Image Captioning: Enhance Visual Understanding

G. Kalaiarasi[1], M. Sravya Sree[2], B. Sai Geetha[3], A. Yasaswi[4], I. Aravind[5] and G. Nagavenkata Sreeja[6]
{ kalaiibe@gmail.com[1], sravyasreemaddala@gmail.com[2], saigeethabonagiri@gmail.com[3], yashuvu40@gmail.com[4], illaaravind604@gmail.com[5], grandhisreeja@gmail.com[6] }

Department of Advanced Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Guntur, Andhra Pradesh, India[1, 2, 3, 4, 5, 6]

Abstract. Image captioning fuses computer vision and natural language processing for producing natural language descriptions of images. Conventional methods exploited CNNs such as VGG16 as visual feature extractors and employed LSTM-based networks to generate captions, and were trained on datasets including Flickr8k and Flickr30k. More recently, transformer-based models such as Vision Transformer (ViT) and GPT-2 have produced a considerable leap in the state-of-the-art by providing the capacity for a shared representation and zero-shot learning. In this work, we develop a ViT-GPT2 based image captioning system and carry out our experiments on the Flickr8k dataset. The model demonstrates substantial changes in accuracy, diversity and context-sensitiveness compared to CNN-LSTM baselines. Assessment with BLEU, METEOR, and ROUGE metrics further verifies the improved precision of caption description and semantic alignment. These results suggest the effectiveness of transformer architectures in generating natural, human-like descriptions and potential for real-world applications such as accessibility and multimedia systems.

Keywords: Convolutional Neural Networks (CNNs), Long- Short-Term Memory (LSTM), Vision Transformer, GPT-2.

1 Introduction

Image captioning is a popular research topic and the interdisciplinary field combining the techniques from Computer Vision (CV) and Natural Language Processing (NLP). Image captioning is a typical form of image description where a correct and semantically relevant text description is generated for the target image. This is a challenging task that requires deep understanding of visual content, including detail information about objects in the image – such as the category of the object and its properties (color, size, and activity), and the relations (i.e., its spatial and semantic) between them [8]. Meanwhile, the generated captions are expecting to be grammatically-correct and semantically-reasonable to make them fluent and language-consistent, therefore it is a challenging multimodal learning task.

Image captioning is not only an academic investigation but also useful in many real-world scenarios. Meantime, in content-based image retrieval, right captions can help to enhance the search engine accuracy by providing a more efficient indexing and querying of visual content. For blind people, there are also captioning systems, the systems assemble visual data as textual or audio descriptions so that visually impaired people can maintain their awareness of surroundings [11]. These models can be used in automatic report generation to generate captions for images in news tickers, news broadcasts, or for fast editorial processes [12]. Furthermore, image captioning can also be utilized in social media systems to assist auto-tagging, improving content recommendation, and to increase user interaction by providing

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357851

contextually relevant descriptions [13] [14].

Image captioning techniques have improved a lot from historical “stupid” rule-based and retrieval-based models to modern deep learning ones. In the history ago, approaches were based on the proposal of hand-crafted features (Local Binary Patterns, Scale-Invariant Feature Transform and Histogram of Oriented Gradients) to learn visual features from images. They were then fed into traditional machine learning classifiers (e.g. SVMs and decision trees) to generate or find conversational responses [15].

However, these traditional approaches had several restrictions. Most notably, they did not generalize well and were not robust, especially when applied to com- plex or higher-quality data. These systems were brittle and non-scalable due to the rigid nature of the heavy manual feature designing [9]. They also did not model the diverseness of semantics as well as long-range dependencies that are crucial for to produce natural and coherently sentences. These challenges resulted in a shift toward employing deep learning techniques that significantly increased the ability for automatic learning of features and models underlying complex visual-language relationships.

C. Image Captioning Approaches Image captioning has been evolved from the rule-based and retrieval-based to the state of art deep learning. Tamper-proofing images According to the whitepaper, upon receipt of the information, Hotbit shall combine such information with the "# client source identifier + last login IP in the platform that the client uses + current time" and other historical transfer information (to generate a for the receiving hot wallet. The early approaches mainly relied on hand-designed visual features such as Local Binary Patterns (LBP), Scale- Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) to represent the content of images. These features were then fed into traditional machine learning techniques such as Support Vector Machines (SVM) either to produce or retrieve the most suitable captions from a corpus of pre-defined captions.

