equation 5. Through these meticulous steps, the text preprocessing module ensures that the input text is refined and optimized for the subsequent stages of the captioning model.

$$caption_{final} = `starttoken' + caption_{old} + `endtoken' \tag{5}$$

D. Image Annotation

Image Annotation played a crucial role for implementing the deep learning system for captioning of images and videos of toys. The procedure was carried out through manual annotation process by employing multiple features and techniques provided by Computer Vision Annotation Tool (CVAT) website [25]. Various functionalities such as bounding boxes, key points, polygons and segmentation masks were utilized to meticulously annotate the dataset for promoting comprehensive understanding of toy images and videos. CVAT offers a user friendly platform which facilitated the smooth

completion of annotation tasks, ensuring a seamless process. Rigorous quality control procedures were implemented to ensure the accuracy and consistency of each annotation, which was critical for the success of the subsequent deep learning model. The format of annotations can be seen in equation 6. The dataset underwent continuous enhancement through iterative procedures, contributing to its quality and diversity. Upon completion, the annotated dataset was exported in a standardized format, streamlining its integration into the training process of deep learning models. This thorough and systematic image annotation module aligned with the research project's goal of developing an innovative assistive technology. It extended deep learning-based image captioning to dynamic content, providing visually impaired children with an inclusive and enriching experience.

$$Annotation =< C_l, X_c, Y_c, W_b, H_b > \tag{6}$$

E. Image Feature Extraction

For the crucial process of feature extraction from images transfer learning technique is utilized which offers improved efficiency for the system. Convolutional Neural Networks (CNNs) such as DenseNet201, VGG16, ResNet50 and ResNet101 are employed for this task which were pre-trained on the ImageNet dataset. A systematic process was used for the adaption of these models where the last layer of classification was removed and custom layers were added to tailor them to fit the dataset.The meticulous integration of pre-trained CNNs and custom layers showcases a sophisticated approach in extracting the visual features, aligning with the goal of creating an inclusive and enriching experience for visually impaired children. VGG16 is used as it provides a balance between the accuracy and computational efficiency on the other hand DenseNet201 offers dense connectivity patterns which excel in capturing fine grained patterns. ResNet50 was employed as it is adept at capturing the intricate details and patterns while ResNet101 offers an additional layer of complexity offering an extended capacity to understand and represent diverse characteristics of the toys in the dataset. The detailed fine-tuned architecture of ResNet101 is shown in Figure 3.

F. Caption Generation

The main objective of the proposed system is to be able to develop captions for the images and videos of toys for children. For this task Long Short-Term Memory (LSTM) is employed for generating sequential captions and Convolutional neural networks for extracting image features. The decoder which consists the LSTM inputs the encoded image features and generates words in a sequential manner. The output produced by the LSTM are represented in the form of word embedding which is a numerical representation. The embedding facilitate a more nuanced understanding of the semantics and context of the words used in the captions. A vocabulary was constructed by compiling all the unique tokens or words from the corpus of textual data which allows effective mapping between words and numerical representations. The conversion of words into numerical representations streamlines the processing by the output layer, resulting in a probability distribution across the vocabulary. This method enables the prediction of the next word in the sequence, hence contributing to the generation of coherent and contextually relevant captions. This module serves as the key component for integrating the visual features with the process of creative meaningful and descriptive captions. Figure 4 shows the diagram of $(t-1)_{th}$, t_{th} and $(t+1)_{th}$ states of an LSTM. LSTM uses various gates to regulate the flow of information within a network like forget gate, output gate, and cell state. In order to decide which information is relevant and irrelevant and should be forgotten lstm uses equation 7.

$$fog_t = \sigma(W_{fog} \cdot X_t + W_{fog} \cdot h_{t-1} + b_{fog}) \tag{7}$$

Equation 8 is utilized to determine which new information should be stored for output and future reference.For the current input Xt and the previous hidden state of (t-1), LSTM h_{t-1} is used. New information is stored in $inter_t$ and calculated using equation 9.

$$i_t = \sigma(W_i \cdot X_t + W_i \cdot h_{t-1} + b_i) \tag{8}$$

$$inter_t = tanh(W_c \cdot X_t + W_c \cdot h_{t-1} + b_c) \tag{9}$$

Equation 10 updates the current cell state for present and future reference in C_t. Equation 11 calculates the output of the LSTM layer, while equation 12 calculates the hidden state for the current timestamp (t) to be used in the next timestamp.

