Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-8, Issue-2, February 2026
DOI: 10.53469/jrse.2026.08(02).13

Image-to-Audio Captioning Systems for Visually Impaired Users: Development and Application

Pham Duc Hau

Amity School Of Engineering & Technology, Greater Noida Amity University, Greater Noida, Uttar Pradesh, India
phamduc@s.amity.edu

Abstract: ***This article provides a comprehensive overview of the evolving landscape of image captioning, with a focus on its applications in accessibility for the visually impaired. It explores the challenges of real - time object recognition, traditional object detection methods, and the transformative impact of deep learning techniques, particularly those employing region proposal object detection algorithms. The paper introduces Vision Voice, a groundbreaking web application that converts text extracted from images into natural - sounding speech. The article details the image processing pipeline, including preprocessing, segmentation, classification, and post - processing stages. It also delves into the mathematical concepts, image preprocessing techniques, and shortcomings of existing models. The study highlights the ResNet - LSTM models significant potential in generating descriptive and contextually coherent image captions, improving the quality of synthesized speech. Moreover, it discusses the future scope of the VisionVoice project, emphasizing the potential for continued advancements in accuracy, hardware capabilities, and the development of full Image - Speech conversion systems. The ultimate goal is to revolutionize accessibility and inclusion, providing visually impaired individuals with better access to information and a higher quality of life.***

Keywords: Object detection, deep learning, image processing, text extraction, speech synthesis, image captioning, ResNet - LSTM, accessibility, visually impaired, future scope

1. Introduction

Object detection plays a pivotal role in the realm of computer vision, autonomous vehicles, industrial automation, and numerous other applications. The continuous and real - time recognition of objects remains a formidable challenge in these domains. While traditional object detection methods have been employed, the advent of deep learning has ushered in a new era of improved accuracy and efficiency in this field. Deep learning techniques, particularly those employing region proposal object detection algorithms, have gained prominence. These methods encompass various approaches such as SPPnet, region - based convolutional neural networks, Fast RCNN, Faster RCNN, among others.

This paper introduces a robust methodology for text extraction from images and its subsequent conversion into speech. While previous research has explored diverse strategies for text recognition and extraction, our study, titled "VisionVoice: An image captioning web application that converts it to audio for the visually impaired, " concentrates on the utilization of image processing techniques for text extraction. The image processing pipeline encompasses essential stages such as pre - processing, segmentation, classification, and post - processing.

In the pursuit of developing a fully automated conversion algorithm, the challenge of image segmentation emerges as a critical component. This process is inherently complex, often requiring manual intervention to achieve optimal results. Segmentation serves two fundamental objectives: first, to partition the image into discernible components for subsequent analysis, and second, to reorganize pixel - level information into higher - level units, enhancing the interpretability of objects within the image.

- Data collection and Pre - Processing: This initial phase typically involves cleaning and enhancingthe input image to improve its suitability for subsequent analysis. Pre - processing steps might include noise reduction, resizing, contrast adjustment, and color normalization. The goal is to prepare the image for accurate feature extraction and segmentation.
- Feature Extraction: Feature extraction aims to identify and quantify relevant characteristics or patterns within the pre - processed image. These features could be edges, textures, shapes, or any other distinctive elements that help distinguish objects or regions of interest. Extracting meaningful features is crucial for subsequent stages of the framework.
- Image Segmentation: Image segmentation involves dividing the image into distinct regions or objects based on the extracted features or other criteria. The goal is to isolate regions that contain text or relevant content. Proper segmentation is fundamental to accurately identify and extract text from the image.
- Model Architecture: Establish an encoder - decoder architecture for the purpose of captioning images. The decoder creates the caption word - by - word once the encoder has processed the image's features.

A. Text to Speech Model

Despite decades of research, creating natural speech from text (text - to - speech synthesis, TTS) is still a difficult task. Different methods have come to dominate the field throughout time. Concatenative synthesis, or the sewing together of small units of previously recorded waveforms, has been state - of - the - art for many years. Following concatenative synthesis, statistical parametric speech synthesis was developed to address many of the boundary artefact problems. This method directly builds smooth trajectories of speech features that are then synthesized by a vocoder. However, compared to human speech, the audio

generated by these algorithms frequently sounds muffled and artificial [4].

