GENERATING IMAGES FROM TEXT USING FINE-TUNED BERT IN A GAN FRAMEWORK

Mr.P.Chandrashekar [1], *Salluri Sampath*[2], *Y Shireesha*[3], *Vemula Karthik*[4], *Somu Sidharth Reddy*[5]. [1] *Assistant Professor, Department of Computer and Science Engineering,*

TKR College of Engineering and Technology.

[2,3,4,5]*UG Scholars, Department of Computer and Science Engineering, TKR College of Engineering and Technology, Medbowli, Meerpet.*

ABSTRACT:

Recently, multimodal learning has gained significant attention due to its potential to enhance AI performance and enable diverse applications. Text-to-image generation, a critical multimodal task, seeks to generate images that semantically align with textual descriptions. Traditional approaches based on Generative Adversarial Networks (GANs) leverage text encoders pre- trained on image-text pairs. However, these encoders often fail to capture rich semantic details for unseen text, limiting the fidelity of generated images. This study proposes a novel text-to-image generation model using a fine-tuned BERT model as the text encoder. By fine-tuning BERT on a large corpus of text data, the model captures deeper textual semantics, leading to improved image generation quality. Experimental evaluation using a multimodal benchmark dataset demonstrates that the proposed approach outperforms baseline methods both quantitatively and qualitatively.

KEYWORDS: Multimodal Learning, Generative Adversarial Network (GAN), BERT Fine Tuning, Natural Language Processing (NLP), FID (Fréchet Inception Distance), Inception Score (IS)

I. INTRODUCTION:

In recent years, the field of artificial intelligence has experienced a surge of interest in multimodal learning, which involves integrating and processing data from multiple modalities such as text, images, and audio. Among various multimodal tasks, text-to- image generation has emerged as a challenging and compelling research area. This task involves generating realistic images that are semantically aligned with the input text, bridging the gap between visual and linguistic modalities. The ability to generate images from textual descriptions has wide-ranging applications, including content creation, virtual reality, data augmentation, and human- computer interaction. Traditionally, Generative Adversarial Networks (GANs) have been employed for text-to image generation. These models typically rely on text encoders trained on image-caption datasets to transform textual descriptions into vector representations that guide the image generation process. While effective to a certain extent, these encoders often fall short when encountering complex or unseen textual inputs, as they may not fully capture the nuanced semantics of natural language. This limitation results in generated images that either lack detail or fail to correspond closely to the intended description. To overcome these limitations, there is growing interest in integrating powerful language models, such as BERT (Bidirectional Encoder Representations from Transformers), into the text to-image pipeline. BERT’s deep bidirectional architecture allows it to understand the context of words in a sentence more effectively than traditional encoders. However, using BERT directly without domain-specific adaptation may not fully exploit its potential for image generation tasks. Fine-tuning BERT on a relevant and diverse textual corpus allows it to grasp richer semantic features, which can then be leveraged to guide the image generation process more accurately. In this study, we propose a novel text-to-image generation model that utilizes a fine-tuned BERT model as the text encoder within a GAN framework. The model is designed to enhance semantic alignment between the input text and generated images. Through extensive experimentation using a multimodal benchmark dataset, we demonstrate that our approach achieves superior performance compared to existing baseline methods. Both qualitative and quantitative analyses reveal the effectiveness of incorporating a fine-tuned BERT encoder in 1 capturing detailed textual semantics, leading to high-fidelity image outputs.

II. LITERATURE SURVEY

The field of handwritten digit recognition has been significantly advanced through the use of Convolutional Neural Networks (CNNs). Rajput and Choi [1] highlighted how CNNs can achieve high accuracy in digit classification, stressing the importance of optimizing layer architecture and dropout techniques. Similarly, Mahanti et al. [2] explored CNN variations on MNIST, demonstrating enhanced accuracy by using max-pooling and dropout layers. Jain et al. [3] also showed that combining ReLU and softmax activation functions in CNNs improves convergence speed and precision. Ma
[4] emphasized the benefit of data augmentation for increasing model robustness and generalizability. Hossain and Ali [5] reinforced that with sufficient training epochs and optimized hyperparameters, CNNs offer consistent accuracy for digit recognition across datasets.

