International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 475-485 www.ijaem.net ISSN: 2395-5252

A Survey on Envisioning the Unseen

Mrs. G. Monika[1], Korapala Sushma[2], Thudimilla Eshwar[3], Poreddy Chaithanya[4]

Assistant Professor, Department of Computer Science (Artificial Intelligence and Machine Learning)1, IV B.Tech Students, Department of Computer Science(Artificial Intelligence and Machine Learning)2,3,4, ACE Engineering College, Hyderabad, Telangana, India.

Date of Submission: 15-03-2024 Date of Acceptance: 28-03-2024

ABSTRACT: In this study, we present a novel deep learning model designed to improve accessibility for the visually impaired by delivering a comprehensive interpretation of images through speech. Our methodology involves the utilization of the Flickr8K dataset, implementing a hybrid CNN-RNN model incorporating CNN and an attention mechanism for generating descriptive image captions. To optimize input for both convolutional and recurrent neural networks, we employ extensive preprocessing of text and image data. Visualization techniques such as word cloud representations provide insights into key dataset features. Leveraging the CNN architecture for model training enhances the extraction of image features. Evaluation involves a comparison between Greedy Search or Beam Search for caption generation, with a focus on the essential BLUE Score metric. The final output is an audio representation obtained by converting generated captions to speech using a text-to-speech library into desired language audio caption. This innovative solution proves to be a valuable tool for fostering accessibility and inclusivity for visually impaired individuals, showcasing the efficacy of deep learning in addressing real-world challenges.

Keywords: deep learning, accessibility, visual impairment, image captioning, CNN-RNN model, attention mechanism, preprocessing, data visualization, BLUE Score metric, text-to-speech library.

I.INTRODUCTION

In the contemporary digital era, the conversion of visual content into text remains a crucial undertaking with applications spanning data analysis, content indexing, and search engines. Deep learning, a subset of artificial intelligence, has emerged as a pivotal force in this transformation by leveraging neural networks and large datasets to extract intricate features automatically. This paradigm shift has significantly elevated the efficacy of translating images into textual descriptions. The integration of deep learning, particularly encoder-decoder architectures, wherein convolutional neural networks (CNNs) extract visual features and recurrent neural networks (RNNs) translate these features into textual descriptions, has yielded substantial performance improvements. The synergy of these networks facilitates the extraction of essential image features, leading to the generation of coherent and natural textual explanations. Attention mechanisms further refine this process by emphasizing specific visual features during text generation, contributing to more meaningful and contextually rich descriptions. The profound impact of deep learning techniques in enhancing image-to-text conversion underscores its importance in realizing consistent and high-quality outcomes.

II.LITERATURE REVIEW

The study focused on a comprehensive exploration of the efficacy of different methodologies within the domain. Through a meticulous examination of relevant research papers, the objective was to evaluate a diverse range of approaches and techniques applied in these specific areas. This investigative process aimed to uncover the subtle intricacies and advancements prevalent within the field.

Kanimozhiselvi et al. [1] In-depth exploration into image captioning involved the application of three distinct CNN architectures: Inception-V3, ResNet50, and Xception models. The research methodology incorporated these architectures for feature extraction, complemented by the integration of Long Short-Term Memory (LSTM) networks for caption generation. The Flickr 8k dataset served as the foundational source for training and evaluation purposes. Among the chosen CNN architectures, the Xception model emerged as the most effective, demonstrating superior performance and achieving the highest accuracy, notably 75% after 50 epochs of training with the Xception + LSTM model. This study underscores the efficacy of the combined approach in image captioning tasks.The investigation into image captioning methodologies unveiled the significance of diverse CNN architectures in conjunction with LSTM networks. The evaluation conducted on the Flickr 8k dataset highlighted the superior performance of the Xception model, showcasing its efficacy in achieving a remarkable 75% accuracy after 50 epochs of training when combined with LSTM. This research contributes valuable insights into the realm of image captioning, emphasizing the pivotal role of architectural choices in enhancing accuracy and overall effectiveness in the task at hand.

