International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Journal Survey on Text to Image Generation Models

Parikshit Hegde

Information Science and Engineering, BMS College of Engineering, Bengaluru, India

ABSTRACT

Text-to-image generation (TIG) involves the use of models capable of processing text input to generate high-fidelity images based on textual descriptions. The roots of text-to-image generation using neural networks can be traced back to the emergence of Generative Adversarial Networks (GANs), followed by autoregressive Transformers. Among generative models, diffusion models stand out as a prominent type, employing a systematic introduction of noises through repeating steps for image generation.

The impressive results achieved by diffusion models in image synthesis have solidified their role as the primary image decoder in text-to-image models, propelling text-to-image generation to the forefront of machine-learning (ML) research. In the era of large models, scaling up model size and integration with large language models have further enhanced the performance of TIG models, yielding generation results nearly indistinguishable from real-world images. This revolutionizes the way we retrieve images.

Keywords: Text to Image Generation, Neural Networks, Generative AI, Machine Learning, Deep Learning, Literature Review, Image Synthesis, Image Generation

Introduction

Text-to-Image Generation (TIG) models represent a transformative intersection of natural language processing and computer vision, offering the capability to convert textual descriptions into visually compelling images. This dynamic field has witnessed remarkable progress over the years, propelled by advancements in deep learning and generative modeling. TIG models address the intriguing challenge of bridging the semantic gap between language and visual content, providing a powerful tool for various applications, including creative content generation, multimedia storytelling, and aiding the visually impaired.

The core principle behind TIG models involves the utilization of sophisticated neural networks, often leveraging architectures like Generative Adversarial Networks (GANs) and Autoregressive Transformers. These models have the ability to understand and interpret textual input, subsequently generating images that align with the provided descriptions. The process involves learning intricate patterns, textures, and contextual details from textual cues, showcasing the potential for these models to revolutionize content creation in diverse domains. As the landscape of TIG continues to evolve, it becomes imperative to conduct a comprehensive review of the existing literature to understand the nuances, challenges, and advancements within this burgeoning field. In this review, we delve into 15 seminal papers that have significantly contributed to the development and refinement of Text-to-Image Generation models.

Related Works

The landscape of text-to-image generation has witnessed evolution, with various approaches enhancing the foundational methods. The initial and straightforward strategy involves a single Generative Adversarial Network (GAN), designed to take a text caption embedding vector as input and subsequently generate a low-resolution image corresponding to the described content.

To improve the quality of text embeddings, researchers have explored the integration of convolutional-recurrent networks. In this approach, the input characters undergo convolutional layers before being processed through a Long Short-Term Memory (LSTM) network. The final output is derived by computing the average hidden unit activation over the entire sequence

Building upon this foundation, subsequent models have sought to refine and extend the capabilities of the basic image-generating GAN. These advancements aim to address challenges such as image resolution, diversity, and semantic coherence in the generated visual content. Researchers continue to explore innovative techniques to push the boundaries of text-to-image generation, opening avenues for enhanced creativity and accuracy in transforming textual descriptions into visual representations.

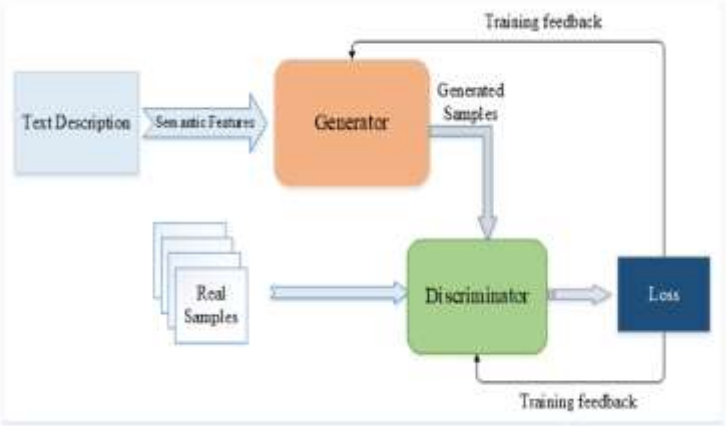

Fig 1: A basic text to image generation model using GAN [16]

Literature Survey

In recent years, there have been significant advancements in the field of text-to-image synthesis, with researchers focusing on improving the controllability, realism, and applicability of generated images. This literature review provides an overview of several key papers addressing diverse challenges in this domain.

