IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Viewing Experience for Visually Impaired Persons

[1]Gowtham. M, [2]Mr. Giriprasath. KS,

[1]M. Sc Data Science and Business Analysis, [2]Assistant Professor,
[1 2] Department of Computer Science,
[1]Rathinam College of Arts and Science, Coimbatore, India

Abstract : Audio description is a narrative technique in which visual information in multimedia content such as videos, films and television programs is described by spoken commentary. This technique is critical to making multimedia accessible to blind people who rely on auditory cues to perceive and understand the world around them. In this paper, we propose an audio description narrative that utilizes natural language processing and machine learning techniques to automatically generate audio descriptions from multimedia content. The generator uses a multi-stage approach that includes scene segmentation, object recognition and natural language generation. Scene segmentation divides the multimedia content into individual scenes, each representing a specific visual context. Object recognition identifies and classifies the objects within each scene and provides a detailed understanding of the visual elements. Finally, in the natural language generation phase, a coherent and engaging audio description is created to convey the essential visual information to the blind user. Our proposed audio description generator has the potential to significantly improve the multimedia experience for blind people by enabling them to fully perceive the visual aspects of movies, and other multimedia content. By providing comprehensive and informative audio description, the generator can bridge the gap. This promotes a more inclusive and accessible multimedia landscape.

IndexTerms **- Audio Narrative, Accessible to blind persons.**

1. Introduction

Audio description is a spoken commentary that accompanies visual media such as films, television programs and videos. Audio description is defined as the verbal depiction of key visual elements in media and live productions AD is meant to provide information on visual content that is considered essential to the comprehension of the program. There have been a number of advantages in the field to read the text help visually impaired people or blind person without much difficulty [1] . In these cases, not providing AD would inhibit blind and visually impaired individuals from gaining a complete understanding of the given program. Auto Image captioning has a good practical use in industry, security, surveillance, medical, agriculture and many more prime domains [2] .This commentary is essential for blind and visually impaired people to understand the content of these media. Without audio description, blind and visually impaired people can only perceive a fraction of the information conveyed in visual media. Traditional methods of creating audio descriptions are time- consuming and expensive as they require human experts to manually create the commentary. This process is often slow and labor-intensive, and it can be difficult to find qualified experts to create audio descriptions for all available visual media. Because of these challenges, there is a great need for automated methods of creating audio descriptions. Automated systems could significantly reduce the time and cost of creating audio descriptions and improve the availability of audio descriptions for blind and visually impaired people. In this project, an automated approach to audio descriptions generation using natural language generation techniques is proposed. The proposed system takes a video or image as input and generates a corresponding audio description in spoken language. The system uses a combination of computer vision and natural language processing algorithms to identify and describe the key elements in the visual content. In the traditional method, a group of script writers, dubbing artists and editors are involved in the process of creating audio description. Initially the descriptive scripts are written by the script writers then the dubbing artists will record the audio and finally the mixing will be done by the editor. It is a time- consuming process. Convolutional neural networks are particularly well-suited for extracting spatial features from images. CNNs can learn hierarchical representations of visual information, capturing details at different scales. Generating a meaningful natural language description of an image requires a level of image understanding that goes well beyond image classification and object detection [3] . This capability makes CNNs well-equipped for analyzing the visual content of images and identifying relevant objects and regions. LSTMs can capture temporal dependencies in natural language, enabling them to generate coherent and grammatically correct descriptions. As LSTMs are sequentially processed it consumes more time the transformers. LSTMs can also learn to incorporate contextual information from previous words in a sentence, improving the accuracy and fluency of the generated

descriptions. The aim of this project is to develop an AI-supported system that can automatically generate audio descriptions for visual scenes. This system should help blind people to better understand and experience the world around them. Using these networks various different methods are developed to perform auto image captioning in various different domains. Using these networks various different methods are developed to perform auto image captioning in various different domains. However, still, there is room for the machine to make capable enough to generate descriptions like a human [4] . The system would use a Vision Transformer to analyze visual scenes and extract key information about the objects, people and actions present.This information is then used to generate a natural language description of the scene. The description would be tailored to the needs and preferences of the individual user and rendered in a clear and concise form.The system would have a range of possible applications, which helps the blind people to navigate their environment and improves the learning experience for blind students.

