International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

AudioScene: Enhancing Visual Independence Through Scene Recognition

Vidhyalakshmi S[1], Dr. J. M. Gnanasekar [2]

[1]Student, Department of Computer science and Engineering Sri Venkateswara College of Engineering Chennai, India.

[2]Professor, Department of Computer science and Engineering Sri Venkateswara College of Engineering Chennai, India.

Emails: *vidhyasubramanian24@gmail.com[1], jmg@svce.ac.in[2]*

Abstract

Visually challenged individuals face numerous challenges in navigating and understanding their surroundings due to their reliance on visual information. This initiative aims to empower them by utilizing cutting-edge technology to enhance their accessibility and independence. By employing deep learning algorithms captions are generated for the image, providing users with information about their surroundings. Leveraging advanced image captioning techniques and datasets like MS COCO, key features are highlighted. The system then delivers this information as audio output to the user, enabling them to navigate with confidence. Ultimately, this innovative solution, AudioScene, offers valuable support to visually impaired individuals, facilitating safer and more informed travel experiences.

Keywords: *AudioScene, Scene recognition, Deep learning, Image captioning, Object positioning, Assistive technology.*

1. Introduction

Scene recognition is essential in computer vision, enabling machines to understand images by categorizing their overall context, such as landscapes or indoor settings [1]. Unlike object recognition, which focuses on specific items, scene recognition interprets broader environmental elements. Real-world scenes present challenges due to their diversity and complexity, including variations in lighting, viewpoints, and objects. Researchers use advanced machine learning, like deep neural networks, to automatically learn features and patterns specific to different scenes. Convolutional Neural Networks (CNNs) are especially effective in extracting detailed features, aiding in understanding intricate image details and relationships. Scene recognition is important in many different disciplines. It improves safety and navigation in autonomous cars by allowing them to understand and adapt to their surroundings. Surveillance systems use scene recognition to detect potential dangers or irregularities in a particular area. Furthermore, in augmented and virtual reality applications, scene recognition helps to provide immersive and contextually relevant user experiences. Despite significant advances, scene identification is a dynamic field with ongoing research to improve algorithmic accuracy and efficiency [2]. As technology advances, the integration of scene recognition into multiple domains has the potential to reshape how robots perceive and interact with the visual environment. This advancement is driving the creation of increasingly intelligent and context-aware technologies, poised to transform multiple sectors and improve the overall experience. Visually impaired individuals encounter numerous obstacles in their daily lives due to their reliance on non-visual senses for information. A significant hurdle they face is the limited accessibility to visual content, which constitutes a substantial portion of online and printed information [3]. This lack of access to visual elements impedes their understanding and interaction with a wide range of content. One major issue is the absence of descriptive information accompanying images. Many online platforms and printed materials fail to provide sufficient textual descriptions, leaving visually impaired individuals without crucial details and context [4]. This limitation significantly limits their ability to grasp visual information that sighted

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

individuals easily comprehend. For example, an image may contain essential information but without proper description, these nuances are lost on those who rely on non-visual means. To address this challenge, efforts must focus on ensuring that visual content is accompanied by detailed textual descriptions [5]. This practice, known as image description, involves providing concise and informative descriptions of images to convey their content and context effectively. Implementing image descriptions can dramatically improve availability of visual content for individuals with vision disorders, which will allow them to access and understand information more efficiently. For example, image recognition technology combined with natural language processing can automatically generate image captions, decreasing the workload on content providers [6]. Furthermore, screen reader software, which is widely used by visually impaired people, can be modified to read image descriptions aloud, allowing users to access visual content more easily. Overall, addressing the lack of accessibility to visual content is crucial for empowering visually impaired individuals and promoting inclusivity. By ensuring that images are accompanied by descriptive information, we can bridge the gap in accessing information and enhance the overall user experience for those with visual impairments [7]. This concerted effort toward accessibility underscores the importance of inclusive design practices in today's digital landscape. The lack of inclusive design in technology and communication aggravates the difficulties faced by visually impaired individuals. Digital platforms often prioritize visual content, leaving visually impaired users unable to fully participate in online activities or access multimedia content. To address this, projects converting captioned images into audio descriptions are crucial [8]. Utilizing technologies like image recognition and natural language processing, these initiatives provide verbal descriptions of visual content, empowering visually impaired individuals to engage more fully with online information and experiences [9]. Such efforts promote inclusivity in the digital sphere, ensuring visually impaired individuals can access and comprehend visual content on equal footing with sighted users. In the proposed system, the aim is to assist visually impaired individuals through a technological solution called AudioScene. This system aims to recognize scenes in images and convert them into meaningful audio descriptions for users with visual impairments [10]. Convolutional Neural Network (CNN) layers are being combined to extract important features from images, and Recurrent Neural Network (RNN) is being used for generating descriptive captions. This allows the visual content and context of scenes captured in images to be understood. Captions are sequentially generated by the RNN based on the relationships between different elements in the scene. Ultimately, AudioScene empowers visually impaired individuals by providing comprehensive audio descriptions of their surroundings, enhancing their awareness and enabling them to navigate their environment more effectively [11]. Through advanced neural network technologies, the goal is to create an immersive and inclusive experience for visually impaired users, bridging the gap in accessing visual information.

