International Journal of Basic and Applied Sciences, 14 (1) (2025) 334-338

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/m9njft54
Research paper

A unified deep learning model for image captioning and text-to-image synthesis

Rose Mary Mathew [1] *, Sujesh P Lal [1], Gowri Ganesh [2]

[1] *Assistant Professor, Department of Computer Applications, Federal Institute of Science and Technology, Angamaly, Kerala, India*
[2] *PG Student, Department of Computer Applications, Federal Institute of Science and Technology, Angamaly, Kerala, India*
**Corresponding author E-mail: rosem.mathew@gmail.com*

Received: March 21, 2025, Accepted: April 27, 2025, Published: May 19 2025

Abstract

Deep learning models have significantly advanced various artificial intelligence tasks, including text-to-image generation and image captioning. However, there remains a semantic gap between textual descriptions and visual representations, which affects the accuracy and coherence of generated images and captions. This paper proposes a novel deep learning model that integrates Stable Diffusion, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks to enhance both text-to-image generation and image captioning tasks. The model employs CNN-LSTM architecture for feature extraction, while Stable Diffusion refines the output iteratively to improve coherence and realism. The approach generates diverse and high-quality images from text inputs and produces accurate captions for images. Experimental evaluations demonstrate the effectiveness of our approach. The image captioning model achieved a BLEU score of 0.89, highlighting its high accuracy. The text-to-image generation results also exhibit substantial improvements in visual realism and semantic alignment. The proposed model offers a robust framework for multimodal AI applications, advancing both content synthesis and understanding in multimedia tasks. These findings underscore the potential of deep learning in bridging the gap between textual and visual modalities, contributing to more effective and versatile AI-driven multimedia solutions.

Keywords: *CNN-LSTM; Image Captioning; Stable Diffusion; Text to Image.*

1. Introduction

The combination of deep learning and multimodal learning has made remarkable strides in various fields, particularly in text-to-image synthesis and image captioning. These tasks are crucial for applications such as content generation, accessibility technologies, and automated multimedia understanding. However, generating realistic images from text and creating accurate captions for images remains challenging due to the semantic gap between textual descriptions and visual representations. To address this challenge, researchers have explored Stable Diffusion models for text-to-image generation and CNN-LSTM architecture for image captioning. Stable Diffusion models iteratively refine images by adding controlled noise, enabling the creation of highly realistic images that align with textual descriptions [1]. On the other hand, Convolutional Neural Networks (CNNs) are effective at extracting visual features from images, while Long Short-Term Memory (LSTM) networks excel in sequence modelling, making them ideal for generating coherent image captions.

In this work, we focus on integrating Stable Diffusion for text-to-image synthesis and CNN-LSTM for image captioning. Our goal is to enhance multimodal learning by combining these techniques into a unified framework, improving upon state-of-the-art methods. Through extensive experiments, we demonstrate the effectiveness of our approach and its potential for bridging the intermodal gap in text-visual information exchange. This methodology has promising applications in multimedia content creation and comprehension, making AI-driven multimodal systems more efficient and versatile.

This work also draws conceptual inspiration from cognitive science and human-computer interaction, particularly in understanding how humans semantically associate language with visual perception. In addition to technical innovation, this work aligns with interdisciplinary research in cognitive psychology, particularly in modeling how humans perceive and interpret visual and linguistic stimuli. Insights into semantic association and perceptual coherence drawn from human cognition can inform future enhancements of the model's interpretability and user-centered design, especially in applications related to accessibility and assistive technologies. The remainder of this paper is organized as follows: Section 2 reviews the related work in this domain, Section 3 details the methodology used in this work, Section 4 presents the results obtained, and Section 5 discusses conclusions and future directions.

2. Related works

Diffusion models have emerged as a transformative approach in generative AI due to their ability to produce diverse and photorealistic images. Stable Diffusion v1.5, for example, demonstrated how textual prompts influence visual outputs through a human-in-the-loop

evaluation, showing that AI systems can rival or even surpass non-experts in distinguishing synthetic from real images [2]. Beyond mere generation, diffusion models also exhibit strong disentanglement capabilities, like Generative Adversarial Networks (GANs)—allowing style modifications without compromising semantic integrity [3]. Further advancements, such as the Lifelong Text-to-Image Diffusion Model (L2DM), addressed the challenge of catastrophic forgetting by incorporating memory supplementation and concept acknowledgment mechanisms, thereby preserving diversity across generated outputs. Experimental evaluations confirmed L2DM's superiority in prompt adaptability, unconditional image synthesis, and diversity under classifier guidance [4].

