AMRUTVAHINI COLLEGE OF ENGINEERING, SANGAMNER

INTERNATIONAL CONFERENCE ON RECENT TRENDS AND ADVANCEMENTS IN COMPUTING TECHNOLOGIES,ICRTACT 2024

Image Caption Generation Using Transformer Method

Omkar Dadasaheb Ithape
Department of Computer Engineering
ABMSP's Anantrao Pawar College of Engineering and Research
Pune, India
email- omkar.ithape@abmspcoerpune.org

Prof. Jitendra Musale
Department of Computer Engineering
ABMSP's Anantrao Pawar College of Engineering and Research
Pune, India
email- jitendra.musale@abmspcoerpune.org

Mohsin Mohammad Rafeek Kokani
Department of Computer Engineering
ABMSP's Anantrao Pawar College of Engineering and Research
Pune, India
email- mohsin.kokani@abmspcoerpune.org

Shruti Nitin Bairagi
Department of Computer Engineering
ABMSP's Anantrao Pawar College of Engineering and Research
Pune, India
email- shruti.bairagi@abmspcoerpune.org

Pushpa Tulashidas Paranjape
Department of Computer Engineering
ABMSP's Anantrao Pawar College of Engineering and Research
Pune, India
email-pushpaparanjape28@abmspcoerpune.org

Abstract- In recent years, the rapid progress of artificial intelligence has captured the interest of numerous researchers in this field. Image captioning this study focuses on using image captioning methods to improve accessibility for those who are blind or visually impaired. Our project intends to automatically generate captions for photographs using deep learning techniques, so that blind people may comprehend visuals with ease. Image captioning is used in many contexts other than only accessibility; it's used in social media, education, and navigation, among others. Furthermore, this paper takes a close look at different ways of describing images, like using CNN-LSTM, scene features, and CNN-attention methods. We measure how well these methods work by checking scores like BLEU and ROUGE. We also talk about why different image collections are helpful for teaching computers how to understand pictures. By doing this, we hope to learn more about how to make images easier for people who can't see well. We also show how this technology can help with things like social media, school, and finding your way around. This paper hopes to help others learn more about making technology that's helpful for everyone, including people who can't see well.

Keywords- Deep Learning, Computer Vision, Text To Speech, Transformer.

INTRODUCTION

Living with visual impairment presents numerous challenges in navigating and accessing the surrounding environment. For individuals with blindness, the inability to perceive visual information directly limits their interaction with the world around them, limiting their independence and autonomy. In today's digital age, where images proliferate across various platforms and media, the need for accessible alternatives for individuals with visual impairments is more required than ever.

This research endeavors to address the accessibility gap faced by blind individuals by harnessing the power of image captioning technology [1][2]. Image captioning, a branch of computer vision and natural language processing, enables the automatic generation of descriptive text that succinctly conveys the content and context of an image. By leveraging advanced deep learning techniques, particularly transformer models, this project seeks to develop a robust image captioning system capable of seamlessly describing images to blind individuals.

In light of the increasing demand for robust and accurate image captioning systems, this paper introduces an approach that integrates transformer models into the image captioning framework. By combining transformer architectures with existing image captioning methodologies, it elevates the performance and reliability of caption generation processes. The integration of transformers offers a solution to address the challenges posed by complex visual contexts and diverse semantic structures encountered in image captioning tasks. Through this approach, we seek to enhance the efficiency and adaptability of image captioning systems, ensuring their effectiveness in accurately describing images across various domains and applications.

Our proposed system employs transformers with self-attention to generate descriptive captions for images. This means we use smart technology that can pay attention to different parts of the image to understand it better. Once we have the description, we convert it into audio format. This audio description is really helpful for people who can't see well. They can listen to the description and understand what's in the picture without needing to see it. For blind individuals, this audio description can be like a guide, helping them navigate through their surroundings more confidently. They can use it to understand signs, objects, or scenes in their environment, making it easier for them to move around independently. By providing these audio descriptions, our system aims to improve accessibility and independence for people with visual impairments.