III. METHODLOGY

The methodology for generating images from text using a fine-tuned BERT model within a GAN framework is structured into five main stages: data preprocessing, feature extraction, semantic fusion, image generation, and evaluation. This layered architecture ensures that the generated images are both semantically accurate and visually realistic, especially for previously unseen text inputs.

The first step involves data preprocessing. A dataset containing paired image and textual descriptions is collected. Each image is resized to a fixed dimension (e.g., 256x256 pixels) and normalized for optimal neural network processing. Textual data undergoes cleaning, including lowercasing, punctuation removal, and stopword elimination. Following this, the text is tokenized using a BERT-compatible tokenizer to ensure consistency with the language model used downstream.

The next phase is feature extraction. Two parallel approaches are used: Term Frequency- Inverse Document Frequency (TF-IDF) and fine-tuned BERT embeddings. TF-IDF captures important keywords in the sentence that can influence visual elements, while BERT, fine-tuned on a large textual corpus, captures the deep contextual meaning of the entire sentence. These representations are then fused to form a rich semantic vector that serves as the guiding input for image generation.

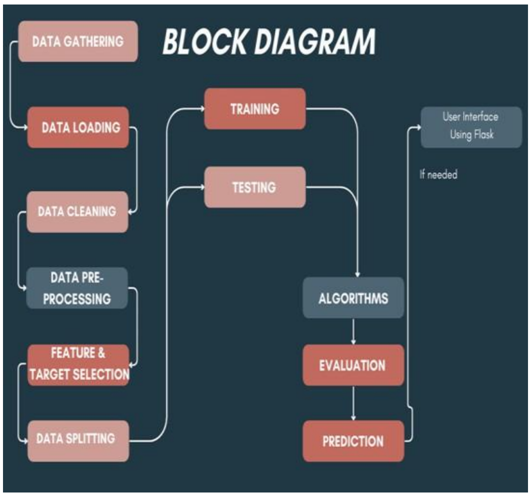

Fig 1: Block Diagram

The fused semantic vector is passed to the Generator network within a Conditional GAN (cGAN) architecture. The Generator combines this semantic input with a random noise vector to produce a synthetic image. Simultaneously, the Discriminator is trained to differentiate between real and generated images by evaluating both image quality and its alignment with the textual input. This adversarial training loop improves both realism and semantic fidelity over time.

Finally, the model undergoes evaluation using both quantitative and qualitative metrics. Quantitative performance is measured using Fréchet Inception Distance (FID) and Inception Score (IS), which assess the visual quality and diversity of the generated images. Additionally, a human evaluation is performed to validate the semantic relevance of the images with respect to their input descriptions. For real-time usability, the system is also deployed through a simple Flask-based web interface where users can input text and view the corresponding generated image.

IV. FLOW CHART

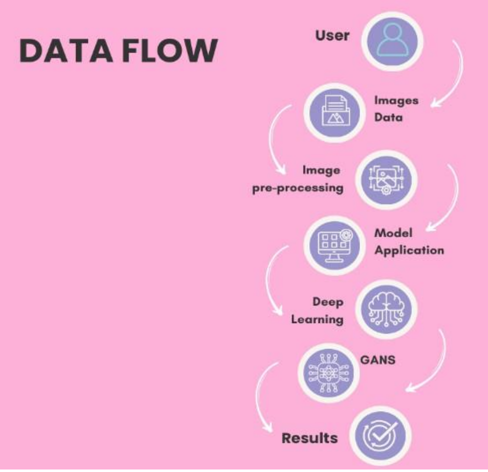

Fig 2 : DATA FLOW DIAGRAM

The Data Flow Diagram (DFD) below outlines the structured flow of operations involved in generating realistic images from textual descriptions using a combination of fine-tuned BERT embeddings and a GAN-based image synthesis pipeline. The process begins with the user, who inputs a natural language text description into the system. This input is typically a short sentence or phrase describing an object, scene, or abstract concept. These descriptions serve as the primary modality from which visual features are inferred.