In parallel, image captioning research has matured through the integration of computer vision and natural language processing. CNN-LSTM architectures remain a dominant paradigm, where CNNs effectively extract spatial features from images, and LSTM networks decode them into coherent, sequential descriptions [5]. Applications range from assistive technology and surveillance to content retrieval. For instance, an automatic captioning model trained on the Flickr8k dataset demonstrated accurate content recognition and description generation, underscoring the practical utility of CNN-LSTM systems in real-world settings [6]. Ongoing research continues to refine feature representation and contextual relevance to enhance semantic coherence in generated captions.

Complementing these architectures are recent transformer-based multimodal models like CLIP and DALL·E 2. CLIP leverages contrastive learning over 400 million image-text pairs to align visual and textual representations in a shared embedding space, enabling zero-shot generalization across tasks [7]. While powerful, it shows performance limitations in fine-grained localization or domain-specific challenges. DALL·E 2 builds on CLIP embeddings using a two-stage diffusion process to generate high-resolution, semantically aligned images from text [8]. Despite its advances in realism and coherence, it may still struggle with spatial complexity or abstract concept rendering.

Despite these strides, existing systems often face limitations: text-to-image models may lack semantic precision or visual consistency, and image captioning models may falter in contextual richness and adaptability [9]. Notably, current frameworks tend to treat these tasks in isolation. This research proposes a unified architecture that leverages the complementary strengths of diffusion models and CNN-LSTM frameworks. While diffusion models offer open-ended visual creativity grounded in large-scale training and attention mechanisms, CNN-LSTM models contribute linguistic structure and contextual coherence. Their integration aims to bridge the semantic gap between modalities, resulting in improved performance for both image synthesis and caption generation. This fusion not only enhances interpretability and robustness but also lays the foundation for more effective multimodal AI systems capable of handling diverse real-world applications.

3. Methodology

This research presents an integrated multimodal framework that synergistically combines text-to-image generation and image captioning to bridge the semantic gap between visual and textual modalities. In the first component, Stable Diffusion is employed to generate high-quality images that accurately reflect the semantic content of textual inputs. In the second component, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are used in tandem to generate descriptive captions based on the visual content of images. The CNN extracts rich spatial features from generated or real-world images, while the LSTM processes these features to produce coherent and contextually relevant captions [6]. By aligning these two tasks within a unified architecture, the framework not only facilitates bidirectional modality translation but also enables mutual enhancement, where improved image generation supports more accurate captioning, and vice versa. The system's effectiveness will be evaluated through comprehensive metrics, including image realism, semantic alignment, and caption accuracy.

3.1. Text-to-image generation using Stable Diffusion

Stable Diffusion models are employed for high-quality text-to-image generation by utilizing a probabilistic diffusion process, inspired by non-equilibrium statistical physics. The method involves a forward diffusion process, which gradually adds noise to input data, and a reverse diffusion process that progressively removes noise to reconstruct meaningful visual content. This architecture, including Variational Autoencoders (VAE), U-Net, and a Text Encoder, allows for efficient image generation.

Several steps are involved in training diffusion models. Figure 1 shows the architecture of text-to-image generation. The first step is image encoding, which converts input images into lower-dimensional representations using an image encoder. The second step is latent space embedding, which refers to a compressed, abstract representation of data that a machine learning model creates internally. This map encodes images into a high-dimensional latent space where transformations occur. In the next step forward, diffusion takes place. Noise is iteratively added to the latent representation until complete distortion. Finally, the denoising process in which models learn to reverse noise additions, reconstructing high-quality images.

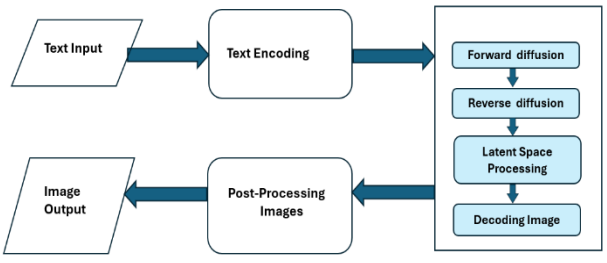

Fig. 1: Architecture for Text-to-Image Generation.