$$C_t = fog_t \cdot C_{t-1} + i_t \cdot inter_t \tag{10}$$

$$out_t = \sigma(W_{out} \cdot X_t + W_{out} \cdot h_{t-1} + b_{out}) \tag{11}$$

$$h_t = out_t \cdot tanh(C_t) \tag{12}$$

G. Object Detection

The research leverages the state of the art YOLOv8 (You Only Look Once) architecture for the task of object detection. YOLO is chosen for its efficiency and accuracy in real time object detection making it well suited for dynamically changing content such as videos. It processes the image in one forward pass. It employs a single neural network to simultaneously predict bounding boxes and class probabilities for all objects in an image or video frame. This methodology ensures real time responsiveness and accuracy which is crucial for providing timely and correct information. YOLO divides the input image into a grid and predict the bounding boxes. This grid based approach enhances the speed and efficiency of the images swiftly. The Object Detection module enhances the assistive technology's effectiveness for visually impaired children by integrating the YOLOv8 framework. To caption a video consisting a toy item the video is first split into individual frames on which then the YOLO algorithm is applied to detect object, if an object is detected then the frame is cropped. After this, image extractor module examines the cropped image and extracts features from it, which are forwarded to the caption generation model which then generated captions based on the features received. This process flow is depicted in Figure 5.

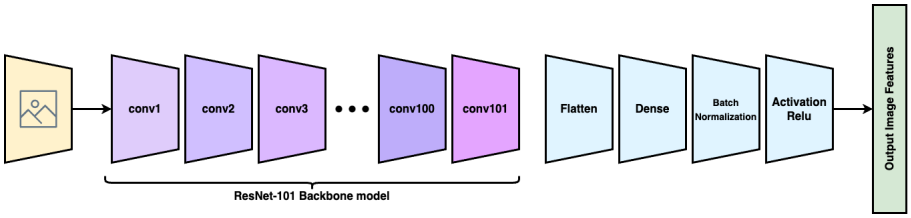


Fig. 3: Framework of ResNet101 for feature extraction from toy images.

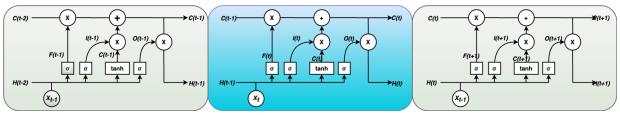


Fig. 4: Detailed representation of the functioning mechanism of the LSTM.

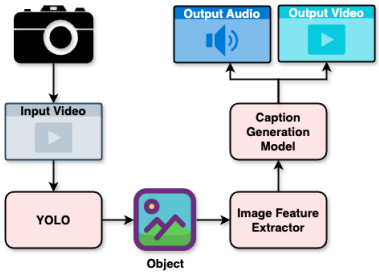


Fig. 5: Visualization of the object detection process using YOLOv8.

H. Audio Generation

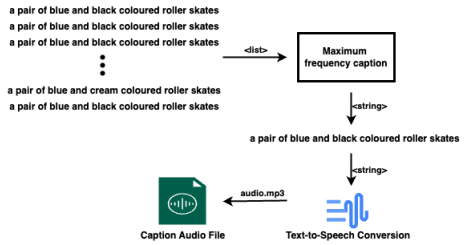


Fig. 6: Illustration of the process of integrating audio output for detected toy objects with Google's text-to-speech technology.

In order to provide greater inclusivity and better experience an audio generation module is utilized which generates the the audio output for the object detected in the video frames. To accomplish this task with precision and finesse, we've leveraged the cutting-edge capabilities of Google's Text-to-Speech technology. It offers high quality and natural sounding voices that closely resemble human speech patterns. For the process of generating audio output first all the captions generated during processing of the input video were stored in a list and then the sentence which has the most number of occurrences were counted. The audio output was generated for the sentence with the maximum number of occurrences. The audio output was generated using English language and in the standard default voice that Text-to-Speech engine offers. The audio file was saved in MP3 format. This process helps generation of captions in the form of audio to help create a invaluable context and enrichment, enhancing the overall play experience for the visually impaired children and provide captions for toy objects. The output audio caption file is then stored separately from the video caption file as shown in Figure 5. For reducing the noise and getting a single sentence which gives a more apt meaning to the video rather than relying on sentences generated every frame equations 13 and 14 was utilized. Using this equation audio output was generated for the sentence with the maximum number of occurrences. This process is illustrated in Figure 6.