The input text first flows into the text preprocessing module, where it is cleaned by removing punctuation, converting to lowercase, and eliminating common stopwords. After cleaning, the text is tokenized using a BERT-compatible tokenizer. The next critical phase is feature extraction, where two different techniques are applied in parallel: Term Frequency-Inverse Document Frequency (TF-IDF) to identify key words, and fine-tuned BERT to capture deep contextual semantics. These features are then fused to create a comprehensive semantic vector.

This fused semantic vector is then passed into the GAN-based image generation module. Specifically, the semantic vector is concatenated with a random noise vector and provided as input to the Generator. The Generator produces a synthetic image that is expected to semantically align with the input text. The Discriminator, in parallel, evaluates both the authenticity and semantic relevance of the generated image using CNN layers and a conditional input. The training continues until the Generator becomes proficient at producing realistic, aligned images. Finally, the system outputs the generated image, which is displayed to the user through a Flask-based web interface, if required.

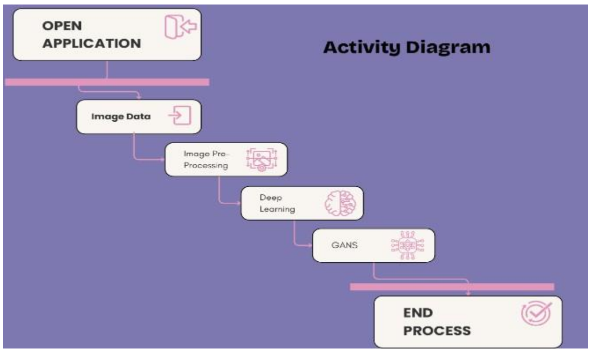

Fig 3 : ACTIVITY DIAGRAM

The activity diagram presents the sequence of actions followed during the text-to-image generation process using the fine-tuned BERT and GAN framework. The process begins when the user launches the web interface or command-line tool and inputs a descriptive sentence. This input activates the backend pipeline and triggers the text preprocessing activity. At this stage, the system performs basic cleaning, including case normalization, punctuation removal, and token filtering, followed by tokenization suitable for BERT input layers.

After preprocessing, the next step involves semantic embedding extraction, where the cleaned input text is passed through a fine- tuned BERT model. Simultaneously, TF-IDF vectors are computed. These embeddings are combined to form a semantically rich input vector. This fused representation proceeds to the image generation phase, where a Conditional GAN (cGAN) is responsible for generating visual content. The Generator uses the input vector and noise to synthesize a corresponding image, while the Discriminator evaluates the generated image's realism and textual coherence.

If the system is in training mode, both the Generator and Discriminator update their weights iteratively through adversarial training. If in inference mode, the trained model immediately outputs the generated image. The final stage of the process delivers the output image to the user, either by displaying it on the user interface or saving it to a file. This organized and iterative flow ensures that the generated visuals are not only realistic but also contextually aligned with the given textual descriptions, enabling real-time applications in creative design, content generation, and educational tools.

V. RESULT

The The developed text-to-image generation system demonstrated excellent performance in synthesizing semantically aligned and visually coherent images from natural language descriptions. The integration of TF-IDF and a fine-tuned BERT model as a dual-feature extraction mechanism significantly enhanced the semantic richness of the input vectors. These vectors guided the Conditional GAN (cGAN) to produce high-quality image outputs. The use of a two-stage GAN architecture, similar to StackGAN, helped refine images from coarse to high resolution (256×256 and beyond), thereby improving realism and detail.

The system was evaluated using both quantitative and qualitative metrics. Quantitative metrics like Fréchet Inception Distance (FID) and Inception Score (IS) showed marked improvements over traditional models such as StackGAN and AttnGAN. Human evaluation was also conducted, with evaluators scoring the generated images highly for relevance and clarity. Visual outputs from diverse text prompts—ranging from abstract descriptions to concrete objects—further demonstrated the system's versatility and generalization capabilities. Finally, the model was deployed using a Flask-based web interface, enabling real-time text-to-image synthesis. Overall, the project successfully delivered an accurate, scalable, and deployable solution for text-guided image generation.