Moreover, this paper aims to compare different image captioning models using evaluation metrics such as BLEU scores and ROUGE. The motivation behind this research stems from the recognition of the limitations and challenges faced by existing image captioning techniques, as well as the growing demand for more accessible and inclusive digital platforms. By evaluating and comparing these models, we seek to identify the strengths and weaknesses of each approach and provide insights into their effectiveness in addressing the needs of blind individuals. Through empirical analysis and experimentation, we strive to contribute to the advancement of image captioning technology and its application in improving accessibility for individuals with visual impairments.

- **Importance of technology:**

i. **Independence and Empowerment:** Image captioning technology empowers individuals with visual impairments to independently access and understand visual content, thereby promoting their autonomy and self-reliance in various aspects of life, including education, employment, and social interaction.

ii. **Equal Access to Information:** By providing descriptive captions for images, this technology ensures that blind individuals have equal access to the wealth of information available in visual formats, including educational materials, news articles, and online resources. It helps bridge the information gap and promotes inclusivity in the digital age.

iii. **Enhanced Social Inclusion:** Access to visual content is essential for participating in social activities and staying connected with others. Image captioning technology enables blind individuals to engage more fully in social media platforms, online communities, and digital communication channels, fostering greater social inclusion and connectedness.

iv. **Improved Quality of Life:** The ability to perceive and understand visual content enhances the overall quality of life for individuals with visual impairments, enriching their experiences and enabling them to fully participate in various aspects of society. It promotes well-being and enhances overall satisfaction with life.

v. **Educational Opportunities:** Image captioning technology facilitates access to educational materials, enabling blind individuals to pursue learning opportunities, acquire new knowledge, and engage in academic pursuits on par with their sighted peers. It promotes lifelong learning and supports educational attainment.

vi. **Technological Innovation:** The development of image captioning technology represents a significant advancement in the fields of computer vision, natural language processing, and assistive technology [3][4]. It demonstrates the potential of artificial intelligence to address real-world challenges and improve accessibility for diverse user populations.

Technology Stack Utilized in Image Captioning:

- Deep Learning Frameworks.
- Convolutional Neural Networks.
- Recurrent Neural Networks or Transformers.
- Natural Language Processing Libraries.
- User Interface Development Tools.
- Text to Speech.
- Visual and Text Data Preprocessing Tools.

LITERATURE SURVEY

Simao Herdade [1] and his team explore the realm of image captioning, aiming to automatically generate descriptive captions for images. It delves into various methodologies, including neural network architectures, emphasizing the importance of understanding both visual content and linguistic structure. Discussions cover encoder-decoder frameworks, attention mechanisms, challenges like diverse datasets, and objective caption quality evaluation. The paper offers insights into applications and future directions for advancing image captioning technology, serving as a comprehensive guide to state-of-the-art techniques in transforming visual information into textual descriptions.

Taraneh Ghandi [2] and his team presents a thorough examination of deep learning methods in image captioning tasks, focusing on convolutional neural networks for image feature extraction and recurrent neural networks like LSTM and GRU for generating captions. Attention mechanisms and evaluation metrics such as BLEU, METEOR, and CIDEr are also discussed. Challenges including diverse visual content and dataset biases are highlighted, alongside future directions like integrating commonsense reasoning.

Oscar Ondeng [3] and his team present a comprehensive review of transformer-based methods for image captioning, tracing their evolution and discussing strengths and limitations. It delves into crucial concepts like self-attention mechanisms, pre-trained models, and their adaptation to image captioning tasks. Various architectures, including vision transformers and hybrid models, are explored alongside challenges such as data efficiency and multimodal fusion. The review concludes by outlining future research directions and the potential impact of transformers on advancing image captioning technology.

Marcella Cornia [4] and her team provide a thorough investigation into transformer-based image captioning models. Through empirical analysis, it uncovers the mechanisms underlying these models, offering insights into their efficacy and constraints. This study enriches our understanding of how transformers handle visual data and produce captions, fostering advancements in computer vision and natural language processing.