[1] VectorFusion: Text-to-SVG by Abstracting Pixel Based Diffusion Models

Jain et al. address the challenge of generating high-quality abstract vector graphics (SVGs) from text captions. While diffusion models have shown promise in image generation, existing models lack the ability to directly generate SVGs. The proposed VectorFusion method introduces a two-phase approach, incorporating a pretrained text-to-image diffusion model and an optimization loop for refining shape parameters. Experimental results demonstrate VectorFusion's capability to generate abstract vector graphics, outperforming CLIP-based methods.

[2] DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Ruiz et al. introduce a novel approach for personalizing text-to-image diffusion models. The focus is on generating photorealistic images of specific subjects in different contexts based on a few input images. The proposed method expands the language-vision dictionary, enabling users to generate novel renditions of subjects guided by simple text prompts. Extensive experiments and user studies validate the effectiveness of DreamBooth in subject-driven generation, establishing it as a pioneering work in this domain.

[3] Shifted Diffusion for Text-to-image Generation

Zhou et al. propose Corgi, a novel diffusion model designed to enhance text-to-image generation. Corgi aims to bridge the gap between image and text modalities, enabling better utilization of pre-trained models like CLIP. The model demonstrates versatility by supporting supervised, semi-supervised, and language-free settings. Extensive experiments highlight the effectiveness of Corgi in various text-to-image generation scenarios, making it a promising advancement in the field.

[4] LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data

Park et al. address challenges in unpaired image-to-translation frameworks by introducing dataset-level annotation. The proposed LANIT framework utilizes candidate textual domain descriptions to specify target domains, reducing the need for per-sample domain labels. LANIT's generator framework, prompt learning, and domain regularization loss contribute to achieving comparable or superior results to existing methods while addressing the challenges of per-sample domain annotation.

[5] GLIGEN: Open-Set Grounded Text-to-Image Generation

Li et al. tackle limitations in large-scale text-to-image generation models by introducing a method that incorporates new grounding conditional inputs. GLIGEN retains text captions but introduces additional modalities like bounding boxes for better controllability. Gated Transformer layers are employed to preserve pretrained model knowledge while integrating new grounding information. GLIGEN demonstrates impressive zero-shot performance and generalization to unseen objects, showcasing its effectiveness in grounded text-to-image generation.

[6] SpaText: Spatio-Textual Representation for Controllable Image Generation

Avrahami et al. propose a novel problem setting for image generation, allowing users to specify both global and local text descriptions. The introduction of a spatio-textual representation, SpaText, enhances controllability by incorporating free-form text descriptions and positions. The method is implemented on state-of-the-art diffusion models and achieves state-of-the-art results in generating images with fine-grained control over scene composition and characteristics.

[7] ReCo: Region-Controlled Text-to-Image Generation

Yang et al. present ReCo, a model that combines the advantages of text-based and layout-based models to achieve precise region control in text-to-image generation. ReCo extends pre-trained models to understand spatial coordinate inputs, allowing users to provide region-controlled text inputs with free-form descriptions and position coordinates. Extensive evaluations demonstrate improved object classification accuracy and detector precision, highlighting ReCo's capability in handling challenging scenes.

[8] Training-Free Location-Aware Text-to-Image Synthesis

Mao and Wang introduce a method for fine-grained control over the location and size of objects in text-to-image synthesis. The proposed approach manipulates the values of cross-attention layers in diffusion models to control the position of individual objects without additional training. Evaluation metrics based on object detectors assess the efficiency of object-wise location-guided generation, demonstrating the effectiveness of the method in achieving user-aligned generation.

