Text-to-Image Generation using Generative Adversarial Network

Chitvan Jamdagni, Vasu Sharma, Jatin Goyal, Divyansh, Baddam Nikhil Kumar Reddy, Payal Thakur

Chandigarh University, Ghauran, Punjab, India - 140301

Corresponding author: Chitvan Jamdagni, Email: 13chitvan@gmail.com

A deep learning model called Text-to-Image Creation with Generative Adversarial Networks (GAN) can generate images from text descriptions. A wide range of applications, including photo-searching, photo-editing, art creation, computer-aided design, image reconstruction, captioning, and portrait drawing, are among the several study fields that it has a significant impact on. Producing realistic visuals consistently under predetermined settings is the most difficult endeavor. Current text-to-image creation algorithms produce images that don't accurately reflect the text. The Caltech-UCSD Birds-200-2011 dataset was used to train the suggested model, and an inception score and PSNR were used to assess its performance. Stage-I and Stage-II make up the proposed StackGAN paradigm. Based on the input written description, Stage-I GAN generates low-resolution images by the method of roughing out the basic shape and colors of the object. By using the Stage-I results and textual descriptions as inputs, together with defect detection and detail addition, Stage-II GAN creates high-resolution and photo-realistic images with fine details.

Keywords: GAN, Generative Adversarial Networks, Deep Learning, Text to Image generation (T2I).

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 209–217. DOI: https://doi.org/10.56155/978-81-955020-7-3-19

1. Introduction

Due to the rising desire for original and individualized visual material, creating graphics from text has become a trend in recent years. With the use of this technology, photo-realistic visuals may be produced from textual descriptions, opening up a world of possibilities for social media, video games, virtual and augmented reality, and other applications. Researchers have created a number of strategies to produce high-quality images that accurately indicate the intended meaning of the textual input thanks to the development of deep learning algorithms and the accessibility of vast datasets. The way we produce and consume visual material is predicted to change as a result of this trend as technology develops and becomes more widely available.

Deep learning tools like Generative Adversarial Networks, for example, can be used to generate images. Advanced neural networks called Generative Adversarial Networks (GANs) pit numerous networks against one another to create images that are incredibly accurate and practically indistinguishable. A generator network generates images, while a discriminator network determines whether they are real or fraudulent using game theory. The generator improves and becomes more realistic as it gains experience, finally reaching convergence when real images may be generated with confidence.

We propose employing Stack Generative Adversarial Networks (Stack GAN) [9] to breaks down the process of creating high-quality images with the use of text descriptions into more manageable sub-processes. Based on text descriptions, our Stage-I GAN creates low-resolution images, which are then enhanced and improved by our Stage-II GAN to create photo-realistic high-resolution images. The final photographs are of excellent quality with a have a wide range of useful uses of machine learning model for predicting future crimes.

2. Machine Learning

Artificial intelligence has a popular subset called machine learning, which describes the ability of information technology systems to solve problems themselves by identifying patterns and similar designs in databases. In other words, machine learning makes it possible to identify IT systems and create suitable solution concepts based on already existing algorithms and datasets. Artificial knowledge is thus created through experience-based machine learning.

Machine learning [1] has a significant influence on many facets of our daily lives, according to the AI community. The representation of the data they are given is necessary for all of these machine learning methods. However, it is exceedingly challenging to extract valuable traits when one wants to apply this expertise to other jobs or sectors. In order to automatically extract important information when doing classification and detection, researchers introduced a novel method called representation learning [2]. Deep learning [3] is a class of representation learning techniques that, by combining a few basic representations, may quickly extract higher-level, more abstract features than previous techniques.

The software needs human intervention in the past to be able to produce solutions on its own. For instance, the systems must be prepared in advance with the necessary algorithms, data, and analytic rules for identifying patterns in the data stock. Following the completion of these two processes, the system can carry out the following functions using machine learning.

3. Objectives of the Research

The main objectives of this research are:

1. Extracting various features for images and corresponding text description.
2. Building GAN model for best generating images with high accuracy.

4. Problem Statement

To create photorealistic images that are semantically compatible with the text descriptions, text-to-image synthesis (T2I) is used. Existing techniques often use conditional generative adversarial networks (GANs) to build an image from noise using phrase embedding, then repeatedly improve the features using fine-grained word embedding. A thorough examination of their produced images exposes a significant flaw: while the overall image generally matches the description, specific image portions or pieces of items are sometimes indistinguishable or inconsistent with words in the sentence, such as "a white crown".
We are using a unique framework called Stack GAN for creating images from text input to solve this issue.