Bai et al. [2] employed a CNN-based generation model, incorporating Conditional Generative Adversarial Networks (CGAN), to produce descriptive image captions. The integration of Multi-modal Graph Convolutional Networks (MGCN) facilitated the establishment of meaningful visual relationships between objects in images. Through extensive experiments on the MSCOCO 2014 dataset, their methodology demonstrated substantial performance improvements compared to prevailing state-of-the-art techniques. The application of CGAN and MGCN highlights the efficacy of combining generative adversarial techniques with graph-based approaches to enhance image captioning quality.Extending the exploration into image captioning, Utilized a CNN-based generation model with Conditional Generative Adversarial Networks (CGAN). The introduction of Multi-modal Graph Convolutional Networks (MGCN) was pivotal for establishing meaningful visual relationships among objects within images. Their experimental evaluation on the MSCOCO 2014 dataset revealed noteworthy enhancements in performance relative to existing state-of-the-art methods. The strategic combination of CGAN and MGCN underscores the effectiveness of integrating generative adversarial techniques with graph-based approaches, contributing to advancements in image captioning quality.

Agrawal et al. [3]In a study by Agrawal et al. [3], a novel model was introduced, employing an encoder-decoder architecture with an attention mechanism. The encoder, integrating a pretrained Convolutional Neural Network (CNN) based on the Inception v3 architecture, adeptly extracted intricate image features. Complementary to convolutional features, Recurrent Neural Networks (RNN) were seamlessly integrated to improve the sequential generation of captions. The pivotal inclusion of the Bahdanau Attention Mechanism played a crucial role, resulting in enhanced performance compared to conventional captioning methods.Expanding upon this research, the study focused on the efficacy of a model designed by Agrawal in the realm of image captioning. The proposed model's innovative approach involved utilizing the Inception v3 architecture within a pretrained Convolutional Neural Network (CNN) as the encoder. This configuration facilitated the extraction of robust image features. The attention-based decoder, coupled with Recurrent Neural Networks (RNN), contributed to the sequential generation of captions, surpassing traditional methods in performance. The incorporation of the Bahdanau Attention Mechanism emerged as a critical factor, underscoring the model's superiority in capturing nuanced relationships between images and generated captions.

Kılıçkaya et al. [4]Addressing image captioning challenges, the Im2Text method, with a focal point on meta-class features, was applied by a research team. The study incorporated the Pascal Sentences dataset, consisting of 1000 images, each associated with five captions from distinct contributors, resulting in a total of 5000 captions. The research team attained a Bleu1 score of 0.0067, showcasing the effectiveness of their methodology in understanding meta-class features and contributing valuable insights to the field of image captioning.In the exploration of image captioning methodologies, a study concentrated on the Im2Text approach, specifically emphasizing meta-class features. The Pascal Sentences dataset, comprising 1000 images, each paired with five diverse captions, totaling 5000 captions, served as the basis for their research. The achieved Bleu1 score of 0.0067 provided notable insights, shedding light on the efficacy of their approach in tackling the intricacies of image captioning with a focus on meta-class features.

Lu et al. [5] In the realm of image captioning, recent endeavors have focused on refining methodologies for generating descriptive captions, particularly in the context of fine art images. A notable contribution involves the implementation of a virtual-real semantic alignment training process, as evidenced in a study utilizing both the MS COCO and ArtCap datasets for model training. The outcomes of this innovative approach are reflected in the model's performance metrics, with a Bleu1 score of 0.508 showcasing proficiency in unigram precision. Additionally, the Meteor performance of 0.1317 indicates a remarkable level of fluency and relevance in the generated captions. These results collectively emphasize the efficacy of the devised approach, addressing the intricate challenge of captioning fine art images.This investigation builds upon the foundation laid by prior research, which elucidates the significance of incorporating a virtual-real semantic alignment training process in the domain of fine art image captioning. Leveraging datasets such as MS COCO and ArtCap during the model training phase has become a prevalent strategy. The observed performance metrics, including the notable Bleu1 score and Meteor performance, substantiate the success of this approach in capturing both precision and fluency in descriptive captions. The effectiveness demonstrated in this study contributes to the evolving landscape of image captioning techniques, particularly within the nuanced task of describing fine art imagery.