The algorithm for Text-to-Image Generation Using Stable Diffusion is specified below:

Step 1: Initialization: Load Stable Diffusion Model

i) Initialize the Stable Diffusion model with predefined parameters.

ii) Load the necessary training data and pre-trained weights.

Step 2: Text Encoding: Text Processing

i) Input: Receive a textual description.

ii) Convert the textual description into a latent representation using a Text Encoder

Step 3: Image Generation Process: Forward Diffusion & Reverse Diffusion

i) Start with Noise: Initialize with a noisy image or random noise.

ii) Add Noise Iteratively
iii) Initialize with Distorted Image
iv) Start with the fully distorted image from the forward diffusion process.
v) Denoise Iteratively
Step 4: Image Output: Generate Final Image
i) Convert the denoised image into a final high-quality image.
The algorithm for text-to-image generation using Stable Diffusion follows a structured process. First, the Stable Diffusion model is initialized with predefined parameters, pre-trained weights, and necessary training data. The input text is then processed through a Text Encoder, converting it into a latent representation. The image generation process begins by initializing a noisy image, either from random noise or a distorted version of an image and iteratively refining it through forward and reverse diffusion steps. Noise is gradually removed in a controlled manner to reconstruct meaningful visual elements. Finally, the denoised image is converted into a high-quality image, completing the text-to-image transformation.

3.2. Caption generation using CNN-LSTM

This study uses a CNN-LSTM hybrid model for image captioning, combining Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for sequential caption generation. The dataset used includes a diverse collection of images curated from sources like Unsplash and Pinterest, with annotations ensuring comprehensive coverage. Additionally, the Flickr8K dataset, comprising 8,000 images with five descriptions each, was utilized. Of these, 6,000 images were used for training, while 2,000 were allocated for validation and testing.
The model consists of two main stages: Feature Extraction and Caption Generation. For Feature Extraction, extracts high-level features from input images, capturing essential visual patterns and semantics. For Caption Generation, the extracted features are fed into an LSTM network, which generates captions word-by-word [10]. This process ensures that the generated captions maintain contextual consistency by leveraging LSTM's ability to handle sequential data. Figure 2 shows the architecture of the caption generation process.

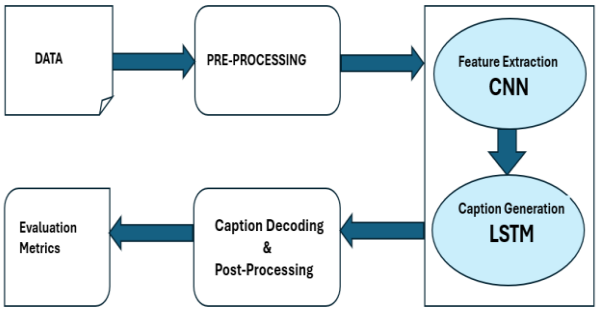

Fig. 2: Architecture for Caption Generation.