Our Stage-I GAN creates low-resolution images from text descriptions, which our Stage-II GAN then enhances and improves to create photo-realistic high-resolution images. High-quality photographs are produced as a result, and they have numerous useful uses.

5. Literature Review

Deep learning concepts and the knowledge of these techniques have made outstanding progress in generating images from textual content.

1. A particular class of neural network called a Variational Autoencoder (VAE) can be trained to learn how to compactly represent complicated data, such as images and sounds. The input data is encoded using probabilistic methods into a latent space, which is a lower-dimensional space. By doing so, they are able to produce fresh data that is comparable to the initial input data. However, the ones of early models [4]–[6] had some extreme obstacles, they did not have a terrific generalization. In 2014,Goodfellow et al. [7] proposed a unique generativemodel, named Generative adverse Networks (GANs)
2. An autoregressive model called Conditional PixelCNN creates images based on inputs like text descriptions or object location constraints. A convolutional neural network is used to model the pixel space's conditional distribution.
3. Laplacian pyramid framework is a method for decomposing images into several scales, where each scale is represented by a different level of the image. This framework is used to create high-resolution images from low-resolution inputs in image processing and computer vision applications. Numerous strategies have been put forth to increase the stability of the training process and deliver outstanding results. It has been claimed that a GAN model based on energy could improve training process stability.
4. AlignDRAW is an image generation approach that produces high-quality images from textual descriptions by aligning the text and image attributes. The model sequentially generates the image while using an attention method for aligning the text and image features. Mansimov et al. created AlignDRAW in 2016, and it has produced realistic images from text descriptions with promising results.
5. Chen, H., Chungcdvsfdf, W., Xu, J. J., Wangsac, G., Qin, Y., & Chauascas, M, create photorealistic images with a resolution of 256x256 using Stack Generative Adversarial Networks (StackGAN) and text descriptions. By breaking the problem down into two smaller issues, the technique introduces Conditioning Augmentation, which enhances diversity and stability. On benchmark datasets, the proposed strategy performs better than cutting-edge techniques. Since StackGAN makes use of a multi-stage GAN architecture, it has an advantage over the mentioned models as it can produce high-resolution images with precise features while keeping textual consistency. Additionally, it creates realistic images that are visually and semantically significant while successfully capturing both low-level along with the high-level characteristics of the image.

6. Proposed Methodology

6.1 Generative Adversarial Network

Through opposed gaining knowledge of, GANs generate information the usage of two neural community fashions: the discriminator and generator. As illustrated in figure 1, whilst the discriminator tells the distinction between actual data and information generated through the generator, the generator takes in random noise (z) and generates information. The discriminator, which has been trained to apprehend the source data, is what the generator targets to misinform with its information.The discriminator's job is to distinguish between genuine and counterfeit samples. Backpropagation [8] and dropout techniques can be used to train both models in this case. Furthermore, GANs do not require approximation inference or Markov chains [17].

The most notable aspect of GANs is it's use of text embedding feature. This textual information about each image representing its features like color, shape of beak, etc. is encoded into a fixed size vector through natural language processing (NLP) or word embedding.

This enables efficient and consistency in training of model with regards to pair of image and it's corresponding information.

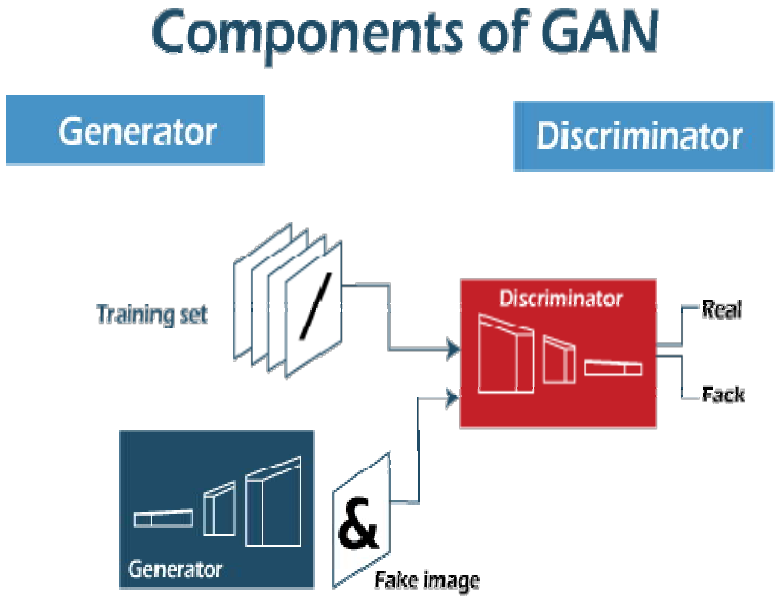

Figure 1. Components

6.2 Stack Generative Adversarial Network

Embedding: Converts the input text's variable length into a vector with a fixed length. A pre-trained character level embedding will be used [14]. Embodiments are employed in StackGAN to transform text descriptions into numerical vectors that can be fed into the generating network. A different neural network that has been trained on a sizable text corpus is used to learn the embeddings as illustrated in figure 2 & 3. The generator network can create visuals that correspond with the text description thanks