Yang et al. [6]Researchers in the field focused their efforts on developing a Human-Centric Caption Model (HCCM) geared towards discerning human behaviors. Their significant contribution involved the introduction of a three-branch hierarchical caption model and the curation of a specialized dataset named Human-Centric COCO (HC COCO). This model heavily relies on intricate feature extraction and interaction mechanisms to generate human-centric captions. Despite showcasing advancements over existing methods, the HCCM exhibited limitations in providing highly detailed captions, suggesting the need for further refinement to capture nuanced aspects of human behavior.In the exploration of human-centric captioning, scholars directed attention to the innovative Human-Centric Caption Model (HCCM). This model, reliant on intricate feature extraction and interaction mechanisms, introduced a three-branch hierarchical caption model and utilized the Human-Centric COCO (HC COCO) dataset. While demonstrating progress over existing methods, the HCCM revealed shortcomings in generating highly detailed captions, indicating a necessity for continued refinement to capture the intricate nuances inherent in human behavior.

Li et al. [7]Image Captioning represents a cross-modal task where the automatic generation of coherent, natural sentences describing image contents poses a significant challenge. Existing methodologies often struggle with inaccurate semantic matching between images and generated captions due to the substantial gap between vision and language modalities. Addressing this issue, the current paper introduces an innovative multi-level similarity-guided semantic matching approach for image captioning. This method effectively combines local and global semantic similarities, extracting fine-grained semantic information from both images and generated captions. A local semantic similarity evaluation mechanism is designed through a comparison of semantic units, complemented by the use of the CIDEr score to capture global semantic similarity. The fusion of these local and global similarities, guided by reinforcement learning theory, enhances model optimization for improved semantic matching. Quantitative and qualitative experiments on the MSCOCO dataset validate the proposed method's superiority in achieving fine-grained semantic alignment between images and generated captions.The research landscape in image captioning has been dominated by the persistent challenge of accurate semantic matching between visual and textual modalities. Existing approaches, grappling with the inherent gap, prompted the development of a novel solution presented in this paper. The proposed method innovatively leverages multi-level similarity guidance, combining local and global semantic measures. Extraction of fine-grained semantic details from images and captions facilitates a local similarity evaluation, supplemented by the global semantic perspective assessed through the CIDEr score. The integration of these two levels is accomplished through reinforcement learning, strategically guiding model optimization for enhanced semantic alignment. Empirical evaluations on the extensive MSCOCO dataset substantiate the effectiveness of the introduced method in achieving precise semantic matching between images and their corresponding captions. Achieving significant metrics such as Bleu1 at 81.2, Bleu4 at 39.0, Rouge_l at 58.9, and CIDEr-D at 128.5.

Jaknamon et al. [8]In the realm of image captioning methodologies, ThaiTC introduces a transformative approach departing from conventional CNN and RNN frameworks. Their methodology, based on Transformers, utilizes the Image Transformer for encoding image features and the Text Transformer for decoding captions, marking a paradigm shift in image captioning architectures. The experimental findings revealed performance variations across distinct datasets, emphasizing the critical role of dataset diversity in assessing model robustness and generalization capabilities.This departure from traditional convolutional and recurrent neural networks is a notable contribution to the evolving landscape of image captioning techniques. ThaiTC's adoption of Transformers showcases a novel methodology, leveraging the Image Transformer for efficient image feature encoding and the Text Transformer for caption decoding. The experimental outcomes, demonstrating performance variations across diverse datasets, underscore the importance of dataset diversity in evaluating the model's robustness and generalization capacities, providing valuable insights for the field of image captioning research.

Krisna et al. [9]The transformation of visual data into textual information holds pivotal significance in digital information retrieval and data analysis. The convergence of visual and textual data through the "image-to-text" process has become a focal point of interest for both researchers and industry professionals. This study delves into the realm of generating text from images, specifically examining the impact of incorporating an attention mechanism into the encoder-decoder framework of the Inception v3 deep learning architecture. The model, trained on the Flickr8k dataset, utilizes Inception v3 to extract image features. The encoder-decoder structure, enriched with an attention mechanism, is employed for next-word prediction, with performance evaluation conducted on the train images from the Flickr8k dataset. The experimental outcomes affirm the model's commendable ability to accurately discern objects within images.In contemporary digital landscapes, the conversion of visual information into textual content holds significant relevance for information retrieval and data analysis. The intriguing intersection of visual and textual data, often referred to as "image-to-text" transformation, has attracted substantial attention from scholars and industry practitioners alike. This article contributes to the ongoing discourse by presenting a study focused on text generation from images. Notably, the investigation seeks to quantify the impact of integrating an attention mechanism into the encoder-decoder architecture of the Inception v3 deep learning model. Trained on the Flickr8k dataset, the Inception v3 model is adept at extracting essential features from images. The utilization of an encoder-decoder structure, enriched with an attention mechanism, facilitates next-word prediction, and the model's performance is rigorously evaluated using the train images from the Flickr8k dataset. The experimental results affirm the model's impressive proficiency in accurately recognizing objects depicted in images.