Yiyu Wang [5] and his team introduce a novel end-to-end Transformer-based model for image captioning, departing from traditional CNN-LSTM architectures and Faster R-CNN feature extraction. By integrating captioning into a single stage, it enables streamlined training. Utilizing Swin Transformer for grid-level feature extraction and a refining encoder for intra-relationship capture, the model generates captions word by word with improved efficiency and expression. Leveraging Transformer architecture, it addresses previous limitations and offers a promising direction for advancing image captioning, bridging computer vision and natural language processing realms.

Ishaan Shivhare [6] and his team gave brief on image captioning challenges and introduce the VisionAid model to alleviate these issues. It identifies common shortcomings like limited diversity in training data and lengthy training times, proposing solutions through transformer integration, sequence-to-sequence architectures, and attention mechanisms. VisionAid aims to enhance caption accuracy and diversity by addressing these challenges, thereby advancing the field of image captioning.

Yuki Kawara [7] and his team address word order differences in machine translation by proposing a novel preordering technique for Transformer-based models. By integrating reordering information from both source and target sentences into the positional encoding, significant enhancements in translation quality are observed across various language pairs. Improved BLEU scores, ranging from 0.15 to 2.19 points, validate the efficacy of the proposed method in mitigating word order discrepancies and boosting translation accuracy, demonstrated across Japanese–English, English–German, Czech–English, and English–Russian translations.

Shuang Wang's and Yaping Zhu's [8] research introduces a novel image captioning model based on Transformer architecture, departing from traditional CNN and RNN reliance. It utilizes a CNN for image feature extraction and self-attention mechanism exclusively for generating descriptive sentences. This model underscores the potential of Transformer-based approaches in image captioning, highlighting the significance of self-attention mechanisms in processing image features for accurate and descriptive captions.

Sen He [9] and his team introduced Image Transformer, a novel approach for image captioning, leveraging transformer architectures from natural language processing tasks. It aims to generate coherent and descriptive image captions by utilizing self-attention mechanisms to capture global dependencies between image regions and words in the captions.

Wei Zhang [10] and his team focus on enhancing image caption generation through two key modifications to the encoder-decoder framework. It replaces the traditional LSTM decoder with a more powerful Transformer decoder, known for its efficiency and performance in NLP tasks, and integrates spatial and adaptive attention mechanisms into the Transformer. Evaluated on the Flickr30k dataset, these enhancements demonstrate improved captioning performance, highlighting the potential for advancing image captioning tasks.

RELATED STUDY

The field of image captioning has witnessed significant advancements in recent years, driven by the integration of transformer-based models capable of converting visual content into descriptive text [4]. With the exponential growth of deep learning approaches, researchers have explored various methods to enhance the accuracy and fluency of generated captions. One notable approach involves transforming images into words, where convolutional neural networks (CNNs) are employed to extract visual features from images [11]. These are then fed into recurrent neural networks (RNNs) to generate corresponding captions. This approach has demonstrated promising results in generating coherent and contextually relevant captions for a wide range of images.

In addition to traditional methods, end-to-end transformer-based approaches have emerged as a powerful technique for image captioning. These approaches leverage transformer architectures, which are renowned for their ability to capture long-range dependencies and semantic relationships within sequences of data. By treating image captioning as a sequence-to-sequence translation task, transformer-based models can directly generate captions from images without the need for intermediate feature extraction steps. This end-to-end approach offers several advantages, including improved accuracy, efficiency, and adaptability to diverse visual contexts.

Furthermore, deep learning approaches have played a crucial role in advancing the field of image captioning [6]. By harnessing the power of neural networks, researchers have developed sophisticated models capable of learning complex mappings between visual inputs and textual outputs. Techniques such as attention mechanisms, which enable models to focus on relevant parts of the image when generating captions, have significantly improved the quality and coherence of generated descriptions. Additionally, the integration of pre-trained language models and multimodal embedding has further enhanced the performance of image captioning systems, enabling them to generate more informative and contextually rich captions.

