Narrating the Unseen: Real-Time Video Descriptions for Visually Impaired Individuals

Nelson Joseph

College of Engineering, The University of Texas at Arlington, Arlington, Texas

Email: nelsonjoseph123@gmail.com

*Abstract*—This paper demonstrates the effective usage and comparison of different image captioning models that includes both pre-trained and manually trained ones in generating captions from the live feed. Also, this research work attempts to combine a CNN type encoder for extracting features from images and a Recurrent Neural Network for generating captions for the extracted features. The CNN encoders used are VGG16 and InceptionV3. The extracted features are then passed to a unidirectional or a bidirectional LSTM for generating captions. The proposed model has used beam search as well as greedy algorithms to generate captions from vocabulary. The generated captions are then compared with actual captions with the help of BLEU scores. The Bilingual Evaluation Understudy score (BLEU) is used to compare how close a given sentence is to another sentence. The BLEU score of captions generated using beam search as well as greedy algorithms are analyzed and compared to see which is better.

Keywords— GPT4Vision, VIT, LSTM, RNN

# Introduction

Camera feeds often present challenges such as varying lighting conditions, different angles, motion blur, and real-time changes, which are not typically a focus in the training datasets of most image captioning models. While CLIP might not be specifically trained on camera feed data, its generalizability and broad training base might make it more capable of handling the diverse and dynamic nature of images from camera feeds compared to other models that are trained on more static and curated datasets.

# ReLated Work

## Related works in the Image Captioning area.

Caption Recommender System is an integral part of understanding the environment, which has various applications (e.g. - subtitle generation, helping visually impaired people to understand their surroundings, storytelling from albums, search using image, etc.). Since many years, many different image caption recommendation approaches have been developed. There have been a lot of contributions from the architecture created by the winner of the ILSVRC. Along with the VGG the research made in the field of natural language translation have helped us continuously in bettering the performance in text generation. Researchers at AI Lab used a Convolution Neural Network for each potential object in the image for producing high-level features of the image. Then a Multiple Instance Learning (MIL) [6] was used for figuring out the best area which matches with each word. This method gave a BLEU score of 22.9% on MS-COCO dataset. The Vinyals came up with a new model called NIC (Neural Image Caption), Show and Tell model [7], which was nothing but an encoder RNN which was given input through a CNN model for computer vision. After this a group of researchers took the NIC model and modified it. They used a technique that makes use of images datasets and their corresponding captions to study the inter-modal correlations between natural language and image data. The model used by those researchers was based on a new combination of CNN around image fields, the LSTM or bidirectional RNN over textual descriptions, and a planned aim of putting the two modals together via bimodal embedding. Flickr 30K, Flickr 8K and MSCOCO were the datasets used by them to achieve these bests in business results. Jonathan further modified their model in 2015 when he suggested an idea of a model related to dense captioning in which the model detects each of the different areas of the image and then suggests a group of captions. Chen Wang also suggested a model which makes use of multiple LSTM networks and a deep CNN in the year 2016. Over a period of time there has been enhancements not only in the captioning models but also in score metrics used for evaluating the accuracy of the models. This project has used the BLEU score for evaluation. BLEU - being a standard evaluation metric adopted by many of the groups. Now, new state of the arts metrics has come like CIDEr which are replacing older metrics like BLEU score, etc. CIDEr was proposed by Vedantam [8].

## Related works associated for visually impaired people

Identify applicable funding agency here. If none, delete this text box.

Web-based applications are built to protect user privacy. The user of this application can switch on the on-demand feature to sacrifice privacy while updating the family on his or her status. They may recognize items around them and comprehend their immediate environment using a low-power Mobile-Net Architecture based on CNN. The experiment’s participants had problems identifying road curbs, changes in the road’s surface, and staircases when it was being carried out outside. The benefits and limitations of each type of soft fabric in inside situations, such as soft cushions and drapes, are also discussed [7]. In [8] an assessment of assistive technology using scores to illustrate how well they can add features, VIP assistive devices are categorized according to their features and operational principles. A quantitative study based on scores is performed to determine these devices’ effectiveness and the possibility of feature augmentation for each category. The idea of expanding the number of items that can be tallied at once was studied in [9] [10] in order to improve the support given to persons who are visually impaired. A voice alarm that can recognize all probable daily multiple objects will also alert consumers to both nearby and far-off objects. The earlier techniques are subject to limitations in terms of accuracy, scene complexity, illumination, and other factors. By providing high precision, the best performance results, and accessible options, the objective is to improve the quality of life for blind or visually impaired people. The application’s User Interface is self-explanatory and easy to use, making it appropriate for blind users [11]. The main goal is to assist a person who is blind indoors by just hearing an audio query about the required object and providing aural feedback, such as the name, distance, and direction of that particular thing. The best depth quality per depth and good accuracy enable a blind person to quickly calculate distance and direction. The results of the tests also show that this technology is easily able to overcome hurdles because it can move around in environments with few limits [12]. The model’s computational efficiency has been enhanced with the help of the architecture of a faster region convolu tional neural network. They removed feature data from the image and used feature mapping to assign a class label. The SSD algorithm’s default boxes used with a multi-scale feature map mitigated this reduction [5]. The Stanford model, which delivers more in-depth infor mation, is utilized to create captions once the visual elements of images are extracted using VGG16. First, picture visual features are extracted using a pre-trained VGG16 model. The Stanford model is then used to create a caption for them [13]. Researchers have developed walking aids that ensure the mobility and safety of people with visual impairments. Any framework for helping people with visual impairments should take into account the study as a strong foundation for a com plete description of the essential elements. The sensors that go into making walking aids may not be used appropriately without undesirable results [14].