WaveNet, a generative model of time domain waveforms, has already been used in certain full - fledged TTS systems and provides audio quality that starts to approach that of authentic human speech. However

2. Literature Review

Convolutional Neural Network - Recurrent Neural Network (CNN - RNN) Based Image Captioning and Convolutional Neural Network - Convolutional Neural (CNN - CNN) Based Image Captioning are two deep learning models used in the approach presented by Liu, Shuang, Bai, Liang, Hu, Yanli, and Wang, Haoran et al. [**1]**. Convolutional neural networks are used for encoding in a CNN - RNN framework, and recurrent neural networks are used for decoding. The images in this case are transformed into vectors using CNN, and these vectors—which are referred to as image features—are then fed as input into recurrent neural networks. The project's actual captions are obtained using the NLTK libraries in RNN's implementation. Only CNN is employed in the CNN - CNN based framework for both the encoding and decoding of the pictures. Here, using the NLTK library, a vocabulary dictionary is employed and mapped with image attributes to obtain the precise word for the provided image. creating the caption that is free of errors. To train the continuous flowing repeatedly repetition of these techniques is undoubtedly slower than the consisting of many models that are offered at the same time of convolution approaches.

The CNN - CNN Model requires less training time than the CNN - RNN Model. As it is sequential, the CNN - RNN Model requires more training time but has lower loss than the CNN - CNN Model. **[2]**

They have employed an encoding decoding model for image captioning in the way put forward by Ansari Hani et al.

The strategy put out by Subrata Das, Lalit Jain, and colleagues is based mostly on how deep learning models are utilized for captioning military images. It mostly makes use of a CNNRNN - based framework. The authors combine Long Short - Term Memory (LSTM) networks with the Inception model to overcome issues with gradient descent during image encoding. An RNN variant called LSTM is effective in detecting long - distance dependencies in sequential data. By keeping track of the words generated so far, LSTM helps in this scenario to generate captions that are cohesive and contextually relevant.

On the other hand, the Inception model is a potent CNN architecture that is renowned for its capacity to collect rich visual characteristics at many scales. The accuracy and usefulness of the generated image captions are enhanced by combining LSTM with the Inception model.

1) *Image Encoding (CNN):* A CNN (typically pre - trained on huge datasets) processes the input image to extract pertinent visual information. These characteristics are shown as a fixed - length vector. [3]
2) *Caption Generation (LSTM):* An LSTM - based language model receives the encoded picture features as its initial input. Word by word, the LSTM network creates the caption, gradually adding context and details from the image attributes.

Convolutional and recurrent neural networks are two different types of neural network topologies, each tailored for activities and data types. **[4]** The following are some key distinctions between RNNs and CNNs:

1) Data type and structure:
RNNs*:* For sequential data, where the sequence of the items is important, RNNs were developed. They are frequently employed for jobs involving speech recognition, time series data, and natural language text.
CNNs*:* CNNs are made for data that can be organized into grids, like 2D or image grids. They are perfect for jobs requiring the extraction of spatial features, like image categorization and object detection.

2) Architecture:
RNNs*:* Because RNNs feature recurrent connections, data can pass from one time step to the next. They can thus be used to model the sequential dependencies in data.
CNNs*:* CNNs are made up of convolutional layers that apply filters to specific local areas of the input data, enabling them to recognize hierarchical features throughout the input grid.

3) Memory and Context:
RNNs*:* Since RNNs can keep track of previous time steps, they are appropriate for jobs requiring context and memory, such language modelling and speech recognition.
CNNs*:* CNNs are primarily concerned with identifying local patterns and characteristics within the data and often lack built - in memory for sequential data.

4) Parameter Sharing:
RNNs: RNNs are able to handle sequences of various lengths since they use the same set of parameters at all time steps. Deep network difficulties with vanishing or expanding gradients can result from this, though.
CNNs: CNNs use convolutional filters that glide across the input grid and share weight. They can effectively learn local patterns because to the weight sharing.