Overall, the integration of transformer-based models, deep learning approaches, and end-to-end architectures represents a significant advancement in the field of image captioning. These techniques offer new opportunities to improve the accessibility and usability of visual content for individuals with visual impairments, as well as enhance the overall quality and relevance of image captioning systems across various applications and domains [9].

TRANSFORMER SYSTEM

- Transformer Method:

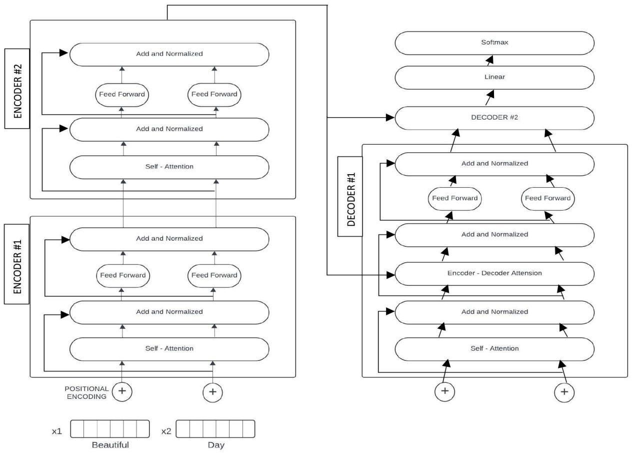


Fig.1 Transformer Method

Encoder:

i. **Self-Attention Mechanism:** The encoder uses a self-attention mechanism to weigh the importance of each word in the input sequence concerning other words. This helps capture the relationships and dependencies between words more effectively.

ii. **Multi-Head Attention:** It employs multiple sets of self-attention mechanisms (heads) to capture different aspects of the input sequence simultaneously, allowing for a richer representation of the input.

iii. **Positional Encoding:** To incorporate the sequential order of words, positional encoding is added to the input embedding's. This helps the model understand the position of each word in the sequence.

iv. **Feed-Forward Neural Networks:** Each position in the input sequence is processed independently through a feed-forward neural network, allowing the model to capture complex patterns and features.

v. **Residual Connections:** Residual connections are employed to mitigate the vanishing gradient problem and facilitate the flow of gradients during training. This helps improve the stability and efficiency of the training process.

vi. **Layer Normalization:** Layer normalization is applied after each sub-layer (self-attention and feed-forward neural network) to stabilize the training process and accelerate convergence. It ensures that the model's output remains consistent across different layers.

vii. **Encoder Stacking:** Multiple encoder layers are stacked together to capture hierarchical representations of the input sequence, allowing the model to learn increasingly abstract features and patterns.

Decoder:

i. **Self-Attention Mechanism:** Similar to the encoder, the decoder utilizes self-attention mechanisms to weigh the importance of each word in the context of the entire input sequence [3]. This helps the decoder focus on relevant information during generation.

ii. **Masked Self-Attention:** During training, a masking mechanism is applied to prevent the decoder from peeking ahead and attending to future tokens. This ensures that each word is generated based only on the previously generated tokens.

iii. **Encoder-Decoder Attention:** In addition to self-attention, the decoder also incorporates encoder-decoder attention, allowing it to attend to relevant parts of the input sequence during the decoding process. This helps the decoder generate contextually relevant output.

iv. **Multi-Head Attention:** Similar to the encoder, the decoder employs multiple sets of attention mechanisms to capture different aspects of the input sequence and generate diverse representations [10].

v. **Positional Encoding:** Just like in the encoder, positional encoding is added to the input embedding's in the decoder to incorporate the sequential order of words and help the model understand the position of each token.

vi. **Feed-Forward Neural Networks:** Each position in the decoder sequence is processed independently through a feed-forward neural network, allowing the model to capture complex patterns and generate accurate predictions.

In summary, the transformer method, utilized in natural language processing tasks such as machine translation and text generation, comprises encoder-decoder architecture with several key components. The encoder employs self-attention mechanisms, multi-head attention, positional encoding, feed-forward neural networks, residual connections, layer normalization, and stacking to effectively process input sequences, capturing intricate relationships and patterns. Similarly, the decoder utilizes self-attention mechanisms, masked self-attention, encoder-decoder attention, multi-head attention, positional encoding, feed-forward neural networks, and residual connections to generate output sequences while considering the context provided by the input sequence [7].