to these embeddings, which collect semantic information about the text. By more accurately capturing the connection between the text and the image, embeddings enhance the quality and diversity of the generated visuals.

Conditional Augmentation: A method employed in generative models to increase the diversity and caliber of generated samples is conditioning augmentation (CA) [13]. In order to help the model acquire a more reliable representation, it entails adding random noise vectors to the conditioning variables during training. In StackGAN, CA is utilized to produce many iterations of the same text description, enabling the generation of a wider variety of images. This aids in addressing the problem of mode collapse, which occurs when a model only produces a small number of comparable outputs.

Stage-1 Generator: A conditional GAN called the stage 1 generator of StackGAN creates a low-resolution image from a provided text description. It produces a 64x64 image from an input of a random noise vector and with the text embedding. The generator is conditioned to maintain fundamental color and shape restrictions while producing images that correspond to the text description provided. The stage II generator then processes the resulting image to improve it and add more information. All things considered, the stage I generator is in charge of producing the initial draft of the image based on the text input.

Stage-1 Discriminator: The resulting low-resolution image and the text embedding are fed into the Stage-1 discriminator of StackGAN, which then applies a sequence of convolutional layers to learn the characteristics of such images [10]–[11]. To differentiate between the generated image and the real image, the discriminator is trained. In terms of both substance and design, it encourages the generator to create images that are comparable to the originals. The output of the discriminator is used to determine the adversarial loss, which is then utilized to modify the parameters of the generator. The Stage I discriminator is crucial in making sure that the generated images adhere to the fundamental color and shape restrictions of the provided text description.

Residual Blocks: Residual blocks are utilized in StackGAN's Stage 1 and Stage 2 generators to enhance the quality of the images that are produced. The network can learn residue factors and features from residual blocks that represent the demarcation between the input and output [13]. This aids in avoiding the issue of vanishing gradients and makes it possible to train deeper networks successfully.

Stage-2 Generator: The low-resolution image created by the Stage 1 generator of StackGAN is improved with more realistic and detailed details by the Stage 2 generator. It is divided into two sections: the first part creates an approximate approximation of the high-resolution image, and the second part uses leftover blocks to add more features to the resulting image. The Stage 2 generator receives a second conditioning vector produced using the Conditioning Augmentation approach in addition to being conditioned on the same text description as the Stage 1 generator. This additional conditioning vector increases the variety of the generated images. The Stage II generator's output is a high-resolution image with realistic details and lots of variability.

Stage-2 Discriminator: The StackGAN Stage II Discriminator is designed to differentiate between actual and manufactured high-resolution images. It takes these high-definition images generated by the Stage II Generator as input and outputs a score indicating how closely the created image resembles the genuine images from the training dataset [15]. The architecture of the Stage II Discriminator is very close to that of the Stage I Discriminator, but it is built to handle high resolution pictures with more detailed features.