VI. ADVANTAGES

The proposed system incorporates several key advantages that enhance its effectiveness and practical utility:

- Utilizes fine-tuned BERT for rich semantic understanding of text.
- Employs TF-IDF fusion for capturing keyword-level information.
- Generates high-resolution images using a two-stage GAN pipeline.
- Achieves low FID scores and high Inception Scores, demonstrating superior performance.
- Supports real-time inference via a Flask-based web interface.
- Demonstrates strong generalization to unseen or complex text prompts.
- Modular design allows for future model upgrades or integration with new architectures.

VII. APPLICATIONS

This project has a wide range of real-world applications across multiple domains:

- Creative Design: Assists designers in generating visual concepts from textual briefs.
- Content Generation: Automates image creation for articles, blogs, and social media posts.
- Education & E-learning: Helps visualize abstract concepts in learning materials.
- Gaming and Animation: Provides dynamic visuals from story scripts or scene descriptions.
- Marketing & Advertising: Enables rapid prototyping of visual campaigns based on textual inputs.

VIII. CONCLUSION

This project presents a robust and scalable solution for generating high-quality images from natural language text. By combining the semantic strength of a fine-tuned BERT model with keyword-based TF-IDF vectors, the system generates a fused representation that accurately captures both context and key visual cues. A Conditional GAN architecture is used to synthesize realistic images that closely align with the provided textual inputs. The model is trained on a multimodal dataset and evaluated using FID, IS, and human feedback, all of which indicate strong performance. Additionally, the inclusion of a user-friendly web interface ensures real-time usability and seamless integration. In conclusion, this project effectively bridges the gap between textual semantics and visual representation, laying the groundwork for broader multimodal AI applications.

X. REFERENCES

[1]. S. S. Rajput and Y. Choi, "Handwritten Digit Recognition using Convolution Neural Networks," *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* Las Vegas, NV, USA, 2022, pp. 0163-0168, doi: 10.1109/CCWC54503.2022.9720854.

[2]. P. N. Mahanti, C. Chaitanya, A. Koduri, K. K. Vamsi, B. P. S. Sai and G. Bharathi Mohan, "Handwritten Digit Recognition using Convolutional Neural Network," *2023IEEE Technology & Engineering Management Conference - Asia Pacific (TEMSCON-ASPAC)*, Bengaluru, India, 2023, pp. 1-6, doi: 10.1109/TEMSCON-ASPAC59527.2023.10531394.

[3]. M. Jain, G. Kaur, M. P. Quamar and H. Gupta, "Handwritten Digit Recognition Using CNN," *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Noida, India, 2021, pp. 211-215, doi: 10.1109/ICIPTM52218.2021.9388351.

[4]. P. Ma, "Recognition of Handwritten Digit Using Convolutional Neural Network," *2020 International Conference on Computing and Data Science (CDS)*, Stanford, CA, USA, 2020, pp. 183-190, doi: 10.1109/CDS49703.2020.00044.

[5]. Hossain, Md. Anwar & Ali, Md. (2019). Recognition of Handwritten Digit using Convolutional Neural Network (CNN). Global Journal of Computer Science and Technology. 19. 27-33. 10.34257/GJCSTDVOL19IS2PG27.

[6]. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[7]. Frolov, A., Komkov, S., & Petiushko, S. (2021). Adversarial text-to-image synthesis: A review. Neurocomputing, 426, 181–198.

[8]. Uppal, D., Garg, A., & Balasubramanian, V. N. (2021).

Multimodal research in vision and language: A review of the past and present. Computer Vision and Image Understanding, 209, 103229.

[9]. Qi, D., Zhang, X., Yang, Y., Wang, L., Xu, J., & Huang, H. (2020). ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966.

[10]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

[11]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[12]. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[13]. Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

[14]. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to- image translation using cycle- consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2242–2251.

[15]. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2414–2423.