






Neural Image Caption Generation with Visual Attention: Enabling Image Accessibility for the Visually Impaired

 Priyanka Agarwal¹,  Niveditha S²,  Shreyanth S³,  Sarveshwaran R⁴,  Rajesh P K⁵

¹MTech in Data Science and Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India

²Department of Biotechnology, Rajalakshmi Engineering College, Thandalam, Chennai, Tamilnadu, India

^{3,4,5}MTech in Data Science and Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India

ARTICLE INFO

Article History:

Accepted: 10 June 2023

Published: 24 June 2023

Publication Issue

Volume 10, Issue 3

May-June-2023

Page Number

562-575

ABSTRACT

The internet is saturated with images that convey messages and emotions more effectively than words alone in today's digital age. Individuals with visual impairments, who are unable to perceive and comprehend these images, face significant obstacles in this visual-centric online environment. As there are millions of visually impaired people around the globe, it is essential to close this accessibility gap and enable them to interact with online visual content. We propose a novel model for neural image caption generation with visual attention to address this pressing issue. Our model uses a combination of CNNs and RNNs to convert the content of images into aural descriptions, making them accessible to the visually impaired. The primary objective of our project is to generate captions that accurately and effectively describe the visual elements of an image. The model proposed operates in two phases. First, a text-to-speech API is utilized to convert the image's content into a textual description. The extracted textual description is then converted to audio, allowing visually impaired individuals to perceive visual information through sound. Through exhaustive experimentation and evaluation, we intend to achieve a high level of precision and descriptivism in our system for image captioning. We will evaluate the performance of the model by undertaking comprehensive qualitative and quantitative assessments, comparing its generated captions to ground truth captions annotated by humans. By enabling visually impaired individuals to access and comprehend online images, our research promotes digital inclusion and equality. It has the potential to improve the online experience for millions of visually impaired people, enabling them to interact with visual content and enriching their lives through meaningful image-based interactions.

Keywords: Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Attention Model, Natural Language Processing (NLP), NLTK, Machine Learning (ML), Deep Learning (DL), Flicker Dataset, gTTS – Google API

I. INTRODUCTION

Visual content has become an integral element of our online experiences in the current digital age. Social media platforms, advertisements, and various online platforms are dominated by images, which effectively convey messages, emotions, and information. However, individuals with visual impairments are severely limited in their ability to perceive and comprehend these visual elements, creating barriers to their access and participation in the digital domain. According to the World Health Organization (WHO), approximately 285 million persons worldwide have visual impairments, including 39 million who are completely blind. To promote inclusion and equal opportunity, it is of the utmost importance to address the accessibility challenges encountered by this sizeable population.

This research paper develops a neural image caption generation model with visual attention to address the issue of image accessibility for the visually impaired. The proposed model combines deep learning and natural language processing techniques to enable visually impaired users to experience the power of internet images via auditory descriptions. This model seeks to bridge the gap between visual content and the auditory perception of individuals with visual impairments by converting image content to speech (Fig 1).

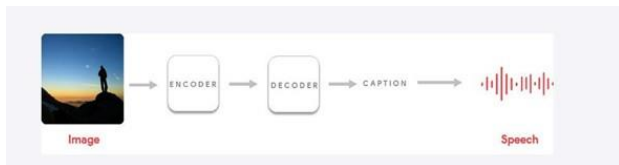


Figure 1: Business Process Flow of Proposed Model

The project includes a convolutional neural network (CNN)-based encoder that extracts salient image features and a recurrent neural network (RNN) decoder that generates textual descriptions. This

model's most important component is the attention mechanism, which replicates the complex cognitive abilities of human visual processing. The attention mechanism enables the model to concentrate on pertinent areas of the image while ignoring redundant information, thereby capturing the essence of the image in the generated captions.

This project's primary objective is to develop a CNN-RNN model with an attention mechanism that can generate accurate and meaningful captions for Flickr8K images. Additionally, the model will use a text-to-speech library to convert these captions into audio, allowing visually impaired individuals to perceive and comprehend the image content via auditory means. The primary objective of this study is to enhance our skills and knowledge in the disciplines of deep learning, natural language processing, and computer vision in order to develop a useful tool for the visually impaired community.

This research has the potential to empower visually impaired individuals to interact autonomously with online visual content. We hope to improve their access to information, amusement, and social interactions in the digital space by providing audio descriptions of images. This research adheres to the universal design principles, fostering equal opportunities for all individuals regardless of their visual abilities.

Traditional methods of assisting the visually impaired, such as braille, provide tactile information but are incapable of effectively representing visual content. Our approach, on the other hand, utilizes advances in deep learning and natural language processing to bridge this divide by converting the image content to an audio format. In addition, the model's attention mechanism optimizes computational efficiency by focusing only on the relevant portions of the image, thereby overcoming the limitations of passing the entire image to each RNN layer.

The scope of this project includes the development of a CNN-RNN based architecture, the construction of a custom model using TensorFlow, evaluation using BLEU score, TensorFlow-based data preparation, a comparison of greedy and beam search strategies, and the incorporation of the attention model. The research will utilize available resources such as an 8GB RAM, 500GB HDD, and platforms such as Google Colab and Google Cloud Platform for development and hosting, respectively.