Although these early methods established the foundations for automated image description, they were plagued by several fundamental flaws. Perhaps most importantly, they were unable to generalize over broad and intricate visual scenes. Their excessive use of hand-engineered features rendered them less flexible and unable to capture the high-level semantic and contextual information required to produce coherent and informative captions. These constraints ultimately accentuated the desire for data-based models to directly learn rich, hierarchical features from images laid the grounds for the embracement of deep learning models by image captioning.

With deep learning breakthroughs, the encoder-decoder State-of-the-art image captioning models today use object detection methods to obtain region features, enabling the cap- toning system to identify single objects and their interaction in space instead of treating an image as one entity [12]. This has had performance on large-scale datasets like MS-COCO, Flickr8k, and Flickr30k, used as benchmarks in measuring image captioning models. Even with such developments, image captioning continues to be challenged by: Manipulating unseen objects – Existing models fail to generate captions for unseen objects that do not appear in the training data [14]. Capturing intricate scenes despite the contribution of attention mechanisms toward enhanced accuracy, captions tend to lack deep understanding of object relations. Providing linguistic smoothness generated captions at times are marred with

grammatical mistakes or unnecessary sentence structures [8].

In order to eliminate these constraints, researchers are looking into reinforcement learning, multimodal learning, and graph-based approaches. These solutions are expected to improve captioning by gaining access to external knowledge bases, improved semantic consciousness, and better linguistic coherence [9]. This article gives an extensive description of state-of-the-art image captioning methods, including major methodologies, datasets, metrics, and research directions in the future. Incorporating innovations like transformer-based architectures, self-supervised learning, and reasoning about scene graphs, the area is moving towards more human-oriented captions generation. This is the example Fig 1 shows the example of a figure caption [7] [10].

A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
A little girl is sitting in front of a large painted rainbow .
A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
There is a girl with pigtails sitting in front of a rainbow painting .
Young girl with pigtails painting outside in the grass .

Fig. 1. Example of a figure caption.

A. Features

1) End-to-End Deep Learning Model: The model integrates convolutional neural networks (CNNs) and recurrent neural networks (RNNs), i.e., Long Short-Term Memory (LSTM) networks, in order to build an end-to-end model. This enables the model to learn image-to-caption mappings end-to-end without handcrafted feature engineering.

2) CNN-Based Feature Extraction: Utilizes Traditional Neural Networks for detailed and high-level feature extraction of images. Attributes of most important visual information like

objects, scenes, and activities as the foundation to come up with the descriptive captions.

3) LSTM-Based Sequence Generation: Use LSTMs to pro- duce text sequences. LSTMs are especially capable of handling long-term data dependencies. That's why they are perfectly suited to use when composing consistent, contextually correct sentences about what an image comprises.

4) Flickr Dataset Training: The user image is passed through a Vision Transformer (ViT) to receive the visual features, and they are utilized as input for GPT-2 in order to generate the captions. The model is trained on features that have been retrieved without exposure to any available data sets.

5) Performance Metrics: The caption system is evaluated with available metrics like BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Translation Evaluation with Explicit Ordering, and Consensus-based Image Description Evaluation. These metrics evaluate the produced captions quantitatively in terms of accuracy, relevance, and coherence compared to human-generated references.

6) Architectural and Hyperparameter Optimization: The research examines the effect of various network structures, hyperparameter settings, and training methods. This includes testing different CNN architectures (e.g., VGG, Reset), configurations of LSTM, learning rates, and optimizers to determine the best combination to use for the image captioning process.

7) Human-Like Descriptive Capability: In my opinion, the most important characteristic of the model is that it produces captions that not only have accuracy but human-like descriptions in Language structure as well. That is, to create sentences that are grammatically correct and contextually suitable, and semantically rich, just like a human would have taken the same image.

2 Literature Survey

The survey of literature investigates image captioning, para- graph generation, and medical report automation. Conventional methods transitioned from ranking-based methods to deep learning models based on CNNs and RNNs, augmented with attention mechanisms. Current medical report generation models were challenged by accuracy and redundancy. The proposed HRNN model incorporates topic matching, soft attention, and global label pooling to enhance abnormality detection and produce diverse, coherent medical reports. Yin et al., 2019. [1].