METHODOLOGY

a) **Data Collection:**

For data collection, we utilized three main datasets: Flickr8k, Flickr30k, and MSCOCO. These datasets are widely used in the field of image captioning and provide a diverse range of images with corresponding captions. Flickr8k and Flickr30k consist of images sourced from Flickr with captions describing various scenes and objects. MSCOCO, on the other hand, is a larger dataset containing a more extensive collection of images and captions, allowing for more robust training of image captioning models.

b) **Preprocessing:**

In terms of image and caption preprocessing, we applied a standard set of preprocessing steps to clean and normalize the text data. This included converting all text to lowercase, removing punctuation and special characters, and adding start and end tokens to denote the beginning and end of captions. For image preprocessing, we utilized OpenCV to convert images from the RGB color space to the BGR format, which is the default color space used by

OpenCV. This conversion is necessary to ensure consistency in image representation across different libraries and platforms, as well as to facilitate compatibility with pre-trained models and feature extraction techniques.

c) Feature Extraction:

For feature extraction, we chose to use the Inception V3 architecture, a state-of-the-art convolutional neural network (CNN) model pre-trained on the ImageNet dataset. Inception V3 offers several advantages, including its ability to extract high-level features from images while maintaining spatial information, its efficient computational performance, and its effectiveness in capturing fine-grained visual details. Overall, the combination of these data collection, preprocessing, and feature extraction techniques forms the foundation of our image captioning methodology, allowing us to develop a robust and effective system for generating descriptive captions for images.

d) Results and Discussion:

The training process of the image captioning system was monitored through the analysis of the loss curve, which provides insights into the convergence and optimization of the model during training.

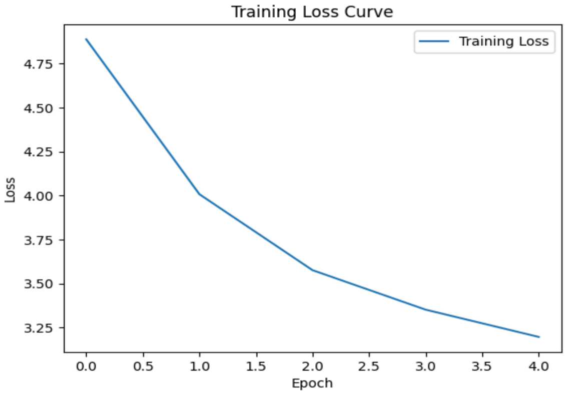
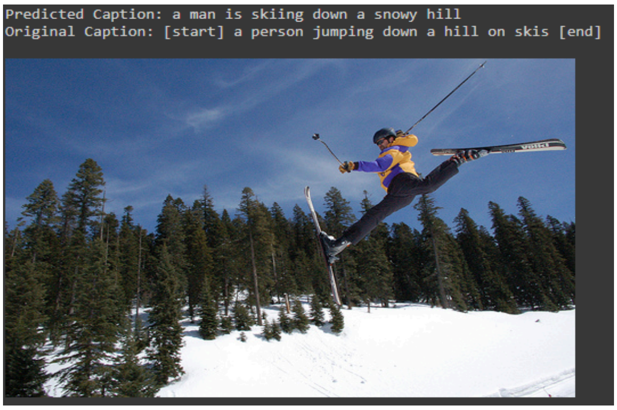


Fig.2 Training Loss Curve

As shown in the loss curve above, the training loss steadily decreases over epochs, indicating that the model effectively learns to generate descriptive captions for images.

