International Journal of Research Publication and Reviews

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

Text Guided Synthesis of Image with Artificial Intelligence

B. Rasagnya, K. Kiranmai

GMR Institute of Technology,Rajam,India.

ABSTRACT—

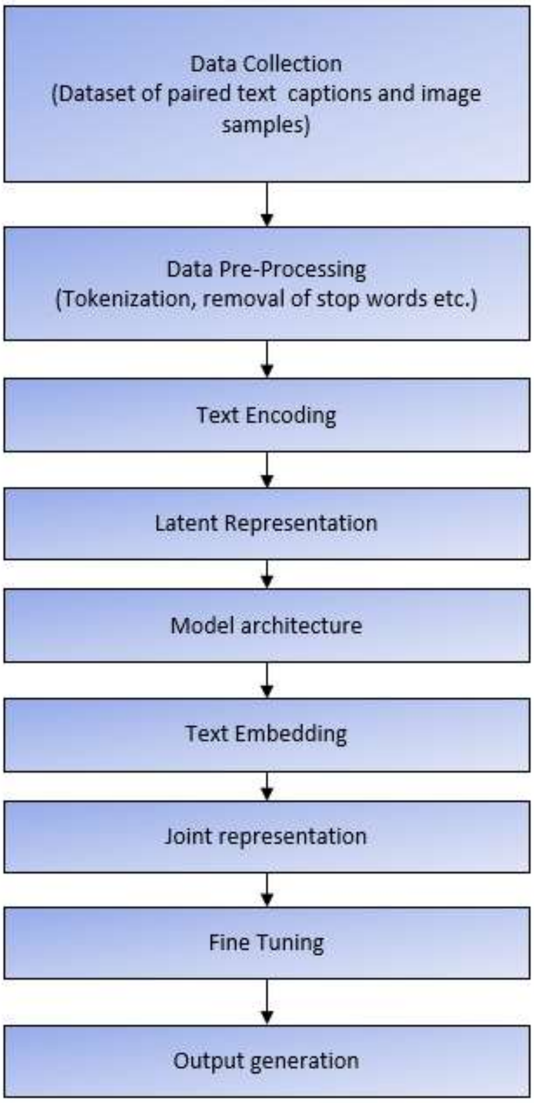
Artificial Intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs. There are many applications of Artificial Intelligence. “Text guided Synthesis of Image” is one of the applications. It creates an image from scratch from a given text description. With these Text-to-Image generation systems, digital artwork is being enhanced. These text-to-image models help us understand how AI systems see and understand our work. Image generation can be created by Autoregressive models, Generator Discriminator Architecture (GAN) and also Vector Quantized Variational Autoencoders (VQVAE Transformers). These methodologies generate finest images when measured by sample quality metrics such as FID, Inception, score and precision. However, the do have drawbacks that make them difficult to scale and apply to new domains. Therefore, a Hybrid approach Diffusion is used in Image Generation. Some of these diffusion models are GLIDE, DALL-E 2, CORGI models etc. Diffusion models are generative models used to generate data similar to the data on which they are trained. They work by destroying Training data via successive addition Gaussian noise and then learn to recover the data by reversing the process. The training Dataset is captioned and the input text is encoded and checked in the latent space to generate a new image. Diffusion models generate quality images with high levels of details, creating realistic images.

Keywords— Artificial Intelligence, Text-to-Image, Diffusion Techniques, FID, Inception, Gaussian noise, Training Dataset.

I. INTRODUCTION

Text-to-image generators have gained immense popularity in recent times, revolutionizing the intersection of Computer Vision and Natural Language Processing. This innovative technology relies on a sophisticated combination of a language model and a generative image model, seamlessly bridging the gap between textual descriptions and visually compelling images. The process begins with a text prompt, which serves as the creative catalyst for these models. The text undergoes Natural Language Processing (NLP), a crucial step that empowers the machine to comprehend and interpret human language effectively. Through this linguistic understanding, the input text is transformed into a latent representation, a nuanced and abstract form of information that encapsulates the essence of the textual description. This latent representation becomes the foundation for the subsequent phase, where generative image models come into play. Noteworthy models in this domain include Generative Adversarial Networks (GAN), Vector Quantized Variational Autoencoders (VQVAE), Corgi, and Diffusion models. These models leverage the latent representation to generate images that vividly align with the given textual input. By exposing the models to diverse datasets during training, they learn to associate specific textual cues with corresponding visual features. This adaptive learning process enables them to capture the intricacies of human language and translate them into visually coherent images. The resulting output from these text-to-image generators is a testament to their creative prowess. They not only reproduce visual elements mentioned in the input text but also demonstrate an aptitude for generating entirely novel and realistic scenes. This breakthrough technology holds immense potential across various industries, from creative arts and design to content generation and virtual world creation. As advancements in both language and image models continue to evolve, the capabilities of text-to-image generators are poised to redefine the landscape of AI-driven creativity.