The training process involves fine-tuning the CNN on the dataset to optimize feature extraction, with the LSTM network trained using a cross-entropy loss function to predict the next word in a sequence based on the extracted image features. Optimization is carried out iteratively using the Adam optimizer with an adaptive learning rate, ensuring convergence over multiple epochs. To mitigate overfitting, batch normalization and dropout techniques are employed. Experimental evaluations assess the model's performance on the Flickr8K dataset, using the standard metric BLEU score. These metrics evaluate the fluency, accuracy, and relevance of generated captions compared to human-annotated references. The results demonstrate the effectiveness of the CNN-LSTM architecture in producing high-quality captions [11].
In-depth analysis explores the impact of dataset diversity, model hyperparameters, and training strategies on caption generation. Areas for future improvement and potential research directions are also discussed. This study contributes to advancing image understanding by integrating CNNs and LSTMs for generating descriptive image captions, with applications in automated image description, such as accessibility tools, image retrieval, and digital content annotation [12]. Each image in the dataset is accompanied by five captions, which undergo preprocessing to ensure consistency and usability [13]. Text preprocessing includes removing punctuation, numbers, and special tokens, converting text to lowercase, and tokenizing the captions using a fixed vocabulary size of 8,464 words. Word frequency analysis helps refine the dataset by identifying the 50 most frequently used terms.
Before images are fed into the model, they undergo a preprocessing pipeline to ensure compatibility with different architectures [14]. For the Xception model, images are resized to 299×299 pixels, while for the VGG16 model, they are resized to 224×224 pixels. The images are also flattened, and their pixel values are normalized to improve model performance. The model is implemented using Keras 2.0, with TensorFlow as the backend for building and training the neural networks. This comprehensive preprocessing ensures that both image data and corresponding captions are properly formatted for model training and evaluation.
Algorithm: Image Caption Generation
Step 1: Preprocessing
1) Image Preprocessing:
i.Resize Images: Adjust the dimensions of images to the required size
ii.Normalize Pixel Values: Scale pixel values to a range
2) Caption Preprocessing
i) Clean Captions: Remove punctuation, numbers, and special tokens from captions.
ii) Convert to Lowercase: Ensure all captions are in lowercase for consistency.
iii) Tokenize Captions: Break captions into individual words or tokens for further processing.
Step 2: Feature Extraction
3) CNN Feature Extraction
i) Feed Image into CNN: Pass the pre-processed image through a pre-trained CNN model
ii) Extract Feature Vector: Obtain the high-level feature vector that represents the image.
Step 3: Caption Generation

4) LSTM Decoding:
i) Initialize LSTM: Use the feature vector extracted from the CNN to initialize the LSTM network.
ii) Generate Captions: Use the LSTM model to generate captions word-by-word based on the image features.
iii) Decoding Techniques: Implement techniques such as beam search or greedy decoding to produce the final caption.

Step 4: Post-Processing

5) Generate Final Caption:
i) Convert Sequence to Text: Transform the sequence of words generated by the LSTM into a coherent and readable caption.
ii) Ensure Coherence: Refine the caption to ensure grammatical correctness and contextual relevance.

Step 5: Evaluation

6) Evaluate Captions:
i) Apply Metrics: Use the evaluation metric BLEU score, to assess the quality of the generated captions.

The image caption generation process begins with a dual stage preprocessing pipeline for both images and text. Images are resized to the appropriate input dimensions for models like VGG16 or Xception and normalized for consistency. Captions are cleaned by removing punctuation, numbers, and special tokens, converting text to lowercase, and tokenizing into words to prepare them for training [15]. A pre-trained Convolutional Neural Network (CNN) is employed to extract high-level visual features from the processed images [16]. These features are then fed into a Long Short-Term Memory (LSTM) network, which generates captions word by word, ensuring contextual coherence. Decoding strategies such as greedy decoding or beam search enhance the fluency and accuracy of the generated sentences. Post-processing is applied to ensure grammatical correctness and semantic clarity in the final captions [17]. Model training involves fine-tuning the CNN and optimizing the LSTM using cross-entropy loss and the Adam optimizer. Techniques like dropout and batch normalization are used to prevent overfitting. Evaluation is conducted using the standard metric BLEU score, which compares the model-generated captions with human annotations for fluency and relevance [18].

This study presents a unified framework combining Stable Diffusion for text-to-image generation with a CNN-LSTM model for image captioning. Images are generated from text and then described through a captioning pipeline that uses CNNs for feature extraction and LSTMs for sequential text generation [19]. The synergy between modules creates a bidirectional system, ensuring consistency between visual and textual data. Evaluated on the Flickr8K dataset using BLEU scores, the model shows strong performance. This cohesive approach supports applications in accessibility, content creation, and semantic search, with future improvements focused on decoding strategies, architecture, and dataset diversity.

4. Results and discussion

This section presents our experimental results, showcasing examples of text inputs and corresponding generated images for text-to-image generation, as well as images and their generated captions for image captioning. From this analysis, we discuss the strengths and weaknesses of the models, highlighting key observations. The stable diffusion-based text-to-image generation module demonstrated promising results during testing. It effectively produced visually appealing images that closely matched the provided textual descriptions. The model performed well with a variety of textual prompts, generating high-quality images. However, some generated images lacked specific details or failed to accurately capture the intended concept, particularly with complex or abstract descriptions.

