Descriptive Audio Model of Images for Blind

Atish Chandra
UCFID-5571476
University of Central Florida
Orlando, Florida, USA

Figure 1

ABSTRACT

In the time of accessibility technology, providing visually impaired individuals with the means to perceive and understand visual content through auditory information remains a critical challenge. This project develops a machine learning model capable of transforming images into descriptive captions, which can subsequently be converted into speech, thereby making visual information accessible to those who are blind or have significant visual impairments. The model employs a sophisticated encoder-decoder architecture enhanced by an attention mechanism, utilizing Convolutional Neural Networks for feature extraction and Recurrent Neural Networks for generating textual descriptions. The attention mechanism enables the model to focus on specific parts of the image relevant to the ongoing descriptive process, mimicking human perceptual focus. Additionally, the implementation of both greedy and beam search strategies in the caption generation process ensures a balance between computational efficiency and contextual accuracy. This paper details the methodology, including the preprocessing steps for handling captions such as tokenization, vocabulary restriction, and sequence padding and discusses the transformative potential of the system in facilitating an inclusive experience for visually impaired users in digital media interaction.

ACM Reference Format:
Atish Chandra. 2024. Descriptive Audio Model of Images for Blind. In . ACM, Orlando, FL, USA, 6 pages.

1 INTRODUCTION

Visual perception is fundamental to how we interact with our environment, influencing everything from navigation to the absorption of information from various media. For individuals with visual impairments, the inability to engage with visual content can be a significant barrier, limiting access to information and reducing overall quality of life. Advances in artificial intelligence and machine learning offer promising solutions to bridge this gap by enabling the conversion of visual data into audible formats that are accessible to the visually impaired.

ACM Conference, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

This project, "Descriptive Audio Model of Images for Blind," aims to harness the power of AI to create a tool that can provide descriptive audio translations of visual content. By developing a machine learning model that can accurately describe images through captions and convert these captions into speech, this initiative seeks to enhance the autonomy and experience of visually impaired individuals, allowing them to "see" through audio descriptions.

The core of the model is built on an encoder and decoder architecture with an attention mechanism. The encoder uses a Convolutional Neural Network to extract features from images, which captures the intricate details and contextual elements of the visual data. The decoder, often implemented as a Recurrent Neural Network , processes these features to generate natural language descriptions of the images. The attention mechanism improves the model's performance by focusing on specific parts of the image during the captioning process, similar to how human attention is selectively directed to salient parts of a visual scene.

Moreover, the model incorporates advanced caption generation techniques such as greedy search and beam search to optimize the relevance and accuracy of the generated descriptions. These methods are crucial for ensuring that the generated captions are both contextually appropriate and coherent with language.

In this report, we elaborate on the methodologies employed in the creation of the "Descriptive Audio Model of Images for Blind" model, including detailed discussions on the preprocessing of captions, feature extraction, and the mechanics of the encoder and decoder framework. Here, we will also explore the practical implications of this technology, highlighting its potential to make digital media, educational content, and everyday visual information more accessible to those with visual impairments. Through this work, we contribute to the ongoing efforts in accessibility technology, aiming to create more inclusive digital environments for all users.

2 RELATED WORK

Speaking personally, it's great to realize that technology is expanding to those sections of people who are really 'kind of in need'. The field of image captioning and its application to assistive technologies for the visually impaired has seen significant research interest, with numerous studies and projects contributing to its development. This section reviews relevant literature and projects that are performed or performing in the same field. Few that are worth to include in this report are:-

(i) MICROSOFT has built an AI tool called as 'PeopleLens', so that visually impaired people can also witness the pictures. In detail, the PeopleLens is a versatile open-AI system that provides essential resources for individuals who are blind or have low vision, enhancing their ability to understand and interact with the social world around them. Recently, it has been particularly useful in educational settings, where it aids blind children in initiating and managing interactions with their peers. This system represents a significant step forward in human-AI interaction, shifting from simple task-based assistance to offering a continuous flow of interactive information. Inspired by ethnographic studies involving Paralympic athletes and spectators, PeopleLens aims to magnify environmental details, thereby helping users to build upon their existing navigational and social skills. It uses a head-mounted augmented reality device in combination with four state-of-the-art computer vision algorithms to *continuously* locate, identify, track, and capture the gaze directions of people in the vicinity. It then presents this information to the wearer through spatialized audio—sound that comes from the direction of the person. The real-time nature of the system gives a sense of immersion in the People Map.