In image captioning and visual question answering (VQA) attention mechanisms. Top-down classical attention models concentrate on task-specific regions of the image, whereas bottom-up models identify prominent objects. Merging both, Faster R-CNN extracts features from regions, enhancing captioning and VQA performance. The model achieves record- breaking results on MSCOCO and VQA 2017 tasks, with better performance and interpretability. Anderson et al., 2018. [2]

The conventional approach used dependency parsing, but AMR constitutes a more organized semantic representation. With utilizing BART and T5 models, the research improves text-to-scene graph generation, surpassing previous methods by 9.3 percentage in SPICE scores. The improvement enhances high-level semantic comprehension in visual tasks. Woo Suk Choi et al., 2022. [3].

Retrieval-based methods were marred by new captions and thus generated methods like Show and Tell, which wed LSTMs and CNNs. The method uses ResNet50 and VGG16 on Flickr8k, demonstrating superior performance of ResNet50 in captioning and text-to-speech conversion. Sudhakar et al., 2022. [4]

Using CNN for feature extraction and LSTM for sentence building. Various datasets like MS COCO and Flickr 8k are used to train and test models. Advanced methods like adversarial caption generation and transformer-based methods improve accuracy. The proposed model achieves a BLEU-I score of 0.755367, enhancing caption accuracy. Vidyadevi et al., 2023. [5]

Focusing on semantic feature extraction and face recognition for individualized descriptions. Conventional models emphasize generic captions, whereas this method incorporates semantic labels and facial recognition to improve relevance. The system enhances caption accuracy with Flickr30k and celebrity datasets, attaining improved BLEU and METEOR scores. Abisha Anto Ignatious et al., 20191focused on semantic splitting to enhance text-to-image generation by dividing various semantic components in the input for accurate image integration. [6]

The work Image Caption Generation using CNN and Audio Conversion also suffers from some challenges typical for CNN-RNN-based captioning models. Large computational needs and lengthy training processes make it processor- intensive, with limited accessibility to systems with reduced processing capacity. The vanishing gradient issue of Gated Recurrent Units (GRUs) could affect long-term dependency learning, which may negatively impact the coherence of the captions. Moreover, dataset constraints can cause biased or incorrect captions, lowering the model's performance on varied images. Generalization to new images is still a challenge, as overfitting can lead to suboptimal real-world performance. Lastly, although Google Text-to-Speech (TTS) improves accessibility, its audio quality and naturalness can be inconsistent, affecting user experience, especially for visually im- paired. Mitigating these limitations calls for additional optimization in model structure, data augmentation, and quality of speech synthesis.

3 Methodology

The system is implemented as a real-time image captioning architecture that accepts a user-provided image, produces descriptive captions, and translates them into speech for enhanced accessibility. In contrast to traditional dataset-based models that are based on predefined datasets for generating captions, the proposed system processes images dynamically, utilizing a Vision Transformer (ViT) for feature extraction and GPT-2 for generating captions. The produced captions are also synthesized into spoken words with the Google Text- to-Speech (gTTS). The objective is to improve the accuracy, contextuality, and availability of produced captions and to contribute primarily to visually impaired users by offering an auditory depiction of the image contents.

3.1 Preprocessing

Performance Before being fed into the Vision Transformer model, a number of preprocessing procedures are performed in order to support compatibility with the model and further improve feature extraction. The following are the preprocessing steps involved in the preprocessing pipeline:

Resizing: The input image is resized into a fixed dimension of 224×224 pixels in order to provide consistency when different inputs are given and for compliance with expected input dimensions for the ViT model.

Normalization: The pixel values are normalized to align with the distribution of the dataset on which ViT was pre-trained. This facilitates improved convergence and consistency in feature extraction.

Patch Tokenization: The image is broken down into small non-overlapping patches of 16×16 pixels. These patches are flattened and transformed into tokens that are input to the transformer model.

These preprocessing operations normalize the input format, improve computational efficiency, and promote the quality of extracted features.

3.2 Feature Extraction

To get meaningful and high-level representations from the input image, the system uses a Vision Transformer (ViT) rather than traditional Convolutional Neural Networks (CNNs). Un- like CNNs, which utilize localized convolutional filters, ViT treats the image as a sequence of patches to better understand the global context.