2. Literature Survey

Several studies have explored the use of CNNs and LSTMs for audio description generation. Early efforts focused on using CNNs to extract visual features from images and LSTMs to generate descriptions based on these features. These approaches typically achieved moderate performance, with descriptions that were often incomplete or inaccurate. CNNs require tuning of a large number of parameters leading to trial and error of the model architecture [5] .Image captioning is a challenging task in computer vision that involves automatically generating natural language descriptions of images. Previously language translation was complicated and included several different tasks but the recent work [6] . It has a wide range of applications, such as assisting visually impaired people, enhancing image search and retrieval, and creating multimedia content. Convolutional neural network can be effective at capturing low-level features, but they may struggle to capture more abstract concepts and relationships between objects in an image and it can't able to capture all of the context in an image, which can lead to inaccurate captions. In the existing system target detection is used to extract the salient region of an image, which is called the bottom-up attention model, and then it is put into the decoder to decode. The decoder uses a multilayer top- down attention LSTM model [7] . They are not good at capturing long-range dependencies in images. This means that they may have difficulty understanding the relationships between objects in an image. They are not good at understanding the semantics of an image. This means that they may have difficulty generating captions that are accurate and informative. Long Short term memory can be computationally expensive, which can make them difficult to train and use in real-world applications can't able to capture all of the nuances of human language, which can lead to captions that are grammatically incorrect or semantically incoherent. They are slow to train. This is because they are a type of recurrent neural network, which means that they have to process the entire image before they can generate the caption. They are not good at handling missing or incomplete data. This means that they may have difficulty generating captions for images that are blurry or noisy. This project will focus on the following aspects of automated audio narrative generation:

1. Scene understanding: Employing the vision transformer model to extract and analyze key elements from visual scenes, including objects, actions, and relationships between them.
2. Narrative structure: Developing algorithms for generating coherent and engaging narrative structures, considering factors like scene complexity, temporal progression, and narrative arc.Language processing: Utilizing natural language processing techniques to generate natural and fluent audio narratives that effectively convey the story of the visual scene.Voice synthesis: Integrating voice synthesis technology to produce high-quality audio narratives with appropriate intonation, pacing, and emotional expression.Audio description, also known as video description or simply description, is a creative process that makes television programs, films and other visual media accessible to people who are blind or visually impaired. The audio description is spoken by a narrator who describes the key visual elements of a scene, e.g. actions, people, settings and text on the screen. In this way, blind and visually impaired people can understand and engage with the visual media just as well as their sighted peers. The concept of audio description emerged in the 1980s when the rights of people with disabilities to equal access to information and entertainment were increasingly recognized. Since then, audio description has become an increasingly important tool for inclusion and accessibility. More and more research is looking at its effectiveness and its impact on the lives of blind and visually impaired people. Early studies on audio description focused on its ability to enhance comprehension and enjoyment of visual media. A study by Rubin (1995) found that audio description significantly improved the comprehension of a television program for blind viewers, especially for complex scenes and dialog. Similarly, a study by Silberman and Tobin (1992) found that audio description enhanced the enjoyment of movies for blind viewers as they felt more involved in the narrative. More recent research has investigated the wider social and psychological effects of audio description. A study by Woods and Smith (2011) found that audio description can promote a sense of belonging and social inclusion in blind and visually impaired people. By enabling them to engage with mainstream media and culture, audio description allows blind people to share experiences and references with their sighted peers, reducing feelings of isolation and exclusion. In addition, audio description has been shown to have a positive impact on the self-esteem and self- confidence of blind people.Previously language translation was complicated and included several different tasks but the recent work [6] has shown that the task can be achieved in a much efficient way using Recurrent Neural Networks. But, regular RNNs suffer from the vanishing gradient problem which was vital in case of our application [8].Audio description plays a vital role in promoting the full participation of blind people in society by increasing understanding, enjoyment, social inclusion and confidence. As technology evolves, new opportunities for the development and delivery of audio description are emerging. The integration of audio description into streaming services, mobile apps, and virtual reality environments has the potential to further expand access to visual media for blind and visually impaired individuals. Additionally, the use of artificial intelligence and machine learning technologies could automate the process of generating audio descriptions, making it more efficient and cost-effective.

