International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

CLIP - Guided Generative Adversarial Network for Semantic Text-to-image synthesis

#1 SK. Himam basha, #2 B. Lakshmidevi

#1 ASSISTANT PROFESSOR # 2 MCA SCHOLAR
DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS
QIS COLLEGE OF ENGINEERING & TECHNOLOGY , ONGOLE
VENGAMUKKAPALEM(V), ONGOLE, PRAKASAM DISTRICT, ANDHRAPRADESH

ABSTRACT

Text-to-image synthesis has emerged as one of the most significant applications of deep learning, enabling the automatic generation of realistic images from natural language descriptions. Traditional Generative Adversarial Networks (GANs) have demonstrated promising capabilities in image generation; however, maintaining semantic consistency between textual descriptions and generated images remains a major challenge. Recent advancements in Contrastive Language–Image Pre-training (CLIP) have provided powerful multimodal representations that effectively capture the relationship between textual and visual information.

This project presents a CLIP-Guided Generative Adversarial Network (GAN) framework for semantic text-to-image synthesis. The proposed approach integrates the semantic understanding capability of CLIP with the image generation strength of GANs to produce visually realistic and textually aligned images. CLIP is employed to evaluate and guide the generator by measuring the similarity between generated images and their corresponding textual descriptions. The generator learns to create images that not only exhibit high visual quality but also preserve the semantic meaning embedded in the input text. The discriminator further enhances image authenticity by distinguishing between real and synthesized samples.

INTRODUCTION

The rapid development of Artificial Intelligence (AI) and Deep Learning has transformed the way machines

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

understand and generate multimedia content. Among various generative tasks, text-to-image synthesis has attracted considerable attention because of its ability to convert human language descriptions into meaningful visual representations. This capability enables computers to create images based solely on textual input, bridging the gap between natural language understanding and computer vision.Text-to-image generation has numerous practical applications, including digital art creation, advertising, game development, virtual reality, educational visualization, medical illustration, and automated content generation. Despite substantial progress in this domain, generating images that are both visually realistic and semantically consistent with the provided text remains a challenging problem. Conventional approaches often struggle to accurately capture complex textual details, resulting in images that fail to

resemble real samples. Through adversarial training, GANs can produce highly detailed and realistic images. However, traditional GAN-based text-to-image reflect the intended meaning of the descriptions.

Generative Adversarial Networks (GANs) have emerged as a powerful framework for image generation. Introduced by Goodfellow et al., GANs consist of two competing neural networks: a generator and a discriminator. The generator attempts to synthesize realistic images, while the discriminator evaluates whether the generated images models frequently encounter issues such as semantic inconsistency, mode collapse, and limited alignment between textual inputs and visual outputs.

LITERATURE SURVEY

1. Generative Adversarial Networks

Authors: Ian Goodfellow et al. (2014)

Description:

Ian Goodfellow and his colleagues introduced Generative Adversarial Networks (GANs), a revolutionary framework for generative modeling. GANs consist of two neural networks, namely the generator and the discriminator, trained simultaneously through an adversarial process. The generator produces synthetic samples, while the discriminator attempts to distinguish between real and generated data. This architecture demonstrated remarkable success in generating realistic images and established the

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

foundation for subsequent advancements in image synthesis. However, the original GAN architecture was not designed to incorporate textual information, limiting its applicability to text-guided image generation.

2. Conditional Generative Adversarial Networks (CGAN)

Authors: Mehdi Mirza and Simon Osindero (2014)

Description:

Conditional GANs extended the traditional GAN framework by introducing auxiliary information such as class labels or textual descriptions into both the generator and discriminator. This modification enabled controlled image generation based on specified conditions. CGANs represented a major step toward text-to-image synthesis by demonstrating that generative models could be directed using external semantic information. Nevertheless, early conditional approaches generated low-resolution images with limited semantic richness.

3. StackGAN: Text-to-Photo Realistic Image Synthesis

Authors: Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, and Dimitris Metaxas (2017)

Description:

StackGAN proposed a two-stage architecture to generate high-resolution images from textual descriptions. Stage-I generated coarse images capturing basic shapes and colors, while Stage-II refined these images by adding realistic details. The model significantly improved image quality and resolution compared with previous methods. Despite these advancements, StackGAN occasionally produced images with semantic inconsistencies when handling complex textual descriptions.

4. AttnGAN: Fine-Grained Text-to-Image Generation with Attentional Generative Adversarial Networks

Authors: Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He (2018)

Description:

AttnGAN introduced an attention mechanism that allowed the model to focus on relevant words in the textual description while generating different regions of an image. This fine-grained alignment improved the correspondence between text and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

image content. The approach achieved state-of-the-art performance on benchmark datasets by generating visually appealing images with enhanced semantic accuracy. However, its 1 CLIP projects both the generated image and the corresponding text description into a shared embedding space. A similarity score is computed to determine how closely the generated image aligns with the textual input.

METHODOLOGY

The methodology of the proposed CLIP-Guided Generative Adversarial Network focuses on generating semantically meaningful images from textual descriptions by integrating adversarial learning with multimodal semantic guidance. The workflow consists of several interconnected stages that collectively improve the quality and relevance of synthesized images.

Step 1: Input Text Acquisition

The process begins by receiving natural language descriptions from users or benchmark datasets. These descriptions specify the characteristics, appearance, colors, objects, and relationships that should be reflected in the generated images.