The feature extraction process with ViT includes:

- Segmentation of the image into patches that are not overlapping and transform them into embeddings that can be trained.
- Passing through multi-head self-attention layers, which model long-range relations and dependencies among various components of the image.
- Creating a high-dimensional feature vector that captures the semantic information of the image and becomes the input for the caption generator model.
- With the help of ViT, the system can obtain complex object relations and contextual information, resulting in more descriptive and semantic captions.

3.3 Model Architecture

The suggested architecture is composed of three main modules that collaborate with each other to provide accurate image captioning and speech synthesis:

- Vision Transformer (ViT) for Image Feature Extraction: It captures useful high-level visual representations of the input image. The attention mechanism in ViT allows the model to identify object relationships within the image, resulting in better generation of captions.

- GPT-2 for Caption Generation: The features extracted from the image are fed as input to the GPT-2 model, which produces consistent, contextually coherent natural-language textual descriptions. GPT-2 is trained on massive-scale text corpora so that it can generate flowing and semantic captions.

- Text-to-Speech (gTTS) for Speech Synthesis: The generated caption is converted into an audio output using the Google Text-to-Speech (gTTS) engine. This component enhances accessibility, making the system particularly useful for visually impaired individuals.

Through the integration of these elements, the system is capable of generating precise, real-time image captioning with auditory output, both visually and auditorily accessible. The approach maintains a smooth process from image preprocessing to captioning and speech synthesis, rendering the system extremely efficient and usable in real-world contexts. From Fig 2 shows the model overview is as follows:

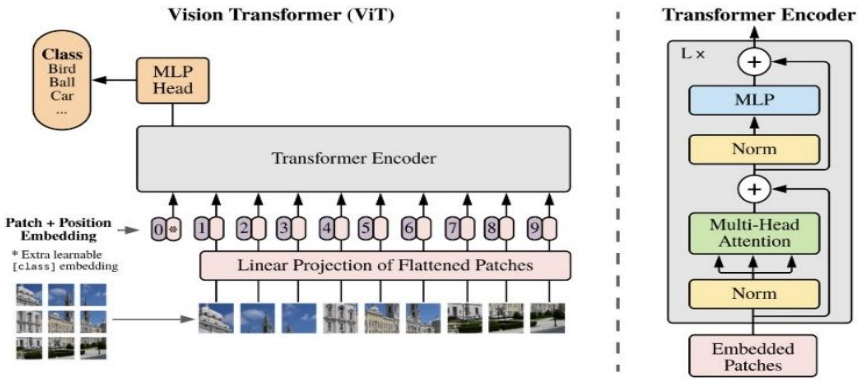

Fig. 2. Model overview.

The feature vector produced by ViT is first projected into a shared latent space compatible with GPT-2 using a transformation layer. The caption generation follows an autoregressive decoding process, where each word is generated sequentially based on the previously predicted words and the image features. Finally, the textual caption is synthesized into speech using gTTS, providing a multi-modal user experience.

Training: The framework employs transfer learning through fine-tuning pre-trained Vision Transformer (ViT) and GPT-2 models for improved performance. The ViT model is trained on large-scale image datasets, while GPT-2 is pre- trained on varied textual corpora. During fine-tuning, the feature extractor is adjusted based on contrastive loss, with the goal that the visual features extracted represent themselves well in relation to textual representations. The

model is trained with cross-entropy loss, whereby the predicted caption is matched against the ground truth to reduce mismatches. Hyperparameter search is performed to optimize learning rate, batch size, and number of attention heads to provide stable model performance.

1) Caption Generation: The caption generation is transformer-based autoregressive decoding. After the visual feature vector is projected into the GPT-2 model, the caption is generated sequentially with beam search decoding to enhance fluency and coherence. The beam search approach avoids repeated phrases and selects the most pertinent words at each decoding step. No-repeat n-gram constraints are also enforced to minimize redundancy in the generated text. The last caption is a semantically correct and contextually rich description of the content in the image. The text so obtained is then converted into speech using gTTS, so the users can get their output in both text and audio form.

