Caption Generation on Images Taken by Visually Impaired People

FAHRI KUMAR
Department of Computer Science, Sri Sairam Engineering College, Tamil Nadu, India

Abstract- Image captioning on images taken by visually impaired people is a area of study that focus to develop computer systems that are able to generate descriptive text for pictures captured by visually impaired individuals. This field of study focuses on image captioning on images taken by visually impaired people. The purpose of this technology is to provide guidance to people who are visually impaired and may require assistance in comprehending the visual content of a picture. Image captioning on photographs captured by visually impaired individuals calls for the application of sophisticated computer natural language processing techniques. These methods involve conducting an analysis of the visual aspects of a picture, such as the objects, colors, and textures that are present, and then using this information to generate a description in natural language that communicates the pertinent information to a user who is visually impaired. A significant move towards increasing accessibility and inclusiveness for people with disabilities is the development of image captioning systems for the images taken by visually impaired people. This is an essential stage in the process. Visually impaired people may benefit from these systems by gaining a greater understanding of their surroundings, being able to navigate their environment more effectively, and having more interaction with the visual world.

Indexed Terms- Caption generation, Image classification, text generation.

I. INTRODUCTION

The field of Computer Vision and Natural Language Processing Image Captioning is a fascinating and rapidly expanding one. It attempts to automatically generate natural language descriptions of an image's content. Image captioning has numerous applications, including assisting the visually impaired to comprehend images. Visually impaired individuals face obstacles when it comes to accessing visual information, which is crucial for daily life. Image captioning can assist in overcoming this obstacle by supplying descriptions of images that visually impaired individuals would not otherwise be able to access. In recent years, there has been a developing interest in image captioning for visually impaired individuals' photographs. These images may contain various forms of visual information, necessitating the use of specialized captioning techniques. The use of these images presents researchers in the field of image captioning with unique challenges and opportunities. The quality of the images is one of the primary obstacles in producing image captions from images captured by visually impaired individuals. These images may be of poor quality, contain noise or blur, or were captured from an unusual angle, making it difficult to derive pertinent information. In addition, visually impaired individuals may not have the same comprehension of visual concepts and vocabulary as sighted individuals, which can further confound the process of image captioning. To overcome these obstacles, researchers have devised a variety of image-captioning techniques. Combining computer vision and natural language processing techniques to analyse the image's content and generate a descriptive caption is one approach. Using crowdsourcing to collect descriptions of images taken by visually impaired individuals from a large group of people is another approach. Unique perspective is one of the benefits of using images captured by visually impaired individuals for image captioning. These images may contain information that sighted individuals would neglect, such as tactile or auditory cues. Additionally, images captured by visually impaired individuals may shed light on the challenges visually impaired individuals face daily.

In conclusion, image captioning on images captured by visually impaired individuals is a significant and expanding field of study. It has the potential to enhance the lives of visually impaired individuals by giving them access to visual information they would not otherwise have. There are challenges associated with this task, such as the quality of the images and visually impaired people's unique understanding of visual concepts and vocabulary, but there are also opportunities for researchers to develop new techniques and gain unique insights into the daily challenges visually impaired people face.

II. METHODOLOGY

LSTM, which stands for "Long Short-Term Memory," is a form of architecture for recurrent neural networks (RNNs) that was developed to solve the disappearing and exploding gradient difficulties that can occur while training regular RNNs on long sequences of data. This architecture was built to handle these problems. LSTMs have found widespread application in a variety of natural language processing and sequence-to-sequence tasks, such as speech recognition, machine translation, and text production. They have the ability to selectively store or forget information over extended periods of time and have been employed for this purpose for quite some time.

The capacity of LSTMs to maintain a "cell state" that can selectively convey information across time steps is the most important aspect of these devices. This enables the network to remember or forget information selectively depending on its requirements. An input vector is provided to the LSTM at each time step, in addition to the previously stored cell state and the hidden state. After that, it devises a set of "gates" to regulate the flow of information by computing them. These gates include the following:

1. The Forget gate is responsible for deciding which pieces of information should be removed from the cell's state.
2. The input gate is responsible for deciding which new information should be added to the current state of the cell.
3. The output gate is what decides which information should be output based on the current condition of the cell.

A sigmoid activation function is used in each gate's implementation. This function produces values between 0 and 1 that indicate the degree to which each gate should be open or closed. In addition, there is something called a "memory cell" in the network, which is responsible for storing information for a longer period of time, as well as a collection of "cell state activations" that are changed at each time step.