II. LITERATURE REVIEW

Significant research effort has been devoted to the production of image captions, with the goal of enabling machines to generate accurate and meaningful descriptions of visual content. This article provides an overview of pertinent studies and approaches in the field of neural image captioning with visual attention, with a focus on the recent developments.

Anderson et al. [1] proposed a bottom-up and top-down mechanism for visual query answering and image captioning. Their method combined image-extracted visual characteristics with attention models to generate informative captions. This study demonstrated the efficacy of attention mechanisms for improving the content of generated captions.

Sharma et al. [2] conducted a thorough examination of image captioning techniques. They investigated various approaches, such as neural networks, attention mechanisms, and reinforcement learning, to shed light on the development of image captioning models. The survey emphasized the significance of attention-based methods and their influence in enhancing caption quality.

The groundbreaking work of Xu et al. [3] introduced the concept of "Show, Attend, and Tell," which combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with attention

mechanisms for the generation of image captions. During the process of caption generation, their model learned to selectively focus on various regions of the image, resulting in improved results.

In the context of image captioning, contrastive learning of visual representations has also been studied. Chen et al. [4] proposed a simple framework for contrastive learning, which leverages large-scale unlabeled image collections to discover meaningful visual representations. This method improved the comprehension of visual content and the efficacy of image captioning models.

Semantic attention has been investigated as a means to enhance the quality of image captioning. Gu et al. [5] proposed a model for image captioning that incorporated semantic information to guide the attention mechanism. By taking into account the semantic significance of image regions, their model generated captions that were more precise and contextually consistent.

Image captioning has also been influenced by reinforcement learning. Zhou et al. [6] presented a deep reinforcement learning-based method for image captioning in which an agent learned to generate captions by optimizing an embedding reward. This technique enabled the model to generate diverse and accurate captions.

Gan et al. [7] centered their research on semantic compositional networks for visual captioning. Combining visual concepts and their relationships, their model utilized a compositional structure to produce captions. This method enabled the production of more descriptive and understandable captions.

The CNN-RNN framework proposed by Wang et al. [8] is applicable to image captioning. Combining CNN for image feature extraction and RNN for caption

generation, their model achieved competitive performance on multi-label classification tasks.

Li et al. [9] conducted an exhaustive survey of visual-to-text techniques, such as image and video captioning. The survey examined numerous approaches, such as CNN-RNN models, attention mechanisms, and reinforcement learning, to shed light on the advancements in visual-to-text generation.

Text-to-image synthesis has been investigated to enhance the performance of image captioning. Hossain et al. [10] investigated the application of text-to-image synthesis techniques for the generation of enhanced image captions. Their approach improved the comprehension and quality of generated captioning by synthesizing visual content based on textual descriptions.

III. METHODOLOGY

Through neural image caption generation with visual attention, this research aims to make images accessible to the visually impaired. The methodology section of this paper describes the systematic approach taken to achieve this objective. This section describes the primary stages involved in data collection and preprocessing, architecture overview, integration of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), integration of attention mechanisms, text-to-speech conversion, and model training and evaluation.

A. Data Collection and Pre-processing

The first step in our methodology is the collection and preprocessing of the data for training and evaluation. For this research, we utilize the Flickr8K dataset, which is a widely used benchmark dataset for image captioning tasks. This dataset consists of 8,000 images, each paired with five captions, resulting in a total of 40,000 caption-image pairs (Fig 2).



The man with pierced ears is wearing glasses and an orange hat .
A man with glasses is wearing a beer can crocheted hat .
A man with gauges and glasses is wearing a Blitz hat .
A man wears an orange hat and glasses .
A man in an orange hat starring at something .



The small child climbs on a red ropes on a playground .
A small child grips onto the red ropes at the playground .
A little girl in pink climbs a rope bridge at the park .
A little girl climbing on red roping .
A child playing on a rope net .

Figure 2: Unveiling the Visual Symphony: A Kaleidoscope of Images for Neural Inspiration

During the data collection process, we ensure that the dataset contains a diverse range of images covering various topics, objects, and scenes. This diversity helps our model learn to generate captions that are applicable to a wide array of visual content.

After obtaining the dataset, we preprocess the data to ensure its suitability for training our neural image caption generation model. The preprocessing steps include the following:

1. Image Preprocessing:

We resize all the images to a uniform size, such as 256x256 pixels, to maintain consistency. We normalize the pixel values of the images to a range of 0 to 1, ensuring that the model can effectively process the visual features.

2. Caption Preprocessing:

We convert all the captions to lowercase to standardize the text and avoid redundancy due to case variations. We tokenize the captions into individual words to create a vocabulary for our model. This step helps in handling the sequential nature of text data and enables the generation of captions word by word. We remove punctuation and special characters from the captions to reduce noise and focus on the essential linguistic information.

3. Dataset Split:

To assess the performance of our model, we split the dataset into training, validation, and testing sets. Typically, we allocate around 70-80% of the data for training, 10-15% for validation, and the

remaining portion for testing. This split ensures that the model is trained on a substantial amount of data while having separate data for evaluation and testing.

By performing these data preprocessing steps, we ensure that the input data is standardized, consistent, and ready to be fed into our neural image caption generation model. The processed dataset serves as the foundation for training the model to learn the relationship between images and their corresponding textual descriptions.

