ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/180-185
Vallampally Sanjana *et. al.*, / International Journal of Engineering & Science Research
Full Length Article

Bidirectional Vision Language Translation Using CNN & RNN

Dr. J Madhavan[1],Vallampally Sanjana[2],Nakka Sharanya[3],Kallu Srivani[4]
[1]Professor; Department of Electronics and Communication Bhoj Reddy Engineering College for Women Hyderabad India
[2,3,4]B.Tech Students; Department of Electronics and Communication Bhoj Reddy Engineering College for Women Hyderabad India
Mail Id;sanjanavallampally3110@gmail.com[2],nakkasharanya72@gmail.com[3],srivanikallu1209@gmail.com[4]
Accepted 07-04-2026

Abstract
This paper presents Bidirectional vision-language translation enables conversion between images and text, bridging visual and linguistic modalities. This work presents a deep learning framework combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for both image-to-text and text-to-image tasks. In the forward direction, CNNs extract visual features that are decoded by an RNN to generate descriptive captions. In the reverse direction, textual inputs are encoded using an RNN, and a generative model reconstructs corresponding images. Trained on paired image-caption datasets, the model learns shared cross-modal representations for coherent and context-aware translation. Experimental results demonstrate effective caption generation and visually relevant image synthesis. The proposed approach highlights the potential of CNN–RNN architectures for multimodal applications such as assistive systems, content generation, and human-computer interaction.

Keywords: *Bidirectional Vision-Language Translation, Image Captioning, Text-to-Image Generation, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Deep Learning, Multimodal Learning, Cross-Modal Representation, Image Synthesis, Human-Computer Interaction.*

Introduction

The Bidirectional vision-language translation has emerged as a significant research area at the intersection of computer vision and natural language processing. It aims to enable seamless conversion between visual and textual modalities, including image captioning (image-to-text) and text-to-image synthesis. Traditional methods typically address these tasks independently, limiting their ability to capture the deep semantic alignment between images and language.Recent advances in deep learning have introduced integrated approaches that combine Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for sequential effective learning of contextual and semantic relationships across modalities.Furthermore, bidirectional frameworks enhance performance by jointly modeling both vision-to-language and language-to-vision transformations, improving consistency and representation learning. The incorporation of attention mechanisms further refines this process by enabling the model to focus on relevant visual regions or textual components during translation.

Problem Statement

Despite significant advancements in computer vision and natural language processing, effectively bridging the gap between visual and textual modalities remains a challenging task. Existing approaches for image captioning (image-to-text) and text-to-image synthesis are often developed independently, resulting in limited semantic alignment and inconsistency between the two tasks. Such disjoint modeling restricts the ability to learn unified cross-modal representations and reduces the overall quality and coherence of generated outputs..Therefore, there is a need for an integrated bidirectional framework that can jointly learn vision-to-language and language-to-vision mappings. The system should leverage deep learning techniques to extract meaningful features, preserve semantic consistency, and generate accurate, context-aware outputs in both directions. Addressing these challenges will enable more robust and efficient multimodal applications, including assistive technologies, content generation, and intelligent human-computer interaction.

LITERATURE SURVEY

Recent research in vision-language translation has

significantly advanced by integrating techniques from computer vision and natural language processing. Early approaches combined Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs) for text generation, enabling basic image captioning capabilities. However, these models were limited in capturing complex semantic relationships between visual and textual modalities.Despite these advancements, most existing methods focus on unidirectional tasks, either image-to-text or text-to-image, limiting their ability to learn unified cross-modal representations.