2. Related Words

In this section, the existing works are reviewed for the development of AudioScene recognition system. Existing image captioning systems frequently struggle to balance logical and fluent use of language with topic richness and accuracy [12]. This is due to the disparity between how we perceive things and how we express them in words. The research proposes a novel paradigm, Text-Guided Generation and Refinement (TGGAR), to address this issue. This technique use guide text to improve the quality of captions. The model is designed as an encoder-decoder system, with a Text-Guided Relation Encoder (TGRE) that learns visual representations that match human visual perception. The model's decoder is divided into two parts: a Generator, which generates the main sentence using a standard LSTM and a Gate on Attention (GOA) module to ensure that it is logical and fluent; and a Refiner, which focuses on the details of the main caption using a caption encoder, an attention-based LSTM, and a GOA module with the guide text. Tests on the MS COCO captioning dataset suggest that this framework performs better than other techniques. This highlights the potential of the TGGAR model to improve image captioning (Wang et al., 2022). SD-RSIC

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

(Summarization Driven Remote Sensing Image Captioning) is a novel approach for creating image captions in the field of remote sensing. This approach consists of three major steps: creating standard captions, summarizing ground-truth captions, and combining summarized captions with standard captions. The first phase generates typical captions using CNNs and LSTMs, both of which are extensively used in image captioning [13]. The second and third steps are the most unique aspects of the suggested approach. In the second step, sequence-to-sequence DNN models are utilized to condense the ground-truth captions and eliminate superfluous information. In the third phase, an adaptive weighting approach is used to accurately merge the summarized captions with the standard captions using image attributes. The proposed approach aims to address difficulties such information deficit and overfitting caused by duplicate information in ground-truth captions, thereby improving the performance of picture captioning models (Sumbul et al., 2021). A novel image captioning framework is developed that creates captions for images within certain subjects. The system employs a cross-modal embedding method to learn the relationships among images, topics and captions. The embedding space is organized hierarchically while retaining the order-embedding strategy, which brings images and captions about the same topic closer together. The MS COCO and Flickr30K datasets produce competitive results in both the caption-image retrieval and caption production tasks. The framework gives users control over producing intended captions for images, which opens up possibilities for exciting applications (Yu et al., 2019). The existing ideas-to-caption system in image captioning confronts issues due to a lack of concepts created by a huge imbalance between positive and negative concept samples, as well as incomplete labelling in training captions due to biased annotation and synonym usage. To address these issues, a technique called Online Positive Recall and Missing Concepts Mining (OPR-MCM) is created. This technique adaptively changes the loss of various samples for online positive recall and uses a two-stage optimisation mechanism for missing concept mining. The goal is to recognise more semantic concepts and write more precise and informative captions for images. During caption generation, it employs an element-wise selection technique to pick the best captions at each time stage. Extensive studies on the MS COCO image captioning dataset reveal that the proposed method outperforms existing competitive methods (Zhang et al., 2019). The difficulty of adapting image captioning systems to new areas due to the labor-intensive and time-consuming process of describing sufficient data is addressed. A unique Multitask Learning Algorithm for Cross-Domain Image Captioning (MLADIC) that optimizes two connected objectives is presented. There are two phases involved: image captioning and text-to-image synthesis. MLADIC employs an encoder-decoder model (CNN-LSTM) for image captioning and a conditional generative adversarial network (C-GAN) for image synthesis. MLADIC uses a two-step strategy to bridge the gap between domains: first, the model is pre-trained to learn the alignment between image and text representations using labelled source domain data, and then the model is fine-tuned using limited image-text pairs and unpaired data from the target domain. MLADIC's performance is evaluated using the MSCOCO dataset as the source domain data and Flickr30k and Oxford-102 as the target domain data, demonstrating significantly superior performance than existing approaches for cross-domain image captioning (Yang et al., 2019). The proposed approach to cross-domain image captioning is novel because it combines two fundamental components: a cross-modal retrieval model and an adaptive image captioning model. The retrieval model works by creating pseudo-image-sentence pairs in the target domain. On the other hand, the adaptive image captioning model is in responsible for shifting information from the source domain to the target domain. This dual-model technique aims to correct disparities between domains, thereby enhancing image captioning accuracy. By using the retrieval model to create pseudo-image-sentence pairs, the system learns about the target domain's linguistic nuances. Meanwhile, the adaptive image captioning system makes certain that the system can adapt its analysis of visual content to the target domain. The basic objective of this method is to reduce the gap across domains,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

