Text-to-Image Generation Using Deep Learning

engineering proceedings

MDPI

Proceeding Paper

Text-to-Image Generation Using Deep Learning [†]

Sadia Ramzan [1,*], Muhammad Munwar Iqbal [1] and Tehmina Kalsum [2]

[1] Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan; munwar.iq@uettaxila.edu.pk

[2] Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan; tehmina.kalsum@uettaxila.edu.pk

[*] Correspondence: sadia.ramzan@students.uettaxila.edu.pk

[†] Presented at the 7th International Electrical Engineering Conference, Karachi, Pakistan, 25–26 March 2022.

Abstract: Text-to-image generation is a method used for generating images related to given textual descriptions. It has a significant influence on many research areas as well as a diverse set of applications (e.g., photo-searching, photo-editing, art generation, computer-aided design, image reconstruction, captioning, and portrait drawing). The most challenging task is to consistently produce realistic images according to given conditions. Existing algorithms for text-to-image generation create pictures that do not properly match the text. We considered this issue in our study and built a deep learning-based architecture for semantically consistent image generation: recurrent convolutional generative adversarial network (RC-GAN). RC-GAN successfully bridges the advancements in text and picture modelling, converting visual notions from words to pixels. The proposed model was trained on the Oxford-102 flowers dataset, and its performance was evaluated using an inception score and PSNR. The experimental results demonstrate that our model is capable of generating more realistic photos of flowers from given captions, with an inception score of 4.15 and a PSNR value of 30.12 dB, respectively. In the future, we aim to train the proposed model on multiple datasets.

Keywords: convolutional neural network; recurrent neural network; deep learning; generative adversarial networks; image generation

check for updates

Citation: 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

Academic Editor: Saad Ahmed Qazi

Published: 29 July 2022

1. Introduction

When people listen to or read a narrative, they quickly create pictures in their mind to visualize the content. Many cognitive functions, such as memorization, reasoning ability, and thinking, rely on visual mental imaging or "seeing with the mind's eye" [1]. Developing a technology that recognizes the connection between vision and words and can produce pictures that represent the meaning of written descriptions is a big step toward user intellectual ability.

Image-processing techniques and applications of computer vision (CV) have grown immensely in recent years from advances made possible by artificial intelligence and deep learning's success. One of these growing fields is text-to-image generation. The term text-to-image (T2I) is the generation of visually realistic pictures from text inputs. T2I generation is the reverse process of image captioning, also known as image-to-text (I2T) generation [2–4], which is the generation of textual description from an input image. In T2I generation, the model takes an input in the form of human written description and produces a RGB image that matches the description. T2I generation has been an important field of study due to its tremendous capability in multiple areas. Photo-searching, photo-editing, art generation, captioning, portrait drawing, industrial design, and image manipulation are some common applications of creating photo-realistic images from text.

The evolution of generative adversarial networks (GANs) has demonstrated exceptional performance in image synthesis, image super-resolution, data augmentation, and image-to-image conversion. GANs are deep learning-based convolutional neural networks

(CNNs) [5,6]. It consists of two neural networks: one for generating data and the other for classifying real/fake data. GANs are based on game theory for learning generative models. Its major purpose is to train a generator (G) to generate samples and a discriminator (D) to discern between true and false data. For generating better-quality realistic image, we performed text encoding using recurrent neural networks (RNN), and convolutional layers were used for image decoding. We developed recurrent convolution GAN (RC-GAN), a simple an effective framework for appealing to image synthesis from human written textual descriptions. The model was trained on the Oxford-102 Flowers Dataset and ensures the identity of the synthesized pictures. The key contributions of this research include the following:

- Building a deep learning model RC-GAN for generating more realistic images.
- Generating more realistic images from given textual descriptions.
- Improving the inception score and PSNR value of images generated from text.