Backpropagation through time (BPTT) is the method that is used to train LSTMs. A number of gradient-based optimisation algorithms, including as Adam can be used to optimize an LSTM after it has been trained. They have been found to be effective at capturing long-term dependencies in sequential data, which is one of the reasons why they are a popular choice for a number of sequence modeling applications.

The fundamental concept behind Show and Tell is to make use of a CNN to extract picture features and an RNN to generate captions based on those features. These two neural networks work in tandem to accomplish this task. The input picture is first encoded using a CNN into a feature vector of a fixed length, and then the RNN is used to decode the feature vector into a caption using the feature vector (Karita, S.,).

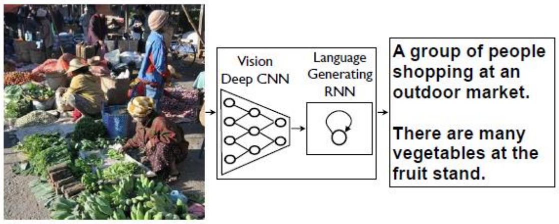
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Figure 1: Working of Show and Tell

The encoder and the decoder are the two primary components that make up the algorithm. The encoder is a convolutional neural network (CNN) that, given an input picture, generates a feature vector that represents the visual content of the image. An RNN known as the decoder is responsible for taking the feature vector that was produced by the encoder and producing a string of words that accurately characterize the image.

The encoder CNN is made up of multiple layers of convolutional and max-pooling operations, which together reduce the dimensionality of the feature map that is produced from the input picture. The last layer of the CNN is a completely connected layer that converts the feature map to a feature vector with a fixed length. The encoder will produce a feature vector as its output. This feature vector will capture the visual content of the picture that was input.

In this project, Resnet-152 architecture was used as an encoder.
The feature vector that was generated by the encoder is used as input by the RNN that acts as the decoder, which then produces a string of words that characterize the image. The recurrent neural network (RNN) is made up of a string of LSTM (long short-term memory) cells that are responsible for producing the word string. The LSTM cells are interconnected in a sequential fashion, with each cell receiving the output of the cell that came before it as an input. The feature vector that was generated by the encoder is used as an input by the very first LSTM cell.

The decoder RNN creates the caption one word at a time, with each word being generated based on the words that came before it as well as the picture that was fed into it. The recurrent neural network (RNN) is responsible for producing a probability distribution over the dictionary of potential words at each time step, after which it chooses the word that is most likely to occur as the output. The probability distribution is determined by computing it based on the output of the LSTM cell that came before it as well as the feature vector that was produced by the encoder.

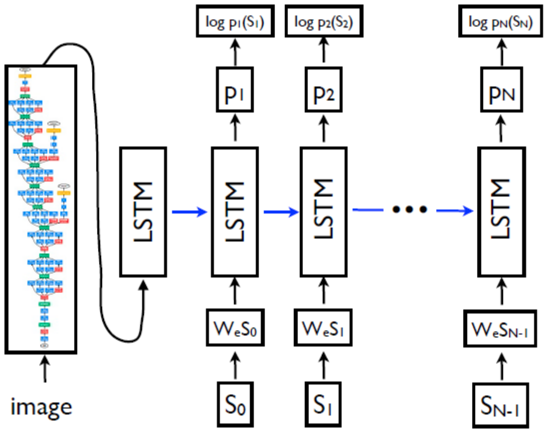

Figure 2: Show and Tell Architecture

During training, the algorithm is taught using a sizable dataset that contains both pictures and the captions that correspond to them. The goal is to reduce, as much as possible, the amount of cross-entropy loss. Backpropagation is used to train the model, with the gradients being transmitted from the output of the decoder RNN to the input of the encoder CNN. Backpropagation is used to train the model.

During testing, the algorithm creates a description for a specific input image by first running the image through an encoder CNN to produce the feature vector. This allows the algorithm to generate a caption for the image. After that, the feature vector is input into the decoder RNN, which causes it to produce the string of words that serve as a description of the picture. The procedure is repeated until either the RNN produces a token signifying the end of the sentence or a maximum length that was previously specified is attained.

Show and Tell is an efficient picture captioning algorithm that makes use of both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to generate captions for images. It has opened up new possibilities for applications such as image retrieval, visual question answering, and many more. It has attained a level of performance that is considered to be state-of-the-art on several benchmark datasets.