B. Architecture Overview

Our proposed architecture for neural image caption generation with visual attention consists of two primary components: an image encoder and a caption decoder. This architecture combines the capabilities of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract image features and generate captions, respectively, and incorporates an attention mechanism to improve the model's ability to concentrate on salient image regions (Fig 3).

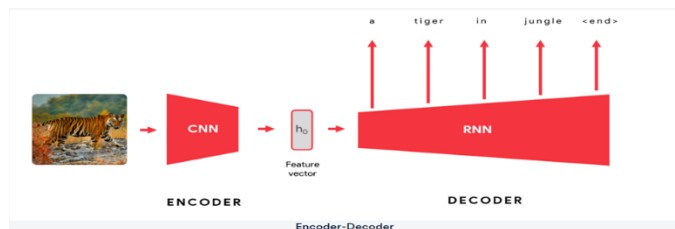


Figure 3: Enhancing Image Accessibility: A Unified Architecture for Neural Image Caption Generation with Visual Attention

The image encoder is essential for capturing the visual information present in the input images. We employ a CNN that has been trained on a large-scale image classification task, such as VGG16 or ResNet. By eliminating the CNN's classification layer, we convert it into a feature extractor. The encoder processes the input image through multiple convolutional layers, pooling layers, and activation functions to produce a

high-level representation of the image's features. This representation of the feature captures essential visual information, including shapes, hues, and textures.

On the other hand, the caption decoder uses RNNs to generate captions based on the extracted image features. Specifically, we employ the Long Short-Term Memory (LSTM) network, a variant of RNN. As it effectively captures long-range dependencies, the LSTM network is ideally adapted for processing sequential data such as language.

We incorporate an attention mechanism into the architecture to facilitate the interaction between the image encoder and caption decoder. The attention mechanism allows the model to dynamically focus on various regions of an image while generating captions, simulating the visual attention process of humans. This mechanism improves the model's capacity to generate accurate and contextually pertinent captions by paying attention to relevant image regions during each step of caption generation.

During the process of caption generation, the attention mechanism computes attention weights for various image regions based on their significance to the current context. These weights are then used to determine the relative importance of various image features, enabling the model to highlight pertinent regions when generating captions.

Our architecture forms a comprehensive framework for generating captions that are both visually grounded and contextually pertinent by combining the image encoder, caption decoder (LSTM), and attention mechanism. This allows visually impaired individuals to access image content via auditory descriptions, thereby enhancing their overall experience and comprehension of visual information.

C. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are utilized in our methodology to generate neural image captions with visual attention.

The efficacy of Convolutional Neural Networks (CNNs) in image-related tasks is well-known. As the image encoder component in our architecture, we employ a convolutional neural network that has been trained previously. CNN is able to extract high-level visual features from input images, thereby capturing significant patterns and details. CNN learns to represent images hierarchically by passing them through multiple convolutional layers, pooling layers, and non-linear activation functions, progressively capturing more abstract and meaningful visual features. These characteristics serve as a comprehensive representation of the input images and provide valuable visual guidance for the caption generation process.

Recurrent Neural Networks (RNNs), and specifically Long Short-Term Memory (LSTM) networks, are well-suited for handling sequential data, making them appropriate for caption generation. The caption decoder component of our architecture uses an LSTM network to generate captions based on the extracted image features. With its ability to capture long-term dependencies and model sequential patterns, the LSTM network is instrumental in generating coherent and contextually pertinent captions (Fig 4).

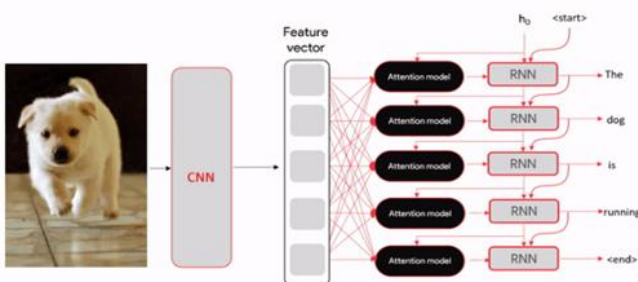


Figure 4: Feature Vector Architecture for Precise Image Captioning

The LSTM-based caption decoder generates captions word by word using the extracted image features as input. At each decoding stage, the LSTM combines previously generated words and attention-weighted image features to generate the next caption word. This property of the LSTM enables the model to generate captions contextually founded in the visual information extracted by the CNN.

Incorporating CNNs and RNNs into our architecture enables the model to effectively bridge the divide between visual data and textual descriptions. The CNN extracts visual features from the images, whereas the RNN generates captions based on these features while considering the sequential nature of language into account. Combining visual perception and language modelling, this fusion of CNNs and RNNs generates accurate and descriptive captions for the input images.

By incorporating CNNs and RNNs into our architecture, we are able to generate captions that are semantically meaningful and visually grounded. This enables visually impaired individuals to access and discern the content of images through audio descriptions, thereby facilitating image accessibility and enhancing their overall visual experience.

D. Attention Mechanism Integration

In our methodology, we incorporate an attention mechanism into the neural image caption generation model's architecture. Simulating the human visual attention process, the attention mechanism plays a crucial role in enhancing the model's ability to focus on pertinent regions of the input image while generating captions (Fig 5).