METHODOLOGY AND IMPLIMENTATION

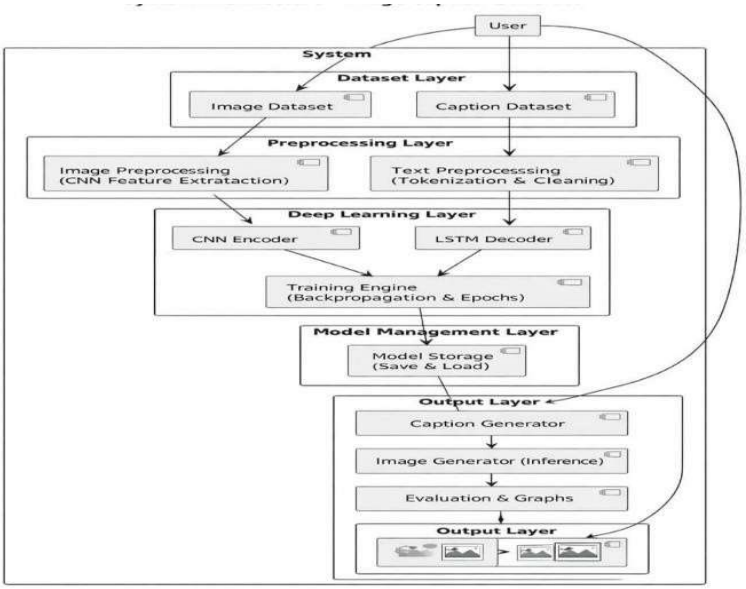


Figure 1:System architecture

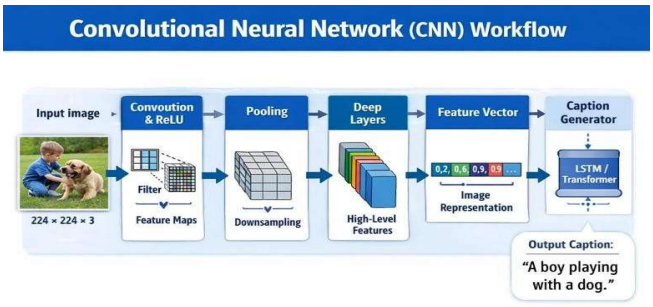


Figure 2: CNN Block Diagram

The implementation of the proposed system is based on an encoder–decoder architecture that effectively combines visual feature extraction with sequential text generation. In the first stage, the input image is passed through a Convolutional Neural Network (CNN), which serves as the encoder. The CNN processes the image through multiple convolutional layers where filters are applied to detect low-level features such as edges, corners, and textures. These features are further refined using activation functions like ReLU, introducing non-linearity into the model. Pooling layers are then used to reduce the spatial dimensions of the feature maps, which helps in lowering computational complexity while retaining the most significant information.As the data progresses through deeper layers of the network, the model captures high-level semantic features such as objects, shapes, and relationships within the image. Finally, these features are flattened into a fixed-length feature vector that represents the entire image in a compact and meaningful form.