$$Num_{cap} = n_{frame} * t_{len} \tag{13}$$

$$Video_{cap} = max_{count}(Num_{cap}) \tag{14}$$

I. Model Training

Prior to model training, the images were processed and resized to 224x224 dimensions, and their color space was converted to BGR from RGB using OpenCV. The names of the images and their corresponding labels were appended to separate lists. For feature extraction, four different CNNs, which were pre-trained on the ImageNet dataset, were experimented with in the training phase. Initially the ResNet101 pre-trained weights on ImageNet dataset were loaded and the fully connected layers at the top of the network were not included to enable addition of custom layer. A Flatten layer was introduced in the network to convert the 2D output to 1D tensor. Following this, a dense layer with 512 units was appended to provide a dense connection for the next layer. Batch normalization was incorporated to normalize the activation and additionally an activation of ReLU was employed to introduce non-linearity. Finally a layer with softmax activation was added for multiclass classification. ResNet50 and DenseNet201 were implemented in the same manner; however, for VGG-16, after loading the pre-trained model, a GlobalAveragePooling2D layer was applied to the output of VGG16 following which a series of dense layers with ReLU activation function were added. Then a dense layer activated by the softmax function to produce class probabilities was used. For generation of captions an LSTM with 256 units was used before which a dropout layer with 0.4 dropout rate was included to prevent overfitting. In the decoder the feature vectors and the output of the LSTM were combined and passed through a dense layer of 256 units with ReLU activation.

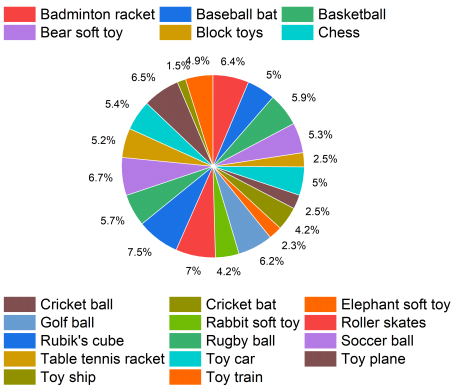


Fig. 7: Visualization of distribution of the different classes present in the collected dataset.

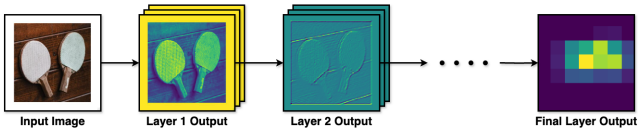


Fig. 8: Intermediate output of the feature extraction model illustrating the outputs of the different stages of the feature extraction.

Another layer with softmax is added to output the probability distribution over the vocabulary for the prediction of the next word. The model was compiled using the Adam optimizer. This enabled prediction of the next word in the sequence and was trained for 30 epochs.

IV. Results

captioning toy objects portrayed in videos and images. In order to comprehend the intricate workings of the system and assess the effectiveness and efficiency of the proposed methodology it is crucial to analyze the following questions:

- How diverse and representative is the collected dataset in terms of different age groups and toy types?
- How the proposed model perceives the input images and analyzes the features?
- How well does the proposed approach perform when weighing the trade-off between accuracy and loss as training progresses?
- How does BLEU score represent the quality and fluency of generated captions?
- How does the proposed approach perform when experimented with different datasets?

In order to develop a system that is appropriate for all age groups of children, it is crucial that the data gathered must include various categories of toys encompassing different age groups and toy types. Figure 7 illustrates that the dataset is diverse across toy categories and environments, including images from inside settings such stuff toys, table tennis, and outdoor settings like badminton and toy plane. This variability guarantees that the model encounters a broad range of visual

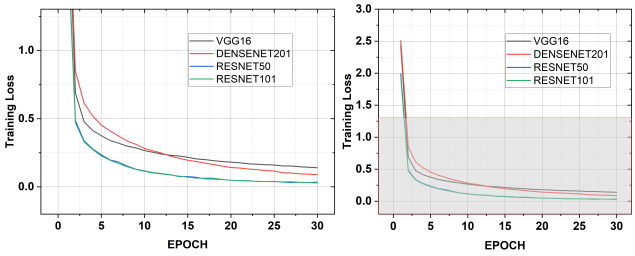


Fig. 9: Analyzing the trends in the training loss for the different proposed approaches.

environments, increasing its ability to adapt across varied scenarios.