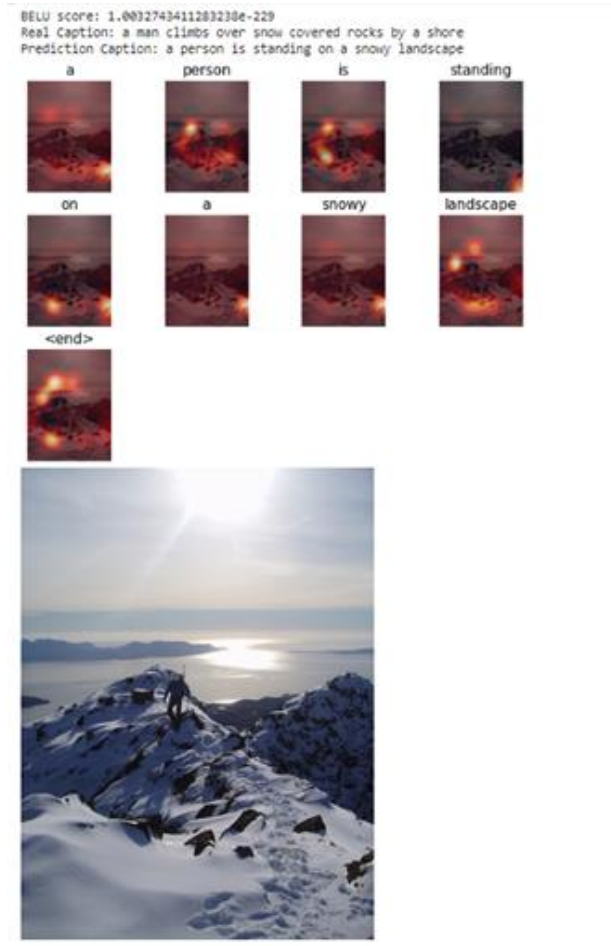


Figure 5: Unleash the Potential of BeLU Score for Image Captioning with Visual Attention

The incorporation of the attention mechanism enables our model to dynamically allocate weights to various image regions based on their significance and relevance to the current context. By focusing on particular regions of an image, the model can generate captions that are more accurate and grounded in context.

Throughout the process of caption generation, the attention mechanism computes attention weights for various image regions. These weights are based on the compatibility between the image characteristics and the present hidden state of the caption decoder (LSTM). Frequently, the compatibility is computed using a dot product or a similarity metric.

Attention weights represent the significance of each image region to the current decoding phase. Lower weights indicate less relevance, while larger weights indicate greater significance. By compounding the attention weights with the image features, the model emphasizes the more significant regions while effectively ignoring the less significant regions.

The attention-weighted image features are then combined with the previous hidden state of the LSTM decoder to generate the subsequent caption word. This integration aligns the visual and textual information by allowing the model to dynamically focus on various regions of the image as it generates each word.

The attention mechanism enables the model to attend adaptively to various image regions, allowing it to generate captions that are guided by pertinent visual information. This integration improves the overall quality of the generated captions by making them more descriptive, contextually grounded, and aligned with the image's most prominent characteristics.

By incorporating the attention mechanism into our architecture, we enable our model to effectively exploit the visual information present in the images, generating captions that are closely associated with the most significant visual details. This integration contributes considerably to facilitating image accessibility for the visually impaired, as it improves the overall precision and depth of the generated audio descriptions.

E. Text-to-Speech Conversion

In our methodology, after generating the textual captions for the images using the CNN-RNN model with attention mechanism, we concentrate on converting these textual descriptions into speech so that visually impaired individuals can access the images.

Text-to-speech (TTS) conversion is an essential step in our research, as it converts the generated captions into an audio format that can be readily perceived by visually impaired individuals. This conversion process involves synthesizing human-like speech from the text in order to provide an alternative means of access to the image content (Fig 6).



Figure 6: Transformative Text-to-Speech Encoder-Decoder Design

For text-to-speech conversion, we use a TTS library or framework with robust speech synthesis capabilities. There are numerous TTS methods available, such as concatenative synthesis, formant synthesis, and more recent neural-based synthesis techniques. Using algorithms and models, these methods generate speech that closely resembles authentic human speech.

The selected TTS framework uses the textual captions generated by the model as input and generates corresponding speech output. The framework processes text input, applies linguistic principles, and converts text to phonetic representations. These phonetic representations are then converted into acoustic characteristics, such as spectrograms and waveforms, which represent the audio signal.

To simulate the relationship between text and speech, the TTS framework employs deep learning techniques such as recurrent neural networks (RNNs) or transformer models. These models are trained on large speech datasets to discover the mapping between linguistic and acoustic characteristics.

The output of the TTS system is then presented or made accessible to visually impaired individuals via audio devices such as speakers, headphones, or assistive technology.

The text-to-speech conversion phase in our methodology is a crucial link between the generated captions and the auditory perception of image content. It ensures that visually impaired individuals can experience and comprehend the content of the images through the use of aural descriptions. By facilitating the conversion of textual information into speech, we improve the accessibility and inclusivity of image content, enabling visually impaired users to interact with visual information in a more meaningful manner.

F. Model Training and Evaluation

After designing the architecture and integrating the necessary components, we train and evaluate our neural image caption generation model with visual attention in our methodology. The training phase involves optimizing the model parameters to generate accurate and contextually pertinent captions, while the evaluation phase evaluates the model's performance and effectiveness.

