Text Pixs

[Alternative Names: Text Vision, Text Visual, Text Render, Text to Image]

Project Proposal

Internal Supervisor
Sir Osama Ahmed Khan

External Supervisor
Mirza Samad Ahmed Baig

Submitted by

Syeda Anshrah Gillani
1337-2021 / IT-12-236

Umema Mujeeb
2396-2021 / IT-21-225

Maheen Ali
1589-2021/ IT-21-234

Department of Computer Science,
Hamdard University, Karachi.

26th/06/2024

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).

Our focus will be on overcoming the challenge of consistently producing realistic images that accurately reflect the provided textual descriptions. Existing algorithms often struggle to align images with text effectively, prompting our development of a deep learning-based architecture: the recurrent convolutional generative adversarial network (RC-GAN). This model integrates advancements in text and image modeling to convert conceptual descriptions into pixelated visual representations.

Initially trained on the Oxford-102 flowers dataset, RC-GAN has demonstrated promising results with an inception score of 4.15 and a PSNR value of 30.12 dB, indicating its ability to generate highly realistic flower images from textual prompts. Moving forward, we plan to expand the model's capabilities by training it on multiple datasets, further enhancing its performance and applicability in diverse contexts.

Keywords:

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

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". 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, 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 Eng. Proc. 2022, 20, 16. https://doi.org/10.3390/engproc2022020016 https://www.mdpi.com/journal/engproc Eng. Proc. 2022,20,16 2 of 6 (CNNs). 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 in 2014, but Reed et al. was the first to use them for text-to-image generation in 2016. Salimans et al. 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. 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. offered a diverse conditional image synthesis model and performed large-scale experiments for different conditional generation tasks. Gao et al. 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. trained a model for generating images from text in an unsupervised manner. Berrahal et al.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 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. Objective

Develop a deep learning model, RC-GAN, for generating realistic images from textual descriptions with improved inception score and PSNR values, focusing on applications in photo-realistic image synthesis and visual content creation.

4. Problem Description

- ### What:
  The challenge of accurately converting textual descriptions into images involves understanding both the linguistic intricacies of the input text and the visual characteristics needed for generating realistic images. Current solutions often face limitations in capturing the full depth of text semantics or producing high-quality visuals. Our research aims to develop a comprehensive understanding of these issues and propose innovative solutions to improve the text-to-image generation process.

- ### Why:
  Advancements in text-to-image generation have significant implications for various domains such as content creation, education, and virtual reality. By

addressing the current limitations in this field, "Text Pixs" aims to provide a robust framework for generating accurate and high-quality images from text, thus enhancing the utility of AI in transforming how we visualize and interact with textual content.

- How:

  This research will involve a detailed study of existing text-to-image generation techniques, identification of gaps and challenges, and the development of novel approaches that integrate state-of-the-art NLP and computer vision methodologies. We will conduct empirical evaluations to assess the performance of our proposed techniques.

5. Methodology

In our upcoming work, we will focus on refining the training details of deep learning-based generative models. We plan to employ Conditional GANs integrating recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to generate meaningful images from textual descriptions. The dataset of flowers and their associated textual descriptions will continue to serve as our foundational data. Our approach will involve preprocessing textual data and resizing images to ensure compatibility for effective image synthesis from text using GANs. We will enhance our methodology by leveraging RNNs to capture contextual information across varying time steps, facilitating robust text-to-image mapping. Additionally, CNNs will be utilized to autonomously extract pertinent features from images. Our strategy includes utilizing RNNs to convert textual descriptions into 256-dimensional word embeddings, concatenated with a 512-dimensional noise vector. During training, we will optimize parameters such as batch size (64), gated-feedback (128), and the integration of noise and textual inputs within the generator. Detailed architectural insights will be depicted in Figure 1 to illustrate our refined model design.