Shambharkar et al. [10]In the contemporary landscape of widespread social media engagement, individuals actively participate in online platforms, sharing images accompanied by diverse captions. Crafting suitable captions is often a laborious task, yet crucial for conveying the essence and significance of the visual content. Recognizing this, a model for an image caption generator becomes pertinent, capable of providing meaningful descriptions for images of varying types and resolutions. This model employs a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) with an encoder-decoder architecture. CNN serves the purpose of extracting essential features from images, while pre-trained ImageNet models are considered, and their performance is benchmarked using the BLEU score metric. Prediction of captions is executed through both beam search and argmax methods, with a comparative analysis between supervised and unsupervised image captioning models. Training and testing are conducted on the Flickr8k and MSCOCO datasets. The potential application of this model within a mobile application holds particular promise, especially for individuals with disabilities, providing them with a valuable tool reliant on text-to-speech functionality for enhanced accessibility.

Feng et al.[11]introduced an innovative model that seamlessly integrates captioning and gaze tracking, establishing a robust connection between captions and eye-tracking patterns. Their dataset comprised 400 training images, 200 validation images, and 400 test images. The model demonstrated a Recall@5 performance of 0.0048, indicating its effectiveness in predicting relevant visual information."Feng et al.[11] introduced a novel model that seamlessly integrates image captions and gaze tracking by leveraging the learned relationship between captions and eye-tracking patterns. The dataset employed for training comprised 400 images, with an additional 200 for validation and 400 for testing. The model demonstrated a Recall@5 performance of 0.0048, indicating its ability to accurately recall relevant information.

Cai et al.[12]In the realm of image captioning, the challenge of generating natural language descriptions persists, necessitating innovative solutions. Noteworthy difficulties include managing uncommon terms, crafting inventive captions, and addressing long-term dependencies. Cai et al.'s multimodal fashion image captioning model demonstrated impressive metrics, emphasizing the integration of visual and textual information. Building upon this, our paper proposes a unique method to tackle image captioning challenges. By combining long short-term memory (LSTM) models with convolutional neural networks (CNNs), our approach leverages pre-trained CNN attributes for caption generation. Mitigating the issue of rare words, we employ a beam search method with a penalty term. Tested on the Flickr8k dataset, our model outperforms contemporary techniques in caption quality and variety, presenting a promising direction for AI-based image captioning with broad applications in image retrieval, visual question answering, and beyond.

Ye et al. [13]In contrast to conventional encoder–decoder models for remote sensing image (RSI) captioning, two-stage RSI captioning methods, incorporating an auxiliary remote sensing task for prior information, have demonstrated enhanced descriptive accuracy. However, existing two-stage approaches treat image captioning and auxiliary tasks independently, leading to time inefficiency and neglecting mutual interference between tasks. Addressing this issue, this study introduces a novel Joint Training Two-Stage (JTTS) RSI captioning method. Employing multilabel classification for prior information, a differentiable sampling operator, replacing non-differentiable sampling, is designed to index multilabel classification outcomes. Unlike prior two-stage methods, JTTS allows joint training, enabling the flow of generated description errors into multilabel classification optimization through backpropagation. Implementation involves approximating the Heaviside step function with a differentiable logistic function for sampling and introducing a dynamic contrast loss function to maintain a margin between positive and negative label probabilities during sampling. An attribute-guided decoder filters multilabel prior information from the sampling operator to enhance caption accuracy. Extensive experiments demonstrate that JTTS attains state-of-the-art performance on RSICD, UCM-captions, and Sydney-captions datasets, showcasing its efficacy in RSI captioning methodologies.