Step 2: Text Encoding Using CLIP

The input text is processed by the pre-trained CLIP text encoder. CLIP converts the textual descriptions into dense semantic embeddings that capture high-level contextual information learned from large-scale image-text pairs.

Step 3: Noise Vector Generation

A random latent noise vector is generated from a predefined probability distribution. This vector introduces diversity into the image generation process, enabling the creation of multiple variations corresponding to similar text prompts.

Step 4: Conditional Image Synthesis

The latent noise vector is concatenated with the CLIP text embeddings and supplied to the generator network. The generator progressively transforms these representations into synthetic images using deep convolutional operations and feature upsampling techniques.

Step 5: Real–Fake Discrimination

The discriminator receives both real images from the dataset and generated images from the generator. It learns to classify whether an image is authentic

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

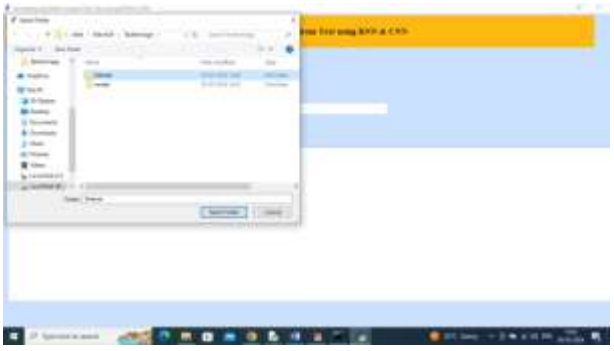
or synthesized, thereby improving the realism of the generated outputs.

Step 6: CLIP-Based Semantic Guidance

The generated images are encoded using the CLIP image encoder. Similarity between image embeddings and corresponding text embeddings is calculated within CLIP's shared representation space. The semantic similarity score acts as an additional supervisory signal during training.

RESULTS

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



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

In above screen dataset loaded and now click on 'Pre-process dataset' button to read and normalize both TEXT and IMAGE features and get below output

In above screen dataset processing completed and now click on 'Generate & Load RNN Model' button to load model and get below page

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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

In above screen in text field I entered some text and then press button to get bellow output

In above screen can see generated image for text 'A girl in pink dress climbing stairs' . Similarly type some text and get outputIn above

screen for given text will get below image

For above sentence we got above image.

Note: For some text we may not get pictures but you can give sentences in any manner from dataset. This algorithms require large amount of training in huge dataset to generate images for all types of questions. While training on large dataset model running out of memory in Google COLAB as well as normal laptops so we trained this model on few images from the dataset.

You can get exact image from all text given in 'samples.txt' file

CONCLUSION

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The development of a CLIP-Guided Generative Adversarial Network for Semantic Text-to-Image Synthesis represents a significant advancement in the field of multimodal artificial intelligence and generative modeling. The proposed framework effectively combines the image generation capability of Generative Adversarial Networks (GANs) with the powerful semantic understanding ability of Contrastive Language–Image Pre-training (CLIP) to address the limitations of conventional text-to-image synthesis methods.

Traditional GAN-based approaches primarily focused on generating visually realistic images but often struggled to preserve the complete semantic meaning embedded within textual descriptions. As a result, generated outputs frequently exhibited semantic inconsistencies, omitted important attributes, or failed to accurately represent complex prompts. By incorporating CLIP as a semantic guidance mechanism, the proposed system ensures that generated images maintain a stronger correspondence with the input text while preserving high visual quality.

The integration of adversarial loss and CLIP-based semantic similarity loss enables the generator to learn both realism and semantic alignment simultaneously. This dual optimization strategy improves the fidelity, diversity, and contextual relevance of synthesized images. Furthermore, the utilization of CLIP's large-scale multimodal knowledge enhances the model's ability to generalize to unseen descriptions and understand complex relationships between language and visual concepts.

REFERENCES

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). "Generative Adversarial Nets." *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pp. 2672–2680.

1. Mirza, M., and Osindero, S. (2014). "Conditional Generative Adversarial Nets." *arXiv preprint arXiv:1411.1784*.
2. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., and Metaxas, D. (2017). "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks." *Proceedings of the IEEE*

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

International Conference on Computer Vision (ICCV), pp. 5907–5915.

3. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1316–1324.
4. Zhu, M., Pan, P., Chen, W., and Yi, Y. (2019). "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5802–5810.
5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763.
6. Esser, P., Rombach, R., and Ommer, B. (2021). "Taming Transformers for High-Resolution Image Synthesis." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883. Author:
7. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). "Zero-Shot Text-to-Image Generation." *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8821–8831.
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.
9. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *Advances in Neural Information*

AUTHOR PROFILE

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Himambasha Shaik is Assistant Professor in the Departmentof Masterof Computer Applications at QIS College ofEngineering and Technology,Ongole, Andhra Pradesh. He earnedhis Master of ComputerApplications (MCA) from Anna University, Chennai.With a strong research background, He has authoredand co-authored research papers published in reputedpeer-reviewed journals. His research interests include Cloud Computing, and Programming Languages. He iscommitted to advancing research and fosteringinnovation while mentoring students to excel in bothacademic and professional pursuits.

STUDENT PROFILE

B.Lakshmidevi A is a postgraduate student pursuing a MCA in the department of computer Applications at QIS College of Engineering & Technology, Ongole autonomous college in prakasam dist. Shecompleted undergraduate degree in MPCs (computer science) from ANU. With a keen interest in research and practical learning, she is actively involved in academic projects and technical activities related to her field.