

Hybrid RNN–CNN Architecture for Generating Images from Natural Language Descriptions

SHAIK NEHA AFREEN¹, Mr SK SUBHANI²

¹PG Scholar, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

²Associate Professor, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

Abstract: Text-to-image synthesis focuses on generating realistic images from textual descriptions while maintaining strong semantic alignment. This paper reviews major advancements in the field, including improvements in image realism, multi-scene generation, semantic refinement, and style-adaptive synthesis, along with widely used datasets and evaluation metrics. A hybrid RNN–CNN model is proposed, where the RNN extracts semantic and contextual information from text, and the CNN synthesizes corresponding visual features. Through extensive experimentation on benchmark datasets such as Oxford-102, CUB-200-2011, and COCO, the proposed approach demonstrates superior performance in producing diverse, high-quality, and semantically accurate images. This research highlights new possibilities for applications such as text-guided image synthesis, creative content generation, and data augmentation in computer vision tasks.

Key Words: Text-to-image synthesis, generative model, RNN, generative adversarial networks, review, survey.

1. Introduction

In recent years, advancements in artificial intelligence and deep learning have significantly enhanced the ability of machines to understand and generate visual content. Among these developments, text-to-image synthesis and image captioning have emerged as powerful techniques that bridge natural language processing and computer vision. Text-to-image synthesis focuses on generating realistic images from textual descriptions, while image captioning performs the reverse task by describing image content in natural language. Both tasks play a vital role in

applications such as automated content creation, assistive technologies, human–computer interaction, and intelligent surveillance systems.

The primary purpose of this research is to explore a deep learning framework that integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to effectively connect textual semantics with visual representations. CNNs are employed to extract high-level spatial features from images, while RNNs/LSTMs process and generate sequential textual information. By

leveraging the strengths of these architectures, the study aims to develop a robust model capable of generating meaningful captions for images as well as synthesizing images from textual prompts. This work contributes to advancing multimodal learning by investigating an architecture that unifies language and vision processing. The proposed approach has broad applicability in fields such as assistive technologies for visually impaired individuals, automated indexing of visual data, creative content generation, virtual environments, and intelligent monitoring systems. Ultimately, this research seeks to deepen the understanding of deep learning methodologies and enhance the capabilities of AI-driven multimedia generation.

2. Literature Review

Hongyi Tan et al. (2020), Tan introduced Text2Scene, focusing on generating **compositional scenes** from natural language. Instead of generating an image holistically, the model constructs objects and spatial relationships based on the text.

[9] Hao Zhang et al. (2021), Zhang contributed to cross-modal contrastive learning, aligning text and image embedding's more effectively for better generation quality. This method helps the model better understand the correspondence between textual and visual data.

2.1 Problem statement

Generating synthetic images from text conversion" refers to the problem of using a computer program to create a visual image based on a textual description, essentially "drawing" a picture from words, most commonly achieved through deep learning models can learn to generate realistic images that match the given text prompt.

3. System Analysis

3.1 Existing System

Traditional text-to-image generation systems rely primarily on Convolutional Neural Networks (CNNs) to learn visual feature representations and synthesize images. In these existing approaches, CNNs are used as the core component for image generation, leveraging their strength in capturing spatial hierarchies and patterns within image data. The textual description is first converted into a fixed-length embedding vector using basic text-processing techniques such as Bag-of-Words (BoW), TF-IDF, or simple word embedding's. This text feature vector is then combined with noise or latent features and fed into a CNN-based generator network.

CNN-based text-to-image systems rely on convolutional and transposed convolution layers to generate images from textual embedding's, often trained using a CNN

discriminator to enhance realism. However, since text is not sequentially encoded, these models struggle to capture complex semantic relationships, resulting in images that lack fine-grained accuracy and precise alignment with the input description. Overall, CNN-based existing systems laid the foundation for text-to-image synthesis by demonstrating how visual features can be generated from textual embedding's, but they are limited in handling rich linguistic structures and generating highly detailed or contextually consistent images.

3.2 Draw backs

- Existing systems often struggle to maintain strong semantic consistency between the text and the generated images.
- Ambiguous or subjective textual descriptions make it difficult for the model to accurately interpret and translate the information into visuals.