Shortcomings of Various Models
There are many flaws in the current paradigm, as we have shown in the literature review. Each existing model has a drawback that reduces the model's effectiveness and accuracy when results are generated. The following are the flaws in all the models that have been observed:

a) CNN - CNN model:
There are many drawbacks to using the CNN - CNN architecture to an image captioning project.

- Redundant Feature Extraction: Redundant feature extraction would result from using two CNNs back - to - back. The first CNN would take picture features out, and the second CNN would effectively take those same elements out again. This redundant information adds to computational complexity without adding valuable data **[7].**
- Higher Cost of Computing*:* Comparing the sequential

running of two CNNs to the usual CNN - RNN technique, the sequential running of two CNNs uses a significant amount more computing power. This may not be practical in real - time applications and can slow down the training and inference processes.

- *Limited Contextual interpretation*: CNNs are primarily developed to extract features from images; they do not, however, naturally capture the sequential or contextual information necessary for natural language interpretation. This restriction is not overcome by using two CNNs in quick succession, which can result in captioning that is less contextually pertinent.
- *Lack of sequential modelling:* Generating captions for images often entails doing it sequentially, word by word. Using two CNNs may not give the essential modelling for sequential caption production because CNNs are not intended for sequential data processing.
- *Overfitting:* Overfitting occurs when two different CNNs trained on the same data become overly specialized in recognizing particular features of the training dataset. Poor generalization to novel, unforeseen imagery may come from this.
- *Training Process Complexity and Difficulty:* Managing the training procedure and loss optimisation for two CNNs in succession might be difficult. Multiple neural networks' training sometimes demands complex design decisions and hyperparameter adjustment.

b) CNN - RNN model:

A popular design for picture captioning is the CNN - RNN (Convolutional Neural Network - Recurrent Neural Network) model, but it has significant limitations and difficulties **[7]**.

- *Lack of Deep comprehension of Content, Context, and Semantics:* CNNs are very good at extracting visual information from images, but they have limited contextual comprehension. Low - level and mid - level features are captured, but
- high - level semantic understanding—which is necessary for producing contextually pertinent captions—might be difficult for them to grasp.
- *Fixed - Length Feature Maps:* Regardless of the complexity or substance of the image, CNNs generate fixed - length feature maps. When encoding images with various levels of detail, this might lead to an information loss. The model's generated captions might not fully capture the essence of the image.
- *Relationships and Interactions:* Relationships and interactions between objects or areas in the image are difficult to represent when using CNNs since they treat each area of the image separately. For the purpose of creating captions that are logical and contextually rich, it is crucial to comprehend how objects connect to one another.
- *Problems Handling Variable - Length Captions:* Sequential models like RNNs and LSTMs generate captions word by word. They frequently call for a fixed - length input vector to start the generating process, which might not work well with sophisticated variable - length captions.
- *Risk of Repetition or Incoherent Text:* RNNs may produce repetitive or illogical text, particularly when trained on extensive datasets with a wide range of linguistic patterns. The model might provide lengthy, incoherent, or overly descriptive captions.

c) CNN - LTSM model:

Although using a CNN - LSTM model for picture captioning provides benefits, there are some downsides as well.

- *Complexity:* Due to the mix of convolutional and recurrent layers, CNN - LSTM models can be computationally expensive and demand large resources for training and inference.
- *Fixed - Length Captions:* CNN - LSTM models frequently produce fixed - length captions, which may restrict their capacity to adjust to images with various degrees of complexity or substance.
- *Poorly repsented:* The generation of accurate captions for unusual or rare objects or scenes that were poorly represented in the training data may be difficult for these algorithms.
- *Overfitting*: To train a CNN - LSTM model properly, large, annotated datasets are necessary, and it might be difficult to avoid overfitting to the training set of data.
- *Lack of Global Context:* Because LSTM models may have a limited comprehension of context and may not adequately capture the global context of the image, captions may lack consistency. **[7]**

3. Proposed Model

When selecting a model, it's critical to consider the unique elements of your dataset, the nature of the photos, and your computational capabilities. To cut down on training time and resource needs, consider the availability of pretrained models and whether transfer learning can be used to your work.