improving the accuracy and relevance of generated image descriptions. By combining the benefits of both models, the system is able to offer more contextually relevant and semantically meaningful captions for images in the target domain. The usefulness of this unique strategy is proven by thorough testing on four publicly available datasets. These trials demonstrate the approach's capacity to increase performance in cross-domain image captioning tasks. Overall, this approach marks a big step forward in the field of image captioning, presenting a potential option for dealing with domain differences and enhancing the quality of image descriptions (Zhao et al., 2021). The focus of the study centres on remote sensing image (RSI) captioning, a task aimed at producing descriptive sentences that encapsulate the content depicted in RSIs. Previous approaches to this task have typically treated each of the five generated sentences independently, potentially leading to the production of ambiguous or disjointed descriptions. In respond to this trouble, the study proposes a unique technique called retrieval topic recurrent memory network (RTRMN). This innovative model is designed to incorporate a set of topic words that capture common information shared among the five sentences, thus fostering coherence and clarity in the generated captions. At the core of the RTRMN model lies a memory network architecture, which leverages the inclusion of topic words as memory cells. By integrating these topic words into the memory network, the model gains the ability to generate sentences that are grounded in a cohesive understanding of the shared information across multiple captions. This approach represents a departure from conventional methods by promoting a more holistic approach to caption generation, wherein the content of each sentence is informed by the overarching themes encapsulated within the topic words. Furthermore, the proposed method offers an additional layer of flexibility through the manual editing of topic words for test images. This feature empowers users with greater control over the caption generation process, allowing them to tailor the output to better suit their specific needs or preferences. Through a series of comprehensive experiments conducted on two caption datasets, the effectiveness of the RTRMN model is thoroughly evaluated. These experiments serve to validate the efficacy of the proposed approach in enhancing the coherence, clarity, and overall quality of captions generated for remote-sensing images (Wang et al., 2020). The study focuses on image captioning, which is an interdisciplinary field where computer vision and natural language processing intersect. Remote sensing images are particularly emphasized in this research, and a novel approach called the multiscale multiinteraction network is introduced. This innovative model is specifically designed to handle the unique characteristics of remote-sensing images, including variations in scale and the distinguishability of targets. By utilizing insights from both CV and NLP domains, the proposed network aims to improve the accuracy and relevance of image captions generated for remote sensing data. In addition to introducing the multiscale multiinteraction network, a comprehensive review of image captioning development is provided. This review covers advances in both natural image captioning, which focuses on frequently seen images, and remote sensing image captioning, that uses images taken with remote sensing technologies. By placing the proposed method within the broader context of image captioning research, the study illuminates the evolution of techniques and strategies used in this vital area of computer vision and NLP. To assess the efficacy of the suggested method, detailed assessments are performed on three distinct remote sensing datasets: RSICD, Sydney-Captions, and UCM-Caption. The study also demonstrates that the proposed model shows improved capabilities in recognizing targets with variations in scale or similar appearances. This highlights the usefulness and robustness of the model in addressing real-world challenges associated with remote sensing imagery. (Wang et al., 2022).