Wang et al. [14]introduced a novel caption transformer (CapFormer) architecture tailored for generating captions for remote sensing images. The CapFormer model demonstrated notable performance enhancements, achieving a Bleu1 score of 66.12 and a Rouge_l score of 49.78. Leveraging transformer technology, the model excelled in capturing contextual relationships within the remote sensing imagery. However, potential challenges may arise in terms of computational intensity, requiring efficient hardware resources for optimal performance.

R. Malhotraet al. [15]introduced a model leveraging ResNet50 for robust image encoding and integrating RNN and LSTM for sentence generation. This innovative approach resulted in commendable performance metrics, including an F1 Score of 77.8, Meteor score of 27.6, and an accuracy rate of 70. The utilization of state-of-the-art technologies such as ResNet50, RNN, and LSTM contributed to the success of the model.

Yang et al. [16]introduced the Context-Sensitive Transformer Network (CSTNet) method, demonstrating notable enhancements over the State-of-the-Art (SOTA) models. The proposed approach achieved impressive metrics, with Bleu1 reaching 81.1, Meteor at 29.4, and Rouge at 59.0, showcasing its superior performance. CSTNet leverages transformer technology for contextual understanding, providing a robust foundation for language processing tasks.

Wang et al. [17]proposed an innovative parallel fusion architecture, combining RNN and LSTM, which demonstrated improved efficiency and outperformed prevailing methodologies. The model achieved notable performance metrics with Bleu1 at 66.7 and Meteor at 16.53 post-training.

Raut D.B., Kshitija Gaikwad, Avantika Gholve, Vaishnavi Karande, Anjali Mulik et al. [18] introduced a novel system incorporating VGG16, a Convolutional Neural Network, and Long Short-Term Memory (LSTM) networks, an extension of Recurrent Neural Networks, for image captioning and audio description generation. Leveraging the "Flickr8k" dataset, the system undergoes stages like image augmentation, feature extraction, text cleaning, tokenization, and LSTM-based caption generation to ensure precise and diverse captions. Experimental results reveal the system's capability to generate accurate and multiple captions and audio descriptions, surpassing contemporary caption generation systems on the Flickr8k dataset. The proposed system holds potential applications in aiding visually impaired individuals in comprehending visual content and enhancing multimedia with detailed captions.

Deepa Mulimani, Prakashgoud Patil, Nagaraj Chaklabbi et al. [19]Image captioning involves generating descriptive sentences for photographs, identifying objects, and locating crucial image parts. Recent advancements enable algorithms to produce contextually accurate, natural language descriptions. This work utilizes

EfficientNetB0, a pre-trained Convolutional Neural Network (CNN), for extracting image visual features. Transformer Encoder and Decoder are employed to construct captions, with training conducted on the Flickr8k dataset. The model's proficiency in understanding and generating text from images is supported by the findings, evaluated using BLEU scores. Furthermore, the model translates image descriptions into text and voice, presenting an ideal approach for visually impaired individuals unable to perceive visuals.

Jasmita Khatal, Prajkta Jadhav, Shraddha Parab , Prof. Rasika Shintre et al. [20] presented a novel real-time image captioning and voice synthesis system, employing a blend of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architecture featuring Long Short-Term Memory (LSTM). The CNN is applied for extracting image features, and the RNN with LSTM is utilized for caption generation, facilitating the instantaneous synthesis of voice corresponding to the generated captions. This innovative approach contributes to the evolving landscape of multimedia processing, addressing the demand for real-time image understanding and auditory accessibility for diverse applications.

Zeynep KARACA , Bihter DAS et al. [21] an innovative system is introduced, incorporating an attention mechanism into the encoder-decoder-based Inception v3 deep learning architecture for image-to-text generation. The proposed methodology entails training the Inception v3 model on the Flickr8k dataset, facilitating the effective extraction of pertinent image features. The encoder-decoder structure, augmented with the attention mechanism, is employed for next-word prediction, and subsequent training on the Flickr8k dataset's images evaluates the model's proficiency in generating descriptive captions for images.