The following is how the rest of the paper is arranged: In Section 2, related work is described. The dataset and its preprocessing are discussed in Section 3. Section 4 explains the details of the research methodology and dataset used in this paper. The experimental details and results are discussed in Section 5. Finally, the paper is concluded in Section 6.

2. Related Work

GANs were first introduced by Goodfellow [7] in 2014, but Reed et al. [8] was the first to use them for text-to-image generation in 2016. Salimans et al. [9] proposed training stabilizing techniques for previously untrainable models and achieved better results on the MNIST, CIFAR-10, and SVHN datasets. The attention-based recurrent neural network was developed by Zia et al. [10]. In their model, word-to-pixel dependencies were learned by an attention-based auto-encoder and pixel-to-pixel dependencies were learned by an autoregressive-based decoder. Liu et al. [11] offered a diverse conditional image synthesis model and performed large-scale experiments for different conditional generation tasks. Gao et al. [12] proposed an effective approach known as lightweight dynamic conditional GAN (LD-CGAN), which disentangled the text attributes and provided image features by capturing multi-scale features. Dong et al. [13] trained a model for generating images from text in an unsupervised manner. Berrahal et al. [14] focused on the development of text-to-image conversion applications. They used deep fusion GAN (DF-GAN) for generating human face images from textual descriptions. The cross-domain feature fusion GAN (CF-GAN) was proposed by Zhang et al. [15] for converting textual descriptions into images with more semantic detail. In general, the existing methods of text-to-image generation use wide-ranging parameters and heavy computations for generating high-resolution images, which result in unstable and high-cost training.

3. Dataset and Preprocessing

3.1. Dataset

The dataset used was Oxford-102 flowers, which include 8189 images of flowers that are all of different species. It has 102 classes; each class consists of 40 to 60 images, and each of the images have 10 matching textual descriptions. In this study, we considered 8000 images for training. This dataset was used to train the model for 300 epochs.

3.2. Data Preprocessing

When the data were collected and extracted initially, it consisted of 8189 images with different sizes of resolutions and corresponding textual descriptions. For normalizing the textual data, we used an NLTK tokenizer, which converted the textual sentences into words. These tokenized lists of words were transformed into an array of caption ids. The images were loaded for resizing to the same dimensions. All training images and testing images were resized to a resolution of 128 × 128. For training purposes, the images were converted into arrays, and both the vocabulary and images were loaded onto the model.

4. Proposed Methodology

This section describes the training details of deep learning-based generative models. Conditional GANs were used with recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for generating meaningful images from a textual description. The dataset used consisted of images of flowers and their relevant textual descriptions. For generating plausible images from text using a GAN, preprocessing of textual data and image resizing was performed. We took textual descriptions from the dataset, preprocessed these caption sentences, and created a list of their vocabulary. Then, these captions were stored with their respective ids in the list. The images were loaded and resized to a fixed dimension. These data were then given as input to our proposed model. RNN was used for capturing the contextual information of text sequences by defining the relationship between words at altered time stamps. Text-to-image mapping was performed using an RNN and a CNN. The CNN recognized useful characteristics from the images without the need for human intervention. An input sequence was given to the RNN, which converted the textual descriptions into word embeddings with a size of 256. These word embeddings were concatenated with a 512-dimensional noise vector. To train our model, we took a batch size of 64 with gated-feedback 128 and fed the input noise and text input to a generator. The architecture of the proposed model is presented in Figure 1.

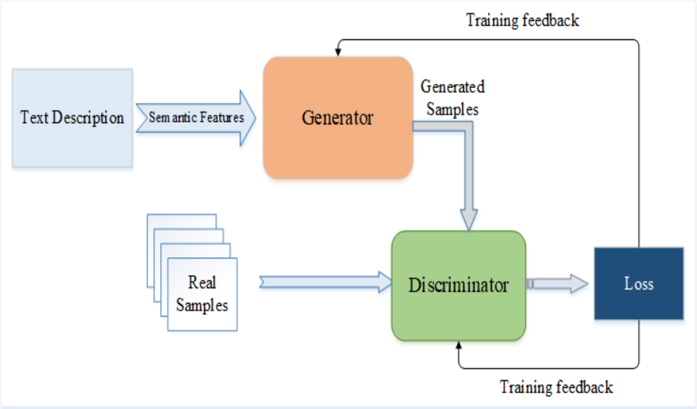

Figure 1. Architecture of the proposed method, which can generate images from text descriptions.