(ii) NVIDIA have also a dedicated page on how their developer has created tool for the blind people. They have named it 'A-Eye for the blind'. In detail, this system is designed to enhance the safety of visually-impaired users while traveling. It involves a hardware setup that includes a Jetson Nano 2GB board, a Raspberry Pi Camera V2, a wifi dongle, a fan, a power bank, and wired headphones. The system captures real-time images of the surroundings, which are processed by a machine learning model that converts 2D images into depth images. This analysis allows the system to provide auditory feedback about obstacles on the user's left, right, or directly ahead. Additionally, the images along with timestamps are securely uploaded to a Firebase database. This allows friends and family to access live images via a dedicated website, helping them monitor the user's safety. This setup not only aids in navigation but also ensures connectivity and support through live monitoring.

Along with these noticeable works of big MNCs, there are lot of ongoing research that can be referred to as the current research and analysis on the same type of work, where with the help of AL and ML applications are being constructed. A few to be mentioned here are:-

(i) Artificial intelligence for visually impaired by Jiaji Wang, Shuihua Wang, Yudong Zhang

(ii) Navigating Eye to Blind People using Machine Learning by Tejaswini B1 , Aishwarya S2 , Anushree R3 , Harshitha R4 , Inchara B R5

(iii) Smart Glass System Using Deep Learning for the Blind and Visually Impaired by Mukhriddin Mukhiddinov and Jinsoo Cho

(iv) Deep learning based object detection and surrounding environment description for visually impaired people by Raihan Bin Islam, Samiha Akhter, Faria Iqbal, Md. Saif Ur Rahman, and Riasat Khan

(v) Bionic Eyes for Visually Impaired Using Deep Learning by Dahlia Sam1 , Jayanthi K2 , I. Jeena Jacob3 , N. Kanya4 , Shekaina Justin

3 APPROACH

Mentioning the overall approach, this project is all about images and their respective captions mapping, modeling, training and then providing audio text to the generated captions.

3.0.1 *Data Loading and Preprocessing*.

(1) Dataset Import and Reading: Initially, the dataset containing images and their corresponding captions is imported into the analysis environment. The data is typically split into two separate variables: one holding the image files (or paths) and the other holding the captions. This separation facilitates specific preprocessing steps that need to be applied independently to the textual and visual data.
(2) Data Visualization: To gain insights into the dataset and verify the integrity of the data, both images and their associated captions are visualized. This step is critical for understanding

the relationship between visual content and textual descriptions, which assists in hypothesizing about the potential challenges in modeling, such as the variety in vocabulary and the complexity of images.

(3) Vocabulary Mapping: To prepare the captions for neural network processing, word-to-index and index-to-word mappings are created. This involves generating a dictionary where each unique word is assigned a specific numerical index. These mappings are essential for converting text data into a numerical format that can be fed into machine learning models, and for translating model outputs (numerical indices) back into human-readable words.

(4) Dataframe Creation: A dataframe is constructed to summarize and structure the dataset, including columns for image paths and captions. This organized format supports more efficient data manipulation and batching during model training.

(5) Word Frequency Visualization: Analyzing the frequency of words in the captions provides insights into the most common vocabulary. Visualizing the top 30 occurring words helps in understanding the dataset's linguistic characteristics, which can guide decisions on data preprocessing steps like removing infrequent words or adding more samples for underrepresented vocabulary.

3.0.2 ***Model Components***.

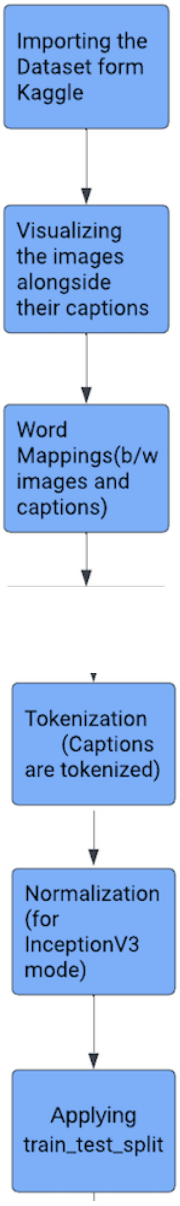
(1) Encoder: The encoder's role is to process the image features extracted by a pre-trained model (like InceptionV3). It consists of a dense layer followed by a dropout layer to reduce overfitting. The dense layer transforms the flattened image features into an embedded space. Activation functions such as ReLU are applied to introduce nonlinearity, enhancing the model's learning capability.