[9] BATINeT: Background-Aware Text to Image Synthesis and Manipulation Network

Morita et al. address challenges in text-to-image synthesis by introducing BATINet, a Background-Aware Text to Image synthesis Network. BATINet aims to generate foreground content that aligns cohesively with a specified background. The architecture comprises three networks, including a Position Detect Network, Generation Network, and Harmonization Network. Extensive experiments on the CUB dataset demonstrate BATINet's ability to generate high-quality images that seamlessly match the given background.

[10] Text-to-Image Generation Grounded by Fine-Grained User Attention

Koh et al. propose TRECS, a sequential generation model for grounded text-to-image synthesis, focusing on longer and more detailed narratives. The model leverages the Localized Narratives dataset, where annotators provide descriptions while pointing with a mouse over images. TRECS integrates a Tagger, Text-to-Image Dual Encoder, Composition of Masks, and Image Synthesis stages to align high-quality images with both language and spatial mouse traces. Evaluations on the COCO portion of Localized Narratives showcase TRECS's superiority over existing methods in terms of realism and image-text alignment.

[11] Towards Language-Free Training for Text-to-Image Generation

Zhou et al. introduce LAFITE, a generative adversarial approach that addresses the challenge of zero-shot text-to-image generation. The key contributions of LAFITE include its versatility, enabling effective performance in language-free, zero-shot, and fully supervised learning settings. Notably, it achieves superior results in zero-shot scenarios compared to state-of-the-art models while reducing the model's trainable parameter size. LAFITE also stands out for its cost-effectiveness, reducing the need for extensive GPU resources and thereby decreasing the associated carbon footprint. The authors achieve language-free training by constructing pseudo image-text feature pairs using the pre-trained CLIP model, showcasing the potential of leveraging existing models for novel tasks. Quantitative validations on various datasets demonstrate the effectiveness of LAFITE in different training schemes, making it a promising solution for text-to-image generation.

[12] DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis

Ruan et al. tackle limitations in existing text-to-image synthesis methods by introducing Dynamic Aspect-aware GAN (DAE-GAN). The paper emphasizes the importance of aspect-level features in textual descriptions and proposes a two-stage generation process involving a novel Aspect-aware Dynamic Re-drawer (ADR). ADR incorporates global and local refinement modules, enhancing both sentence-level and aspect-level features. The causality study demonstrates the interpretability of DAE-GAN, providing insights into how the model leverages aspect information for image refinement. The proposed framework exhibits superior performance through extensive qualitative and quantitative evaluations, showcasing its effectiveness in generating nuanced and detailed images from textual descriptions.

[13] DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis

Tao et al. [13] present DF-GAN, a one-stage text-to-image backbone that directly synthesizes high-resolution images. The novel Target-Aware Discriminator and Deep Text-Image Fusion Block (DFBlock) contribute to improved semantic consistency and effective text-image fusion. DF-GAN outperforms existing models in generating realistic and text-consistent images, addressing limitations in current text-to-image synthesis models. The one-way output strategy in the discriminator and the introduction of DFBlock for deep fusion further enhance the model's performance. Extensive experiments and benchmarking on challenging datasets demonstrate the superiority of DF-GAN, positioning it as a compelling solution for generating high-quality images from textual descriptions.

[14] Txt2Img-MHN: Remote Sensing Image Generation From Text Using Modern Hopfield Networks

Xu et al. focus on the relatively unexplored domain of remote sensing text-to-image generation with Txt2Img-MHN. The hierarchical prototype learning approach leverages Hopfield Networks to generate realistic remote sensing images based on given text descriptions. By emphasizing the learning of representative prototypes instead of diverse joint feature representations, Txt2Img-MHN outperforms existing methods. The paper provides a comprehensive performance comparison with VQVAE and VQGAN and explores zero-shot classification as a relevant evaluation criterion for remote sensing data. The contributions of Txt2Img-MHN extend beyond text-to-image generation, with potential applications in simulated urban planning and data augmentation for labeled samples.