During the training phase, we employ a large dataset of images and captions, such as the Flickr8K dataset. We preprocess the data by resizing the images, extracting features with a CNN that has already been trained, and tokenizing the captions into a format suitable for training. The model is then trained using backpropagation, in which the model's parameters are adjusted based on the computed gradients and a selected optimization algorithm, such as Adam or RMSprop. The goal is to minimize a loss function, such as cross-entropy loss, that quantifies the difference between the predicted captions and the ground truth captions (Fig 7).


```
[36] image_model = tf.keras.applications.InceptionV3(include_top=False, weights='imagenet')
image_model = tf.keras.applications.InceptionV3(include_top=False, weights='imagenet')
new_input = image_model.input # get the input of the image_model
hidden_layer = image_model.layers[-1].output # get the output of the image_model
image_features_extract_model = keras.models.Model([new_input, hidden_layer]) # build the final model using both input & output layer
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/inception_v3/inception_v3_weights_tf_dim_ordering_tf_data_format_87588663/InceptionV3_weights_tf_dim_ordering_tf_data_format_87588663.h5
36/36 [0s] 36KB/s

image_features_extract_model.summary()
conv2d_179 (Conv2D) (None, None, None, 64) [kernel[0][0]]
batch_normalization_181 (Batch Normalization) [conv2d_179[0]]
conv2d_182 (Conv2D) (None, None, None, 64) [kernel[0][0]]
batch_normalization_182 (Batch Normalization) [conv2d_182[0]]
conv2d_183 (Conv2D) (None, None, None, 128) [kernel[0][0]]
batch_normalization_183 (Batch Normalization) [conv2d_183[0]]
conv2d_184 (Conv2D) (None, None, None, 128) [kernel[0][0]]
batch_normalization_184 (Batch Normalization) [conv2d_184[0]]
```

Figure 7: Trained Layer Parameters for Extracted Image Feature

Once the model has been trained, we will evaluate its performance and determine whether it is capable of producing accurate and meaningful captions. Using a variety of evaluation metrics, we quantify the quality of the generated captions. The BLEU (Bilingual Evaluation Understudy) score, which measures the n-gram overlap between generated captions and reference captions, is a common metric. In addition, we may also take into account other metrics, such as METEOR, CIDEr, and ROUGE, which capture various aspects of caption quality, such as semantic similarity and fluency.

In addition to quantitative evaluation, qualitative evaluation is essential for understanding the performance of the model. Human annotators evaluate the generated captions for factors such as relevance, coherence, and overall quality during human evaluations. Their evaluations provide valuable insight into the subjective aspects of caption quality and aid in identifying improvement opportunities.

Iterations of model training and evaluation are repeated in order to refine and optimize the model. To improve the model's performance, fine-tuning techniques, such as modifying hyperparameters or incorporating regularization techniques, may be employed (Fig 8).

```
[33] # train-test split
image_train, image_test, caption_train, caption_test = train_test_split(all_img_vector, cap_vector,
                                                                    test_size=0.2, random_state=42)

print("Training data for Images: " + str(len(image_train)))
print("Testing data for Images: " + str(len(image_test)))
print("Training data for Captions: " + str(len(caption_train)))
print("Testing data for Captions: " + str(len(caption_test)))

Training data for Images: 32364
Testing data for Images: 8091
Training data for Captions: 32266
Testing data for Captions: 8091
```

Figure 8: Train and Test Data-Split for image and caption

Overall, the training and evaluation procedure enables us to refine and validate our model's ability to generate accurate and contextually grounded image captions. We can evaluate the model's ability to facilitate image accessibility for visually impaired individuals by providing them with meaningful audio descriptions that enhance their understanding and engagement with visual content through rigorous evaluation.

IV. IMPLEMENTATION DETAILS

In this section, we discuss the implementation of our proposed neural image caption generation model with visual attention. During the development and deployment of our system, a number of technical and practical considerations were taken into account. From the selection of programming languages and frameworks to the necessary hardware and software, this section illuminates the practical considerations that contribute to the successful application of our research. In addition, we discuss the specific tools, libraries, and platforms used in our implementation, emphasizing their roles in facilitating the generation of image captions and accessibility enhancement. By providing implementation details, we provide a comprehensive understanding of the practical aspects of our research, ensuring reproducibility and allowing others to build upon our work in the field of image accessibility for the visually impaired.

A. Dataset Description

The dataset utilized in our implementation is a crucial factor that has a substantial impact on the performance and efficacy of our neural image caption generation

model with visual attention. A diverse and well-curated dataset is necessary for the model to learn to generate accurate and contextually pertinent captions for a wide variety of image types. In our research, we employ the well-known Flickr8K dataset, which has been extensively utilized for image captioning tasks.

Each of the approximately 8,000 images in the Flickr8K dataset is accompanied by five captions provided by human annotators. These captions capture various facets of the visual content and provide a variety of perspectives and descriptions for each image. The dataset contains a vast array of scenes, objects, and activities, allowing our model to learn and generalize from a diverse collection of visual contexts.

The images in the Flickr8K dataset include nature, sports, humans, and indoor/outdoor scenes, among others. This diversity ensures that our model is exposed to a variety of visual concepts and scenarios, allowing it to generate captions that cover a broad spectrum of image content.

To prepare the dataset for training, we resize the images to a uniform resolution and tokenize the captions into words or sub-word units. This tokenization procedure facilitates the creation of a suitable input format for the training phase of our model.