(2) Attention Model: The attention mechanism is crucial in focusing on different parts of the image during different stages of the caption generation process. It uses a set of weights to learn which areas of the image are most relevant to predicting the next word in the caption. This model computes a context vector by applying a softmax function to the scores derived from the combination of image features and the previous hidden state of the decoder. This context vector essentially represents the focused area of the image for a particular timestep in the caption generation.

(3) Decoder: The decoder uses the context vector provided by the attention model along with the previous word's embedding to generate the next word in the caption. It consists of an embedding layer for the captions, a GRU (Gated Recurrent Unit) layer for maintaining the sequence context, and two dense layers where the first dense layer helps in transforming the GRU output to a suitable shape, and the second dense layer or say output layer predicts the next word in the vocabulary. The decoder is where the sequential data processing happens, and it iteratively predicts each word of the caption based on the previous word and the context vector until a stop signal, like an end of sequence token is generated.

3.0.3 ***Model Training***. Then the training process happens which includes setting hyperparameters such as batch size, embedding dimensions, and the number of units in LSTM layers. Training involves feeding the images and their corresponding captions into the model and optimizing a loss function to improve the caption predictions.

3.0.4 ***Converting text to Audio***. Then by transforming the generated captions into spoken words is done by using the Google Text-to-Speech (gTTS) library. This conversion of text to audio is particularly valuable in applications aimed at assisting visually impaired users, as it provides an auditory representation of the visual content, thereby making digital content more accessible.



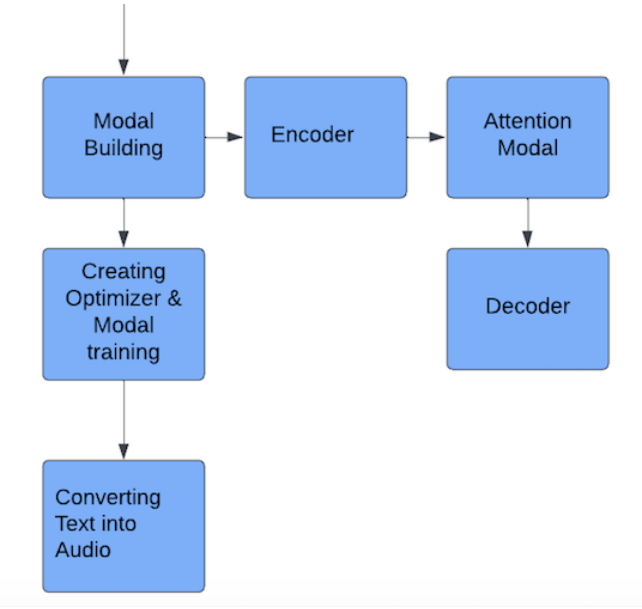


Figure 2: Dataflow explaining the modal

4 EXPERIMENT

Under this section, elaboration of the techniques and process will be mentioned, like how the whole modal comes into play. So, mentioning section wise:-

Data Acquisition: The data is sourced from Kaggle, namely the "eye-for-blind" dataset which contains images and corresponding captions. The dataset is directly downloaded to a cloud environment rather than being stored locally. The source link has been attached under references section.

Data Preparation: Images are stored in a zip file and then extracted to a specific directory.

The captions associated with images are processed, where each caption is paired with its corresponding image. Libraries and Environment: Extensive use of TensorFlow and Keras for model building.

Other Python libraries used include Pandas for data manipulation, Matplotlib for visualization, and NLTK for text processing.

Preprocessing: The captions are tokenized using Keras's Tokenizer class.They are tokenized by splitting them based on spaces and various other characters specified by filters. This step helps in breaking down sentences into individual words or tokens, forming the basic units for model input.