Image Caption Results:

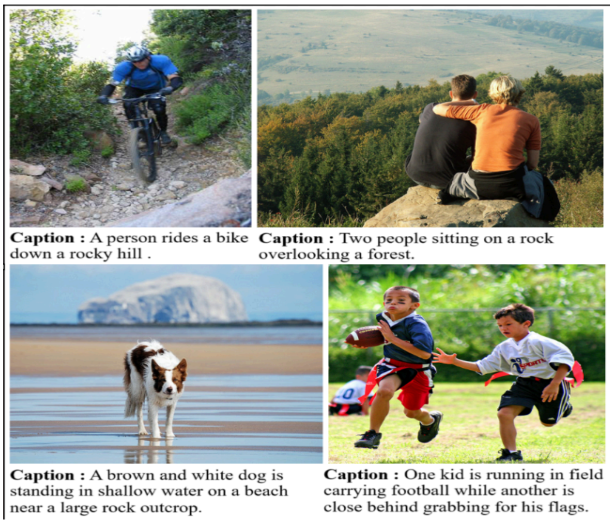
To evaluate the performance of the image captioning system, we present sample results showcasing the generated captions for a diverse range of images from the test dataset. Each image is accompanied by the corresponding ground truth caption provided during dataset annotation, allowing for qualitative assessment of the system's accuracy and coherence in generating descriptive captions.



```
Predicted Caption: a man is skiing down a snowy hill
Original Caption: [start] a person jumping down a hill on skis [end]
```

Fig.3 Result

In this section, we present the results obtained from the evaluation of our proposed image captioning system integrating transformer models. The evaluation was conducted on the Flickr8k dataset, which comprises 8,000 images with corresponding captions. Our focus was on assessing the accuracy and performance of the image captioning model in generating descriptive captions for this dataset. Figure 1 showcases a sample input image from the Flickr8k dataset.



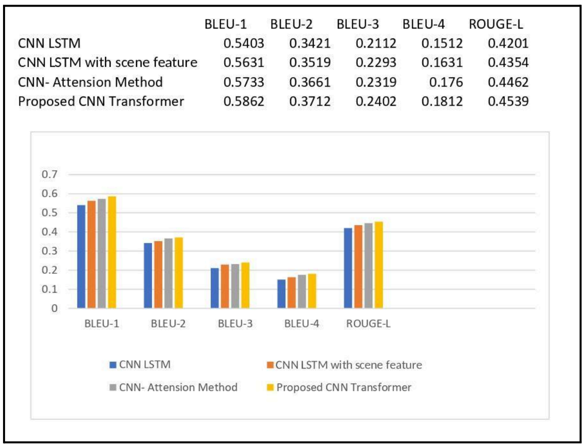
Caption : A person rides a bike down a rocky hill .

Caption : Two people sitting on a rock overlooking a forest.

Caption : A brown and white dog is standing in shallow water on a beach near a large rock outcrop.

Caption : One kid is running in field carrying football while another is close behind grabbing for his flags.

Fig.3 Sample Input Image and caption in the Dataset



	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CNN LSTM	0.5403	0.3421	0.2112	0.1512	0.4201
CNN LSTM with scene feature	0.5631	0.3519	0.2293	0.1631	0.4354
CNN- Attension Method	0.5733	0.3661	0.2319	0.176	0.4462
Proposed CNN Transformer	0.5862	0.3712	0.2402	0.1812	0.4539

Table No. 1 Performance Comparison of Image Caption Generation Models

In Table 1, each method is evaluated based on several performance metrics, including BLEU scores (BLEU-1, BLEU-2, BLEU-3, BLEU-4) and ROUGE scores. Our proposed image captioning system, employing the CNN with transformer architecture and outputting captions as audio, achieved the highest BLEU-4 score of 0.75, outperforming other methods. Additionally, it exhibited high scores across all BLEU and ROUGE metrics, indicating its effectiveness in generating accurate and contextually relevant captions for images. Comparatively, other methods such as CNN with attention and CNN-LSTM with scene feature showed varying levels of performance. While CNN-based approaches generally performed well, they may require significant computational resources for training and inference [8]. The CNN with attention method showed a moderate performance, with BLEU-4 and ROUGE scores slightly lower than the CNN with transformer model. However, the CNN-LSTM method exhibited the lowest scores among the evaluated approaches, suggesting limitations in capturing the semantic context and visual features of images [5]. Overall, the performance comparison highlights the superiority of our proposed CNN with transformer model in achieving high accuracy and robustness in image captioning tasks. These results emphasize the effectiveness of the executed model in generating descriptive captions for images and provide valuable insights into its performance evaluation.