[15. Text-to-Image Generation via Semi-Supervised Training

Ji et al. propose a semi-supervised approach for text-to-image synthesis to overcome challenges related to expensive labeled data. The Modality-Invariant Semantic-Consistent Module (MiSc) is introduced to bridge the gap between image and text modalities, facilitating the use of both labeled and unlabeled data in the training process. Experimental validations on MNIST and the Oxford 102 flower dataset demonstrate the effectiveness of the proposed method compared to traditional supervised approaches. The extension of the approach to image translation showcases its adaptability and versatility.

4. Conclusion

In Conclusion, the reviewed literature reflects the continuous evolution of TIG models, driven by advancements in deep learning, generative modeling, and the integration of large-scale language models. As TIG continues to push boundaries, it opens up new possibilities for content creation, human-computer interaction, and innovative applications across diverse fields. This review serves as a valuable resource for researchers, practitioners, and enthusiasts seeking to understand the state-of-the-art in Text-to-Image Generation and explore future directions in this exciting and rapidly advancing field.

References

[1] Jain, A. Xie and P. Abbeel, "VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 1911-1920, doi: 10.1109/CVPR52729.2023.00190.

[2] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 22500-22510, doi: 10.1109/CVPR52729.2023.02155.

[3] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen and J. Xu, "Shifted Diffusion for Text-to-image Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 10157-10166, doi: 10.1109/CVPR52729.2023.00979.

[4] J. Park et al., "LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 23401-23411, doi: 10.1109/CVPR52729.2023.02241.

[5] Y. Li et al., "GLIGEN: Open-Set Grounded Text-to-Image Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 22511-22521, doi: 10.1109/CVPR52729.2023.02156.

[6] O. Avrahami et al., "SpaText: Spatio-Textual Representation for Controllable Image Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 18370-18380, doi: 10.1109/CVPR52729.2023.01762.

[7] Z. Yang et al., "ReCo: Region-Controlled Text-to-Image Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 14246-14255, doi: 10.1109/CVPR52729.2023.01369.

[8] J. Mao and X. Wang, "Training-Free Location-Aware Text-to-Image Synthesis," 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 995-999, doi: 10.1109/ICIP49359.2023.10222616.

[9] R. Morita, Z. Zhang and J. Zhou, "BATINeT: Background-Aware Text to Image Synthesis and Manipulation Network," 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 765-769, doi: 10.1109/ICIP49359.2023.10223174.

[10] J. Y. Koh, J. Baldridge, H. Lee and Y. Yang, "Text-to-Image Generation Grounded by Fine-Grained User Attention," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 237-246, doi: 10.1109/WACV48630.2021.00028.

[11] Y. Zhou et al., "Towards Language-Free Training for Text-to-Image Generation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 17886-17896, doi: 10.1109/CVPR52688.2022.01738.

[12] S. Ruan et al., "DAE-GAN: Dynamic Aspect-aware GAN for Text-to-Image Synthesis," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 13940-13949, doi: 10.1109/ICCV48922.2021.01370.

[13] M. Tao, H. Tang, F. Wu, X. Jing, B. -K. Bao and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 16494-16504, doi: 10.1109/CVPR52688.2022.01602.

[14] Y. Xu, W. Yu, P. Ghamisi, M. Kopp and S. Hochreiter, "Txt2Img-MHN: Remote Sensing Image Generation From Text Using Modern Hopfield Networks," in IEEE Transactions on Image Processing, vol. 32, pp. 5737-5750, 2023, doi: 10.1109/TIP.2023.3323799.

[15] Z. Ji, W. Wang, B. Chen and X. Han, "Text-to-Image Generation via Semi-Supervised Training," 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, 2020, pp. 265-268, doi: 10.1109/VCIP49819.2020.9301888.

[16] Ramzan, S.; Iqbal, M.M.; Kalsum, T. Text-to-Image Generation Using Deep Learning. Eng. Proc. 2022, 20, 16. https://doi.org/10.3390/engproc2022020016