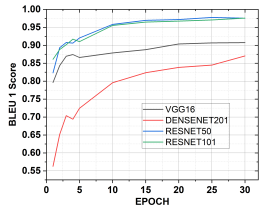
Figure 8 demonstrates the way the model processes the input image. Filters are used to the input image to identify and quantify edges, as well as detect objects.

Figure 9 depicts the graph created while training various proposed techniques, including VGG16, DenseNet201, ResNet50, and ResNet101. It can be observed that there is a significant reduction in the loss value before epoch 5, which is relevant because during the initial epochs, the model begins to learn patterns from the data, resulting in a decrease in loss over subsequent training as the model continues to adapt to the data. Upon close examination, it becomes evident that ResNet101 significantly reduces the loss above 10% while training on comparing with other suggested methods. ResNet101 outperforms other models in this regard, having the lowest training error.

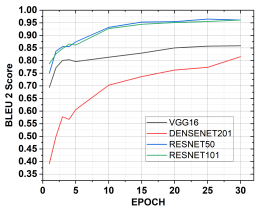
To evaluate the efficacy and efficiency of the proposed approach, an assessment metric is required to be established. The BLEU score was utilized in the study to assess performance and the similarity between the generated text and the reference toy captions. The BLEU score is calculated by comparing the n-grams in the produced text and the reference text, then counting the number of matches. The performance of the technique was assessed using the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 metrics in the study. Equation 15 shows the equation for calculating the BLEU score.

$$BLEU - n = BP \times exp(w_n \times \frac{\sum_{clipped_count}}{\sum_{count}}) \quad (15)$$

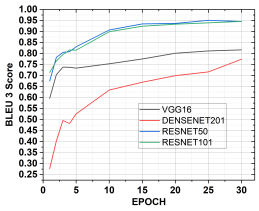
The 1-gram, 2-gram, 3-gram, and 4-gram BLEU scores were computed using the NLTK corpus_bleu functionality. A BLEU score of more than 0.6 is regarded as acceptable. This measure will evaluate whether or not the suggested method produces toy image captions that closely correspond with the reference captions.Following fine-tuning and integration with the caption generating system, the pre-trained CNN models VGG16, DenseNet201, ResNet50, and ResNet101 were trained for 30 epochs each. Figure 10a, 10b, 10c and 10d shows the graphs for BLEU score vs Epochs for different proposed approaches. Figure 10a demonstrates that the starting BLEU-1 scores for ResNet50 and ResNet101 are 0.8229 and 0.8603, respectively. These scores show an upward trend with increasing epochs,



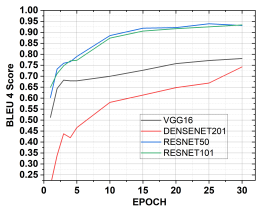
a: Depicting the BLEU-1 score values vs Epochs for different proposed approaches.



b: Depicting the BLEU-2 score values vs Epochs for different proposed approaches.



c: Depicting the BLEU-3 score values vs Epochs for different proposed approaches.



d: Depicting the BLEU-4 score values vs Epochs for different proposed approaches.

Fig. 10: BLEU-1 and BLEU-2 analysis of the proposed approach with respect to the existing approaches

Fig. 11: Example image captioning results obtained from the proposed approach for different classes of toys.