II. Related Work



Text-to-image generation is a complex process involving the transformation of textual descriptions into visual content. The initial phase requires a substantial dataset of text-image pairs for training, followed by meticulous preprocessing of textual data, including tokenization and stop word removal. Concurrently, relevant features are extracted using word embeddings converting words into numerical vectors. Images undergo resizing and pixel normalization for neural network integration. Model architecture, often variants of GANs or VAEs, is crucial. Text embedding maps features into a shared space, creating a joint representation for nuanced relationships between textual and visual domains. The training phase involves minimizing a loss function to generate images aligning with textual descriptions. Evaluation metrics, such as perceptual quality and diversity, guide fine-tuning by adjusting hyperparameters or modifying architecture. In the inference phase, the model uses learned associations to generate images from text, followed by optional post-processing for enhancements. Ethical considerations, including bias mitigation and avoiding harmful content, are vital. Text-to-image generation is a synthesis of linguistic and visual understanding, with each step contributing to the model's ability to transform text into vivid visual representations.

III. LITERATURE SURVEYS

Shifted diffusion for text-to-image generation:

The innovative integration of the Shifted Diffusion technique within the Corgi model represents a significant stride in realistic image generation and the establishment of robust image embeddings. This novel approach primarily targets enhancing the authenticity of generated images, with a specific focus on improving realism. One of the noteworthy features of this model is its facilitation of semi-supervised learning, effectively bridging the gap between image and text modalities to yield images of superior quality and effectiveness. While the model exhibits complexity and acknowledges certain biases, it demonstrates proficiency through training on the COCO and CUB datasets. Evaluation metrics, such as the FID score and Inception score, attest to its performance. However, the model's quest for practical application necessitates additional training to further refine image realism, emphasizing the ongoing evolution and potential of this advanced technology.

Muse: Text-to-image generation via masked generative transformers:

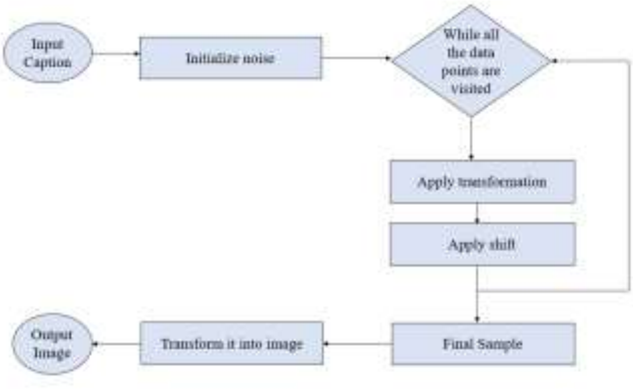
MUSE, a cutting-edge transformer model, spearheads State-Of-The-Art image generation through a masked modelling task in a discrete token space. This innovative approach involves pre-trained Large Language Models serving as encoders, enabling the model to learn predictive patterns for randomly masked image tokens. The effectiveness of MUSE transcends both qualitative and quantitative dimensions, validated through evaluation metrics like FID score and CLIP score. Despite its overall efficacy, MUSE encounters challenges in handling multi-word phrases and multi-cardinalities. This limitation highlights areas for potential refinement to ensure a broader application spectrum. Nevertheless, MUSE's remarkable image generation capabilities make it a promising candidate for image editing applications, thanks to its high-quality outputs. The model's proficiency in balancing both qualitative and quantitative aspects underscore its potential impact on advancing the landscape of state-of-the-art image synthesis techniques.