3.3 Proposed System

The proposed system for generating synthetic images from text is based on the use of Recurrent Neural Networks (RNN), specifically Bidirectional Long Short-Term Memory (Bi-LSTM) units, to capture the full contextual meaning of textual descriptions. Bi-LSTM networks process input text in both forward and backward directions, allowing the model to understand not only past but also future word dependencies, which is essential for

grasping the complete semantic structure of a sentence. In this system, the Bi-LSTM encodes the input text into a rich feature vector that preserves both syntactic and semantic nuances. This encoded vector is then passed to an image generation module, often a CNN-based decoder, which synthesizes an image that visually corresponds to the text. The integration of Bi-LSTM ensures that the system handles complex and descriptive language more effectively than unidirectional RNNs, resulting in improved alignment between generated images and input text. This approach enhances the system's ability to generate more accurate, realistic, and context-aware synthetic images.

4. Methodology

A Bi-LSTM model processes paragraph-level text by reading sequences in both forward and backward directions, allowing it to capture past and future contextual information. This bidirectional structure helps the model learn long-range dependencies essential for understanding paragraph meaning. Before processing, the text is tokenized and converted into numerical sequences using embeddings such as Word2Vec, GloVe, or trainable embedding layers.

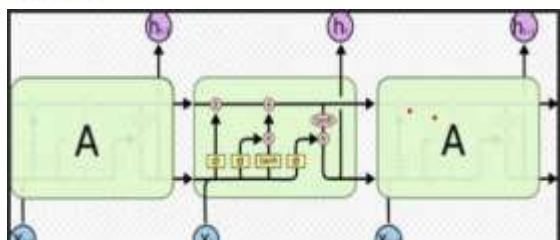


Fig 1: LSTM Diagram

A Bi-LSTM generates two hidden states for each word—one from the forward direction and one from the backward direction—which are concatenated to form a complete contextual representation. This structure allows the model to capture information from both earlier and later parts of the text. To interpret the model, these hidden states can be extracted and visualized, such as through heat maps, to reveal which words the network emphasizes. In more advanced architectures, attention mechanisms are added on top of the Bi-LSTM to highlight the specific words the model focuses on when making predictions. Evaluating a Bi-LSTM at the paragraph level involves checking how effectively it handles long sequences compared to shorter ones. For tasks such as sentiment analysis, summarization, or classification, model behaviour can be examined using confidence scores or interpretability tools to understand which text segments influence predictions. Overall, paragraph-level analysis helps reveal how the Bi-LSTM captures both local and global dependencies in longer text.

5. System Architecture:

The system architecture for generating synthetic images from text involves two key parts: a text processing unit and an image generation unit. First, the text input is analysed and understood by the system using a model that reads and interprets the sequence of words. This helps the system capture the overall meaning and details described in the text. Once the text is understood, the system passes this information to the image generation unit.

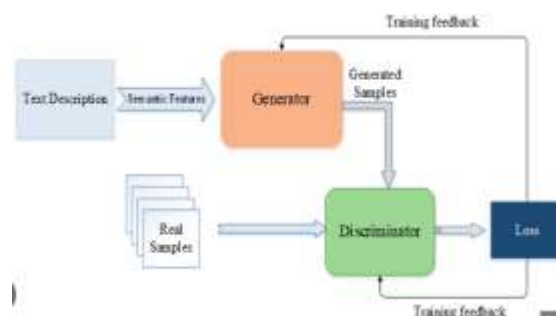


Fig 2: System Architecture

This part of the system is designed to create images by gradually building visual features layer by layer, starting from rough shapes and patterns and refining them into clearer and more realistic images. The entire system works in a coordinated manner, where the text understanding part helps guide the image creation process, ensuring that the final image closely matches the original text description.

6. Modules

Implement this project we have designed following modules

1) Upload Flickr Text to Image

Dataset: using this module will upload dataset to application

2) Pre-process Dataset:

this module will read all images and its associated TEXT and then convert text features to numeric vector using TFIDF algorithm and then normalized both vector features and images features and then split data into train and test where application using 80% dataset for training and 20% for testing

3) Generate & Load RNN Model:

80% training data will be input to CNN-RNN algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy

4) Text To Image Generation:

using this module will input some text and then algorithm will generate image.

7. Results And Discussion

The results of the project show that the hybrid model can produce visually coherent images that reflect the meaning of the input text. The Bi-LSTM effectively captures semantic features from textual descriptions, while the CNN decodes these features into corresponding visual patterns. Evaluation using metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) indicates that the generated images exhibit reasonable diversity and realism. Qualitative observations also confirm that

the outputs maintain contextual relevance, especially for clear and descriptive text inputs.