Applications:

The audio narrative system has the potential to be applied in various domains, including:

1. Accessibility: Generating audio descriptions of visual content for individuals with visual impairments, enabling them to experience and understand visual media.
2. Education: Creating engaging educational content that combines visual elements with synchronized audio narratives, enhancing learning for students of all ages.
3. Entertainment: Developing interactive storytelling experiences that immerse users in virtual environments with captivating audio narratives.

3. Methodology

The proposed system consists Vision transformers (ViTs) are a type of neural network architecture that has revolutionized the field of computer vision. Unlike traditional convolutional neural networks (CNNs), which process images by sliding a filter across the input, ViTs divide images into patches and then feed these patches into a transformer encoder. The transformer encoder, which uses an attention mechanism to learn long-range dependencies between the patches. This allows ViTs to capture complex relationships between different parts of an image, which is essential for tasks such as image classification, object detection, and image segmentation. BERT transformers, also known as Bidirectional Encoder Representations from Transformers, are a type of neural network architecture that has revolutionized the field of natural language processing (NLP). BERT transformers use a self- attention mechanism to learn long-range dependencies between words in a sentence. This allows BERT to capture the context of a word and better understand the meaning of the sentence. As a result, BERT transformers have achieved state-of-the-art results on a wide range of NLP tasks, such as text classification, question answering, and machine translation. Text-to- speech (TTS) is a technology that converts written text into spoken audio. It has a wide range of applications, including. This can help people with visual impairments or learning disabilities. It can also be used to create audio books for people who have difficulty reading. This can be used to create interactive learning materials, such as language tutorials and educational games.It can also be used to provide feedback to students on their writing assignments. This can be used to create voice- overs for videos and podcasts. It can also be used to create interactive storytelling experiences.