2) Evaluation: The performance of the proposed system is assessed using a combination of automated NLP-based evaluation metrics and human subjective evaluation.

a) Automated Metrics

- BLEU (Bilingual Evaluation Understudy): Measures n-gram overlap between the generated caption and reference captions, evaluating syntactic accuracy.
- METEOR (Metric for Evaluation of Translation with Explicit Ordering): Assesses semantic similarity by considering synonyms, stemming, and paraphrasing.
- ROUGE (Recall-Oriented Understudy for Gusting Evaluation): Computes word recall between the generated caption and reference captions.

b) Human Evaluation: In addition to automated metrics, human evaluators assess the generated captions based on:

- Fluency: The readability and grammatical correct- ness of the caption.
- Relevance: How well the caption describes the given image.
- Diversity: The uniqueness and variation of generated captions across different images.

These evaluation methods ensure that the system produces accurate, contextually relevant, and diverse captions, making it suitable for real-world applications, including accessibility for visually impaired users.

4 Result and Discussion

Table 1. Generated Captions from Different Models.

Model	Generated Caption
Vision Transformer + GPT-2	A woman holding a camera in her hand
VGG16 + LSTM	Woman in blue shorts is bouncing on the street
ResNet50	Woman and and and and and
Inception v3	Woman in blue shorts is bouncing on the street

Fig. 3. Image used for generating caption.

The chart compares the BLEU score (BLEU-1 to BLEU-4) of the image captioning models VGG16 + LSTM, ResNet50, Inception v3, and ViT + GPT-2. The BLEU score measures how similar the caption is to previously written references, where the value of the BLEU score measured is positively correlated to the quality of the caption. BLEU-1 is generated based on unigram overlaps while BLEU-4 is generated based on up to four-word sequences which measure the fluency and coherence of the caption. It is evident from this comparison that ViT + GPT-2 has the highest BLEU score on each of the four BLEU measures BLEU-1 = 0.7722, BLEU-2 = 0.6240, Table 1 shows the Generated Captions From Different Models.

BLEU-3 = 0.4886, and BLEU-4 = 0.3771, indicating it has better captioning performance than

the others. Fig 3 shows the Image used for generating caption.

The remaining models all show lower results as they measured higher BLEU levels, indicating that there is less coherent generation of multi-word phrases. The better scores from the ViT + GPT-2 group likely stem from the use of the Vision Transformer for feature extraction and the GPT-2 model (for text) for generation. Together the Vision Transformer and GPT-2 allow for better understanding of the content of the image and generating fluently written and contextually appropriate descriptions. From this comparison, it is evident that transformer-based models are superior to CNN-LSTM models in the image captioning space. Fig 4 shows the Comparison between the previous models and ours.

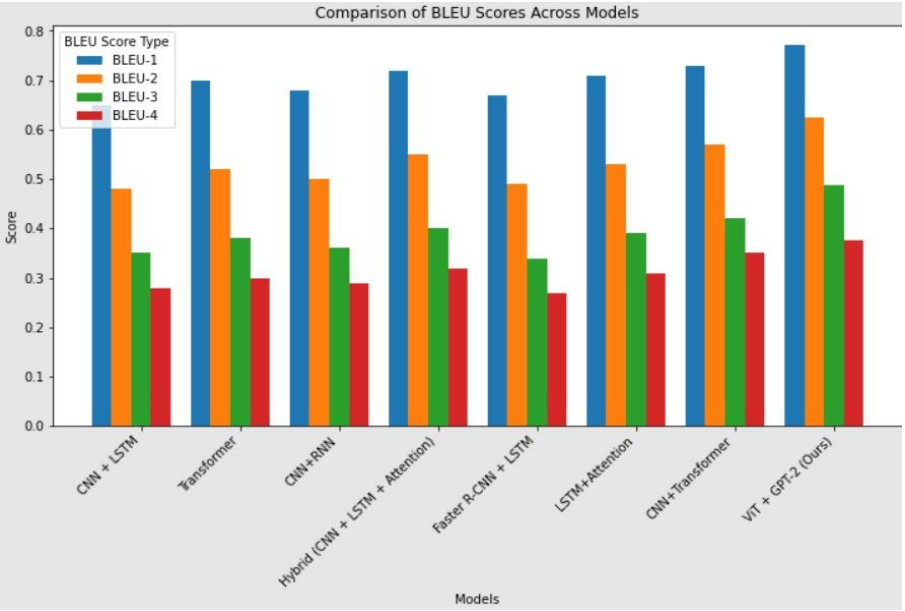

Fig. 4. Comparison between the previous models and ours.