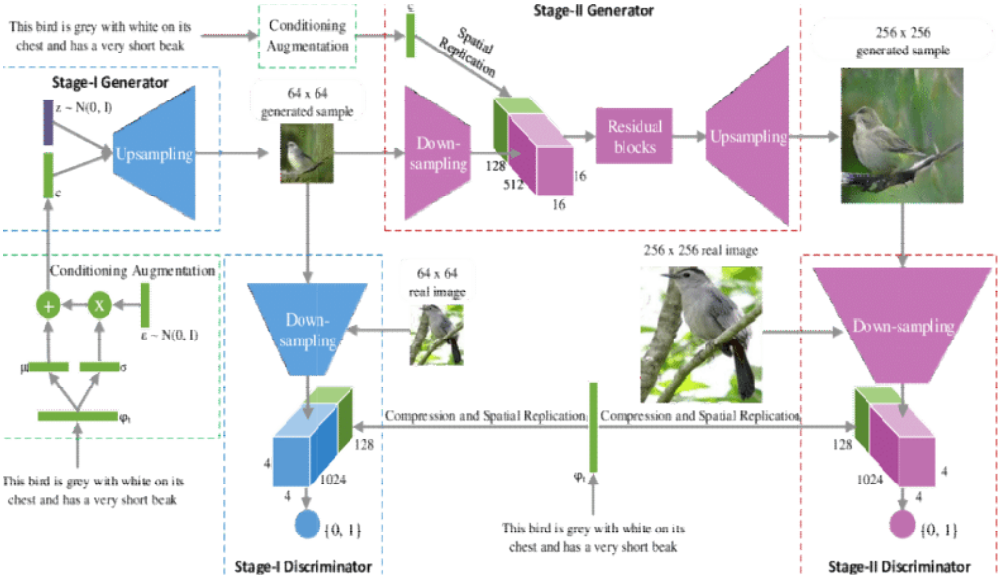

Figure 2. Generator

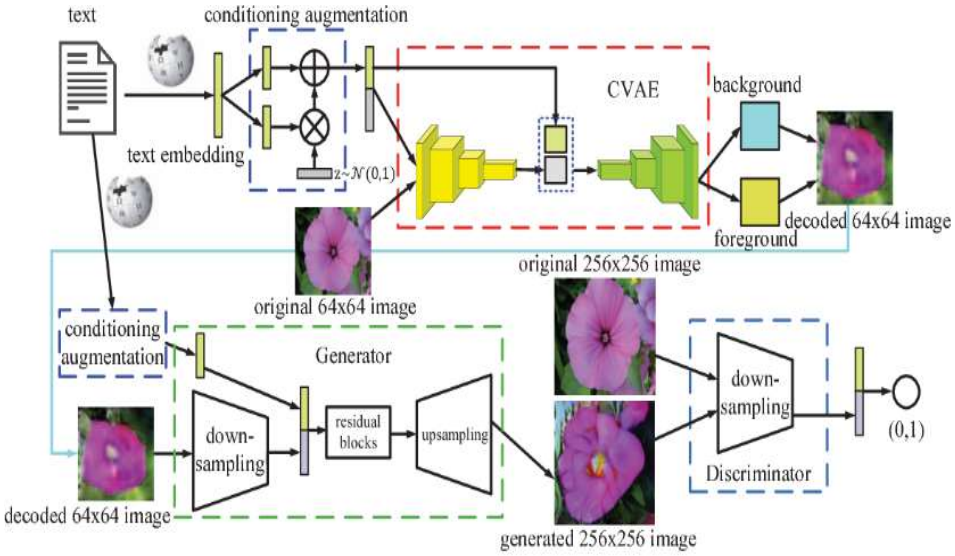

Figure 3. Image generation

7. Results

The following resources were used in the project's implementation:
1. The Jupyter notebook software
2. NVIDIA Cuda platform
3. Tensorflow software.

By training the model, the aforementioned functions and procedures were carried out, and photos were produced. The goal of the project is to employ Stack Generative Adversarial Network to produce photo-realistic images from textual descriptions [16]. The Deep Convolutional Generative Adversarial Networks (DCGAN) was proposed by Radford et al. [12] as a result of the fact that CNN (Convultion based Gans) is superior to MLP in extracting image data.

The model was trained in two steps, as stated in the recommended approach. Stage 1 ran for 28 epochs in 8 hours. Stage 2, which needed more high performance computing power as compared to Stage 1, took 15 hours to complete for 5 epochs.

For each third epoch in both stages, 10 photos were preserved. As a result, we had 85 photos stored after 28 epochs of stage 1.