Table 1: Image Captioning BLEU Score Performance

Model	BLEU Score
CNN-LSTM	0.89

The CNN-LSTM-based image captioning module showed strong performance in generating captions for diverse images. The generated captions were generally relevant and accurately described the content of the images. However, in some cases, the model made incorrect assumptions or overlooked essential details, leading to captions that were overly generic or inaccurate. The performance of the image captioning model was evaluated using BLEU (Bilingual Evaluation Understudy) scores, which measure the overlap between generated captions and reference captions. A higher BLEU score indicates better performance. Table 1 shows the BLEU Score of the caption generated model. The model performs well in generating captions that align with human-provided references. However, BLEU scores do not account for creativity or contextual depth, which remain areas for further improvement.

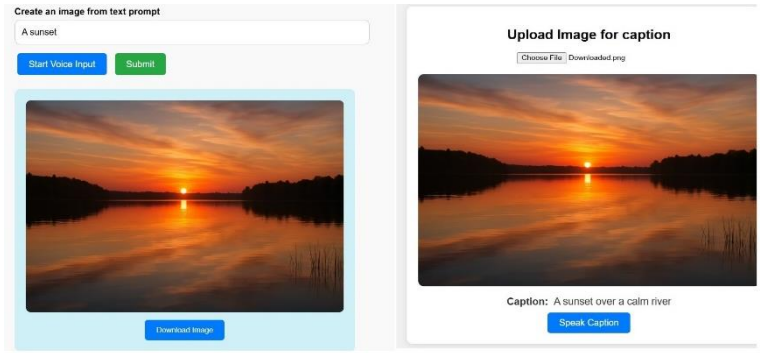

Fig. 3: Example of Image Generation and Caption Generator.

To enhance user experience, we implemented a download functionality, allowing users to save and share generated images. This feature was particularly beneficial for buyers who wished to store images from both tools. Additionally, a text-to-speech capability was integrated, enabling users to listen to generated captions instead of reading them. This functionality provided accessibility for individuals with auditory disabilities or those who preferred auditory feedback. A comparison of the two modules revealed distinct strengths and weaknesses. The stable diffusion-based text-to-image generation module excelled at producing aesthetically pleasing images but struggled with generating visuals from complex and abstract text prompts. On the other hand, the CNN-LSTM-based image captioning module generated meaningful

captions but occasionally produced generic or inaccurate descriptions. Overall, both models contributed effectively to the project goal of enabling users to generate images from text and captions from images. However, further refinement and optimization are necessary to enhance the quality and accuracy of outputs. Fig.3 represents the sample outputs of the work.

In addition to its technical and practical implications, this work acknowledges the need for addressing ethical considerations related to synthetic media, including the potential for misuse in generating misleading or harmful content. Policymakers must consider frameworks for the responsible deployment of AI-generated media, ensuring transparency, accountability, and the mitigation of bias in multimodal systems.

5. Conclusion

This paper presents a significant advancement in multimodal AI, particularly in text-to-image generation and image captioning. By integrating Natural Language Processing (NLP) and computer vision, the developed system effectively synthesizes images from text and generates accurate captions for images. The model selection, training, and evaluation processes have resulted in improved user satisfaction, enhanced image quality, and increased caption accuracy. Future improvements could focus on refining model architectures, leveraging better computational resources, and exploring novel attention mechanisms. Techniques like continual learning, domain adaptation, and multi-scale representations can further enhance system performance The project highlights the emerging potential of multimodal AI in enhancing content creation, accessibility, and human-computer interaction, to improve communication between humans and machines; however, its broader impact remains contingent on further validation across diverse datasets and real-world scenarios. From a technical standpoint, model compression and optimization can make the system deployable on mobile and edge devices. Advanced attention mechanisms, such as self-attention and multi-head attention, may improve the relationship between textual and visual elements. Adversarial training can mitigate mode collapse and enhance diversity in generated outputs. Domain adaptation techniques can tailor models for specific applications, improving relevance and usability. To ensure adaptability, continuous learning methods can help the system evolve with changing datasets and user preferences. Privacy-preserving techniques like federated learning and differential privacy can safeguard user data while supporting collaborative model improvements. Additionally, transfer learning and few-shot learning can improve adaptability with minimal labelled data. Multi-scale and hierarchical models can further enhance the understanding of complex textual and visual relationships. By exploring these innovations, the project seeks to push the boundaries of multimodal AI, paving the way for more advanced and efficient systems in text-to-image generation and image captioning.