Dt2i: Dense text-to-image generation from region descriptions:

Employing DT2I and DTC-GAN for image generation based on regional descriptions, this model adopts a conditional generation approach, emphasizing semantic image-text matching. Notably, it excels in generating realistic images, even for complex scenes, enhancing overall visual intuition. Leveraging semantically rich regions ensures a harmonious match between the generated images and their descriptions, mitigating the risk of mismatching. The model's commitment to fidelity ensures that the generated output consistently aligns with and fulfils all specified attributes. However, challenges arise in achieving a consistently realistic output, revealing a potential area for improvement. Evaluation metrics, including FID score, Inception score, CLIP score, and precision score, provide a comprehensive assessment of the model's performance. This analysis underscores the model's strengths in semantic matching and attributes satisfaction while also identifying areas for refinement to enhance the overall realism of the generated images.

IV. METHODOLOGIES

Shifted Diffusion -



Encoder:

- used to convert textual descriptions to numerical/vector representation that is used as input for generative model

Decoder:

- transform the representation generated by encoder into coherent image
- can be either a diffusion model or a generative adversarial network (GAN)
- bridges the gap between abstract information in encoded text & concrete visual representation of image
- maps the encoded text to realistic image via large set of paired examples

Prior model

- to generate target CLIP image embedding
- done through sequential sampling process
- allows for controlled and iterative image synthesis
- image refinement

Methodology:

- initialization:

initial point: x0

- noise level:

sequence of noise level **σ_t** for each step ‘t’

determine the amount of noise to be added at each step

- diffusion steps:

add noise to current data point “xt” to find next data point “x{t+1}”

$x_{t+1} = x_t + \sigma_t * \eta_t$

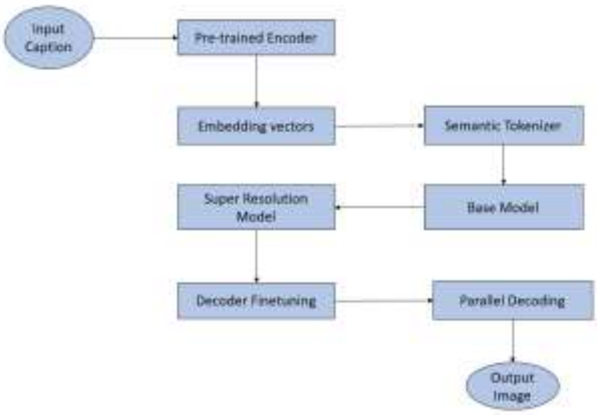
- reversibility:

noise can also be subtracted

$x_t = x_{t+1} - \sigma_t * \eta_t$

- generation: start from random initial points and perform reverse process to remove noise until you reach final generated data point
- training:
  - model is trained to estimate the noise levels
  - goal is min the diff btw generated and real data

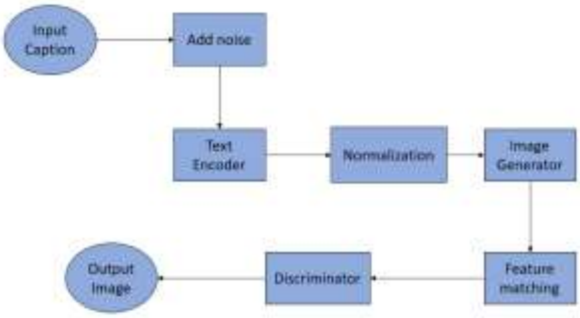
Masked Generative Transformers –



- Pre trained text encoder:
  - Pretrained LLM and Frozen T5-XXL encoder
  - input caption is passed through encoder – results in embedding vectors
- Semantic Tokenizer using VQGAN:
  - Contains encoder and decoder with quantisation layer
  - A convolution layer is present that supports encoding images with different resolutions
- Base Model:
  - Randomly mask fraction of image tokens & replace with special tokens
  - produces output tokens corresponding to 16x16 latent resolution and 512x512 spatial resolution.
- Super resolution model:
  - low resolution tokens are passed via series of transformation layers & resulting o/p is concatenated with text that is extracted from text prompt.
  - Translates lower resolution to higher resolution
- Decoder Finetuning:
  - Improves visual quality without re training any model components.
  - Adding more resolution layers

- Variable masking rate:
  - Train our model based on Cosine scheduling (i.e., chooses a certain fixed fraction of the highest confidence masked tokens that are to be predicted at that step)
- Classifier Free guidance:
  - Enables negative prompting so that features associated with negative prompt are removed
- Iterative parallel decoding at inference:
  - Reduce the sampling steps of diffusion model