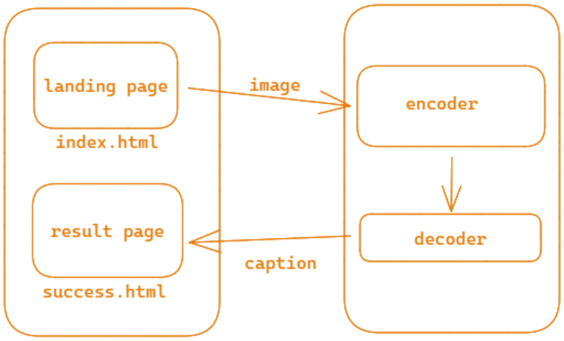

Figure 3: Architecture of entire Project

After the model has been trained using the dataset, the following steps were performed in order to host it on a local server utilizing the FLASK API:
1. Create a Flask application: The first step in this process is to construct a Flask application. Flask is a web application framework that was developed in Python. It simplifies the process of developing and deploying web applications. Using the Flask package found in Python, it is possible to build a Flask application.

2. Create an API endpoint: Following the development of a Flask application, it is necessary to establish an API endpoint. An application programming interface endpoint (API endpoint) is a URL that clients can use to send messages to a server. In this instance, the API endpoint will be used to send a picture to the server in order to receive a caption as a response from the system.
3. Define the image captioning function: The image captioning function is the most important part of the application programming interface (API). It requests a picture from the user and then outputs a caption in response to that image. Importing the trained model and creating a function that accepts an image as input, preprocesses the image, runs it through the model, and returns the generated caption are both required steps in the process of defining the image captioning function.
4. Define the API route: After the image captioning function has been defined, the next step is to specify the API route. Clients will navigate to the specified URL using the API route in order to upload a picture to the server. The app.route decorator of the Flask framework can be used to specify the API route.
5. Handle the image upload: The server is responsible for taking care of the image upload whenever a client sends an image to be processed by the API route. Handling the submission of images can be done with the help of the Flask request object.
6. Return the generated caption: Give the client the description that was automatically generated after the image has been uploaded and processed by the image captioning function. This step must be completed after the image has been given to the client. Using the jsonify function provided by Flask, it is possible to return the generated description in the form of a JSON object.
7. Run the Flask application: after it has been created and the API endpoint has been defined The Flask application needs to be started after it has been created and the API endpoint has been specified. The Flask application can be executed by utilizing the run function of the Flask package.

A Flask API can be incorporated with an image captioning algorithm and hosted on a local server if the aforementioned steps are followed in order. Because of this, implementing and testing the picture captioning algorithm on one's own machine is a simple process.

III. TRAINING

Encoder Architecture:

The ResNet152 model can serve as the foundational neural network architecture in this scenario, having undergone pre-training on the ImageNet dataset. It is possible to substitute the ultimate fully connected layer of ResNet152 with an attention mechanism and a language model. This approach can be utilised to train the model to produce captions.

RESNET152:

The ResNet152 architecture is a convolutional neural network that has been designed by Microsoft Research for the purpose of image classification and other related computer vision tasks. The architectural design was first presented in 2015 as an extension of the ResNet lineage of models, denoting "Residual Network".

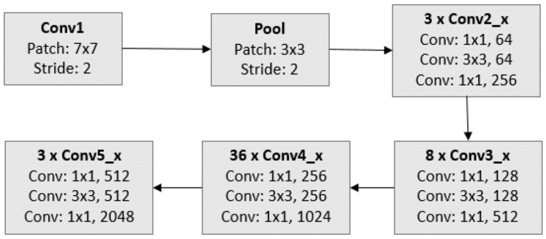

Figure 4: Resnet152 architecture

The ResNet152 model is composed of 152 layers, which renders it deeper than its predecessors, namely ResNet50 and ResNet101. The enhanced depth of the model enables it to apprehend intricate features in images, thereby potentially enhancing its efficacy in tackling demanding computer vision assignments.

The ResNet architecture's primary innovation lies in its utilization of residual connections, which are commonly referred to as skip connections. Residual connections facilitate the direct transmission of information from one layer to another, circumventing multiple intermediary layers. The aforementioned technique serves as a preventive measure against the issue of vanishing gradients, which may arise when gradients diminish in magnitude to an extent that they are unable to propagate through multiple layers of a deep neural network.

The ResNet152 architecture employs bottleneck blocks, which incorporate residual connections. These blocks comprise three convolutional layers: a 1x1 convolution layer for input dimensionality reduction, a 3x3 convolution layer for feature extraction, and a 1x1 convolution layer for output dimensionality expansion. The residual connection facilitates the bypassing of the 3x3 convolution layer, thereby enabling the input to be directly added to the output of the bottleneck block.

IV. RESULTS

To check the efficiency of our trained model on the backend, we design a jupyter notebook. This notebook takes a random image from a sample folder and generates captions relevant to the image picked. This process helps us infer our trained model.