3. Proposed Word for Audio Scene Recognition

The system under consideration aims to create AudioScene, a specialized technological tool designed specifically for individuals with visual impairments. This pioneering system emphasizes scene recognition, converting images accompanied by captions into elaborate audio descriptions. Leveraging sophisticated image captioning

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

technology, AudioScene furnishes visually impaired users with extensive details regarding their surroundings through auditory means. The principal objective is to enhance accessibility and self-sufficiency for individuals with visual impairments, granting them a heightened comprehension of the visual components in their environment through inventive audio-based solutions. The main objective is to empower visually impaired individuals through an artificial vision system built on deep learning techniques. To achieve this, we start by gathering the Microsoft Common Objects in Context (MS COCO) dataset, a renowned resource in computer vision and image processing. Developed by Microsoft Research, MS COCO contains a vast collection of diverse images, each accompanied by multiple descriptive captions. Before training, the dataset undergoes essential pre-processing to ensure it works smoothly with deep learning algorithms. This step is crucial for preparing the dataset to be effectively utilized by the model. After pre-processing, feature extraction is carried out using the EfficientNet algorithm. This algorithm is selected for its ability to extract important features from images, providing a strong foundation for the artificial vision system's subsequent stages. The captions are mapped to the images. Each image is mapped to five captions. The captions are loaded from the annotations file and mapped to the corresponding image from the images file. Mapping captions with images is a crucial step in tasks such as image captioning, where the goal is to associate descriptive text with visual content. This process involves creating a structured relationship between images and their captions, allowing a machine learning model to understand the associations and generate meaningful descriptions for new, unseen images. The following paragraphs elaborate on the process of mapping captions with images. After gathering the dataset, each image is matched with its corresponding caption, establishing a direct link between visual and linguistic data. This pairing is crucial for training supervised machine learning models, which learn to connect visual features with language patterns using the dataset examples. During training, the model learns the connections between visual content and descriptive language using the paired images and captions. Convolutional Neural Networks (CNNs) are typically used to obtain visual aspects from images, whereas Recurrent Neural Networks (RNNs) handle the sequential information present in captions (Figure 1).