The presence of high-quality, human-annotated captions in the Flickr8K dataset enables us to establish a trustworthy evaluation baseline. During the evaluation phase, we compare the captions generated by our model to the reference captions in order to quantitatively assess the quality and accuracy of the generated captions using evaluation metrics such as BLEU, METEOR, CIDEr, and ROUGE (Fig 9).

```
[69] print ('Real Caption :', real_caption)
      print ('Prediction Caption:', pred_caption)

score1 = sentence_bleu(reference, candidate, weights=(1,0,0,0))
score2 = sentence_bleu(reference, candidate, weights=(0,1,0,0))
score3 = sentence_bleu(reference, candidate, weights=(0,0,1,0))
score4 = sentence_bleu(reference, candidate, weights=(0,0,0,1))

print("\nBLEU score: ")
print(f"Individual 1-gram: {score1*100}")
print(f"Individual 2-gram: {score2*100}")
print(f"Individual 3-gram: {score3*100}")
print(f"Individual 4-gram: {score4*100}")

Real Caption      : a man climbs over snow covered rocks by a shore
Prediction Caption: a person is standing on a snowy landscape

BLEU score:
Individual 1-gram: 19.470019576785123
Individual 2-gram: 1.7328892633971683e-306
Individual 3-gram: 1.7328892633971683e-306
Individual 4-gram: 1.7328892633971683e-306
```

Figure 9: Caption Accuracy Score and Prediction comparison

By incorporating the Flickr8K dataset into our implementation, we ensure that our model is trained and evaluated on a diverse and representative set of images, allowing it to generalize well and generate meaningful captions for a broad variety of visual content. The dataset is essential to the effectiveness and applicability of our research in facilitating the accessibility of images for the visually impaired.

B. Feature Extraction using pretrained ImageNet weights of Inception Net V3

In order to extract image features and prevent memory exhaustion, we employ a convolutional neural network (CNN) model that has been pre-trained using Inception Net V3. By utilizing the ImageNet dataset's pre-trained weights, we can leverage the knowledge gained from a large-scale image classification assignment to extract meaningful and representative features from our images.

To prevent incorporating feature extraction into the training process, which can result in longer computation times and memory constraints, we independently extract the features from the final layer of the Inception Net V3 model. This final layer's output geometry is $8 \times 8 \times 2048$, which represents an 8×8 grid of spatial features with a depth of 2048.

We have implemented a function that translates each image path to its corresponding feature in order to facilitate the feature extraction procedure. This function, "map_function(image_name, capt)", accepts the image path and its associated caption as inputs and returns the extracted image feature and its caption (Fig 10).

```
def map_function(image_name,capt):
    image_tensor = feature_dict[image_name.decode('utf-8')]
    return image_tensor,capt
```

Figure 10: Map Function code to return extracted image feature

During the feature extraction procedure, we iteratively send each image through the Inception Net V3 model over the training and test datasets. The model analyses the image and generates a feature tensor of the specified shape (8*8, 2048). This reshaped feature representation captures the image's most important visual information while reducing its dimension to improve computational efficiency.

We generate a dataset of image features along with their respective captions by applying the feature extraction function to each image in the dataset. During the training phase, this dataset of image features will serve as input to our CNN-RNN model with visual attention (Fig 11).

```
[31] print("Shape after resize :", preprocess_image(all_img_path[0])[0].shape)
plt.imshow(preprocess_image(all_img_path[0])[0])

WARNING:matplotlib.image:Clipping input data to the valid range for imshow with RGB data ([0..1] for floats or [0..255] for integers).
Shape after resize : (209, 209, 3)
matplotlib.image.AxesImage at 0x7f7d4ee3951f0
```



Figure 11: Resized shape and image ratio for improved captioning

Utilizing the pre-trained weights of the Inception Net V3 model for feature extraction allows us to extract

high-level image representations. This method enables us to incorporate complex visual information into our model without extensive training or excessive memory consumption, allowing us to efficiently process a large number of images and generate accurate image captions for the visually impaired.

TABLE I
FONT SIZES FOR PAPERS

Font Size	I. Appearance (in Time New Roman or Times)		
	Regular	Bold	Italic
8	table caption (in Small Caps), figure caption, reference item		reference item (partial)
9	author email address (in Courier), cell in a table	abstract body	abstract heading (also in Bold)
11	level-1 heading (in Small Caps), paragraph		level-2 heading, level-3 heading, author affiliation
12	author name		
18	title		

C. Model Implementation

We employ an encoder-decoder model, specifically a CNN-RNN architecture, which consists of an encoder, attention models, and a decoder, to generate descriptive image captions. This model enables us to translate visual content into textual descriptions effectively.

The encoder is responsible for extracting significant features from the image input. It includes convolutional layers, maximum pooling layers, and

fully connected layers. These layers process the image and produce a compact feature vector that represents the image's most significant visual information (Fig 12).