Semantic information from the textual description was used as input in the generator model, which converts characteristic information to pixels and generates the images. This generated image was used as input in the discriminator along with real/wrong textual descriptions and real sample images from the dataset.

A sequence of distinct (picture and text) pairings are then provided as input to the model to meet the goals of the discriminator: input pairs of real images and real textual descriptions, wrong images and mismatched textual descriptions, and generated images and real textual descriptions. The real photo and real text combinations are provided so that the model can determine if a particular image and text combination align. An incorrect picture and real text description indicates that the image does not match the caption. The discriminator is trained to identify real and generated images. At the start of training, the discriminator was good at classification of real/wrong images. Loss was calculated to improve the weight and to provide training feedback to the generator and discriminator model. As soon as the training proceeded, the generator produced more realistic images and it fooled the discriminator when distinguishing between real and generated images.

5. Results and Discussion

In this section, the experimental analysis and generated images of flowers are presented. The training of the proposed model was performed on an Nvidia 1070 Ti GPU, 32 GB memory and windows 10 operating system. The weights of the generator and discriminator were optimized using an Adam optimizer, the mini-batch size was 64, and the learning rate was 0.0003. The ground truths from the dataset and the images generated from the input textual descriptions are shown in Figure 2. For evaluating the performance of the proposed model, the inception score (IS) and PSNR values are calculated. Inception scores capture the diversity and quality of the generated images. PSNR is used for calculating the peak signal-to-noise ratio in decibels among two images. The quality of the original and produced images is compared using this ratio. The PSNR value increases as the quality of the created or synthesized image improves. PSNR value is calculated using following equation:

$$\text{PSNR} = 10\text{log}10\ (\text{R}2/\text{MSE}) \tag{1}$$

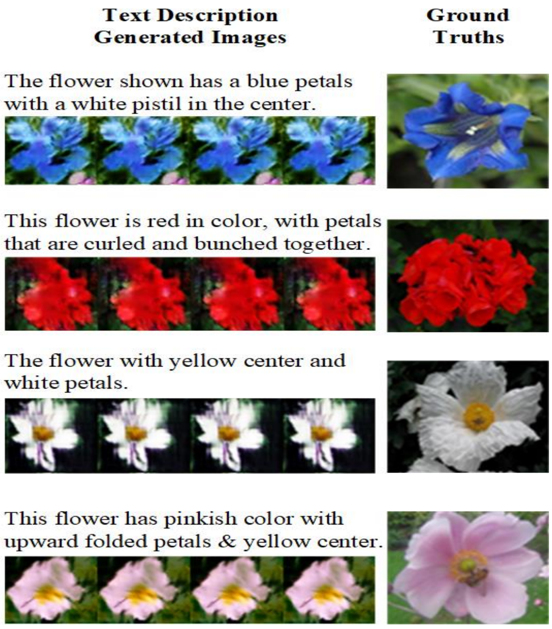
Text Description
Generated Images
Ground
Truths
The flower shown has a blue petals with a white pistil in the center.
This flower is red in color, with petals that are curled and bunched together.
The flower with yellow center and white petals.
This flower has pinkish color with upward folded petals & yellow center.

Figure 2. Input textual descriptions and resultant generated images with ground truths.

To validate the proposed approach, the results are compared with those of existing models including GAN-INT-CLS, StackGAN, StackGAN++, HDGAN, and DualAttn-GAN on the Oxford-102 flowers dataset. This performance comparison in terms of inception score is shown in Table 1 and that of the PSNR value is shown in Table 2. These results show that our model is capable of generating more unambiguous and diverse photos than the other models.