Table – 1. The comparison of literature reviews

S.no	**Year**	**Technique/Methodology**	**Pros**	**Cons**	**Blue-1**
1	2022	CNN architectures: Inception-V3, ResNet50, Xception and LSTM networks	Xception model exhibited superior performanceand Achieved a remarkable 75% accuracy with Xception + LSTM after 50 epochs.	Limited discussion on potential limitations or challenges encountered in the methodology.	52.77
2	2021	CNN-based generation model, CGAN, MGCN	Significant performance improvements, Meaningful visual relationships established and Enhanced quality of image captioning	Computational complexity may increase with the integration of CGAN and MGCN.	72.67
3	2021	Encoder-decoder architecture, Inception v3 CNN, Recurrent Neural Networks (RNN), Bahdanau Attention Mechanism.	Superior performance compared to traditional methods.	The use of both CNNs and RNNs in the model architecture may increase computational complexity and training time.	
4	2014	Im2Text method, Pascal Sentences dataset and Meta-class features	Comprehensive exploration of meta-class features	Limited discussion on scalability and Dependency on Pascal Sentences	0.0067

				dataset	
5	2022	Virtual-real semantic alignment, MS COCO dataset, ArtCap dataset.	High Bleu1 score , proficient unigram precision; Meteor performance (0.1317), demonstrating fluency and relevance in captions.	Limited insight into potential challenges or drawbacks in the provided information.	0.508
6	2022	Detailed feature extraction and interaction mechanisms, three-branch hierarchical caption model, Human-Centric COCO dataset.	Noteworthy improvements over existing methods in generating human-centric captions.	Limitations in generating highly detailed captions for human behaviors.	
7	2021	Leverages semantic matching and local semantic similarity measurement mechanisms to establish correlations between images and captions.	High-performance and Integration of semantic similarities	Challenges in scalability or adaptability and Dependency on the quality.	81.2
8	2022	Transformer architectures, specifically the Image Transformer for encoding image features and the Text Transformer for decoding captions.	Enhanced Attention Mechanism and Parallelization.	Data Sensitivity and Complexity.	67.9
9	2022	Conditional GAN, Inception v3 encoder, Bahdanau Attention mechanism-based GRU decoder.	Effective handling of distorted images and Successful generation of accurate captions for rainy-noisy images	Potential complexity in implementing and fine-tuning the GAN-based architecture.	
10	2021	Beam-search based CNN+RNN architecture	Exhibits improved caption diversity and enhances the likelihood of selecting a more contextually relevant and accurate caption.	computational complexity of the beam-search process.	
11	2022	Leveraged advanced machine learning techniques, incorporating both captioning and gaze tracking mechanisms for a comprehensive analysis of visual data.	Seamless integration of captions and gaze tracking, robust recall performance.	Limited dataset size, potential scalability challenges.	
12	2022	Multimodal integration, attention mechanisms, deep learning.	Impressive performance metrics, effective integration of visual	Specific limitations	46.5

			and textual information.		
13	2022	Joint Training and Two-Stage Process	Efficient caption generation for remote sensing images.	Potential complexity in implementing the joint training and two-stage process.	0.8696
14	2022	Novel caption transformer (CapFormer) architecture tailored for generating captions for remote sensing images	Improved performance, contextual relationship capture.	Computational intensity, requires efficient hardware.	
15	2022	ResNet50, RNNand LSTM	Effective image encoding, successful sentence generation, high F1 Score, Meteor score, and accuracy.	Specific limitations or drawbacks not explicitly mentioned in the provided information.	66.12
16	2023	Transformer, Context-Sensitive Transformer Network (CSTNet)	Superior performance compared to SOTA models, impressive Bleu1, Meteor, and Rouge scores.	Limited information provided; further evaluation across diverse datasets may be necessary for comprehensive validation.	81.1
17	2023	Parallel Fusion RNN+LSTM Architecture	Enhanced efficiency and Superior results compared to dominant approaches.	Specific performance details beyond Bleu1 and Meteor scores not provided.	66.7
18	2021	VGG16, RNN and LSTM	Rich visual feature extraction using VGG16 and Enhanced caption generation through the combined use of RNN and LSTM.	Potential computational complexity with VGG16and Possible challenges in handling diverse image characteristics within the Flickr8 dataset.	
19	2023	Utilized EfficientNetB0 for image feature extraction, Transformer Encoder and Decoder for caption generation on Flickr8k dataset.	Effective image feature extraction and Utilizes Transformer architecture for caption	Dependency on pre-trained models may limit adaptabilityand Performance may be influenced by	89.5