In the preprocessing stage for images in the machine learning project, two critical steps are undertaken to prepare the images for effective processing by the InceptionV3 model and those are 'Resizing' and 'Normalization'. By resizing and normalizing the images, the preprocessing ensures that the images are suitably formatted and standardized for optimal processing by the neural network, which is designed to work with specific input sizes and value ranges.

Model Architecture: Utilizes the InceptionV3 model pretrained on ImageNet for feature extraction from images. The output from the last convolutional layer of InceptionV3 is used, which provides a rich feature set for each image. A custom model is built on top of these features, likely for the task of generating captions. This model includes LSTM layers to handle the sequential nature of language and may also include dense layers for prediction.

Training and Testing: The dataset is split into training and testing sets to evaluate the performance of the model. This is crucial for understanding how well the model generalizes to new, unseen data.

Performance Metrics: Although specific metrics aren't detailed in your summary, typical metrics for such tasks include BLEU scores for evaluating the quality of generated captions relative to the true captions.

Evalution the data with Greedy Search: As we know the algorithm of Greedy algorithm, it fits for this project. So, code snippet is included in the project where it will evaluate the outcomes of the project where it will check that image is coming with the respective captions. Visualization and Debugging: Images and captions are visualized to ensure data integrity and to provide insight into the dataset's characteristics.

Comparison of this model with other baseline models, whose references are attached in the 'References' section.

4.0.1 *Baseline and Objectives:*

- Previous or say the other baseline projects are operates primarily as a digital accessibility tool. It focuses on making static visual content, like images on websites accessible by generating text based captions. This aligns with enhancing digital inclusion, allowing visually impaired users to understand and interact with content that would otherwise be inaccessible.
- This project are designed as a real time assistive device. It translates the visual world into auditory descriptions, helping visually impaired users navigate and interact with their immediate environment. This approach addresses a broader spectrum of daily challenges, including mobility and social interaction, by providing continuous audio feedback about one's surroundings.

4.0.2 *Technological Approaches:*

- This project utilizes an established dataset to train a deep learning model on a predefined task, like caption generation. The use of InceptionV3, a sophisticated image recognition model, indicates a reliance on strong, feature-rich deep learning algorithms to interpret images and generate relevant captions. The neural network's training on image-caption pairs emphasizes the model's capacity to link visual elements with descriptive text.
- The other baseline projects integrates more complex systems involving not only software but also hardware components like Raspberry Pi, cameras, and a real-time operating system. The use of convolutional neural networks extends beyond basic object recognition to include dynamic interaction features such as emotion detection, text recognition, and even celebrity recognition using Microsoft's Vision API. This multifaceted approach is geared towards a comprehensive interpretation of visual cues, which are then converted into audio output through devices like Alexa.

4.0.3 *Outcomes and Functional Impact:*

- This projects aims to produce a scalable solution that can be implemented across various platforms where images need captions, potentially benefiting a wide range of applications from educational resources to social media. The outcome is primarily in text form, which could be used directly on websites or read aloud by standard text-to-speech software.
- The other baseline projects, in contrast, delivers a direct, interactive experience tailored to the needs of the visually impaired in real-time scenarios. The system not only identifies objects but also interprets complex scenes and conveys this information through speech, enabling users to "hear" their environment. This immediate feedback is crucial for tasks requiring real-time decision-making, such as navigating busy streets or avoiding obstacles.

4.0.4 *Methods and Implementation:*

- This project is more software centric, with its success heavily reliant on the accuracy of the machine learning models and the quality of the training data. It's a more controlled application where improvements can be directly linked to enhancements in model training and data preprocessing.
- The other baseline projects necessitates a holistic approach, combining software efficiency with hardware reliability. It involves real-world testing and optimization to ensure the system remains robust and responsive in various environmental conditions. Additionally, the integration of third party APIs for specific functionalities introduces a layer of complexity in maintaining and updating the system to keep it functional and relevant.

5 CONCLUSION AND FUTURE WORK

This project successfully implemented a machine learning model that uses the InceptionV3 architecture for feature extraction from images, which is integrated with an LSTM network to generate descriptive captions. This approach leveraged a comprehensive dataset from Kaggle, consisting of various images paired with corresponding textual descriptions. The application of this model has shown promise in automating the generation of accurate and contextually relevant captions, which can significantly enhance accessibility for the visually impaired by providing textual interpretations of visual content.