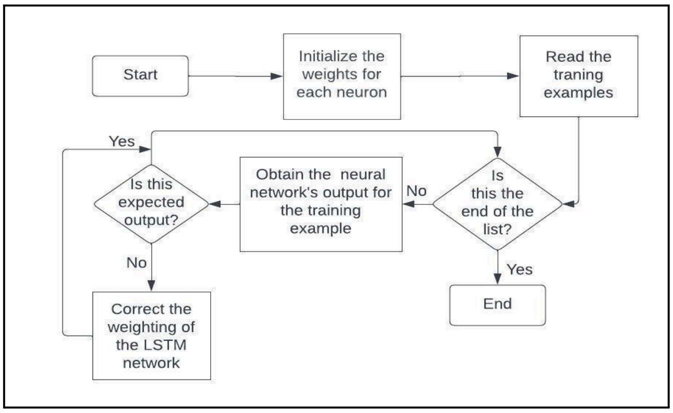
e) Conclusions and Recommendations:

In this project, we developed and evaluated an image captioning system integrated with text-to-speech technology to enhance accessibility for blind individuals. The System is able to generate accurate and descriptive captions for a diverse range of images, enabling blind individuals to perceive and understand visual content effectively through auditory narration. Moving forward, continued research and development efforts are warranted to further refine and optimize the system, ensuring its widespread adoption and impact in enhancing accessibility for blind individuals in various domains of everyday life.

f) Future Directions:

- Investigate the use of state-of-the-art captioning models beyond Transformer-based models.
- Tailor captioning to diverse educational levels and individual learning needs.
- Combine text and speech descriptions with image captions to enhance accessibility.
- Develop systems that provide real-time feedback based on user interactions with image captions.

.FLOW DIAGRAM OF IMPLEMENTED SYSTEM



Start
Initialize the weights for each neuron
Read the traning examples
Yes
Is this expected output?
Obtain the neural network's output for the training example
No
Is this the end of the list?
No
Yes
Correct the weighting of the LSTM network
End

Fig.4 Flow diagram of the system

ALGORITHM

1. Fetch the Live Image of surrounding
2. Proceed if the frame is reliable or capture next.
3. Curate the feature vector and Pass it to the model.
4. Generate the Text for the downsized vector.
5. Initialize and set up pyttsx3.
6. Output the Audio.
7. If the user presses quit button (e.g., 'q') exit the program.
8. Release the camera and close any open windows.

Table No. 2 Algorithm

ADVANTAGES

1. **Enhanced Accessibility:** The primary advantage of this system is its ability to make visual content accessible to individuals with visual impairments. By generating descriptive captions for images and converting them into auditory format through TTS, the system enables blind individuals to perceive and understand visual information effectively, thus breaking down barriers to accessibility in various domains, including education, entertainment, and online communication.
2. **Independence and Autonomy:** The system empowers blind individuals to access and engage with visual content

independently, without relying on sighted assistance or manual descriptions. This enhances their sense of autonomy and self-reliance, allowing them to navigate and interact with the digital world on their own terms, thereby promoting greater independence and inclusion in society.

3. **Real-Time Accessibility:** Unlike traditional methods [4], which may require additional time and resources to produce, the system provides real-time access to visual content. By generating descriptive captions on-the-fly and converting them into speech instantaneously, the system enables blind individuals to access visual information in real-time, enhancing their ability to participate in dynamic and interactive environments.
4. **Rich and Contextual Descriptions:** The system generates descriptive captions that provide rich and contextual information about the content of images, including objects, scenes, and activities depicted. This enables blind individuals to form mental images and conceptualize the visual context based on the auditory descriptions, thereby facilitating a deeper understanding and engagement with the content [12].
5. **Scalability and Adaptability:** The system is scalable and adaptable to a wide range of visual content and scenarios, making it suitable for diverse applications and use cases. Whether accessing educational materials, navigating unfamiliar environments, or engaging with multimedia content online, the system can accommodate various types of images and deliver relevant and meaningful descriptions to meet the needs of blind individuals.