			construction.	the quality and diversity of the training dataset.	
20	2021	CNN+RNN architecture, LSTM for sequential modelling, Real-time image captioning and Voice synthesis.	Real-time captioning and voice synthesis, Comprehensive image understanding, Utilization of LSTM for effective sequential modelling.	Potential complexity in the combined CNN+RNN architecture, Computational intensity, especially with real-time processing	
21	2023	Inception v3, Encoder-Decoder Structure, Attention Mechanism	Improved image-to-text generation through attention mechanismand Effective utilization of Inception v3 for feature extraction	Potential complexity in implementing and tuning attention mechanism and Dependence on the quality and representativeness of the training dataset	55.0
	The Proposed	**CNN** + GRU with Attentionmechanism into desired language audio caption	Effective feature extraction, Enhanced sequential generation,Improved focus on relevant image parts, Reduced noise		7**8.9**

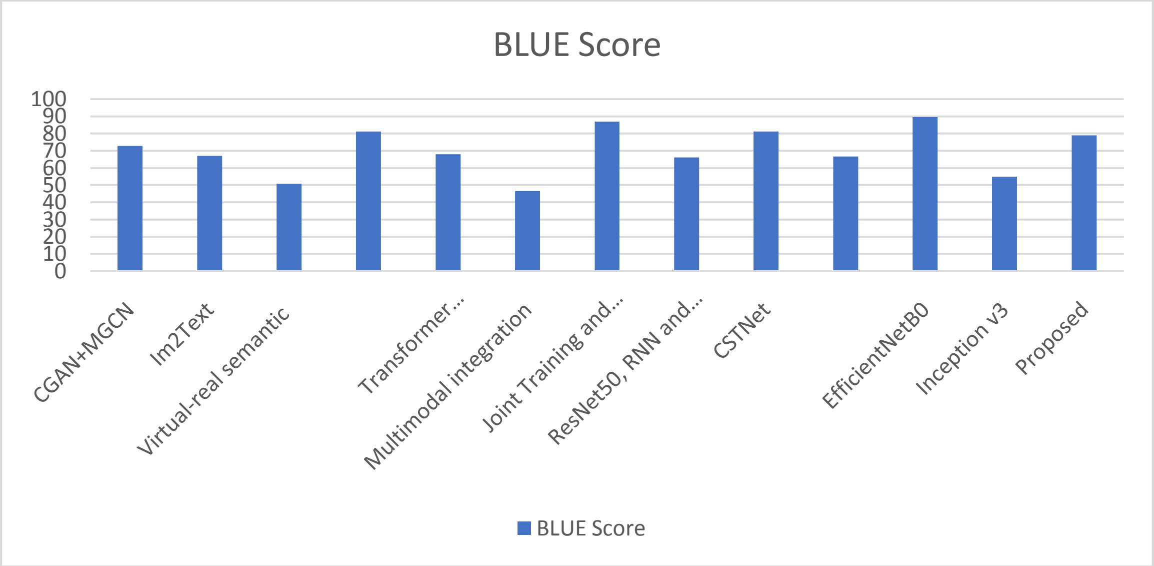

Fig 1- Accuracy(BLUE score) for different methodologies

III.CONCLUSION

In conclusion, our project presents a groundbreaking solution aimed at empowering visually impaired individuals to comprehend visual content. Through the integration of a CNN-RNN model, featuring CNN as the encoder and LSTM-based attention mechanisms, we've developed a system that converts images into descriptive captions, subsequently transformed into audio using a text-to-speech library. This innovative approach not only serves as the eyes for the blind, providing detailed auditory information about the visual world but also holds immense potential to significantly improve independence and understanding for individuals with visual impairments.

ACKNOWLEDGEMENT

We would like to thank our guide **Mrs. G Monika** for her continuous support and guidanceAssistant Professor, Department of CSE (Artificial Intelligence & Machine Learning), and **Mr. SHASHANK TIWARI,** Assistant Professor, Project Coordinator and we profoundly thank **Dr. KAVITHA SOPPARI**, head of the Department CSE(Artificial Intelligence & Machine Learning) for her guidance and continuous support.

REFERENCES

[1]. C. S. Kanimozhiselvi, K. V, K. S. P, and K. S, "Image Captioning Using Deep Learning," in 2022 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740788.

[2]. C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, "Boosting convolutional image captioning with semantic content and visual relationship," Displays, vol. 70, p. 102069, Dec. 2021, doi: 10.1016/j.displa.2021.102069.