The methodology involved preprocessing both the images means resizing and normalizing to fit InceptionV3 input requirements and captions, doing tokenization and padding for it. The model's training was executed on a substantial dataset, ensuring the learning phase was robust, encompassing a wide variety of image contexts and descriptions. Evaluations based on test data demonstrated that the model could generalize well to new images, showcasing the effectiveness of the convolutional neural networks combined with LSTM in handling such complex tasks.

Future Work. While the current model performs satisfactorily, there is always room for improvement and expansion. Future directions could include:

(1) Improving Model Accuracy:
   - Enhanced Preprocessing: Further refinement of image and caption preprocessing could improve model performance. For example, advanced techniques in natural language processing could be used to handle nuances in language more effectively.
   - Hyperparameter Tuning: Experimenting with different architectures and tuning the hyperparameters of the LSTM and CNN could yield better results in terms of accuracy and efficiency.
(2) Dataset Expansion:
   - Diverse Data Sources: Incorporating more diverse datasets could help the model learn a wider array of objects and scenarios, thus enhancing its ability to generate relevant captions across more contexts.
   - Multilingual Support: Expanding the dataset to include captions in multiple languages could pave the way for a more universally applicable model.
(3) Real-time Application:
   - Integration into Real-time Systems: Adapting the model for real-time captioning of live images or videos could significantly broaden its applicability, making it a valuable tool for real-time accessibility aids for the visually impaired.
   - Mobile and Web Integration: Developing mobile apps or web services that utilize the captioning model could make it accessible to a broader audience, providing on-the-go captioning services.
(4) Interdisciplinary Applications:
   - Educational Tools: The technology could be adapted for educational content, where it can provide descriptive captions for educational videos or images in textbooks.
   - Assistive Technologies in Other Domains: Extending the model to assist in other domains such as navigation, where it could describe environments or provide contextual information about nearby locations or objects.
(5) Ethical and Societal Implications:
   - Bias Mitigation: Further research should be directed towards identifying and mitigating any biases in the model, ensuring that the generated captions are fair and equitable across different demographics and contexts.
   - User Privacy and Data Security: Implementing robust measures to protect user data, especially when dealing with potentially sensitive visual information.

6 AUTHORS

author - Atish Chandra
email - at598954@ucf.edu

7 REFERENCES

(1) Kaggle Dataset: Ritesh Patil. (Year). *Eye for Blind Dataset.* Available at Kaggle: https://www.kaggle.com/riteshpatil8998/eye-for-blind-dataset (Accessed: date).
(2) TensorFlow and Keras:
   - Martín Abadi, et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. https://www.tensorflow.org/
   - François Chollet, et al. (2015). *Keras.* https://keras.io

(3) InceptionV3 Model: Christian Szegedy, et al. (2016). *Rethinking the Inception Architecture for Computer Vision*. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
(4) Natural Language Toolkit (NLTK): Steven Bird, Edward Loper, and Ewan Klein (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
(5) Research on Deep Learning for Caption Generation: Oriol Vinyals, et al. (2015). *Show and tell: A neural image caption generator*. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
(6) Jiaji Wang, Shuihua Wang, Yudong Zhang, 'Artificial intelligence for visually impaired', 2023, https://www.sciencedirect.com/science/article/pii/S0141938221000246
(7) Tejaswini B , Aishwarya S , Anushree R , Harshitha R , Inchara B R5, 'Navigating Eye to Blind People using Machine Learning', 2023, https://ijarsct.co.in/Paper10645.pdf
(8) Mukhriddin Mukhiddinov and Jinsoo Cho,'Smart Glass System Using Deep Learning for the Blind and Visually Impaired', 2021, https://www.mdpi.com/2079-9292/10/22/2756
(9) Raihan Bin Islam, Samiha Akhter, Faria Iqbal, Md. Saif Ur Rahman, and Riasat Khan, 'Deep learning based object detection and surrounding environment description for visually impaired people', 2023, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10360
(10) Dahlia Sam , Jayanthi K , I. Jeena Jacob , N. Kanya , Shekaina Justin, 'Bionic Eyes for Visually Impaired Using Deep Learning', 2023, https://eudl.eu/pdf/10.4108/eai.16-5-2020.2304024