reaching 0.9754 and 0.9758 for ResNet50 and ResNet101, respectively. This implies that both models demonstarte comparable performance with ResNet101 illustrating more effective learning with a higher BLEU-1 score. Furthermore, Figure 10b shows that DenseNet201 is initially unable to capture bi-gram, which accounts for word pairs that occur together, and learning becomes stationary at the 27th epoch. ResNet50 and ResNet101 demonstrate similar performance for BLEU-2 which is similar to performance with BLEU-1. Figure 10c displays that the performance of VGG16 remains relatively unchanged after 15th epoch due to which it can be said it is not able to capture the 3-gram precision hence resulting in a lower value for BLEU-3. ResNet101 is learning better as compared to other models with respect to BLEU-4 as evident from Figure10d although the rate of increase in BLEU-4 score slows down after 10th epoch. Table II provides a full breakdown of the evaluation's results based on BLEU scores for the collected dataset. ResNet101 demonstrated superior performance compared to the other models, as indicated by its BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.975825, 0.961004, 0.946121, and 0.933815, respectively. Additionally, it is worth mentioning that ResNet50 exhibited acceptable performance, as evidenced by its BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.975492, 0.961104, 0.945395, and 0.93103, which are notably similar to those of ResNet101. In order to determine the effectiveness of the approach it is crucial to compare it with different datasets hence the proposed methodology was experimented with Flickr8k dataset and the BLEU score results are shown in Table III. It can be noted that there is a 2.61% increase as compared to approach [10], a 3.3% increase in comparison to [26] and a 36.53% increase as compared to [27].

TABLE II: Comparative experiments with various proposed methods based on BLEU score metric.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16	0.907449	0.858897	0.816732	0.780725
DenseNet	0.870305	0.815885	0.77397	0.742267
ResNet50	0.975492	0.961104	0.945395	0.93103
ResNet101	0.975825	0.961004	0.946121	0.933815

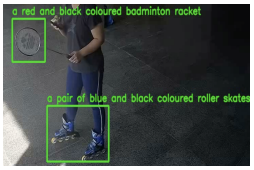
TABLE III: BLEU score comparison for Flickr8k.

Approach	BLEU score
[10]	0.53356
[26]	0.53
[27]	0.401
Proposed	0.547511

First the approach was experimented to generate captions for images of different toy classes, the results of the image captioning are displayed in Figure 11 explaining the actual and the predicted captions generated by the system. Moving forward the developed system was experimented for video detection of toy images. Figure 12a, 12b, 12c, 12d display the output video screenshots for different scenarios.

V. Conclusion

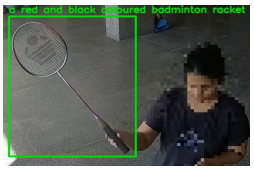
This research presented an approach to address the distinctive challenges faced by visually impaired children in accessing and engaging with educational and recreational materials particularly toys. The dataset utilized in the research was sourced from websites. The dataset consisted of 581 copyright free images which were augmented to form a total set of 3,581 images. The images were mapped with captions written manually in English language to describe the items of the images. For making the dataset ready for training the images went under image preprocessing techniques to maintain uniformity across all the images. The textual data was also preprocessed to maintain a format across the data. Image annotation, image features extraction and object detection along with caption generation techniques were used to construct the proposed model. The feature extraction module leverages various pretrained CNN architectures on the ImageNet dataset, including VGG, DenseNet201, ResNet50, and ResNet101, along with custom layers. The research employed language model LSTM



a: Video caption output for badminton racket and roller skates.

b: Video caption output for bear soft toy.

c: Video caption output for roller skates.



d: Video caption output for badminton racket.

Fig. 12: Video captioning output for different categories

and word embedding to sequentially generate captions of the images. For extending the image captioning to video captioning YOLOv8 was utilized to integrate static image captioning system to caption the dynamic contents of a video. Additionally audio output for the generated caption for the video was also generated and stored as MP3 file. The evaluation metrics employed for the study was BLEU score which is used to evaluate the quality of the machine generated translations by comparing them to reference translations. It was noted that ResNet101 outperformed the other approaches by obtaining the following values of BLEU-1, BLEU2, BLEU-3 and BLEU-4 scores of 0.975825, 0.961004, 0.946121, and 0.933815 respectively. In the future, other approaches can be used for feature extraction and caption generation with employing a larger dataset to increase the performance of the system.

REFERENCES

[1] N. Panchal and D. Garg, "Image captioning: A comprehensive survey, comparative analysis of existing models, and research gaps," in *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2023, pp. 1120–1127.

[2] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4651–4659.

[3] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic image caption generation using deep learning," *Multimedia Tools and Applications*, vol. 83, pp. 5309–5325, 2024. [Online]. Available: https://doi.org/10.1007/s11104202315555y

[4] A. Z. Al-Jamal, M. J. Bani-Amer, and S. Aljawarneh, "Image captioning techniques: A review," in *2022 International Conference on Engineering & MIS (ICEMIS)*, 2022, pp. 1–5.