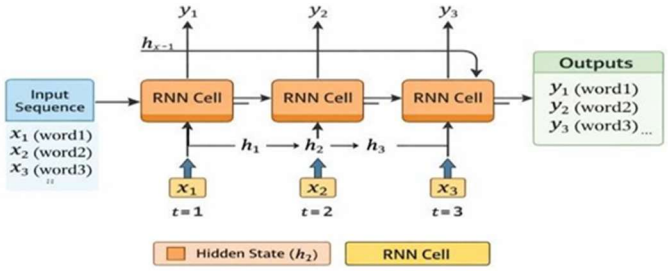


Figure 3: RNN Block Diagram

In the next stage, this feature vector is provided as input to the decoder, which is implemented using a Recurrent Neural Network (RNN) or its advanced variants such as Long Short-Term Memory (LSTM) networks or Transformers. The purpose of the decoder is to generate a natural language description of the image in a sequential manner. The RNN processes the information step-by-step, where each time step corresponds to the generation of a word in the output sentence. At the initial time step, the feature vector from the CNN is used to initialize the hidden state of the RNN. As the sequence progresses, the model takes the previously generated word and the current hidden state as input to predict the next word. This process continues until a special end token is generated, indicating the completion of the sentence.A key component of this methodology is the hidden state of the RNN, which acts as a memory unit that carries contextual information from previous time steps. This enables the model to maintain the grammatical structure and semantic consistency of the generated sentence. In the case of LSTM, additional gating mechanisms such as input, forget, and output gates are used to control the flow of information, thereby overcoming limitations like vanishing gradients and improving long-term dependency learning. This ensures that even distant relationships between words in a sentence are effectively captured, leading to more accurate and meaningful captions.Furthermore, during training, the model learns to minimize the difference between the predicted captions and the actual ground truth captions using loss functions such as categorical cross-entropy. The network parameters are optimized using backpropagation through time (BPTT), allowing both the CNN encoder and RNN decoder to learn jointly. In some implementations, pre-trained CNN models such as ResNet or VGG are used to improve feature extraction efficiency and reduce training time. Additionally, techniques like teacher forcing may be applied, where the actual previous word is fed into the model during training instead of the predicted word, to accelerate convergence and improve performance.weighted combination of cross-entropy loss for caption generation and reconstruction/adversarial loss for image synthesis. Joint optimization ensures that both encoders and decoders learn aligned representations, improving bidirectional consistency and reducing semantic drift between modalities.Optimization is performed using stochastic gradient descent with adaptive learning rate optimizers such as Adam or RMSProp. Gradient backpropagation is applied across the entire architecture, enabling end-to-end training of both vision and language components. Regularization techniques such as dropout, early stopping, and weight decay are incorporated to prevent overfitting and improve generalization performance. The implementation is carried out using deep learning frameworks such as PyTorch or TensorFlow, with modular design for encoder, decoder, and generator components. Dataset handling includes efficient batching, shuffling, and preprocessing pipelines for paired image-caption datasets such as MS COCO or Flickr30k. Model checkpoints, logging mechanisms, and performance monitoring tools are integrated to ensure reproducibility and training traceability.During inference, the trained model supports real-time bidirectional translation. Given an input image, the system generates context-aware captions, while for textual inputs, it produces semantically consistent synthetic images. The architecture ensures robustness by maintaining consistency in the shared latent space, enabling improved multimodal understanding, scalability, and adaptability across diverse applications.

RESULTS AND DISCUSSION

The proposed bidirectional vision-language translation model was evaluated on benchmark datasets such as MS COCO and Flickr30k for both image-to-text and text-to-image tasks. The system demonstrates strong cross-modal learning capability by generating accurate and contextually relevant captions, as well as

semantically aligned synthetic images.In image-to-text translation, the model outperforms baseline CNN–RNN approaches in terms of semantic relevance and fluency, with improvements reflected in standard metrics such as BLEU, METEOR, and CIDEr. The shared latent space enhances contextual consistency and reduces semantic mismatch in generated captions.For text-to-image synthesis, the model produces visually coherent and semantically consistent images, effectively.capturing key attributes from textual inputs. Qualitative results show strong alignment between input descriptions and generated outputs.