Table 1. Performance comparison of state-of-the-art methods vs. the methodology presented here by inception score.

Ref.	Models	Inception Score
Reed et al. [8]	GAN-INT-CLS	2.66 ± 0.03
Zhang et al. [16]	StackGAN	3.20 ± 0.01
Zhang et al. [17]	StackGAN++	3.26 ± 0.01
Zhang et al. [18]	HDGAN	3.45 ± 0.05
Cai et al. [19]	DualAttn-GAN	4.06 ± 0.05
Proposed Method	RC-GAN	4.15 ± 0.03

Table 2. PSNR value of generated images.

Ref.	Model	PSNR Value
Proposed Method	RC-GAN	30.12 dB

6. Conclusions and Future Work

In the fields of computer vision and natural language processing, text-to-image generation is a hot topic these days. For producing visually realistic and semantically consistent images, we presented a deep learning-based model (RC-GAN) and described how it works in the confluence of computer vision and natural language processing. The model was trained by text encoding and image decoding. Extensive experiments on the Oxford-102 flowers dataset demonstrated that the proposed GAN model generates better-quality images compared with existing models by providing the best recorded IS. The performance of our proposed method was compared with that of state-of-the-art methods using IS. Our model achieved an IS of 4.15 and a PSNR value of 30.12 dB on the Oxford-102 flowers dataset. We want to expand this work by training the model on a variety of datasets for image generation.

Author Contributions: The authors confirm contribution to the paper as follows: study conception, design and data collection: S.R.; analysis and interpretation of results: S.R., M.M.I., T.K.; draft manuscript preparation: S.R., M.M.I. All authors have read and agreed to the published version of the manuscript.

Funding: The Authors do not received funding for this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset is publicly available.

Conflicts of Interest: The authors are anonymously declared that they have no conflict of interest.

References

1. Kosslyn, S.M.; Ganis, G.; Thompson, W.L. Neural foundations of imagery. *Nat. Rev. Neurosci.* **2001**, *2*, 635–642. [CrossRef] [PubMed]
2. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
3. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
4. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
5. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
6. Kim, P. Convolutional neural network. In *MATLAB Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 121–147.
7. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
8. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. *arXiv* **2016**, arXiv:1605.05396.
9. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
10. Zia, T.; Arif, S.; Murtaza, S.; Ullah, M.A. Text-to-Image Generation with Attention Based Recurrent Neural Networks. *arXiv* **2020**, arXiv:2001.06658.
11. Liu, R.; Ge, Y.; Choi, C.L.; Wang, X.; Li, H. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16377–16386.
12. Gao, L.; Chen, D.; Zhao, Z.; Shao, J.; Shen, H.T. Lightweight dynamic conditional GAN with pyramid attention for text-to-image synthesis. *Pattern Recognit.* **2021**, *110*, 107384. [CrossRef]
13. Dong, Y.; Zhang, Y.; Ma, L.; Wang, Z.; Luo, J. Unsupervised text-to-image synthesis. *Pattern Recognit.* **2021**, *110*, 107573. [CrossRef]
14. Berrahal, M.; Azizi, M. Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques. *Indones. J. Electr. Eng. Comput. Sci.* **2022**, *25*, 972–979. [CrossRef]
15. Zhang, Y.; Han, S.; Zhang, Z.; Wang, J.; Bi, H. CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis. *Vis. Comput.* **2022**, 1–11. [CrossRef]

16. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
17. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [CrossRef] [PubMed]
18. Zhang, Z.; Xie, Y.; Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6199–6208.
19. Cai, Y.; Wang, X.; Yu, Z.; Li, F.; Xu, P.; Li, Y.; Li, L. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network. *IEEE Access* **2019**, *7*, 183706–183716. [CrossRef]