Below are some of the images generated in stage 1

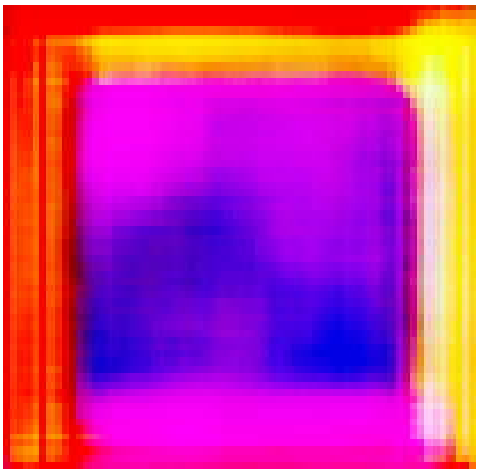

Figure 4.1

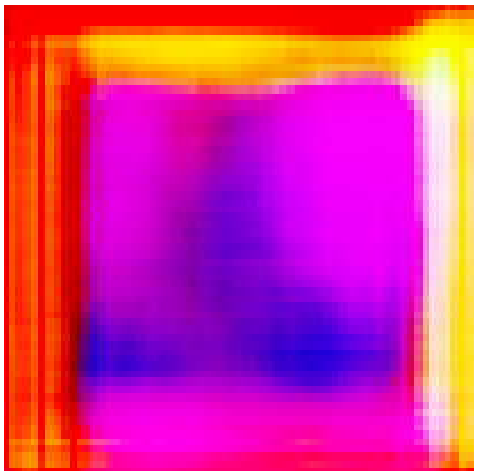

Figure 4.2

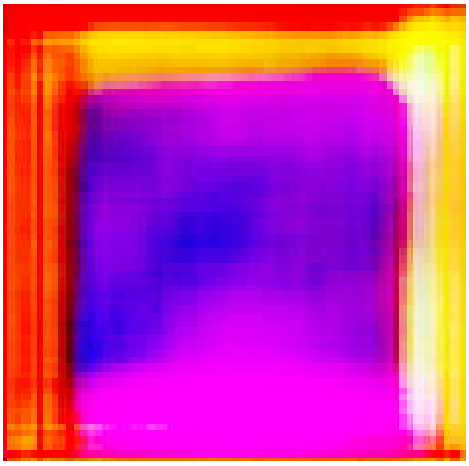

Figure 4.3

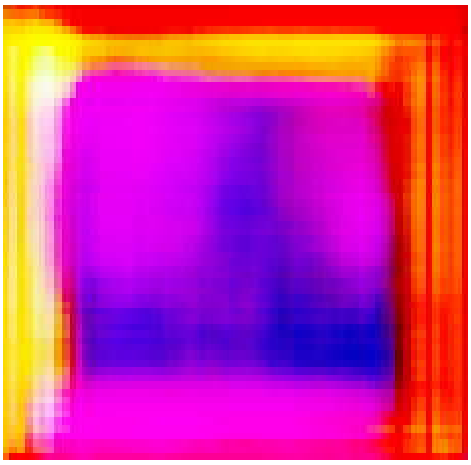

Figure 4.4

8. Conclusion

In conclusion, the suggested technique for creating photorealistic images from text by combining Conditioning Augmentation with Stack Generative Adversarial Networks (StackGAN) displays promising results. Compared to other text-to-image generative models, Stage-I and Stage-II GANs enable the production of images with greater resolution and more photorealistic features. This method offers a lot of potential for use in a variety of industries, including virtual reality, assistive communication devices, and interior design. The use of StackGAN through Conditioning Augmentation can produce increasingly more spectacular outcomes in creating realistic images from text descriptions as technology develops and more complicated datasets become accessible.

9. Limitations

Numerous intricate factors, some of which are related to the software, the dataset accuracy, and designing might have an impact on producing accurate imagery.

Accurate images are based on two elements: Good data are more crucial than a good model. Despite having a large dataset, the features it offers are inadequate for the first round of image generation and since we had some hardware limitations and time restrictions we only went forward with the first round of image generation whose results are shown above.

The color scheme was matched correctly to some extent where we could generalize an outline of the desired text for the first round.

10. Future Work

Increasing data: More data, such as on geography, animals, flowers, demographic, can aid the image generation

Concentrating on specific type of text for example if a animal image needs to be generated the dataset should be rigorous and detailed.

Employing a hybrid of oversampling and under-sampling approaches and going for multiple rounds of image generation after the first round preferably upto 5 rounds of generation for super accurate results.

References

[1] McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12
[2] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal. Crime Analysis using K-Means Clustering
[3] HARDI, M. PATEL, RIPAL PATEL, Enhance Algorithm to Predict a Crime Using Data Mining
[4] Riya Rahul Shah, Crime Prediction Using Machine Learning
[5] Dataset: https://data.cityofchicago.org/
[6] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, vol. 5, Apr 2009, pp. 448–455. [Online].
[7] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[8] P. Smolensky, "Information processing in dynamical systems: foundations of harmony theory," Cambridge, MA: MIT Press, 1986.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.
[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, Aug 2013.
[11] K. P. Murphy, "Machine learning: a probabilistic perspective," Cambridge, MA: MIT Press, 2012.
[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533–536, 1986.
[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in International Conference on Learning Representations, 2014.
[14] H. Akaike, "Fitting autoregressive models for prediction," Annals of the Institute of Statistical Mathematics, vol. 21, no. 1, pp. 243–247, Dec 1969.
[15] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, 2016, pp. 1747–1756. [Online
[16] J. Donahue, P. Kr?henb´l´zhl, and T. Darrell, "Adversarial feature learning," in International Conference on Learning Representations, 2017.
[17] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," in International Conference on Learning Representations, 2017.