Due to its capacity to successfully capture both sequential context and sophisticated visual elements at the same time, the ResNet - LSTM model provides a clear advantage over existing picture captioning designs. As a deep convolutional neural network, ResNet excels in extracting detailed, hierarchical visual features that allow it to recognize minute details in images. The ResNet - LSTM model can produce captions that are contextually coherent in addition to being visually descriptive when paired with LSTM, a recurrent neural network. It is a great choice for tasks requiring the merging of spatial and sequential comprehension since it produces captions that are both aesthetically accurate and contextually significant, which is frequently necessary for accurately expressing the information of images. **[6]**

Alternative designs, such as CNN - RNN or CNN - CNN, may, however, have difficulty striking a good balance between feature extraction and sequential modelling. Although CNN - CNN models are good at capturing geographical characteristics, they may lack sequential context, which makes captions less intelligible. On the other hand, CNN - RNN models could struggle to capture minute visual features, which might lead to captions that are less illustrative. Because it can successfully achieve a synergy that improves the overall quality and significance of output captions, the ResNet - LSTM model is positioned as a strong option for picture captioning

3.1 Mathematics Involved

Using the FLICKR 8K dataset and the ResNet - LSTM model, image captioning uses a number of mathematical ideas, terminology, and calculations. The following essential components are frequently employed when creating image captions:

a) (Objective Function) Loss Function

Cross - Entropy Loss (Log - Likelihood Loss) is a popular loss function that is used to quantify how different predicted captions are from the actual captions. It measures the discrepancy between the actual words in the captions and the expected word probabilities

$$\text{Cross-Entropy Loss} = -\sum_i y_i \log(\hat{y}_i)$$

b) Optimization

SGD: Stochastic Gradient Descent It is a typical optimization approach to iteratively adjust the model's weights during training in order to reduce the loss function.

$$\theta \leftarrow \theta - \alpha \frac{\partial L}{\partial \theta}$$

c) LSTM - based recurrent neural networks

LSTM Cell: To regulate the information flow inside the network across successive time steps, LSTM units are defined by a collection of mathematical operations, such as forget gates, input gates, and output gates. LSTM variants (such as vanilla LSTM and GRU) have different LSTM operation formulas. These procedures include sigmoid activations, tanh activations, and element - wise multiplications.

d) Focus Mechanisms

Calculating Attention Weights: In attention mechanisms, you compute attention weights to assess the significance of various aspects of the image when coming up with each word in the caption. Mathematical techniques like dot products, softmax, and weighted summation are used to calculate these weights.

e) Embeddings in Word

Embeddings such as Word2Vec or GloVe express words as vectors in a continuous space. Mathematical techniques such as dot products and cosine similarity are used to calculate how similar two - word vectors are to one another.

3.2 Image Preprocessing

Particularly for computer vision problems, image preprocessing is a common step in machine learning and deep learning applications. It entails converting the raw photos into a format that the model can accept. Some of the image preprocessing techniques are:

- *Resizing:* is the procedure of altering an image's dimensions, either by scaling it up or down or by cropping it to a specific size. In order to make the photographs work with the input geometry of the model or to lower the processing cost of huge images, resizing is frequently required. For instance, ResNet models require photos with a dimension of 224x224x3 as input12; therefore, any image that does not fit this requirement must be scaled before being sent to the model. Several libraries, including OpenCV, PIL, and TensorFlow, can be used for resizing.
- *Color Conversion:* The process of converting an image's color space, such as from RGB (red, green, blue) to BGR (blue, green, red), or from grayscale to RGB, is known as color conversion. Sometimes color conversion is necessary to simplify the visual representation or to match the color space that the model was trained on. For instance, since BGR is the default color space for some TensorFlow models1

4. Conclusions

In summary, the study on "VisionVoice - An image captioning web application that converts it to audio for the visually impaired" marks a significant advancement in the pursuit of accessibility and inclusivity. This study successfully established the viability of producing descriptive captions for photographs, answering a crucial need for the community of people who are visually impaired. It did this by utilizing the ResNet - LSTM model with the FLICKR 8K dataset. The quality of synthesized speech was further improved by the incorporation of cutting - edge tools like Tacotron 2 and WaveNet as a vocoder, guaranteeing that the information provided in image captions is not only correct but also presented in a natural and understandable way.