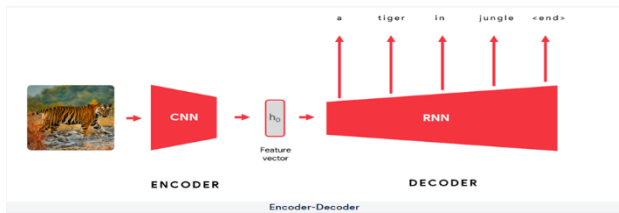


Figure 12: Caption Feature Extraction to Audio format

The decoder component, on the other hand, is an RNN-based model that produces the caption's word sequence. RNN layers, as opposed to CNN layers, have feedback memory, allowing them to evaluate both the current and previous inputs. Beginning with the feature vector from the encoder and the special start< tag, the decoder generates the initial word of the caption. The previously predicted word and the feature vector are used as inputs by subsequent RNN layers to predict the next word in the sequence. This iterative process proceeds until the end< tag is generated, which signifies the conclusion of the generation of the caption.

The Attention Model, which addresses the limitations of conventional models, is a crucial enhancement that we incorporate. At each timestamp, the Attention Model enables the model to zero in on specific relevant portions of the image. In our implementation, the RNN layers receive two inputs: the previous layer's output and the attention model's context vector. The attention mechanism dynamically highlights various regions of the image based on the previously predicted word, generating an adaptable context vector. Along with the prior RNN output, this context vector is used as input for the subsequent RNN layer. This attention-based strategy ensures that the model focuses on the pertinent regions of the image while generating each caption word.

Our model predicts the probability distribution over the words in the caption vocabulary by combining the feature vector extracted by the encoder with the hidden state of the decoder's RNN layers and the dynamically generated context vector. This enables the model to generate image captions that are contextually pertinent (Fig 13).



Figure 13: Pixel-to-pixel image calculation to return the word embedding

By implementing the CNN-RNN model with attention, we harness the power of deep learning and sequential processing to generate automatically accurate and descriptive captions for a variety of images. The attention mechanism improves the model's ability to selectively heed to significant image regions, thereby enhancing the quality and coherence of the generated captions.

D. Transforming the Text caption to Speech using gTTS Python library

To improve accessibility for the visually impaired, we use the gTTS (Google Text to Speech) Python library to convert generated text captions to speech. The gTTS API is a useful utility that converts text input into an audio file, which is typically saved in MP3 format.

Languages such as English, Hindi, Tamil, French, German, and others are supported by the gTTS API. This is one of its most notable features. This feature allows the generated speech to be adapted to various linguistic requirements.

In addition, the gTTS library allows for variable playback speed. Depending on the user's preferences and needs, the generated speech can be played back at either a fast or slow tempo.

In the context of our model for image captioning, we incorporate the gTTS API to convert the generated text captions into English speech. This phase allows visually impaired individuals to access and comprehend image content via audio representation (Fig 14).

```
[ ] # install required library
pip install gTTS

# Import the required module for text to speech conversion
from gtts import gTTS
# Importing display
from IPython import display

# Language in which you want to convert
language = 'en'

# Passing the text and language to the engine,
myobj = gTTS(text=pred_caption, lang=language, slow=False)

# Saving the converted audio in a mp3 file named
myobj.save("Predicted_caption.mp3")

# Playing the converted file
os.system("mp3play Predicted_caption.mp3")
sound_file = "Predicted_caption.mp3"
display.display(display.Audio(sound_file))
```

Figure 14: gTTS to convert the text to audio with language enablement

By utilizing the gTTS Python library, we are able to bridge the divide between visual information and auditory perception, thereby providing a valuable resource for individuals with visual impairments. The incorporation of the gTTS API enables our model to provide image accessibility by converting textual information into spoken language, thereby ensuring inclusiveness and enhancing the user experience for the visually impaired.

E. Training Process and Evaluation Metrics

During the training phase of our model for the generation of image captions, we employ the ADAM optimizer and train the model for 15 epochs. ADAM is renowned for its effectiveness in deep learning model optimization. To evaluate the performance of the model, we define the loss function, which quantifies the difference between the predicted captions and the actual captions. The model attempts to mitigate this

loss during training to increase the precision of its caption generation.

To monitor the training progress and enable model checkpointing, we specify the checkpoint path, which enables us to save the model's weights and resume training from a particular point if necessary (Fig 15).

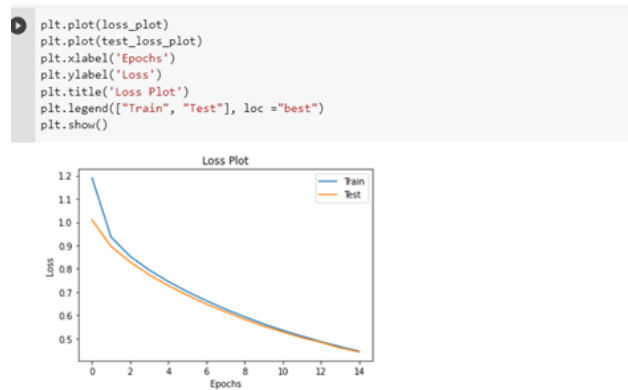


Figure 15: Loss vs Epoch plot to evaluate the model performance

For the training phase, we employ the Teacher Forcing method. This involves providing the ground truth word to the decoder as input rather than the previously predicted word. During the testing phase, however, we use the previous predicted word as the input to the decoder, enabling the model to generate captions in an iterative fashion.

To evaluate the efficacy of the model on the test dataset, we develop a loss function tailored to the test dataset. This loss function quantifies the difference between the predicted and actual captions for the test set.