References

[1] Papa, L., Faiella, L., Corvitto, L., Maiano, L., & Amerini, I. (2023). On the use of stable diffusion for creating realistic faces: From generation to detection. *Proceedings of the 2023 IEEE International Workshop on Biometrics and Forensics (IWBF)*, 1–6. https://doi.org/10.1109/IWBF57495.2023.10156981.

[2] Dhariwal, P., & Nichol, A. (2021). *Diffusion models beat GANs on image synthesis*.

[3] Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., & Chang, S. (n.d.). *Uncovering the disentanglement capability in text-to-image diffusion models*.

[4] Sun, G., Liang, W., Dong, J., Li, J., Ding, Z., & Cong, Y. (2024). Create your world: Lifelong text-to-image diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*, 6454–6470. https://doi.org/10.1109/TPAMI.2024.3382753.

[5] Sairam, G., Mandha, M., Prashanth, P., & Swetha, P. (2022). Image captioning using CNN and LSTM. *Proceedings of the 4th Smart Cities Symposium (SCS 2021)*. https://doi.org/10.1049/icp.2022.0356.

[6] Amritkar, C., & Jabade, V. (2018). Image caption generation using deep learning technique. *Proceedings of the 2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–4. https://doi.org/10.1109/ICCUBEA.2018.8697360.

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *38th International Conference on Machine Learning (ICML)*.

[8] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*.

[9] Rohitharun, S., Reddy, L. U. K., & Sujana, S. (2022). Image captioning using CNN and RNN. *Proceedings of the 2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*. https://doi.org/10.1109/ASIANCON55314.2022.9909146.

[10] Gorkar, T., Kale, V., Jagdale, V., Tarte, Y., & Battalwar, S. (2023). Image caption generator using deep learning. *International Research Journal of Modernization in Engineering Technology and Science*, *5*(12).

[11] Han, S.-H., & Choi, H.-J. (2020). Domain-specific image caption generator with semantic ontology. *Domain-Specific Image Caption Generator with Semantic Ontology.*, 526–530. https://doi.org/10.1109/BigComp48618.2020.00-12.

[12] Napa, K. K., Dhamodaran, V., Mohan, A., Laxman, K., & Yuvaraj, J. (2019). Detection and recognition of objects in image caption generator system: A deep learning approach. *Proceedings of the 2019 International Conference on Advanced Computing and Communication Systems (ICACCS*. https://doi.org/10.1109/ICACCS.2019.8728516.

[13] Yang, Z., Liu, Q., & Liu, G. (2020). Better understanding: Stylized image captioning with style attention and adversarial training. *Symmetry*, *12*(12). https://doi.org/, 12(12), 1978. https://doi.org/10.3390/sym12121978.

[14] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning (ICML), 10347–10357. https://doi.org/10.1109/ICCV48922.2021.00010.

[15] HSyed, U., & Subbarao, M. (2024). CaptionCraft: VGG with LSTM for image insights. *Proceedings of the IEEE International Conference on Emerging Technologies (ICET)*, 1–5. https://doi.org/10.1109/ICTEST60614.2024.10576172.

[16] Sharma, H., & Padha, D. (2024). Neuraltalk+: Neural image captioning with visual assistance capabilities. *Multimedia Tools and Applications*, 1–29. https://doi.org/10.1007/s11042-024-19259-9.

[17] Bansal, P., Malik, K., Kumar, S., & Singh, C. (2023). EfficientNet-based image captioning system. *Proceedings of the DICCT*. https://doi.org/10.1109/DICCT56244.2023.10110117.

[18] Latimier, A., Peyre, H., & Ramus, F. (2020). *A meta-analytic review of the benefit of spacing out retrieval*. https://doi.org/10.31234/osf.io/kzy7u.

[19] L. A. A. Ignatious, S. Jeevitha, M. M. and M. H. (2019). A semantic driven cnn-lstm architecture for personalised Image caption generation. *11th International Conference on Advanced Computing (ICoAC)*, 356–362. https://doi.org/10.1109/ICoAC48765.2019.246867.