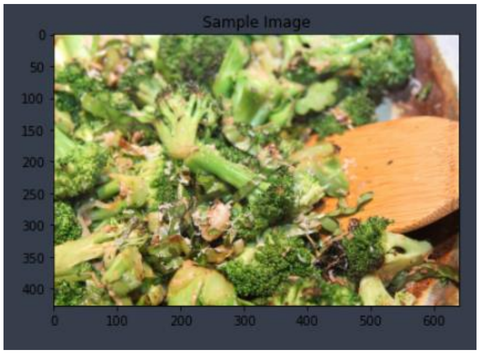

Figure 5: Inference to check the model on the back end

PREDICTED CAPTION: Bowl of Broccoli and carrots on a table

The process of inference for a model designed for image captioning during a show and tell activity entails the input of an image into the model that has undergone training, followed by the generation of a corresponding caption. After the generation of the caption, it is possible to transmit it to a text-to-speech (TTS) mechanism for the purpose of transforming it into an auditory form.

The playback of the audio file can be executed through diverse techniques, contingent upon the specific demands of the project. The audio output of the system can be transmitted through either local speakers or headphones connected to the host device, or alternatively, it can be remotely streamed over the internet to a separate device.

In summary, the integration of show and tell image captioning alongside a text-to-audio feature has the potential to offer individuals with visual impairments a robust mechanism for obtaining visual information. The conversion of image captions into audio format has the potential to enhance the accessibility and inclusivity of visual content, while also offering a more immersive experience for all users.

CONCLUSION

In conclusion, a show-and-tell picture captioning project that includes a text-to-audio feature for the benefit of those who are visually impaired can be a useful tool for promoting accessibility and inclusion if it is implemented appropriately. This is because the text-to-audio option allows individuals who are visually impaired to hear what is being shown on the screen. Even for people who are unable to view the photos themselves, it is possible to write captions for photographs that are accurate and insightful by making use of a deep learning model such as ResNet152. This is the case even for those people who are unable to view the images directly. Because of this function, individuals who are completely blind or who have eyesight that is significantly impaired will be able to read the captions. It is now possible to transform these captions into an audio format as a result of the implementation of a text-to-audio capability. As a direct consequence of this, people with visual impairments such as blindness or low vision can now view captions that were previously out of their reach.
This initiative has the potential to be useful in a wide number of disciplines, including, to name just a few of those fields: education, the entertainment business, and social media. It is feasible to give a more exciting experience for all users, regardless of how well they see, by making it easier to access visual content. This is possible by making it more accessible. As a consequence of this, it will be possible to deliver an

experience that is more immersive.

It is possible that in the future, research will be conducted on topics such as the accuracy and efficiency of the model for photo captioning, as well as the development of text-to-speech systems that are more complex. It is possible for the dataset to be expanded to include a greater variety of image kinds and captions, and it is also possible for this technology to be connected with other forms of assistive technology in order to construct an accessibility solution that is more all-encompassing. In addition to that, there is room in the dataset for further items to be added.

If the initiative also includes a text-to-audio capability, it is conceivable for a show-and-tell image captioning effort to have a significant positive impact on the lives of visually impaired individuals and to contribute to the construction of a society that is more accessible and inclusive. This will be the case if the initiative has both features.

REFERENCES

[1] Lu, D. and Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), pp.823-870.

[2] Manakkadu, S., Joshi, S.P., Halverson, T. and Dutta, S., 2021, December. Top-k User-Based Collaborative Recommendation System Using MapReduce. *In 2021 IEEE International Conference on Big Data (Big Data)* (pp. 4021-4025). IEEE.

[3] Weng, Q., Lu, D. and Schubring, J., 2004. Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. *Remote sensing of Environment*, *89*(4), pp.467-483.

[4] Manakkadu, S. and Dutta, S., 2024. Ant Colony Optimization based Support Vector Machine for Improved Classification of Unbalanced Datasets. *Procedia Computer Science*, *237*, pp.586-593.

[5] Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S. and Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote sensing of environment*, *115*(5), pp.1145-1161.

[6] Manakkadu, S. and Dutta, S., 2024. Efficient Feature Clustering for High-Dimensional Datasets: A Non-Parametric Approach. *Procedia Computer Science*, *237*, pp.576-585.

[7] Weng, Q., 2009. Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends. *ISPRS Journal of photogrammetry and remote sensing*, *64*(4), pp.335-344.

[8] Manakkadu, S. and Dutta, S., 2022. ACO based Adaptive RBFN Control for Robot Manipulators. *arXiv preprint arXiv:2208.09165*.

[9] Weng, Q., 2002. Land use change analysis in the Zhujiang Delta of China using satellite remote sensing, GIS and stochastic modeling. *Journal of environmental management*, *64*(3), pp.273-284.