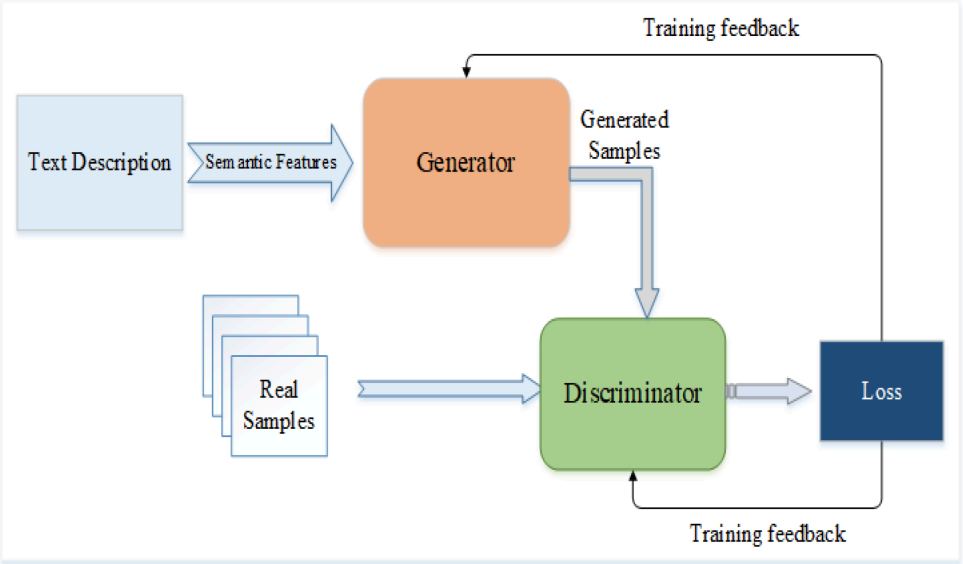

Figure 1. Architecture of the proposed method, which can generate images from text descriptions.

In our upcoming work, we will utilize semantic information extracted from textual descriptions as input to our generator model. This process converts characteristic details into pixel-level information, enabling the generation of realistic images. These generated images will be evaluated using a discriminator, which also receives inputs of real and mismatched textual descriptions alongside genuine dataset images. Our approach involves training the discriminator on a sequence of distinct pairs: real images with correct textual descriptions, incorrect images with mismatched texts, and generated images with authentic textual descriptions. By training on these varied pairings, the discriminator learns to distinguish between real and generated images effectively. Throughout training, we will calculate losses to refine model weights and provide feedback to both the generator and discriminator. As training progresses, our goal is to enhance the generator's ability to produce increasingly realistic images, ultimately challenging the discriminator's ability to differentiate between real and generated samples.

i. **In-Scope:**

- Research and development of new methodologies for text-to-image generation.
- Evaluation and analysis of proposed techniques.
- Compilation and preprocessing of large datasets for training and testing.

ii. **Out-of-Scope:**

- Development of a commercial application or user interface.
- Customization for specific industry applications.
- Real-time image generation capabilities.

6. Feasibility Study

i. Risks Involved:

- Technical Complexity: Developing accurate and efficient models requires significant technical expertise. To mitigate this, we will rely on a strong foundation of existing research and iterative testing.
- Resource Constraints: Training complex models may demand substantial computational resources. We plan to use cloud-based services to manage these requirements effectively.

ii. Resource Requirement:

- Computing Resources: Access to GPUs and cloud-based computational services for model training.
- Software: Python, TensorFlow or PyTorch, NLP libraries (Hugging Face Transformers), GAN frameworks.
- Data: Large and diverse image-caption datasets for effective training and evaluation.

7. Solution Application Areas

- **Content Creation:**
  Provides tools for generating visual content from textual descriptions, aiding writers, designers, and marketers.

- **Education:**
  Enhances educational materials by visualizing complex concepts described in text, improving comprehension and engagement.

- **Entertainment:**
  Assists in visualizing scenes from textual descriptions, benefiting authors, scriptwriters, and game developers.

- **Virtual Reality:**
  Enables the creation of immersive environments from textual descriptions, enhancing virtual and augmented reality experiences.

8. Tools/Technology

- **Hardware:**
  1) High-performance GPUs
  2) Cloud computing platforms (e.g., AWS, Google Cloud).

- **Software:**
  1) Python
  2) TensorFlow/PyTorch, NLP libraries (e.g., Hugging Face Transformers)
  3) GAN frameworks.

- **Data Sources:**
  Publicly available datasets such as COCO and Flickr30k

9. Responsibilities of the Team Members

- **R**: Responsible
- **A**: Accountable
- **C**: Consulted
- **I**: Informed

Project Deliverable Activity	Supervisors	Syeda Anshrah Gillani	Umema Mujeeb	Maheen Ali
Literature Review	C,I	A, R	I	I
PRD, User stories	C, A	R, A	C, I	C, I
Data Collection & Preprocessing	C, I	C, I	R	R
AI Model Development / Fine-Tuning	C, I	R, A	R, I	R, I
FE & BE Development	C, I	I, A	R	R
Final Version Release & Production Deployment	C, I	R, A	I, R	I, R
FYP Report	C, A	R, A	R, I	R, I