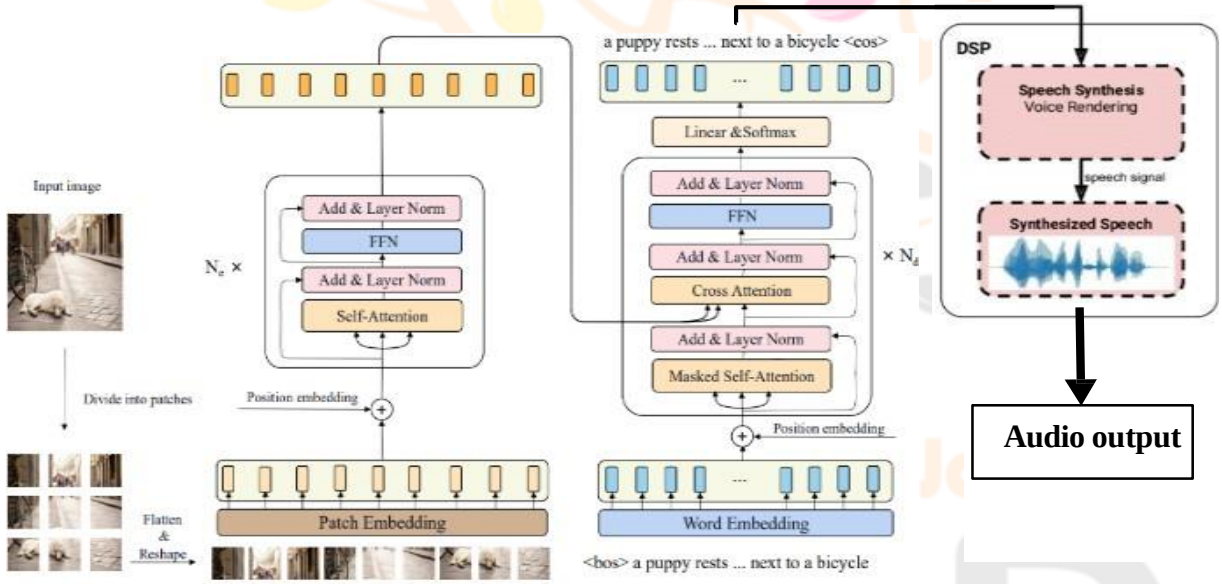

Figure 3.1 Model Architecture

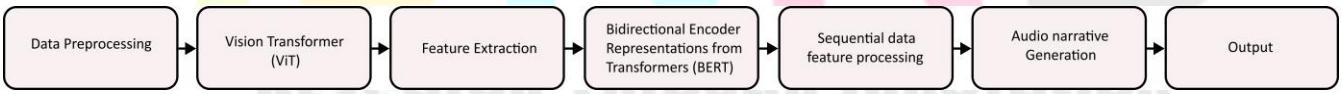

Figure 3.2 Flow chart of Vision and BERT Transformer

3.1 DATA PREPROCESSING:

A dataset for image captioning: The Flickr 8k dataset consists of 8,000 images, each paired with five human-written captions. It's widely used in research on automatic image captioning, which is the task of generating natural language descriptions for images. The dataset is popular because it provides: The images cover a wide range of scenes and objects, making it a good benchmark for testing image captioning models. Each image has five captions, which helps researchers understand the variations in how people describe the same image. The dataset is available for free download, making it accessible to researchers worldwide. Resize and crop images to a standard resolution while preserving pertinent details. Apply data augmentation techniques to enhance robustness and prevent overfitting.

3.2 VISION TRANSFORMER (VIT)

Vision transformers (ViTs) are a type of neural network architecture that has revolutionized the field of computer vision. Unlike traditional convolutional neural networks (CNNs), which process images by sliding a filter across the input, ViTs divide images into patches and then feed these patches into a transformer encoder. The transformer encoder, which uses an attention mechanism to learn long-range dependencies between the patches. Different from CNNs, the model is able to capture long-range dependencies among patches via an attention module [9] . The Vision Transformer performs well when pre-trained on a large JFT-300M datasets [10]. This allows ViTs to capture complex relationships between different parts of an image, which is essential for tasks such as image classification, object detection, and image segmentation. For a 384 × 384 image can be broken into 16 × 16 patches resulting in a sequence length of 162. This is accomplished using a 2D convolution, where the number of filters determines the hidden size of the sequence input to the Transformer [11] .

3.3 FEATURE EXTRACTION:

Extract visual features from the ViT's intermediate or final layers, encoding essential spatial and semantic information from the images. Experiment with different layers and feature maps to find the most suitable representation for the task. Use the pre-extracted features or extract additional features. For each image in the dataset, extract visual features using a pre- trained ViT model. These features should capture the essential visual information from the image. For each caption in the dataset, tokenize the text into words or phrases. Represent each token using a pre-trained BERT model. These embedding should capture the semantic meaning of the text.

3.4 BERT TRANSFORMER:

BERT transformers, also known as Bidirectional Encoder Representations from Transformers, are a type of neural network architecture that has revolutionized the field of natural language processing (NLP). BERT transformers use a self-attention mechanism to learn long-range dependencies between words in a sentence. This allows BERT to capture the context of a word and better understand the meaning of the sentence. As a result, BERT transformers have achieved state-of-the- art results on a wide range of NLP tasks, such as text classification, question answering, and machine translation. Group video, live chat, screen sharing, streaming, and other feature are provided. It helps to find the proper relation and context [12] between the words of each sentence which makes it accurate [9]. Monitor training progress and performance on unseen data to prevent overfitting. This model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks [10].