DTC-GAN –



- Generator:
  - Conditioned on layout of region description region is described by bounding box
  - predict affine transformation parameters γ and β at each generator layer to modulate the visual features
  - γ and β parameters are Layout-Aware & Text-Sensitive
- Discriminator:
  - consists of multiple ResBlocks
  - heads are used for adversarial training to encourage realism on the image and region level
- Regional triplet loss:
  - trained to distinguish real from generated pairs
  - encourage semantic image-text matching and penalize mismatching pairs
- Regional DAMSM Loss:
  - computes the similarity between an image and global sentence using an attention mechanism
  - a pre-trained BERT text encoder to map matching image regions and text features
  - leads to better alignment between individual regions and corresponding captions
- Multi-Modal Region Feature Matching (MMRFM):
  - task is an inherent one-to-many mapping problem between input conditions and output images
  - we use the resulting region features after projection-based conditioning and minimize the distance between corresponding real and generated image-text features at the region level

V. Conclusion

In the work of text-to-image synthesis, we studied various generative models. Each of the model adapts some techniques to generate images. However, all the models are trained and evaluated based on some performance metrics such as Quality, Effectiveness, FID, Inception, Cosine similarities etc. And the novel technique used in our study is **Corgi** model which adapts shifted diffusion for Image generation. In this technique, our model adapts Gaussian distribution as its initialization for sampling process that makes the parameters learn-able and leads to better results. Our method is based on Semi-supervised learning which obtains better FID than the supervised learning. Our model leads to better image text alignment and better-quality images. The images generated using this technique implies better embedding results. Shifted diffusion leads to better qualitative and quantitative results. Our shifted diffusion model is pre-trained and can be directly plugged into any domain without any further training or fine tuning.

VI. REFERENCES

- Zhou, Y., Liu, B., Zhu, Y., Yang, X., Chen, C., & Xu, J. (2023). Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10157-10166).
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, *35*, 36479-36494.
- Mishra, P., Rathore, T. S., Shivani, S., & Tendulkar, S. (2020, February). Text to image synthesis using residual gan. In *2020 3rd International conference on emerging technologies in computer engineering: Machine learning and internet of things (ICETCE)* (pp. 139-144). IEEE.
- Chen, Z., Chen, L., Zhao, Z., & Wang, Y. (2020, July). AI illustrator: Art illustration generation based on generative adversarial network. In *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)* (pp. 155-159). IEEE.
- Khan, M. Z., Jabeen, S., Khan, M. U. G., Saba, T., Rehmat, A., Rehman, A., & Tariq, U. (2020). A realistic image generation of face from text description using the fully trained generative adversarial networks. *IEEE Access*, *9*, 1250-1260.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.
- Chang, H., Zhang, H., Barber, J., Maschinot, A. J., Lezama, J., Jiang, L., ... & Krishnan, D. (2023). Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.
- Tan, H., Liu, X., Yin, B., & Li, X. (2022). DR-GAN: Distribution regularization for text-to-image generation. *IEEE Transactions on Neural Networks and Learning Systems*.
- Frolov, S., Bansal, P., Hees, J., & Dengel, A. (2022, September). Dt2i: Dense text-to-image generation from region descriptions. In *International Conference on Artificial Neural Networks* (pp. 395-406). Cham: Springer Nature Switzerland.
- Zhao, J., Zheng, H., Wang, C., Lan, L., & Yang, W. (2023). MagicFusion: Boosting Text-to-Image Generation Performance by Fusing Diffusion Models. *arXiv preprint arXiv:2303.13126*.
- Xu, Y., Yu, W., Ghamisi, P., Kopp, M., & Hochreiter, S. (2022). Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. *arXiv preprint arXiv:2208.04441*.
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- Kim, Y., Lee, J., Kim, J. H., Ha, J. W., & Zhu, J. Y. (2023). Dense Text-to-Image Generation with Attention Modulation. *arXiv preprint arXiv:2308.12964*.
- Gallego, V. (2022). Personalizing text-to-image generation via aesthetic gradients. *arXiv preprint arXiv:2209.12330*.
- Zhu, W., Wang, X., Lu, Y., Fu, T. J., Wang, X. E., Eckstein, M., & Wang, W. Y. (2023). Collaborative Generative AI: Integrating GPT-k for Efficient Editing in Text-to-Image Generation. *arXiv preprint arXiv:2305.11317*.