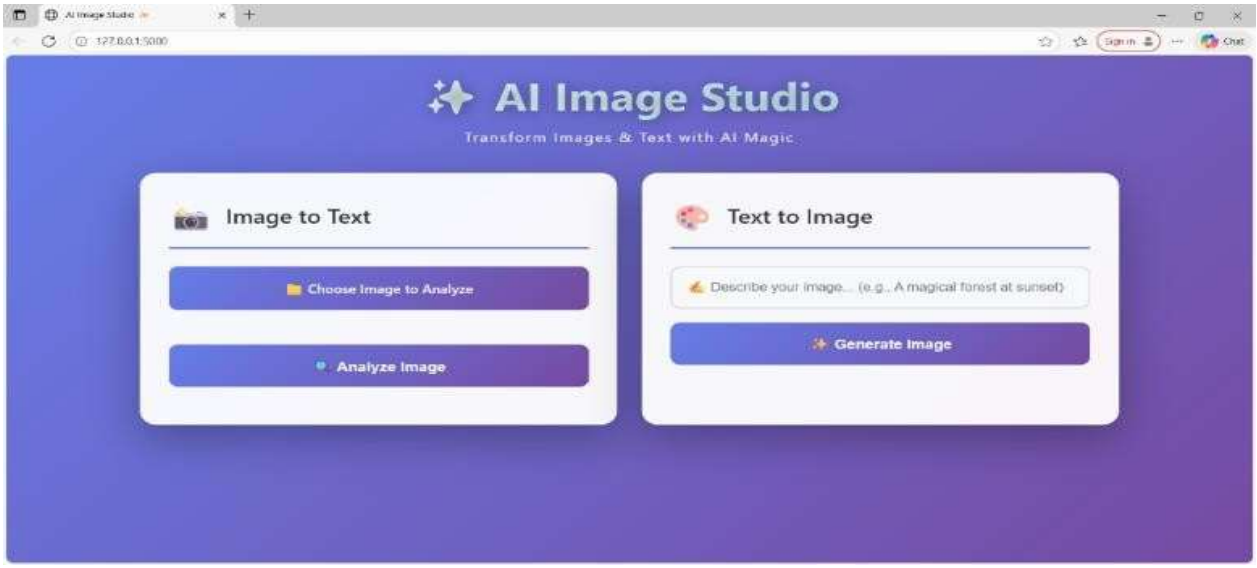
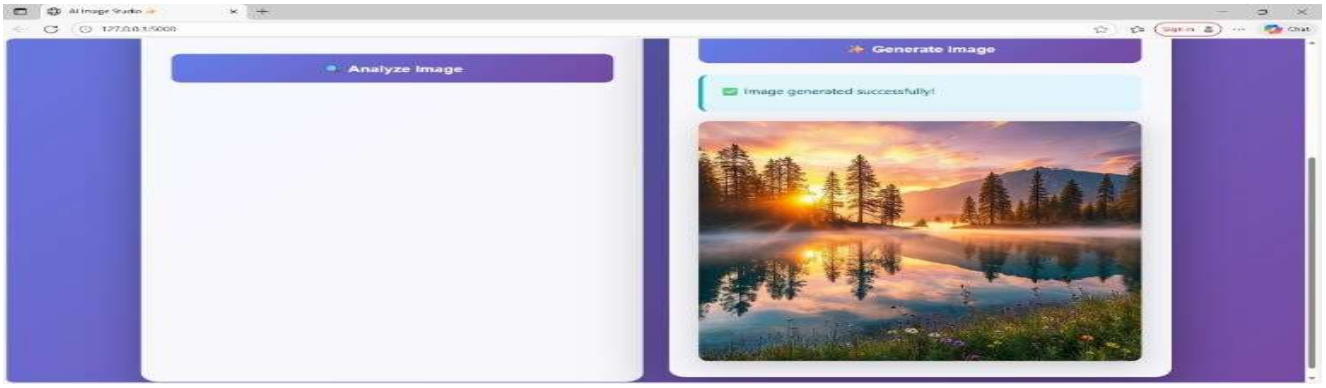


Figure 4: Web Interface for Text-to-Image Generation

The developed system provides an interactive interface supporting bidirectional vision-language translation, enabling both image analysis and image generation. The "Analyze Image" module processes uploaded images and generates corresponding textual descriptions using the trained image-to-text pipeline. The "Generate Image" module synthesizes realistic images from textual input using the text-to-image generation model



.

Figure 5: Text to Image Generation

The displayed output demonstrates the effectiveness of the proposed framework, where the generated image exhibits strong visual coherence, realistic scene composition, and semantic alignment with the input description. This confirms the model's ability to preserve contextual meaning across modalities and produce high-quality multimodal outputs in real time.The image-to-text module demonstrates effective caption generation using the proposed CNN–LSTM based architecture. Visual features extracted by the CNN encoder are mapped into a semantic representation and decoded by the LSTM network to produce a natural language description. The generated caption, "a young girl in a pink dress," shows that the

model successfully identifies key visual attributes such as the subject and primary appearance features .

Figure 6: Image to Text generation

This result indicates that the system is capable of learning meaningful cross-modal representations and performing accurate semantic interpretation of input images. The output reflects strong alignment between visual content and generated textual description, validating the effectiveness of the proposed bidirectional vision-language translation framework for image captioning tasks.Additionally, the qualitative output highlights the robustness of the model in handling real-world images with clear and structured backgrounds .