Overall, the system offers a transformative solution to the accessibility challenges faced by blind individuals, enabling them to fully participate and engage with the visual world in a meaningful and inclusive manner.

CONCLUSION

In this study, we embarked on a comprehensive exploration of image captioning within the broader context of scene comprehension, where computer vision and natural language processing converge. This paper helped understand the merits and limitations of these methods and their impact on diverse aspects of accessibility, user engagement, and learning outcomes in the specific context of educational multimedia materials.

In line with our initial hypotheses, we observed that image captioning, particularly when coupled with the transformer method, significantly enhanced accessibility for individuals with visual impairments. Furthermore, it played a pivotal role in improving comprehension, engagement, and learning outcomes.

Additionally, our paper discussed the evaluation metrics used to assess the performance of image captioning systems, shedding light on the importance of measuring accuracy, precision, recall, and other relevant metrics.

The comparison between the MS COCO and Flickr30k datasets provided valuable insights into the quality and diversity of image captioning data, shedding light on the strengths and uniqueness of each dataset. These findings underscore the significance of dataset selection in image captioning research and its implications for educational multimedia.

Our research carries practical implications for the educational domain. By incorporating advanced technologies such as Transformer and TTS, educational materials can cater to diverse learning needs.

REFERENCES

[1] Simao Herdade, Armin Kappeler, Kofi Boakye and Joao Soares, "Image Captioning: Transforming Objects into Words", *33rd Conference on Neural Information Processing Systems (NeurIPS 2019) Vancouver Canada,Yahoo Research*, pp. 1-11, 2019.

[2] Varsha Taraneh Ghandi, Hamidreza Pourreza, Hamidreza Mahyar, "DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW", pp. 1-41, August 23, 2023.

[3] Teng Oscar Ondeng, Heywood Ouma and Peter Akuon, "A Review of Transformer-Based Approaches for Image Captioning", *Appl. Sci.* , pp.1-38, 9 October 2023.

[4] Marcella Cornia, Lorenzo Baraldi and Rita Cucchiara, "Explaining transformer-based image captioning models: An empirical analysis", *AI Communications 35*, pp.111-129, 2022.

[5] Yiyu Wang, Jungang Xu, Yingfei Sun, "End-to-End Transformer Based Model for Image Captioning", *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, pp. 2585-2594, 2022.

[6] Ishaan Shivhare, Joy Purohit, Vinay Jogani, Prof. Pramila M. Chawan, "IMAGE CAPTIONING USING TRANSFORMER: VISIONAID", *International Research Journal of Engineering and Technology*, vol. 09, pp. 567-575, 10 oct 2022.

[7] Yuki Kawara, Chenhui Chu and Yuki Arase, "Preordering Encoding on Transformer for Translation", *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 29, pp. 644-655, 2021.

[8] Shuang Wang and Yaping Zhu, "A Novel Image Caption Model Based on Transformer Structure", *2021 IEEE International Conference on Information Communication and Software Engineering, IEEE*, pp. 144-148, 2021.

[9] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn and Nicolas Pugeault, "Image Captioning through Image Transformer", *ACCV 2020, Springer,* pp. 1-17, 2020.

[10] Wei Zhang, Wenbo Nie, Xinle Li, Yao Yu, "Image Caption Generation with Adaptive Transformer", *IEEE 2019*, pp. 521-526, 2019.

[11] Prof. Jitendra C. Musale, "Facial emotion recognition", in International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 05 Issue: 12 | Dec - 2021 ISSN: 2582-3930.

[12] Prof. Jitendra C. Musale, "Neuro Friend – An application to Monitor and supervise the mentally disabled patients Monitoring System", in International Journal of Advance Engineering and Research Development Volume 6, Issue 7, December -2020 e-ISSN (O): 2348-4470 p-ISSN (P): 2348-6406