[3]. V. Agrawal, S. Dhekane, N. Tuniya, and V. Vyas, "Image Caption Generator Using Attention Mechanism," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Jul. 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[4]. M. Kılıçkaya, E. Erdem, A. Erdem, N. İ. Cinbiş, and R. Çakıcı, "Data-driven image captioning with meta-class based retrieval," in 2014 22nd Signal Processing and Communications Applications Conference (SIU), Apr. 2014, pp. 1922-1925, doi: 10.1109/SIU.2014.6830631.

[5]. Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," Neurocomputing, vol. 490, pp. 163-180, Jun. 2022, doi: 10.1016/j.neucom.2022.01.068.

[6]. Z. Yang, P. Wang, T. Chu, and J. Yang, "Human-Centric Image Captioning," Pattern Recognition, vol. 126, p. 108545, Jun. 2022, doi: 10.1016/j.patcog.2022.108545.

[7]. J. Li, N. Xu, W. Nie, and S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching," Visual Informatics, vol. 5, no. 4, pp. 41-48, Dec. 2021, doi: 10.1016/j.visinf.2021.11.003.

[8]. T. Jaknamon and S. Marukatat, "ThaiTC:Thai Transformer-based Image Captioning," in 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAINLP), Nov. 2022, pp. 1-4, doi: 10.1109/iSAINLP56921.2022.9960246.

[9]. A. Krisna, A. S. Parihar, A. Das, and A. Aryan, "Endto-End Model for Heavy Rain Image Captioning," in 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Dec. 2022, pp. 1646-1651, doi: 10.1109/ICAC3N56670.2022.10074181.

[10]. P. G. Shambharkar, P. Kumari, P. Yadav, and R. Kumar, "Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, pp. 857-864, doi: 10.1109/ICICCS51141.2021.9432245.

[11]. Y. Feng, K. Maeda, T. Ogawa, and M. Haseyama, "Human-Centric Image Retrieval with Gaze-Based Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp. 3828-3832, doi: 10.1109/ICIP46576.2022.9897949.

[12]. C. Cai, K.-H. Yap, and S. Wang, "Attribute Conditioned Fashion Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp.

1921-1925, doi: 10.1109/ICIP46576.2022.9897417.

[13].] X. Ye et al., "A Joint-Training Two-Stage Method For Remote Sensing Image Captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, doi: 10.1109/TGRS.2022.3224244.

[14]. J. Wang, Z. Chen, A. Ma, and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Jul. 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199.

[15]. R. Malhotra, T. Raj, and V. Gupta, "Image Captioning and Identification of Dangerous Situations using Transfer Learning," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Mar. 2022, pp. 909-915, doi: 10.1109/ICCMC53470.2022.9753788.

[16]. Xin Yang et al., "Context-Aware Transformer for image captioning," Neurocomputing, vol. 549, p. 126440, 2023, doi: 10.1016/j.neucom.2023.126440.

[17]. Raut D.B., Kshitija Gaikwad, Avantika Gholve, Vaishnavi Karande, Anjali Mulik*4 et al .” IMAGE CAPTION GENERATOR WITH VOICE” International Research Journal of Modernization in Engineering Technology and Science, Volume:05/Issue:05/May-2023, e-ISSN: 2582-5208.

[18]. Rachel Calvin, Shravya Suresh et al. “Image Captioning using Convolutional Neural Networks and Recurrent Neural Network”, 2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.941800

[19]. Deepa Mulimani, Prakashgoud Patil, Nagaraj Chaklabbi et al. “Image Captioning using CNN and Attention Based Transformer” 2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), Data Sci-ence and Intelligent Computing Techniques, 157–166. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-14

[20]. Jasmita Khatal , Prajkta Jadhav , Shraddha Parab , Prof. Rasika Shintre et al. “Real Time Image Captioning and Voice Synthesis using Neural Network” International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 01 | Jan 2021, e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[21]. Zeynep KARACA , Bihter DAS. et al, “From Pixels to Paragraphs: Exploring Enhanced Image-to-Text Generation using Inception v3 and Attention Mechanisms “, DUJE (Dicle University Journal of Engineering) 14:4 (2023), Doi: 10.24012/dumf.1340656.