Figure 2. RACI Chart

10. Planning

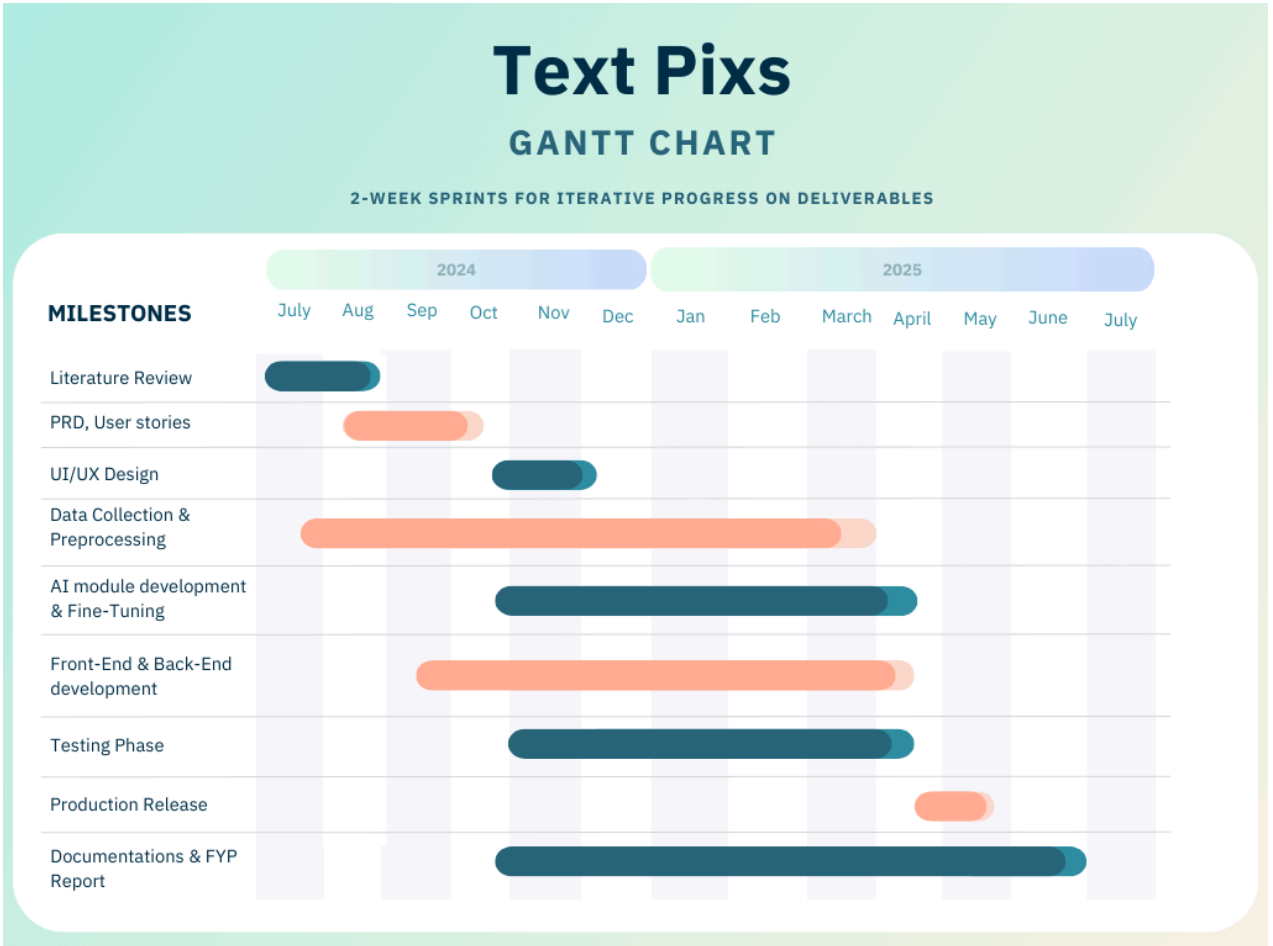

Figure 3. GANTT Chart

11. References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[2] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. arXiv preprint arXiv:1609.07093, 2016.

[3] M.-M. Cheng, S. Zheng, W.-Y. Lin, V. Vineet, P. Sturgess, N. Crook, N. J. Mitra, and P. Torr. Imagespirit: Verbal guided image parsing. ACM Transactions on Graphics (TOG), 34(1):3,
2014.

[4] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems, pages 1486–1494, 2015.

[5] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 5706–5714, 2017.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and superresolution. In Proceedings of European Conference on Computer Vision, pages 694–711. Springer, 2016.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[9] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Advances in Neural Information Processing Systems, pages 3581–3589, 2014.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In Proceedings of European Conference on Computer Vision, pages 740 755. Springer, 2014.

[11] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov. Generating images from captions
with attention. arXiv preprint arXiv:1511.02793, 2015.

[12] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In Advances in Neural Information Processing Systems, pages 42–51, 2018.

[13] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4467–4477, 2017.

[14] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In Proceedings of International Conference on Image Processing (Cat. No. 03CH37429), volume 1, pages 253–256. IEEE, 2003.

[15] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016.

[17] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In Advances in Neural Information Processing Systems, pages 217–225, 2016.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.

[19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pages 2234–2242, 2016.

[20] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681, 1997.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-Ucsd Birds-200-2011 dataset. 2011.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning, pages 2048–2057, 2015.

[25] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1316–1324, 2018.

[26] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
pages 21–29, 2016.

[27] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 5907–5915, 2017.

[28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8):1947–1962, 2018.

[29] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang. Progressive attention guided recurrent network for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and
Pattern Recognition, pages 714–722, 2018.

[30] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. MDNet: A semantically and visually interpretable medical image diagnosis network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6428–6436, 2017.

[31] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In Proceedings of European Conference on Computer Vision, pages 597–613. Springer, 2016.