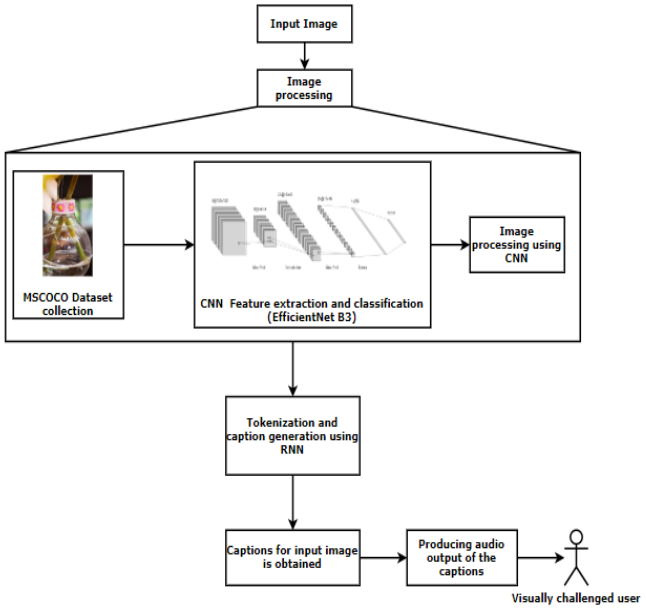

Figure 1 Architecture of The Proposed System

EfficientNet stands out for its efficiency in extracting important features from images with different levels of complexity. Its strength lies in finding a balance between model size and performance, making it adaptable to diverse datasets like COCO. This adaptability is vital for accurately recognizing various scenes and objects within large-scale datasets. The architecture of EfficientNet is designed to capture hierarchical and abstract features effectively. It consists of multiple layers that progressively learn from simple to complex patterns. This hierarchical feature learning is crucial for understanding the detailed information present in visual data. Additionally, EfficientNet incorporates techniques like batch normalization for stable and faster training. Overall, EfficientNet's architecture and techniques contribute to its effectiveness in feature extraction and performance optimization for artificial vision systems. To further enhance the accuracy of caption generation, the system employs sophisticated image captioning techniques. These techniques, which include tokenization methods, enable the algorithm to break down textual input into

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

smaller units (tokens). This process facilitates better comprehension and processing of the textual information by the algorithm, ultimately leading to more precise and relevant captions. At the heart of the system lies the training of the model using Recurrent Neural Network (RNN), renowned for its ability to handle sequential data effectively. RNNs have internal memory, enabling them to remember previous inputs, which is ideal for sequential tasks. The Bahdanau attention algorithm is added to enhance the caption generation process. Using this attention mechanism, the model may focus on certain sections of the image when constructing textual descriptions, improving the process. Especially useful with sequential data, the Bahdanau attention algorithm helps the model choose and pay attention to different areas of the input image while generating captions. This improves the quality and importance of the captions that are created. This approach, which is based on attention, changes where the model focuses on the features of the input image, making RNN-based image captioning tasks more efficient and effective. In image captioning, the Bahdanau attention technique allows the model to effectively focus on certain regions of the input image when providing textual descriptions. Unlike fixed attention methods, Bahdanau's approach calculates attention weights for each element of the input sequence during each decoding step. This adaptability ensures that the model accurately describes the key features of the image in contextually relevant captions. During training, the model learns attention weights, enabling it to prioritize important image regions for generating specific words in the caption. This adaptability is crucial for image captioning, where the relevance of different visual elements may change during caption generation. By incorporating Bahdanau attention into the RNN image captioning process, the model can focus on intricate details, improving caption quality. When an image is captured, the trained model recognizes scenes and predicts their content. These predictions are then converted into audio output, providing visually impaired users with information about object positions and distances. This real-time audio description system enhances their confidence when navigating independently. The artificial vision system follows a step-by-step procedure that includes dataset collection, pre-processing, feature extraction with EfficientNet, and model construction with RNN and Bahdanau attention. From predicting scenes to providing real-time audio descriptions, this system empowers visually impaired individuals during solo travels. When an image is captured by the camera, the processor identifies and predicts the objects within the image. Additionally, it predicts the most prominent objects present. Through tokenization, the machine learns to associate various scenes with descriptive captions. Following this prediction, the system proceeds to convert the generated captions into audio output using gTTS (Google Text-to-Speech). Once the captions are prepared, the gTTS library from Google becomes useful. This library helps convert text into speech in a convenient way. To start, the gTTS library is added to the project environment. Then, the captions are put into the gTTS function, along with details like language, speed, and pitch. The gTTS function then changes the text into speech and creates an audio file with the spoken captions. Users can adjust different parts of the speech, like the language and how fast it speaks, to make the audio fit their needs better. This adaptability ensures that the speech is easy to understand and use. This speech can be added into different systems, like assistive tools for people with vision problems, educational programs, or apps with multimedia, to make the user experience better for everyone. Overall, using gTTS to change captions into speech offers a smooth way to change written words into spoken ones. With this technology, developers can make things easier to understand, get users more involved, and create more engaging experiences in many different apps and platforms. This audio output provides users with auditory cues, facilitating their ability to discern the identified objects within the image.