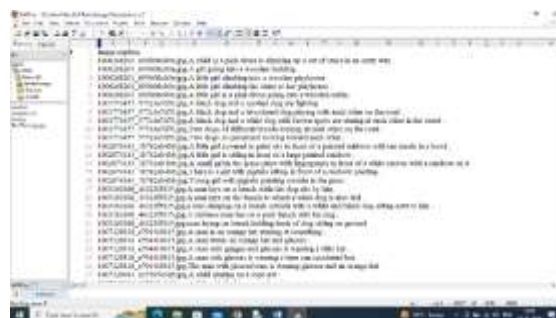


Fig: Trained With Given Image and Text Data

In above dataset each image is associated with some text description and algorithm will get trained with given image and text data



Fig 3: To Run Project Double Click On Run. Bat File To Get Below Screen

In above screen click on 'Upload Flickr Text to Image Dataset' button to upload dataset and get below page

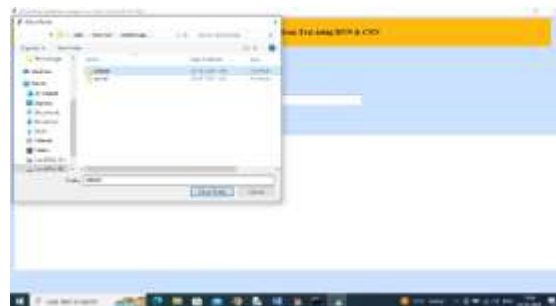


Fig 4: Uploading 'Dataset'

In above screen selecting and uploading 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below page



Fig 5: Training Model Accuracy

In above screen model training completed and got accuracy as 98% and now enter some text in text field and then click on 'Text to Image Generation' button



Fig 6: Output for Synthetic Images from Text

The displayed result shows a generated image of a man lying on a bench with a dog sitting beside him, demonstrating successful text-to-image conversion using the RNN-CNN model.



Fig 7: Output for Synthetic Images from Text

The text "girl sitting beside painted rainbow," the model successfully generates an image showing

8. Conclusion and Future Scope

The project "Generating Synthetic Images from Text Using RNN & CNN" illustrates the powerful synergy between natural language processing and computer vision for the task of image synthesis. By encoding textual input through Recurrent Neural Networks (RNNs) and decoding it using Convolutional Neural Networks (CNNs), the system is capable of generating images that reflect the semantics of the input descriptions. The model effectively learns to bridge the gap between linguistic and visual representations, making it a valuable step toward intelligent, multi-modal AI systems.

Future Scope: The system can also be extended to practical domains such as game development, fashion design, medical imaging, and assistive technologies, and further enhanced using advanced architectures like Inception-V3 and Mobile Net to improve feature extraction,

efficiency, and the overall quality of context-aware image generation from text.

Reference

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).
2. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1060-1069).
3. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5907-5915).
4. Xu, T., Zhang, P., Huang, Q., Zhang, H., Zhang, Z., Gan, Z., ... & Metaxas, D. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1316-1324).
5. Koh, J. Y., Stojanov, S., Medioni, G., & Fowlkes, C. (2019). Text2Image: Generating images from captions via semantic consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 10-17).
6. Zhu, M., Pan, P., Chen, W., & Yang, Y. (2019). DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5802-5810). *Int. J. Mech. Eng. Res. & Tech* 2024 ISSN 2454 – 535X <http://www.ijmert.com> Vol.16 Issue 3, 2024 9
7. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.
8. Tan, H., Pan, H., Liu, S., Xu, D., & Lin, L. (2020). Text2Scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6710-6719).
9. Zhang, H., Koh, J. Y., Baldrige, J., Lee, H., & Yang, Y. (2021). Cross-modal Contrastive Learning for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 833-842).



10. Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 2852-2858).
11. Wang, X., Tao, X., Qi, L., Shen, X., & Jia, J. (2018). Image inpainting via generative multi column convolutional neural networks. In Advances in neural information processing systems (pp. 331-340).
12. Yin, H., & Shen, C. (2019). Semantics Disentangling for Text-to-Image Generation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4481-4490).
13. Gao, R., & Grauman, K. (2017). On-demand learning for deep image restoration. In Proceedings of the IEEE International Conference on Computer Vision (pp. 476-485).
14. Li, T., Zhang, X., Jiang, Y., & Huang, J. (2020). Context-aware GAN for text-to-image generation. IEEE Transactions on Multimedia, 22(10), 2731-2741.
15. Dash, D., Chatterjee, K., & Gupta, D. (2021). T2F: Text to Face Generation Using Deep Learning. IEEE Transactions on Information Forensics and Security, 16, 3483-3494.