This study highlights the enormous potential for improvements in accessibility technologies and image captioning as we look to the future. There is a definite route towards obtaining even greater levels of caption generation accuracy with the continuous development of hardware capabilities and deep learning models, thereby boosting the user experience for people who are visually impaired. The goal of extending this paradigm to complete Image - Speech conversion offers the prospect of revolutionizing how the blind access and comprehend information, promoting greater independence and enhancing their daily life. In essence, "VisionVoice" is a ground - breaking initiative to close the gap between auditory and visual comprehension, opening up the world of information to everybody.

5. Future Scope

Future study on "VisionVoice - An image captioning web application that converts it to audio for the visually impaired" has a huge potential for accessibility and technological breakthroughs. Despite the obstacles in the way of exact caption production at the moment, this research opens the door to intriguing possibilities in the years to come.

First off, as technology develops, especially in terms of hardware capabilities and deep learning models, it is anticipated that the accuracy of creating image captions would considerably increase. The continual advancements in artificial intelligence and machine learning present opportunities for improving image identification, caption generating algorithms, and enabling robots to comprehend visual content more accurately. This development will result

in more accurate and insightful image descriptions for people who are blind or visually impaired.

Additionally, there is tremendous promise in the idea of expanding this model to develop a full Image - Speech conversion system that speaks image captions. With real - time access to information and a better awareness of their surroundings, such a device may be a priceless aid for those with visual impairments. Modern Text - to - Speech (TTS) technologies will be combined with image captioning models to provide a seamless and accessible multimedia experience that will improve the quality of life for the blind and visually impaired community.

In conclusion, despite certain difficulties now, the "VisionVoice" project's future seems bright. There is a good chance that accurate picture caption generation will become more and more feasible with continued improvements in hardware and deep learning models. The ultimate objective of developing this technology to enable Image - Speech conversion has the power to revolutionize inclusion and accessibility, having a significant and profound effect on the lives of those who are visually impaired. acknowledgement

.

References

[1] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran, “Image Captioning Based on Deep Neural Networks”, MATEC Web Conf. Volume 232, 2018.

[2] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, “Image Caption Generating Deep Learning Model”, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 09, 2021.

[3] H. Agrawal et al., "nocaps: novel object captioning at scale, " 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp.8947 - 8956, doi: 10.1109/ICCV.2019.00904.

[4] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017, February 7). *Comparative study of CNN and RNN for natural language processing*. arXiv. org. https: //arxiv. org/abs/1702.01923

[5] Jiao, Y., Gabrys, A., Tinchev, G., Putrycz, B., Korzekwa, D., & Klimkov, V. (2021, February 15).

[6] *Universal neural vocoding with parallel WaveNet*. arXiv. org. https: //arxiv. org/abs/2102.01106

[7] Donahue, Jeffrey, et al. ”Long - term recurrent convolutional networks for visual recogni - tion and description. ” Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[8] S. Takkar, A. Jain and P. Adlakha, "Comparative Study of Different Image Captioning Models, " 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp.1366 - 1371, doi: 10.1109/ICCMC51019.2021.9418451.

[9] Anitha Kumari, K., Mouneeshwari, C., Udhaya, R. B., Jasmitha, R. (2020). Automated Image Captioning for Flickr8K Dataset. In: Kumar, L., Jayashree, L., Manimegalai, R. (eds) Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications.

[10] AISGSC 2019 2019. Springer, Cham. https: //doi. org/10.1007/978 - 3 - 030 - 24051 - 6_62

[11] P. Taylor, “Text - to - Speech Synthesis”, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[12] J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database, ” in Proc. ICASSP, 1996, pp.373–376.

[13] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gib - ¨ iansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, “Deep voice: Real - time neural text - to - speech, ” CoRR, vol. abs/1702.07825, 2017.