4. Experimental Analysis

The model architecture incorporates the Bahdanau attention mechanism, comprising a convolutional neural network (CNN) for extracting image features and an RNN with Bahdanau attention for generating captions. This attention mechanism enables the model to focus on various image parts while creating captions, enhancing its contextual understanding. During training, batches of image-caption pairs are

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

presented to the model. The CNN extracts feature from images, while the RNN with Bahdanau attention processes these features to generate captions sequentially. The model's predictions are compared to ground truth captions, and a loss is computed using an appropriate objective function, typically cross-entropy loss. The loss plot is a useful tool for determining how well the model is learning during training. It assists practitioners in determining if the model fits the training data too closely (overfitting) or does not learn enough from the data (underfitting). Overfitting implies that the model performed well on training data but badly on fresh data, whereas underfitting shows that the model did not learn enough from the training data. Regularly checking the loss plot allows practitioners to make adjustments to improve model performance, such as tweaking hyperparameters or modifying the architecture. Once the model has reached satisfactory performance on a validation set, it can be deployed to generate captions for new images confidently (Figure 2).

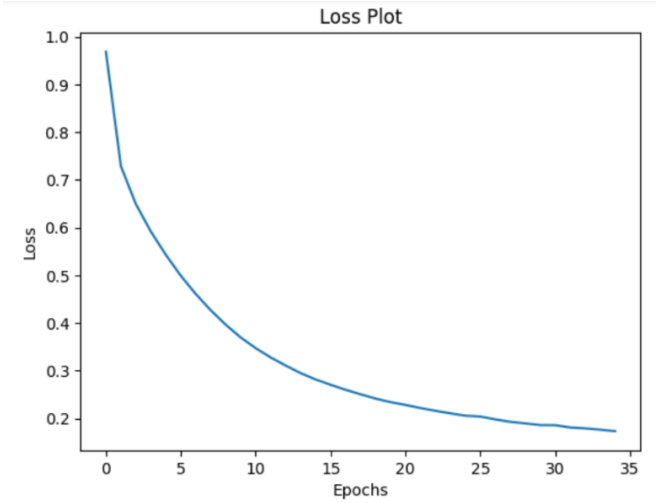

Figure 2 Loss Plot

5. Results

When an image is given to the system, it quickly figures out what objects are in the picture. It does this by looking at the image and deciding which objects are the most important. This important step helps make sure that the descriptions of the image are accurate and helpful. After the objects are recognized and predicted, the system delivers an audio output to help visually impaired people. This audio is made by changing the words in the picture's description into speech using a tool called gTTS. By changing the captions into audio output, the system makes it easier for people with vision problems to understand their environment more conveniently. The results obtained are given below (Figure 3).

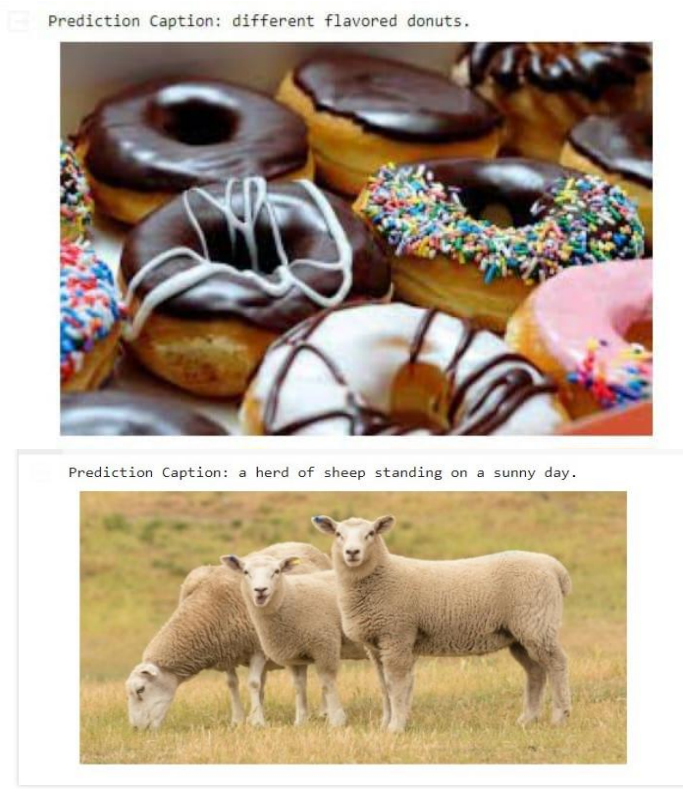

Figure 3 Image Mapped with Captions

Conclusion

In the effort to improve accessibility and inclusivity for people with visual impairments, advanced technology plays a crucial role. This model is specifically designed to find objects in images, making it easier for visually impaired people to understand images through text. The system uses complex techniques for adding captions to images, along with methods like tokenization. These methods show how dedicated the system is to making sure the captions are accurate and detailed, improving the overall experience for users. The system's main job is to quickly find and predict objects in images. By carefully looking at the images, the system's processor can identify and separate objects, which helps create detailed descriptions of what's in the images. This step is important because it ensures that