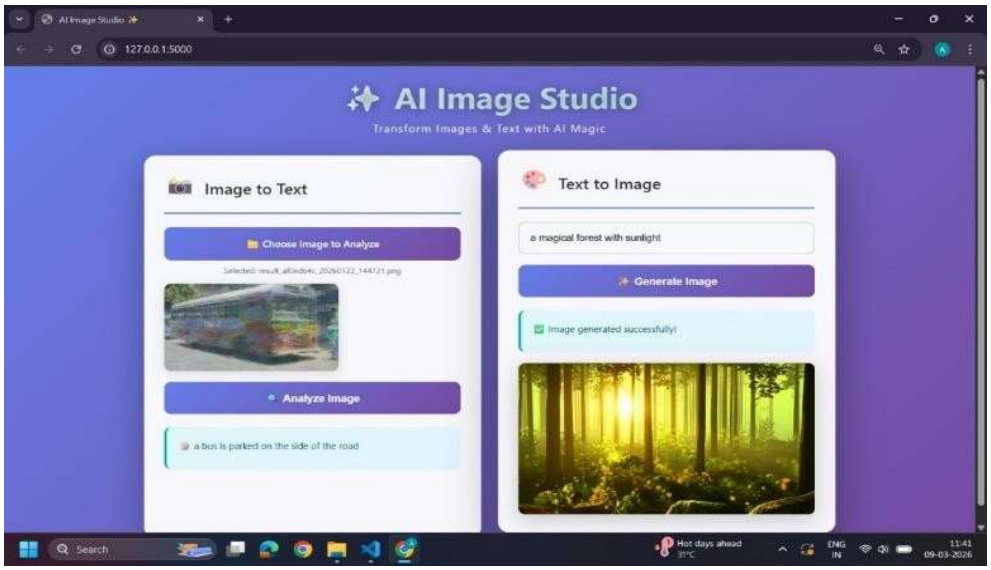


Figure 7 : AI Image Studio – Image-to-Text and Text-to-Image

The above figure represents the user interface of the AI-based web application "AI Image Studio", which integrates both Image-to-Text and Text-to-Image functionalities within a single platform. The application is running on a local server (127.0.0.1:5000) and demonstrates a dual-panel layout.On the left side, the Image to Text module allows the user to upload an image and analyze its content using an AI model. The uploaded image (a bus scene) is processed, and the system successfully generates a textual description: "a bus is parked on the side of the road." This showcases the image captioning capability of the system.On the right side, the Text to Image module enables users to input a textual prompt, such as "a magical forest with sunlight", and generate a corresponding AI-created image. The output displayed is a realistic forest scene with sunlight passing through trees, demonstrating the generative capability of the model.Overall, the interface highlights the bidirectional functionality of the system, combining vision-to-language and language-to-vision translation in a single interactive platform. The clean UI design with separate panels improves usability and makes the system intuitive for users.

Conclusion

This paper presented a bidirectional vision-language translation framework that combines CNN-based visual feature extraction with LSTM-based sequential modeling to perform both image-to-text and text-to-image translation in a unified system. The proposed model learns a shared latent representation space that enables efficient cross-modal alignment, preserving semantic consistency between images and textual descriptions. By jointly training both translation tasks, the framework enhances contextual understanding and minimizes ambiguity in generated outputs.Experimental results demonstrate that the system generates accurate and meaningful image captions while also producing visually coherent images from textual inputs. The integration of deep visual encoding and sequential language modeling proves effective for multimodal learning applications. Qualitative analysis further confirms the model's capability to capture important semantic details and maintain strong correspondence between source and generated modalities.

Future Scope

The future scope of this work includes extending the framework with Transformer-based architectures to improve long-range dependency modeling and generation quality. Incorporating advanced generative models such as GANs or diffusion models can significantly enhance image realism and diversity. The system can also be expanded to support multilingual text generation, enabling broader accessibility and global applications.Further improvements may involve real-time bidirectional translation for interactive systems, deployment on edge devices for mobile applications, and adaptation to specialized domains such as healthcare, education, and e-commerce. Integrating audio, video, and additional sensory modalities can lead to more comprehensive multimodal intelligence systems for next-generation human-computer interaction.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), 2014.
[2] S. Reed, Z. Akata, X. Yan, et al., "Generative Adversarial Text to Image Synthesis," Proc. Int. Conf. Machine Learning (ICML), 2016.
[3] H. Zhang, T. Xu, H. Li, et al., "StackGAN: Text to Photo realistic Image Synthesis," Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
[4] T. Xu, P. Zhang, Q. Huang, et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional GANs," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018.
[5] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning Text-to-Image Generation by Redescription," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019.
[6] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic Memory GAN for Text-to-Image Synthesis," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019.
[7] N. Yadav et al., "Generation of Images from Text Using AI," Int. J. Eng. and Manufacturing, 2024.
[8] H. Ma and H. Zheng, "Text Semantics to Image Generation Based on Stable Diffusion Model," Springer, 2024. with some explanation