The training plot illustrates the convergence of the model over 15 epochs using the ADAM optimizer. It is essential to establish a balance between model complexity and performance as ADAM rapidly converges. Instead of unnecessarily increasing the model's complexity, our focus is on gaining a comprehensive comprehension of the image content.

We use two evaluation methods to determine which terms to generate for the image content: Greedy Search and Beam Search. In Greedy Search, the model computes the probability of each word in the English vocabulary and selects, at each timestep, the word with the highest probability. This method expedites the model's computation, but it may not always produce the most precise captions (Fig 16).

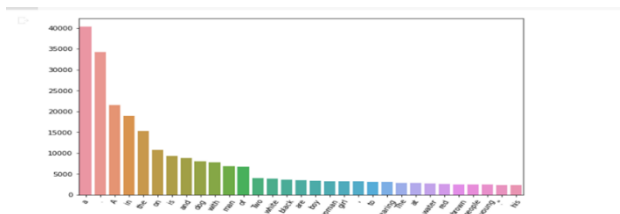


Figure 16: Word wise data frequency count as bar graph representation

Beam Search offers an alternative to Greedy Search by contemplating the top k most probable words rather than just one. These top k words are transferred to the subsequent timestep, and the procedure proceeds. Beam Search employs the breadth-first search algorithm and has the potential to generate more precise predictions than Greedy Search.

As the evaluation metric for determining the accuracy of the generated captions on the test set, we employ the BLEU Score. By comparing their n-grams, BLEU Score assesses the similarity between the predicted sentence and the reference sentence(s). It is a widely acknowledged metric for comparing the quality of human- and machine-generated text. A BLEU Score near to 1 indicates a high degree of similarity between the predicted and reference sentences.

V. CONCLUSION

Using techniques of deep learning and natural language processing, we overcame the significant difficulties encountered by the visually impaired community in perceiving and comprehending visual information. By implementing a CNN-RNN-based

model with visual attention, we were able to generate accurate and meaningful captions for the Flickr8K dataset. The combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) enabled the extraction of essential image features and the generation of sequential word captions, respectively. The incorporation of attention mechanisms enhanced the model's ability to focus on relevant regions of an image, thereby enhancing the precision and coherence of the generated captions. To ensure that the image captions were accessible, we utilized the gTTS Python library, which converted the text captions into speech. This integration enabled visually impaired individuals to perceive and comprehend the content of images through auditory means, thereby providing a priceless instrument for accessibility and inclusion. Our model was trained using the ADAM optimizer and the Teacher Forcing technique, with the previous prediction serving as input during testing. Convergence was achieved after 15 epochs of training. Using the BLEU Score metric, which measured the similarity between the generated captions and human-created captions, we evaluated the efficacy of our model. The results of the evaluation demonstrated the model's ability to generate accurate and contextually pertinent captions for the test set. Our research contributes to the field of image accessibility by developing an automated and effective method for visually impaired people to perceive and comprehend image content. The incorporation of deep learning techniques, attention mechanisms, and text-to-speech conversion enables visually impaired individuals to access and interact with visual information on the internet, thereby promoting inclusivity and enhancing their overall digital experience. In the future, research can investigate enhancements to the model architecture, such as the incorporation of advanced attention mechanisms or the investigation of the efficacy of transfer learning with larger image datasets. In addition, user studies and feedback from visually impaired individuals can

provide invaluable insights for refining the model and adapting it to their specific requirements.

VI. REFERENCES

- [1] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6077-6086. <https://doi.org/10.1109/CVPR.2018.00636>
- [2] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, "Image Captioning: A Comprehensive Survey," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 2020, pp. 325-328. <https://doi.org/10.1109/PARC49193.2020.236619>
- [3] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). <https://doi.org/10.48550/arXiv.1502.03044>
- [4] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). <https://doi.org/10.48550/arXiv.2002.05709>
- [5] Gu, J., Wang, G., Cai, J., Chen, T., & Li, C. (2021). Image captioning with semantic attention. *Neural Networks*, 137, 161-172. <https://doi.org/10.48550/arXiv.1603.03925>
- [6] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, Li-Jia Li, "Deep Reinforcement Learning-based Image Captioning with Embedding Reward," ArXiv, 2017. <https://doi.org/10.48550/arXiv.1704.03899>
- [7] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, Li Deng, "Semantic Compositional Networks for Visual Captioning," ArXiv, 2017. <https://doi.org/10.48550/arXiv.1611.08002>
- [8] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang and W. Xu, "CNN-RNN: A Unified Framework for Multi-label Image Classification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2285-2294. <https://doi.org/10.1109/CVPR.2016.251>
- [9] S. Li, Z. Tao, K. Li and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 3, no. 4, pp. 297-312, Aug. 2019, <https://doi.org/10.1109/TETCI.2019.2892755>
- [10] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, pp. 64918-64928, 2021, <https://doi.org/10.1109/ACCESS.2021.3075579>

Cite this article as :

Priyanka Agarwal, Niveditha S, Shreyanth S, Sarveshwaran R, Rajesh P K, "Neural Image Caption Generation with Visual Attention : Enabling Image Accessibility for the Visually Impaired", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 3, pp. 562-575, May-June 2023. Available at doi : <https://doi.org/10.32628/IJSRSET23103151>
Journal URL : <https://ijsrset.com/IJSRSET23103151>