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 06 June 2024
Page No: 1776- 1783
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0244

users get accurate and helpful information about the images. All of these efforts come together to create audio descriptions of images, helping visually impaired users understand what's in the images. By using a tool called gTTS, the system turns written captions into audio output, making it easier for visually impaired people to access visual content. This process ensures that visually impaired individuals can understand images more easily, making technology more inclusive and improving overall user engagement. In summary, the integration of the deep learning model, along with advanced image captioning techniques, marks a significant step forward in accessibility technology. This system shows how technology can empower visually impaired individuals, making their lives richer and promoting a more inclusive society. In the future, helping visually impaired individuals with advanced technology has a lot of potential for new ideas and making a big difference.

References

[1]. Wang, D., Hu, Z., Zhou, Y., Hong, R., & Wang, M.. (2022). A Text-Guided Generation and Refinement Model for Image Captioning. https://doi.org/10.1109/tmm.2022.3154149

[2]. Sumbul, G., Nayak, S., & Demir, B.. (2021). SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning. 59(8). https://doi.org/10.1109/TGRS.2020.3031111

[3]. Moganapriya, C., et al. "Dry machining performance studies on TiAlSiN coated inserts in turning of AISI 420 martensitic stainless steel and multi-criteria decision making using Taguchi-DEAR approach." Silicon (2021): 1-14.

[4]. Kaliyannan, Gobinath Velu, et al. "Development of sol-gel derived gahnite anti-reflection coating for augmenting the power conversion efficiency of polycrystalline silicon solar cells." Materials Science-Poland 37.3 (2019): 465-472.

[5]. Velu Kaliyannan, Gobinath, et al. "An extended approach on power conversion efficiency enhancement through deposition of ZnS-Al 2 S 3 blends on silicon solar cells." Journal of Electronic Materials 49 (2020): 5937-5946.

[6]. Sathishkumar, T. P., et al. "Investigation of chemically treated randomly oriented sansevieria ehrenbergii fiber reinforced isophthallic polyester composites." Journal of Composite Materials 48.24 (2014): 2961-2975.

[7].

[8]. Yu, N., Hu, X., Song, B., Yang, J., & Zhang, J.. (2019). Topic-Oriented Image Captioning Based on Order-Embedding. 28(6). https://doi.org/10.1109/TIP.2018.2889922

[9]. Zhang, M.-X., Yang, Y., Zhang, H., Ji, Y., Shen, H. T., & Chua, T.-S.. (2019). More is Better: Precise and Detailed Image Captioning Using Online Positive Recall and Missing Concepts Mining. 28(1). https://doi.org/10.1109/TIP.2018.2855415

[10]. Yang, M., Zhao, W., Xu, W., Yabing, F., Zhao, Z., Chen, X., & Lei, K.. (2019). Multitask Learning for Cross-Domain Image Captioning. 21(4).https://doi.org/10.1109/TMM.2018.2869276

[11]. Zhao, W., Wu, X., & Luo, J.. (2021). Cross-Domain Image Captioning via Cross-Modal Retrieval and Model Adaptation. 30. https://doi.org/10.1109/TIP.2020.3042086

[12]. Wang, B., Zheng, X., Qu, B., & Lu, X.. (2020). Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning. 13. https://doi.org/10.1109/JSTARS.2019.2959208

[13]. Wang, Y., Zhang, W., Zhang, Z., Gao, X., & Sun, X.. (2022). Multiscale Multiinteraction Network for Remote Sensing Image Captioning. 15. https://doi.org/10.1109/jstars.2022.3153636