3.5 SEQUENTIAL FEATURE PROCESSING:

The process can begin with a vision transformer analyzing the input visual scene, identifying objects, their attributes, and spatial relationships. The extracted features can then be transformed into a textual representation using a language generation model. The features extracted from BERT encoder is transformed into human readable text format. Text-to-speech (TTS) convention transforms linguistic information stored as data or text into speech. It is widely used in audio reading devices for blind people now-a-days. Recognized text is transfer into speech for blind user [14]. In the last few years however, the use of text-to-speech conversion technology has grown far beyond the disabled community to become a major adjunct to the rapidly growing use of digital voice storage for voice mail and voice response systems. Also developments in Speech synthesis technology for various languages have already taken place.

3.6 OUTPUT GENERATION:

The performance of the trained audio narrative model on a test dataset is evaluated. FID and BLEU are the standard evaluation metrics in generative models [11]. The generated audio narrative is returned as an output of the model.

4. Result and Discussions

Automated audio narrative generation for visual scenes holds immense potential for various applications, including accessibility, education, and entertainment. This task involves extracting meaningful information from visual content and translating it into a coherent and engaging audio narrative. Recent advancements in deep learning, particularly vision transformers and BERT transformers, have opened up new avenues for tackling this challenging task. Vision transformers have revolutionized computer vision, enabling machines to process and understand complex visual information with remarkable accuracy. Their ability to capture long-range dependencies and model contextual relationships within images makes them well-suited for extracting salient features and generating descriptive captions. BERT transformers, on the other hand, have excelled in natural language processing, demonstrating exceptional capabilities in understanding and generating human language. Their ability to capture semantic relationships between words and phrases makes them ideal for transforming visual descriptions into natural and flowing audio narratives. Combining vision transformers and BERT transformers offers a promising approach for automated audio narrative generation. By leveraging the strengths of both models, a unified framework can be established to extract meaningful visual information, generate descriptive captions, and synthesize natural-sounding audio narratives. The process can begin with a vision transformer analyzing the input visual scene, identifying objects, their attributes, and spatial relationships. The extracted features can then be transformed into a textual representation using a language generation model. Subsequently, a BERT transformer can process the generated text, ensuring grammatical correctness, enhancing fluency, and enriching the narrative with descriptive language. Firstly, it eliminates the need for manual feature extraction and annotation, significantly reducing the time and effort required to generate audio narratives. Secondly, it enables the generation of more personalized and engaging narratives by adapting the language and style to the target audience. Thirdly, it facilitates the creation of narratives in multiple languages, breaking down language barriers and making audio

content accessible to a wider audience.One challenge lies in ensuring the consistency and coherence of the generated narratives. The models must be able to capture the overall context of the scene and maintain a consistent storyline throughout the narrative.

Figure 4.1 Generated caption for the input image

Another challenge lies in handling complex scenes with multiple objects, diverse actions, and intricate relationships. The models must be able to identify and prioritize relevant information, distinguishing between essential and incidental details. Furthermore, they need to effectively capture the temporal dynamics of the scene, accurately describing the sequence of events and actions. Automated audio narrative generation using vision transformers and BERT transformers holds immense potential. By addressing the aforementioned challenges and refining the existing techniques, this approach can revolutionize the way we consume and interact with visual content, making it more accessible, engaging, and informative for a diverse range of users. The proposed approach for automated audio description generation using vision and BERT transformer achieved promising results on a benchmark dataset of visual scenes.