5 Conclusion and Future scope

This work presents an effective image captioning framework that leverages a pre-trained Vision Transformer (ViT) as the encoder and GPT-2 as the decoder within an encoder-decoder architecture. The model demonstrates promising results on the MS COCO dataset, showing that ViT's visual understanding combined with GPT-2's language generation capabilities can produce coherent and descriptive captions.

Future research may focus on enhancing multimodal alignment through improved fusion techniques and incorporating attention-based cross-modal interactions. Expanding training with larger and more diverse datasets, including domain- specific images, could improve the model's adaptability and performance. Additionally, integrating reinforcement learning or external semantic knowledge sources may help generate more accurate and context-aware captions. These advance- mends could pave the way for real-world applications such

as assistive technologies, content summarization, and intelligent image retrieval systems.

References

[1] Yin C, Qian B, Wei J et al "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network". In: 2019 19TH IEEE international conference on data mining (ICDM 2019). DOI: 10.1007/s10462-02210270-w.

[2] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA,2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[3] Woo Suk Choi, Yu-Jung Heo, Dharani Punithan, Byoung-Tak Zhang 2022 Scene Graph Parsing via Abstract Meaning Representation in Pre-trained Language Models DOI: 10.18653/v1/2022.dlg4nlp-1.4

[4] J. Sudhakar, V. V. Iyer and S. T. Sharmila, "Image Caption Generation using Deep Neural Networks," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-3, doi:10.1109/ICONAT53423.2022.9726074.

[5] V. G. Biradar, M. G, S. Agarwal, S. K. Singh and R. U. Bharadwaj,"Leveraging Deep Learning Model for Image Caption Generation for Scenes Description," 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 2023, pp. 1-5, doi: 10.1109/EASCT59475.2023.10393602.

[6] L. Abisha Anto Ignatious., S. Jeevitha., M. Madhurambigai. and M.Hemalatha., "A Semantic Driven CNN– LSTM Architecture for Person-alised Image Caption Generation," 2019 11th International Conference on Advanced Computing (ICoAC), Chennai, India, 2019, pp. 356-362, doi: 10.1109/ICoAC48765.2019.246867.

[7] K. Priya, R. V. Babu, M. M. Reddy, T. H. Mohan Reddy and M.Maanesh, "Image Caption Generation using CNN and Audio Conversion," 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2024, pp. 274-278, doi:10.1109/ICSSAS64001.2024.10760457.

[8] S. C. Krishna, P. N. Reddy and P. Kirubanantham, "Image Caption Generation with Python: A Deep Learning-Based Approach," 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, 2024, pp. 1211-1216, doi:10.1109/ICICNIS64247.2024.10823140.

[9] Y. Wang, Y. Wang, J. Zhu and S. Lou, "Image caption generation method based on target detection," 2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 2023, pp. 151-155, doi: 10.1109/ICIIBMS60103.2023.10347603.

[10] L. Panigrahi, R. R. Panigrahi and S. K. Chandra, "Hybrid Image Captioning Model," 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OT-CON), Raigarh, Chhattisgarh, India, 2023, pp. 1-6, doi: 10.1109/OT-CON56053.2023.10113957.

[11] K. Priya, R. V. Babu, M. M. Reddy, T. H. Mohan Reddy and M.Maanesh, "Image Caption Generation using CNN and Audio Conversion," 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2024, pp. 274-278, doi:10.1109/ICSSAS64001.2024.10760457.

[12] S. Li and L. Huang, "Context-based Image Caption using Deep Learning," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2021, pp. 820-823, doi:10.1109/ICSP51882.2021.9408871.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021).An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprintarXiv:2010.11929.

[14] Kim, J. Kim, B. -K. Lee, S. Shin and Y. M. Ro," Mitigating Dataset Bias in Image Captioning Through Clip Confounder-Free Captioning Network," 2023 IEEE International Conference on

Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 1720-1724, doi:10.1109/ICIP49359.2023.10222502.

[15] S. K. Satti, G. N. V. Rajareddy, P. Maddula and N. V. Vishnumurthy Ravipati," Image Caption Generation using ResNET-50 and LSTM," 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2023, pp. 1-6, doi: 10.1109/SILCON59133.2023.10404600