*Outcome of Literature Survey and research gaps*

Many visually impaired people around us are facing numer ous challenges. One of the most common and complicated difficulties is when they have to navigate themselves in an unfamiliar environment. After reviewing the literature and conducting a survey, we decided to create an object detection summarization for visually impaired people. So that they can visualise a summary of the objects around them in their native language (braille). In addition, voice modulation is being used to recognize the objects. As a result, object detection applications using machine learning techniques are employed to aid or assist them.

# Research Problem

Recent technological developments have focused on creat ing practical systems to improve living standards. In order to help visually impaired people recognize objects, we can use machine learning. These people use their senses to learn about their surroundings by touching, hearing, and smelling, but they are limited in how much of their environment they can comprehend because they cannot use their eyes. Therefore, we decided to design an application called Image/video summa rization in text/speech for visually impaired people to assist them so that they can record images of their surroundings and identify objects in them.

## Contributions of this paper.

With gathering much information in the Existing system, many electronic devices like walking sticks and alternative devices were created to assist blind people. However, the disadvantage of these systems is that they use hardware for development which is expensive and can’t be afforded by all. This hardware is heavier for blind people to carry. This drawback motivated the proposal of an affordable and portable system to find objects and narrate for the visually impaired and assist them in their surroundings. The main Objectives are:

* Smartphones have sensors integrated with cameras and speakers, thus reducing the development cost of application.
* Server and portable android application to detect the objects.
* Evaluate the accuracy of the proposed method with the existing methods.

# Framework and System Design

System design is designing a system’s architecture, parts, and interfaces to satisfy the users’ needs. Fig. 2 shows the proposed model’s various phases, and each of their function ality is described below.

* Server: In this phase, after starting Computer. The server is used for sharing the information with other computers to detect the image and deliver the output to Android via Wi-Fi. The objects detected is from the trained set of the caffe model.
* wifi: Image from the server is being detected by wifi.
* Android Phone: After being detected by the server, this android device detects the image and voice.
* Select IP Address: In this phase, the android phone selects the IP Address, and the IP address of the phone and Computer should be the same.
* Initialize Video Camera: As soon as, the android phone and Computer is connected to the same IP address, the video camera can be initialized.
* Caffe Model: It is a machine learning model with an image classification or image segmentation model trained using caffe.
* Object Detection: The objects trained from the caffe model can be detected and sent to the server.
* Image: The objects which are trained from the caffe model are detected as soon as the video starts streaming.
* Stream: The video can be streamed as soon as the android phone is detected with the same IP Address.
* Image to speech: The object detected is converted to image as well as to speak.

# Results and Discussion

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

# CONCLUSION AND FUTURE SCOPE

We put forth a system that can be utilized to help blind per son comprehend their surroundings by describing the nearby things. Our proposed model will undoubtedly benefit visually impaired people. More emphasis has been placed on object detection and converting images to text and speech, ultimately aiding in object recognition. As an outcome, the above information gives clarity of what’s been attained in assisting the visually impaired. The life quality of blind persons is improved by image/video summarization applications so that they can replace darkness. The proposed work’s future focus will be on recognizing several items in a view more accurately and quickly. Any sort of thing having a high frame rate can be characterized by this system’s extension. The text-to-speech module was likewise created at a forward-thinking clip. Self-trained models can be used in place of pre-trained models. The model can be trained to visual features that the user sees regularly. As a result, it may be tailored to the user’s unique demands and assures safer navigation. With the addition of a facial recognition capability, the application may be trained to remember details about those most intimately associated with the user, making it easier for them to distinguish between friends and strangers.

##### References

1. OriolVinyals, AlexanderToshev, SamyBengio, DumitruErhan. ShowandTell:ANeuralImageCaptionGenerator. arXiv:1411.4555v2 [cs.CV] 20 Apr 2015.
2. Alexey Dosovitskiy , Lucas Beyer , Alexander Kolesnikov , Dirk Weissenborn , Xiaohua Zhai , Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. arXiv:2010.11929v2 [cs.CV] 3 Jun 2021
3. Md Milon Islam, Muhammad Sheikh Sadi, Kamal Z Zamli, and MdManjur Ahmed. Developing walking assistants for visually impaired people: A review. IEEE Sensors Journal, 19(8):2814–2828, 2019.
4. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.K. Elissa, “Title of paper if known,” unpublished.
5. Sunit Vaidya, Naisha Shah, Niti Shah, and Radha Shankarmani. Real time object detection for visually challenged people. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pages 311–316. IEEE, 2020.
6. Sadia Zafar, Muhammad Asif, Maaz Bin Ahmad, Taher M Ghazal, Tauqeer Faiz, Munir Ahmad, and Muhammad Adnan Khan. Assistive devices analysis for visually impaired persons: a review on taxonomy. IEEE Access, 2022.
7. Jawaid Nasreen, Warsi Arif, Asad Ali Shaikh, Yahya Muhammad, and Monaisha Abdullah. Object detection and narrator for visually impaired people. In 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pages 1–4. IEEE, 2019.
8. Chaitra C1, Chennamma1, Vethanayagi R1, Manoj Kumar M V1, Prashanth B S1, Snehah H R1, Likewin Thomas3, Shiva Darshan S L2. Image/Video Summarization in Text/Speech for Visually Impaired People. 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon).
9. Sahil Takkar, Anshul Jain, Piyush Adlakha. Comparative Study of Different Image Captioning Models. Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE Xplore Part Number: CFP21K25-ART
10. “Flickr 8k Data | Illinois.” https://forms.illinois.edu/sec/1713398 (accessed Mar. 05, 2021).
11. Derek Tam, Colin Raffel, Mohit Bansal. Simple Weakly-Supervised Image Captioning via CLIP’s Multimodal Embeddings. UNCChapel Hill {dtredsox, craffel, mbansal}@cs.unc.edu