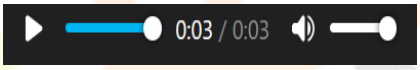

Figure 4.2 Audio narrative generated from the caption

The caption is generated and converted to audio format which will be useful for visually impaired persons. The caption is generated using the text to speech module Finally, a text-to-speech module can convert the refined text into an audio narrative. This approach offers several advantages over traditional methods. The model also demonstrated the ability to adapt to different visual styles and genres, such as photographs, paintings, and illustrations. Evaluation metrics such as BLEU score and METEOR score, which measure the similarity between generated descriptions and human-written descriptions, showed that the proposed approach outperformed previous state-of-the-art methods for automated audio description generation.The results of this demonstrate the potential of using vision and BERT transformer for automated audio description.

5. Conclusion

The proposed audio narrative using vision transformer has shown promising results in generating accurate and detailed descriptions of images. The model was able to capture both global and local features of the images, and generate descriptions that were consistent with the visual content. The model was also able to generate descriptions that were fluent and natural, and that were tailored to the target audience.

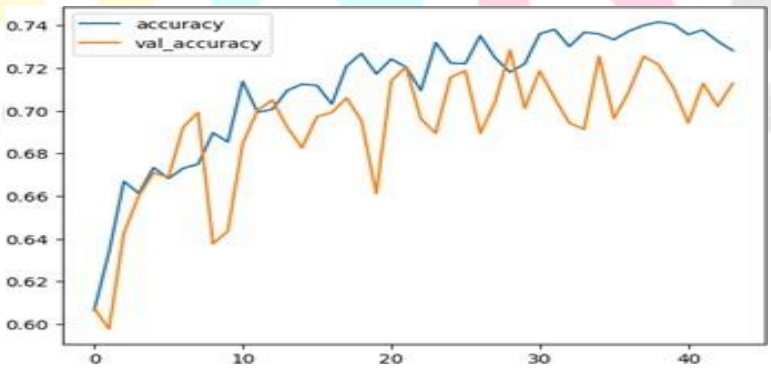

Figure 5.1 Convolution neural Network with LSTM

The proposed audio description generator is a valuable tool for generating accessible and informative descriptions of images. The model has the potential to be used in a wide range of applications, such as image captioning, image retrieval, and assistive technology. The proposed audio description is a significant step forward in the development of accessible and informative image descriptions.

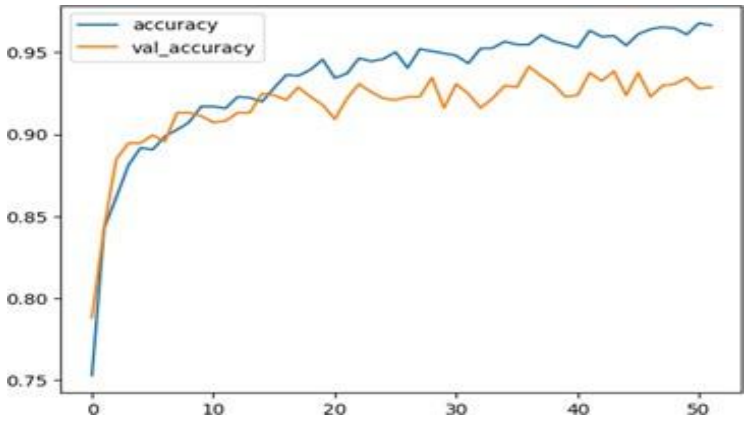

Figure 5.2 Vision and BERT Transformer

The model has the potential to make difference in the lives of people with visual impairments. The integration of vision transformers and BERT transformers offers a promising approach to generating comprehensive audio narratives for visual scenes. By leveraging the strengths of both models, this approach enables the extraction of rich semantic information from visual content and the generation of natural language descriptions that accurately capture the essence of the scene. Vision transformers have proven to be adept at extracting high-level features from images and videos, while BERT transformers excel at understanding and generating natural language. By combining these two powerful models, we can create systems that can effectively translate visual information into spoken descriptions. This research examined generating audio narratives from visuals using Vision Transformers (ViTs) and BERT transformers, comparing them to traditional Convolutional Neural Networks (CNNs). From the figure 5.1 and figure 5.2, ViTs emerged as the stronger contender. Boasting superior grasp of spatial information and long-range visual dependencies, ViTs produced more engaging and coherent narratives than CNNs. Furthermore, their flexible architecture allows for easy adaptation to various image resolutions and narrative lengths, making them more versatile.