[5] Stacy, "Why toys are important for a child's development? – creating compassionate kids," 01 2023. [Online]. Available: https://creatingcompassionatekids.org/ why-toys-are-important-for-a-childs-development/

[6] B. K. Lizeth Sustaita Delgado, "Study of the experience of children's toys for low-vision and blindness - industrial designers society of america," https://www.idsa.org/education- paper/study-of-the-experience-of-childrens-toys-for-low-vision-and-blindness/, 2023, [Accessed 03-05-2024].

[7] D. Sharma, C. Dhiman, and D. Kumar, "Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey," *Expert Systems with Applications*, vol. 221, p. 119773, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423002749

[8] A. D. Shetty and J. Shetty, "Image to text: Comprehensive review on deep learning based unsupervised image captioning," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, 2023, pp. 1–9.

[9] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Inception recurrent convolutional neural network for object recognition," *Machine Vision and Applications*, vol. 32, p. 28, 2021. [Online]. Available: https://doi.org/10.1007/s00138020011573

[10] C. Amritkar and V. Jabade, "Image caption generation using deep learning technique," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–4.

[11] P. Singh, C. Kumar, and A. Kumar, "Nextlstm: a novel lstmbased image captioning technique," *International Journal of System Assurance Engineering and Management*, vol. 14, pp. 1492–1503, 2023. [Online]. Available: https://doi.org/10.1007/s13198023019567

[12] R. A. Ahmad, M. Azhar, and H. Sattar, "An image captioning algorithm based on the hybrid deep learning technique (cnn+gru)," in *2022 International Conference on Frontiers of Information Technology (FIT)*, 2022, pp. 124–129.

[13] M. Sarkar, S. Biswas, and B. Ganguly, "A hybrid transfer learning architecture based image captioning model for assisting visually impaired," in *2023 IEEE 3rd Applied Signal Processing Conference (ASPCON)*, 2023, pp. 211–215.

[14] T. Kamruzzaman, S. Kamruzzaman, and A. Zaman, "A deep learning approach for bangla image captioning system," in *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 2021, pp. 1–6.

[15] A. K. Poddar and D. R. Rani, "Hybrid architecture using cnn and lstm for image captioning in hindi language," *Procedia Computer Science*, vol. 218, pp. 686–696, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S1877050923000492

[16] G. Hoxha, F. Melgani, and J. Slaghenauffi, "A new cnn-rnn framework for remote sensing image captioning," in *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, 2020, pp. 1–4.

[17] C. Cai, K.-H. Yap, and S. Wang, "Attribute conditioned fashion image captioning," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1921–1925.

[18] M. Humaira, S. Paul, M. Jim, A. Ami, and F. Shah, "A hybridized deep learning method for bengali image captioning," *International Journal of Advanced Computer Science and Applications*, vol. 12, p. 698, 02 2021.

[19] V. Atliha and D. Šešok, "Comparison of vgg and resnet used as encoders for image captioning," in *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 2020, pp. 1–4.

[20] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance," *Pattern Recognition Letters*, vol. 119, pp. 229–237, 2019. [Online]. Available: https://www.sciencedirect.com/science/article /pii/S0167865517303811

[21] Pexels, "Free stock photos," Pexels.com, 2023. [Online]. Available: https://www.pexels.com/

[22] Pxhere, "Free images and free stock photos - pxhere," Pxhere.com, 2018. [Online]. Available: https://pxhere.com/

[23] Pixahive, "100% free stock photos with cc0 license - pixahive.com," PixaHive, 2022. [Online]. Available: https://pixahive.com/

[24] Unsplash, "Beautiful free images and pictures," Unsplash, 2022. [Online]. Available: https://unsplash.com/

[25] Cvat, "Cvat," www.cvat.ai, 2023. [Online]. Available: https://www.cvat.ai/

[26] S. S. Aote *et al.*, "Image caption generation using deep learning technique," *Journal of Algebraic Statistics*, vol. 13, p. 22602267.

[27] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, and P. Gupta, "Deep image captioning using an ensemble of cnn and lstm based deep neural networks," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 5761–5769, 2021.