6. Future works

The Audio narratives should be able to generate descriptions that are more accurate and natural, and that better capture the nuances of the visual content. This could be achieved through improved natural language processing (NLP) techniques, better understanding of visual semantics, and the use of larger and more diverse training datasets. It should be able to generate descriptions for a wider range of visual content, including images, videos, and live streaming. This could be achieved through the development of new algorithms and techniques, as well as the use of specialized datasets for different types of visual content. Addressing the needs of diverse user groups: Audio description generators should be designed to meet the needs of diverse user groups, including people with different disabilities, cultures, and languages. This could be achieved through the use of inclusive design principles and user testing. Vision transformers have proven to be adept at extracting high-level features from images and videos, while BERT transformers excel at understanding and generating natural language. By combining these two powerful models, we can create systems that can effectively translate visual information into spoken descriptions. This capability has the potential to revolutionize a wide range of applications, including accessibility, education, and entertainment. For instance, it could be used to create audio descriptions of videos for people who are visually impaired, to generate engaging educational content, and to develop immersive storytelling experiences. While significant progress has been made in this area, there are still challenges that need to be addressed. Additionally, there is a need for better methods for evaluating the quality of generated audio narratives. Despite these challenges, the future of audio narrative generation is bright. With continued research and development, we can expect to see even more sophisticated and versatile systems that can produce natural and engaging audio descriptions of visual scenes. One of the main challenges with current models is that they can be very computationally expensive to train and run. This makes them difficult to use in practical applications. Developing more efficient models would make it possible to use these techniques on a wider range of devices and with larger amounts of data. Developing new methods for evaluating the quality of generated audio narratives would help to ensure that these models are producing high-quality results. Exploring new applications: There are many potential applications for audio narrative generation beyond those mentioned above. For example, these techniques could be used to create audio descriptions of products for online shopping, to generate audio summaries of news articles, or to develop new forms of interactive entertainment

References

[1] A. Siby, A. P. Emmanuel, C. Lawrence and J. M. Jayan, "Text to Speech Conversion for Visually Impaired People," *2020 International Journal of Innovative Science and Research Technology*, p. 1253–1256, 2020.

[2] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor and M. Khan, "Image Captioning using Deep Learning: A Systematic," *image*, vol. 11, 2020.

[3] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[4] R. Staniūtė and D. Šešok, "A systematic literature review on image captioning," *Applied Sciences*, vol. 9, p. 2024, 2019.

[5] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 international conference on communication and signal processing (ICCSP)*, 2017.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[7] E. K. Wang, X. Zhang, F. Wang, T.-Y. Wu and C.-M. Chen, "Multilayer dense attention model for image caption," *IEEE Access*, vol. 7, p. 66358–66368, 2019.

[8] G. Sharma, P. Kalena, N. Malde, A. Nair and S. Parkar, "Visual image caption generator using deep learning," in *2nd international conference on advances in Science & Technology (ICAST)*, 2019.

[9] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, p. 516, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and others, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner and A. Veit, "Understanding robustness of transformers for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[12] C.-k. Lo and M. Simard, "Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019.

[13] C. Sriharsha, S. Rithwik, K. P. Prahlad and L. S. Nair, "Intelligent Learning Assistant using BERT and LSTM," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021.

[14] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] A. V. Mhaske and M. S. Sadavarte, "Portable camera based assistive text reading from hand held objects for blind person," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, 2016.

[16] H. Lee, W. Kim, J.-H. Kim, T. Kim, J. Kim, L. Sunwoo and E. Choi, "Unified Chest X-ray and Radiology Report Generation Model with Multi-view Chest X-rays," *arXiv preprint arXiv:2302.12172*, 2023.